



HAL
open science

Contributions to the Design and Training of Transformers in Computer Vision

Alaaeldin Mohamed Elnouby Abdallah Ali

► **To cite this version:**

Alaaeldin Mohamed Elnouby Abdallah Ali. Contributions to the Design and Training of Transformers in Computer Vision. Computer Science [cs]. INRIA Paris; ENS Paris - Ecole Normale Supérieure de Paris; PSL University, 2023. English. NNT: . tel-04477587

HAL Id: tel-04477587

<https://inria.hal.science/tel-04477587>

Submitted on 26 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à École Normale Supérieure

Contributions to the Design and Training of Transformers in Computer Vision

Soutenu par

Alaaeldin Ali

Le 11 Juillet 2023

École doctorale n°386

**Sciences Mathématiques
de Paris-Centre**

Spécialité

Informatique

Préparée au

Willow - Meta AI

Composition du jury :

Tinne TUYTELAARS
KU Leuven

Président du jury
Examinatrice

Paolo FAVARO
University of Bern

Rapporteur

Cees SNOEK
University of Amsterdam

Rapporteur

Alexei EFROS
University of California Berkeley

Examineur

Natalia NEVEROVA
Meta AI

Membre Invitée

Ivan LAPTEV
Inria and DI/ENS

Directeur de thèse

Hervé JEGOU
Meta AI

Co-directeur de thèse

Abstract

Transformers have revolutionized representation learning across modalities, achieving state-of-the-art results in natural language processing, computer vision, speech, and beyond. This thesis explores the potential of Transformer models for computer vision. We propose architectural innovations to overcome their limitations, developing sample-efficient self-supervised pre-training methods, and advancing multimodal learning with Transformers. First, we propose Cross-Covariance Attention to reduce the quadratic complexity of self-attention achieving similar performance as vision transformers with lower memory footprint and computational cost, enabling the application of vision transformers to higher-resolution images. We then investigate self-supervised pre-training for vision transformers. We propose SplitMask, a denoising autoencoding method based on masked image modeling. Unlike joint embedding methods, SplitMask does not require large-scale pre-training datasets and can be applied to diverse visual data. SplitMask matches the performance of joint embedding methods when pre-trained on datasets two orders of magnitude smaller, highlighting its improved sample efficiency. Moreover, we apply masked image modeling to neural image compression in the form of an improved entropy model yielding a strong rate-distortion performance and enabling the compression of images to the size of a short SMS or tweet. Finally, we propose ImageBind, a method for learning a shared embedding space across six modalities. ImageBind leverages the abundance of images and text on the web to enable transfer to modalities with scarce annotations like depth, thermal, audio, and IMU. In summary, this thesis demonstrates the potential of Transformers for computer vision through architectural innovations, new self-supervised objectives, and multimodal knowledge transfer. The methods proposed in this thesis push the boundaries of transformers in vision by enhancing their scalability and generality, enabling more sample-efficient representation learning, and facilitating transfer across modalities.

Keywords : vision transformers, self-supervised learning, weakly supervised learning, multimodal learning, image compression

Résumé

Les transformateurs ont révolutionné l'apprentissage de la représentation dans de nombreuses modalités, obtenant des résultats de pointe dans le traitement du langage naturel, la vision par ordinateur, la parole et bien d'autres domaines. Cette thèse explore le potentiel des modèles de transformateurs pour la vision par ordinateur. Nous proposons des innovations architecturales pour surmonter certaines de leurs limites. Nous développons des méthodes de pré-entraînement auto-supervisé efficaces en termes d'échantillons, et considérons l'utilisation de ces transformateur dans un contexte d'apprentissage multimodal.

Dans un premier temps, nous proposons l'attention à covariance croisée pour réduire la complexité quadratique de l'attention d'origine et obtenir des performances similaires avec une empreinte mémoire et un coût de calcul moindres, ce qui permet d'appliquer les transformateurs de vision à des images à plus haute résolution. Nous étudions ensuite le pré-entraînement auto-supervisé pour les transformateurs de vision. Nous proposons SplitMask, une méthode de débruitage automatique basée sur la modélisation d'images masquées. Contrairement aux méthodes de plongements conjointes, SplitMask ne nécessite pas d'ensembles de données de pré-entraînement à grande échelle et peut être appliqué à diverses données visuelles. SplitMask est aussi performant que les méthodes de plongements conjoints lorsqu'il est entraîné sur des ensembles de données deux fois plus petits, ce qui met en évidence l'amélioration de l'efficacité d'apprentissage avec peu de données. En outre, nous appliquons la modélisation d'images masquée à la compression d'images neuronales sous la forme d'un modèle entropique amélioré. Cela permet d'obtenir de bonnes performances en matière de débit-distorsion dans les régimes où la compression d'image est extrême, tels la taille d'un SMS ou d'un tweet. Enfin, nous proposons ImageBind, une méthode d'apprentissage d'un espace de plongement partagé entre six modalités. En résumé, cette thèse démontre le potentiel des transformateurs pour la vision par ordinateur grâce à des innovations architecturales, de nouveaux objectifs auto-supervisés et un transfert de connaissances multimodal. Les méthodes proposées dans cette thèse repoussent les limites des transformateurs en vision en améliorant leur passage à l'échelle et leur généralité, en permettant un apprentissage de la représentation plus efficace en termes d'échantillons, et en facilitant le transfert entre les modalités.

Mots clés : transformateurs de vision, apprentissage auto-supervisé, apprentissage faiblement supervisé, apprentissage multimodal, compression d'image.

Acknowledgements

I would like to express my deepest gratitude to the many people who have made this thesis possible. First and foremost, I would like to thank my wife Sara for her unwavering support throughout my academic journey. Her encouragement, patience, and love have been my pillars of strength, and I could not have done this without her. I am deeply grateful to my parents for their constant support and encouragement. Their sacrifices and belief in me have been instrumental in shaping my academic and personal life. I owe them a debt of gratitude that can never be repaid. I would also like to thank my lovely sisters for their support and encouragement, especially during the tough times. Their love, guidance, and motivation have been invaluable to me.

I would like to express my sincere appreciation to my PhD advisors Hervé, Ivan, and Natalia for their guidance and support throughout my PhD journey. I feel truly fortunate to have had such exceptional mentors who dedicated their time, expertise, and energy to help me achieve my goals and have advocated for me countless times. In particular, I cannot overstate my gratitude to Hervé for his exceptional mentorship. He taught me to be ambitious, to never settle for easy wins, and most importantly, how to be a kind and compassionate leader. His unwavering support and guidance were instrumental in my success during the PhD program. I will forever be grateful for his mentorship and the impact he has had on my professional as well as personal development. I also want to thank Ivan for his support and guidance during the PhD program. His expertise and insights were invaluable in shaping the direction of my research and helping me navigate the challenges of the program. He consistently challenged me to ask the hard questions, pushing me to grow and excel in my research. Finally, I want to express my deep appreciation to Natalia, who played a crucial role in my PhD journey by presenting me with the opportunity and believing in my potential. Her guidance and mentorship during the first half of the program were invaluable, and she continued to be a great mentor throughout the program and hopefully beyond. I am also very grateful to the members of the jury. It is a great pleasure to present and defend my PhD work in front of researchers I admire and have been looking up to ever since I started my research journey. I would like to thank the rapporteurs, Cees Snoek and Paolo Favaro, and the examiners, Alexei Efros and Tinne Tuytelaars, for dedicating the time to review and provide feedback for the manuscript and my PhD work.

I can't express enough gratitude to my collaborators, who've played a crucial part in my academic journey. Working with Gautier was one of the highlights of my PhD program. He was an amazing collaborator, and we had some of the most fun and engaging projects together. Our discussions were always insightful, and I learned a great deal from working with him. I am grateful to Ishan, his ability to approach complex problems with clarity and creativity has been truly remarkable. From my interactions with Ishan, I have come to see him as a gold standard

Chapter 0. Acknowledgements

for post-PhD research career success. I feel fortunate to have had the opportunity to work and learn from him during my PhD program. I also want to thank Rohit who was a very dedicated and inspiring collaborator. His contributions were invaluable and contributed significantly to the quality of our research. I want to thank Jakob for his mentorship and for providing valuable feedback and insights that helped me to shape my research. Our weekly discussions were always constructive and thought-provoking, and I learned a great deal from his expertise and experience. I would like to thank Matthew for being a great collaborator and supporter, he has a great work ethic and strong drive that was especially helpful when projects did not go as planned. I wish to thank Edouard, Armand, Piotr, Matthijs and Gabriel whom presence was very influential and inspiring for me throughout my PhD journey. I want to thank Federico and Joao, the amazing interns that I had the pleasure of working with. Finally, I would like to thank my fellow PhD students, Hugo, Mathilde, Adrien, Mohamed, Baptiste, Charlotte, Roberto, Pierre, Lina, Badr, Guillaume and Alex.

Once again, I would like to express my deepest gratitude to all these individuals who have contributed to the completion of this PhD thesis. I could not have done this without their support, guidance, and encouragement.

Table of Contents

Abstract	i
Résumé	ii
Acknowledgements	iii
Table of Contents	v
1 Introduction	1
1.1 Challenges	2
1.1.1 Computational Complexity	2
1.1.2 Scalable and Sample Efficient Representation Learning	3
1.1.3 Multimodal Learning	3
1.2 Outline and Contributions	3
1.2.1 Taming the quadratic complexity of vision transformers	3
1.2.2 Sample efficient self-supervised pre-training with Transformers	4
1.2.3 Masked Image modeling for improved image compression	4
1.2.4 ImageBind for learning a shared embedding space for six modalities	5
1.3 Publications	5
2 Background	7
2.1 Modern Architectures in Computer Vision	7
2.1.1 Convolutional Neural Networks	7
2.1.2 Self-Attention and Transformers	8
2.1.3 Efficient Self-Attention	9
2.1.4 CNNs Augmented with Self-Attention	10
2.1.5 Vision Transformers	11
2.2 Self-Supervised Learning	12
2.2.1 Autoencoders	12
2.2.2 Pretext tasks	12
2.2.3 Joint embeddings methods	13
2.3 Vision-Language pre-training	15
2.4 Lossless Compression	15
3 Cross Covariance Image Transformers	16
3.1 Introduction	17
3.2 Related work	19
3.3 Method	20

Table of Contents

3.3.1	Background	20
3.3.2	Cross-covariance attention	22
3.3.3	Cross-covariance image transformers	24
3.4	Preliminary study on Vision Transformers (ViT)	26
3.4.1	Impact of resolution versus patch size	26
3.4.2	Approximate attention models in ViT with DeiT training	27
3.4.3	Training and testing with varying resolution	27
3.5	Experimental evaluation	28
3.5.1	Image classification	28
3.5.2	Object detection and instance segmentation	32
3.5.3	Semantic segmentation	34
3.5.4	Implementation details	35
3.6	Conclusion and Future Work	35
4	Sample Efficient Self-Supervised Pre-training with Vision Transformers	36
4.1	Introduction	37
4.2	Related Work	39
4.3	Analysis	40
4.3.1	Sample Efficiency	41
4.3.2	Learning using non object-centric images	42
4.3.3	Tokenizers	42
4.4	Methodology	44
4.4.1	SplitMask	44
4.4.2	Encoder-Decoder Architecture	44
4.4.3	Global Contrastive Loss	45
4.5	Experiments	45
4.5.1	Datasets	45
4.5.2	Dense Prediction	47
4.5.2.1	Object detection and Instance Segmentation	47
4.5.2.2	Semantic Segmentation	47
4.5.3	Image Classification	47
4.5.4	Pre-training using ImageNet	49
4.5.5	Ablations	49
4.5.5.1	SplitMask vs BEiT	49
4.5.5.2	Encoder-Decoder vs BEiT	50
4.5.5.3	Overfitting during pre-training	51
4.5.6	Implementation Details	51
4.6	Conclusion	52
5	Image Compression with Product Quantized Masked Image Modeling	53
5.1	Introduction	54
5.2	Related work	56
5.3	Product Quantized Masked Image Modeling	57
5.3.1	High-level architecture: PQ-VAE	58
5.3.2	Image entropy model	59
5.3.3	Training the PQ-MIM	61
5.4	Experiments	62
5.4.1	Experimental setup	62
5.4.2	Main experimental results	64
5.4.3	Analysis and Ablations	65

5.4.4	Limitations	67
5.5	Conclusion	68
6	Learning a shared embedding space for six modalities with ImageBind	70
6.1	Introduction	71
6.2	Related Work	72
6.3	Method	73
6.3.1	Preliminaries	73
6.3.2	Binding modalities with images	74
6.3.3	Implementation Details	75
6.3.3.1	Inference implementation details	75
6.4	Experiments	76
6.4.1	Emergent zero-shot classification	77
6.4.1.1	Zero-shot evaluation details	77
6.4.2	Comparison to prior work	78
6.4.3	Few-shot classification	79
6.4.3.1	Few-shot evaluation details	80
6.4.4	Analysis and Applications	80
6.4.4.1	Qualitative evaluation details	81
6.4.5	Pretraining details	82
6.4.6	Datasets and Metrics	82
6.5	Ablation Study	84
6.5.1	Scaling the Image Encoder	85
6.5.2	Training Loss and Architecture	85
6.6	Discussion and Limitations	88
7	Conclusion	89
7.1	Summary of Contributions	89
7.1.1	Reducing the quadratic complexity of Vision Transformers	89
7.1.2	Masked Image Modeling as a powerful tool	90
7.1.3	Learning representations for data-scarce modalities	91
7.2	Open Problems and Future Directions	92
7.2.1	Efficient Attention for Video Data	92
7.2.2	Improving the scalability of Self-supervised Learning	92
8	Résumé Substantiel	94
8.1	Défis	95
8.1.1	Complexité computationnelle	96
8.1.2	Apprentissage de la représentation à grande échelle et efficace en termes d'échantillons	96
8.1.3	Apprentissage multimodal	96
8.2	Présentation et contributions	97
8.2.1	Maîtriser la complexité quadratique des transformateurs de vision	97
8.2.2	Pré-entraînement auto-supervisé efficace avec Transformer	97
8.2.3	Modélisation d'images masquées pour une meilleure compression d'images	98
8.2.4	ImageBind pour l'apprentissage d'un espace d'intégration partagé pour six modalités	98
	Appendix	100

Table of Contents

A	Additional results for XCiT	100
A.1	More XCiT models	100
A.2	Image Retrieval	100
References		104

Chapter 1

Introduction

Deep learning has had a profound impact on the field of computer vision. This impact has revolutionized the way we perceive, understand, and interact with the digital world. The rapid progress has driven the development of new algorithms and architectures that have continuously advanced the state of the art in computer vision tasks. The advancements in deep learning have outperformed classical computer vision methods, and have enabled a wide range of applications such as autonomous driving [Chen, 2015], image and video generation [Ramesh, 2022a; Singer, 2022], and medical image analysis [Shin, 2016], to name but a few. One of the key factors contributing to the success of deep learning in computer vision has been the innovation in network architectures [He, 2016; Szegedy, 2015; Hu, 2018; Wang, 2018]. These innovations have resulted in models that can better learn and represent complex patterns from large-scale data. Moreover, they have enabled the development of models that can learn hierarchical representations, which are crucial for understanding the structure in visual data.

Convolutional Neural Networks (CNNs) [Fukushima, 1980; LeCun, 1989] have long been considered the go-to architecture for computer vision tasks. CNNs are characterized by their inductive biases and sample efficiency, which enable them to learn and generalize effectively from a limited number of training examples. Their ability to learn local features and robustly represent spatial hierarchies has made them the dominant architecture for vision tasks. The success of CNNs can be attributed to their ability to exploit the spatial structure inherent in images, allowing them to efficiently learn meaningful representations. Despite their effectiveness, the same inductive biases and design restrictions that make CNNs efficient can potentially limit their generality as the amount of training data increases. For example, CNNs are known to have a strong bias towards texture [Geirhos, 2018], which can limit their ability to generalize to new and diverse image datasets.

The landscape of deep learning began to change with the introduction of the Transformer architecture [Vaswani, 2017b]. Originally designed for machine translation and natural language processing (NLP) tasks, Transformers have proven to be highly effective in capturing long-range dependencies and complex patterns in sequential data. The key innovation in the Transformer architecture is the self-attention mechanism, which allows the model to dynamically weigh the importance of different input elements in the context of the entire sequence. The success of Transformers in NLP tasks led researchers to explore their potential in the realm of computer

vision. Initially, self-attention mechanisms were integrated into existing CNN architectures, resulting in hybrid models [Wang, 2018] that combined the strengths of both approaches. The next logical step in this evolution was the development of Vision Transformers (ViT) [Dosovitskiy, 2021a], which marked a significant departure from traditional CNN-based approaches. ViT models are composed entirely of self-attention and Multi-Layer Perceptron (MLP) modules, completely discarding convolutional layers. Despite this radical shift in architecture, ViTs have achieved state-of-the-art results on a wide range of computer vision tasks, surpassing the performance of their CNN-based counterparts. Moreover, Vision Transformers exhibit different properties compared to CNNs, including higher robustness towards occlusions and perturbations as well as lower bias towards texture [Naseer, 2021]. These properties make ViTs a promising candidate for pushing the generalization performance for a wide range of vision tasks.

Transformers have opened up a vast landscape of possibilities for computer vision research. The expressiveness and generality of Transformer models make them highly adaptable to a wide range of tasks and domains. Moreover, their ability to model long-range dependencies enables capturing global context and complex relationships within visual data more effectively than CNNs. In this thesis, we delve deeper into the world of Transformers, exploring their potential, and applications in various computer vision tasks. We will investigate the factors that have contributed to their success and how to adapt them better for vision applications. Furthermore, we explore how the unification of architectures across modalities leads to innovative approaches for providing supervision to modalities with scarce data by leveraging strong vision models.

1.1 Challenges

Transformers pave the way to exciting new possibilities for learning stronger visual representations via supervised and unsupervised learning as well as opening the door for more homogenous approaches for multimodal learning. In this manuscript, we address several challenges related to these topics.

1.1.1 Computational Complexity

One significant challenge posed by the Transformers, as introduced by Vaswani *et al.* [Vaswani, 2017b], is the quadratic complexity with respect to the sequence length. The self-attention mechanism employed by transformers involves computing pairwise interactions between every element in the input sequence, resulting in a complexity of $\mathcal{O}(N^2)$, where N is the sequence length. In the context of computer vision, this complexity poses a challenge for the direct application of transformers to high-resolution images, which is often required for tasks such as object detection, semantic segmentation, and image compression. To overcome this limitation, there is a need for architectural innovations that can reduce the computational cost of transformers for large-resolution image tasks.

1.1.2 Scalable and Sample Efficient Representation Learning

Transformers excel at capturing long-range dependencies and can learn hierarchical representations effectively. Similar to their transformative impact on representation learning in Natural Language Processing with approaches like BERT [Devlin, 2018] and GPT [Radford, 2018], Transformers have the potential to revolutionize self-supervised pre-training for vision tasks. However, to fully realize this potential, it is critical to develop more sample-efficient methods that can scale more easily in terms of data and compute compared to existing joint embedding approaches. The expressive power of the Transformer architecture, when used in conjunction with more generic and less biased denoising objectives, has the potential to deliver the efficiency and scalability required to substantially improve self-supervised pre-training of vision models.

1.1.3 Multimodal Learning

As a universal architecture, the Transformer model has seen successful application across various modalities. This versatility opens the door for shared design principles and components that can facilitate multimodal learning. By using Transformers as a common framework for multimodal learning, we can develop methods that effectively integrate information from diverse modalities and enable seamless transfer of knowledge between different tasks and domains. However, while there may be an abundance of data for some modalities, others may suffer from severe data scarcity. In such cases, the knowledge transfer enabled by using a common framework like Transformers can have a significant impact. By leveraging the strengths of different modalities and transferring knowledge learned from one domain to another, we can improve the performance of models in domains where data scarcity is a major issue.

1.2 Outline and Contributions

The manuscript begins with a discussion of the background and relevant literature in Chapter 2. We then delve into the details of each of our four main contributions in the subsequent chapters as outlined below.

1.2.1 Taming the quadratic complexity of vision transformers

In Chapter 3, we delve into the computational complexity of vision transformers [Dosovitskiy, 2021a]. Specifically, we identify that the self-attention operation, which lies at the heart of the Transformer architecture [Vaswani, 2017b], displays a quadratic growth rate with respect to image resolution. This can render vision transformers computationally prohibitive, particularly when tasked with processing higher-resolution images, as is frequently encountered in essential downstream computer vision applications such as semantic segmentation and object detection. To address this challenge, we propose a novel alternative formulation of the self-attention operation, which we refer to as *Cross-Covariance Attention* (XCA) [El-Nouby, 2021c]. Cross-covariance attention exhibits a linear complexity with respect to input size and can be seamlessly integrated as a replacement for self-attention in vision transformers. We introduce

XCiT, a novel architecture for computer vision with XCA at its core. We demonstrate that XCiT offers significant enhancements in terms of memory consumption and throughput while retaining the strong performance of vision transformers.

1.2.2 Sample efficient self-supervised pre-training with Transformers

The Transformer architecture has sparked innovation in self-supervised pre-training in Natural Language Processing, leading to the development of influential models such as BERT [Devlin, 2018] and GPT [Radford, 2018]. In computer vision, joint embedding methods [Chen, 2020c; He, 2020; Caron, 2021] have emerged as dominant due to their strong off-the-shelf performance and competitive performance with supervised methods for finetuning. However, joint embedding methods suffer from some limitations, such as their dependence on hand-crafted data augmentation techniques tailored for ImageNet [Deng, 2009b]. Moreover, these methods can be challenging to scale in terms of model size, as pointed out by [Chen, 2021b]. Motivated by the success of BERT in NLP, in Chapter 4 we investigate the efficacy of denoising autoencoding methods when used in conjunction with vision transformers for self-supervised pre-training. Firstly, we propose SplitMask [El-Nouby, 2021a], a novel self-supervised pre-training method based on masked image modeling. Secondly, we find that SplitMask offers improved sample efficiency and can be employed to learn robust representations using datasets orders of magnitude smaller than those required by joint embedding methods. Additionally, we observe that SplitMask is well-suited for training using a wider range of visual data as it is not biased towards a specific distribution of images, unlike joint embedding methods that are typically biased towards object-centric images of ImageNet.

1.2.3 Masked Image modeling for improved image compression

Neural image compression has emerged as a promising alternative to traditional codecs due to its superior perceptive and psychovisual image quality. Typically, neural image compression systems consist of three main components: a deep neural encoder and decoder, which learn a mapping and its inverse between pixel and latent representations of the image; a quantizer, which maps the continuous latent representations to a set of discrete symbols; and an entropy model, which exploits redundancies in the set of discrete symbols to reduce the final length of the bitstream. In Chapter 5, we propose (1) Adopting XCiT for designing the encoder and decoder; (2) Employing product quantization in place of vector quantization for better scaling in bit-rates; (3) A novel entropy model for neural image compression based on masked image modeling. Our method, PQ-MIM [El-Nouby, 2023], achieves strong rate reduction compared to simple frequency-based baselines while providing a significant speedup compared to autoregressive methods that can be prohibitively slow for high-resolution images. As a result, PQ-MIM enables extreme compression of images to a size of a short tweet or SMS.

1.2.4 ImageBind for learning a shared embedding space for six modalities

Transformers have emerged as a versatile architecture that achieves exceptional performance across various modalities, such as text [Devlin, 2018; Radford, 2018], images [Dosovitskiy, 2021a; Touvron, 2020], video [Gedas Bertasius, 2021; Tong, 2022], audio [Xu, 2022], and graphs [Yun, 2019], among others. This presents exciting possibilities for developing multimodal systems with shared components and similar designs. However, the predominant approach for training these modalities is supervised learning, which relies on learning a mapping from the sensory inputs to a set of categorical labels. Unfortunately, supervised learning is constrained by the challenge of efficiently and scalably collecting annotations. To overcome this challenge, weakly supervised learning has emerged as an alternative paradigm that relies on collecting large amounts of data with noisy labels that can be easier to acquire. This paradigm has shown immense success for learning visual representations [Joulin, 2016; Radford, 2021; Zhai, 2022] powered by extremely large-scale collections of image and text pairs scraped from the open web. While there is no shortage in images accompanied by textual descriptions on the open web, generating analogous collections for other modalities of interest, such as audio, thermal images and depth images, is considerably more challenging. In Chapter 6 we address this issue introducing ImageBind, a novel method for training encoding of six different modalities in a shared latent space leveraging the strong performance of existing vision and language models.

The manuscript concludes with Chapter 7 summarizing our key findings and insights, discussing the limitations of our approaches and future work.

1.3 Publications

The following publications are included in whole or in part within this manuscript:

- **Alaaeldin El-Nouby**, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek and Hervé Jegou. “XCiT: Cross-Covariance Image Transformers”. In *Conference on Neural Information Processing Systems (NeurIPS), 2021*. The code is available at <https://github.com/facebookresearch/xcit>. (see Chapter 3)
- **Alaaeldin El-Nouby***, Gautier Izacard*, Hugo Touvron, Ivan Laptev, Hervé Jegou, Edouard Grave. “Are Large-scale Datasets Necessary for Self-Supervised Pre-training?”. *Under Review*. (see Chapter 4)
- **Alaaeldin El-Nouby**, Matthew J. Muckley, Karen Ullrich, Ivan Laptev, Jakob Verbeek and Hervé Jégou. “Image Compression with Product Quantized Masked Image Modeling”, In *Transactions of Machine Learning Research (TMLR), 2023* (see Chapter 5)
- Rohit Girdhar*, **Alaaeldin El-Nouby***, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, Ishan Misra*. “ImageBind: One Embedding Space To Bind Them All”. To presented at *Conference on Computer Vision and Pattern Recognition (CVPR), 2023* as a **highlighted paper**. The code is available at <https://github.com/facebookresearch/imagebind>. (see Chapter 6)

Additionally, the publications listed below were a result of research conducted during the PhD program, but are not directly included within the thesis:

- Rohit Girdhar*, **Alaaeldin El-Nouby***, Mannat Singh*, Kalyan Vasudev Alwala*, Armand Joulin and Ishan Misra*. “OmniMAE: Single Model Masked Pretraining on Images and Videos”. To be presented at *Conference on Computer Vision and Pattern Recognition (CVPR), 2023*. The code is available at <https://github.com/facebookresearch/omnivore/tree/main/omnimae>.
- **Alaaeldin El-Nouby**, Natalia Neverova, Ivan Laptev, Hervé Jégou. “Training vision transformers for image retrieval”. arXiv 2021.
- Ben Graham, **Alaaeldin El-Nouby**, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, Matthijs Douze. “LeViT: a Vision Transformer in ConvNet’s Clothing for Faster Inference”. *International Conference on Computer Vision, 2021*. The code is available at <https://github.com/facebookresearch/LeViT>
- Hugo Touvron, Matthieu Cord, **Alaaeldin El-Nouby**, Jakob Verbeek, Hervé Jégou. “Three things everyone should know about Vision Transformers”. *European Conference on Computer Vision, 2022*.

Chapter 2

Background

2.1 Modern Architectures in Computer Vision

In this section, we provide an overview of modern architectures in computer vision starting from Convolutional Neural Networks to Transformers passing by hybrid models that combines components from both. Additionally, we will summarize multiple efforts that aimed to reduce the computation complexity of Transformers. Finally, we delve into the most recent family of vision models that were developed after and motivated by Vision Transformers.

2.1.1 Convolutional Neural Networks

Convolutional Neural Networks, in their current form, were first proposed by LeCun *et al.* [LeCun, 1989]. Convolutional layers operate by applying a set of learnable kernels to an input. Each kernel has dimensions that are relatively small, but slides across the full width and height of the input computing the dot product between the weights of the kernel and the input at every spatial position, allowing it to effectively capture local patterns. The work of LeCun *et al.* [LeCun, 1989] demonstrated that the kernel weights can be effectively trained using backpropagation. The LeNet-5 network [LeCun, 1998] was subsequently designed for digit recognition and marked the inception of CNNs. LeNet-5 was primarily composed of alternating convolutional and pooling layers, followed by a fully connected layer for output classification. The fundamental architecture of CNNs remained unchanged for years until the introduction of AlexNet [Krizhevsky, 2012]. AlexNet significantly increased the depth and width of CNNs, adopted Rectified Linear Units (ReLU) as activation functions and popularized data augmentation as a regularization technique. The success of AlexNet marked the start of the deep learning revolution with many innovations in the CNN design to follow. VGGNet [Simonyan, 2014] and GoogLeNet [Szegedy, 2015] further built upon the foundation laid by AlexNet. VGGNet demonstrated that the depth of a CNN is a crucial factor for its performance, while GoogLeNet introduced the inception module, which allowed for increased network depth and width without an explosion in computational cost.

The introduction of ResNet [He, 2016] was another important milestone in the evolution of CNNs in particular and deep learning in general. ResNets introduced residual connections, which helped solve the vanishing gradient problem, allowing for the training networks with hun-

dreds of layers. With CNNs demonstrating a dominant performance for various vision tasks, the efficiency and scalability of network architectures then became the focus of research. MobileNet [Howard, 2017] and EfficientNet [Tan, 2019] were designed to be lightweight and efficient while maintaining high performance. MobileNet introduced depthwise separable convolutions, while EfficientNet used a systematic approach to scale up CNNs, considering depth, width, and resolution. RegNets [Radosavovic, 2020] introduced the idea utilizes a design space exploration approach to identify network architectures that offer excellent trade-offs between computational cost and accuracy. Most recently, NfNets Brock *et al.* [Brock, 2021] have been introduced as another step forward in efficient network design. NfNets employ a normalization-free design with model scaling, and they have achieved state-of-the-art performance on ImageNet while being more efficient than existing models. To sum up, the architecture of CNNs has retrospectively come a long way from the early days of LeNet, with each new development leading to improved performance and efficiency.

2.1.2 Self-Attention and Transformers

The attention mechanism was initially proposed by Bahdanau *et al.* [Bahdanau, 2014] for sequence-to-sequence models, especially in neural machine translation tasks. By dynamically distributing the model’s “*attention*” over different parts of the input sequence based on their relevance to the current task. Attention improved upon fixed-length context vectors and enabled more effective learning of long-range dependencies. A significant evolution in the self-attention mechanism was achieved with the advent of the Transformer model by Vaswani *et al.* [Vaswani, 2017b]. The Transformer discarded the traditional recurrent and convolutional layers and relied solely on attention mechanisms for processing the input. This approach, which allows parallel computation and captures long-range dependencies effectively, has resulted in remarkable improvements in performance on several NLP tasks.

The self-attention mechanism allows each token in the input sequence to compute a weighted sum of all tokens, including itself. In the simplest form, it’s defined by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

where:

- Q is the matrix of queries;
- K is the matrix of keys;
- V is the matrix of values;
- d_k is the dimension of the key.

The self-attention mechanism is applied to the input sequence, where Q , K , and V are linear projections of the same input

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V. \quad (2.2)$$

In this setting, each token in the sequence attends to all other tokens to compute its representation. However, in practice the self-attention mechanism uses multiple sets of learned linear transformations, resulting in the *Multi-Head* attention as proposed by Vaswani *et al.* [Vaswani, 2017b]. This allows the model to focus on different types of information. For h heads, multi-head attention is defined as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_O \quad (2.3)$$

where each head i is computed as:

$$\text{head}_i = \text{Attention}(QW_{Q_i}, KW_{K_i}, VW_{V_i}) \quad (2.4)$$

Here:

- W_{Q_i} , W_{K_i} , and W_{V_i} are the parameter matrices for each head
- W_O is the output projection.

The Transformer model since its inception has been generalized to other modalities and tasks. One of its most notorious applications has been language understanding and generation tasks, such as BERT [Devlin, 2018] and GPT [Radford, 2018]. BERT uses the Transformer’s bidirectional self-attention to encode a rich understanding of contextual relationships between words, demonstrating state-of-the-art performance on several NLP benchmarks. GPT, on the other hand, uses an autoregressive causal Transformer to generate text, illustrating the potential for large-scale language models. Transformers have also been extended to non-textual modalities. In the realm of speech recognition and synthesis, models like Conformer [Gulati, 2020] have integrated self-attention into a hybrid architecture, combining the best of convolutional and self-attention mechanisms to deal with time-series data effectively. Self-attention and Transformers have evolved from being solutions to specific problems in machine translation to cornerstone architectures across different modalities. Their potential for capturing complex patterns and long-range dependencies, coupled with their suitability for parallel computation, position them as a driving force in the evolution of machine learning models.

2.1.3 Efficient Self-Attention

While Transformers have achieved state-of-the-art performance in various domains, they often come with costly computational and memory requirements due to the quadratic complexity of self-attention with respect to the input sequence length. Many variants of the self-attention operation has been proposed to address this particular issue. An overview of such methods is summarized in Figure 2.1. One family of efficient Transformer models are Sparse Transformers [Child, 2019], which aim to reduce the computational complexity by limiting the number of attended positions in self-attention. Models like the Longformer [Beltagy, 2020] and BigBird [Zaheer, 2020] introduced novel sparse attention patterns that significantly reduce complexity while maintaining high performance. Another line of research, referred to as Kernelized Transformers, leverages kernel methods to approximate the attention mechanism. These meth-

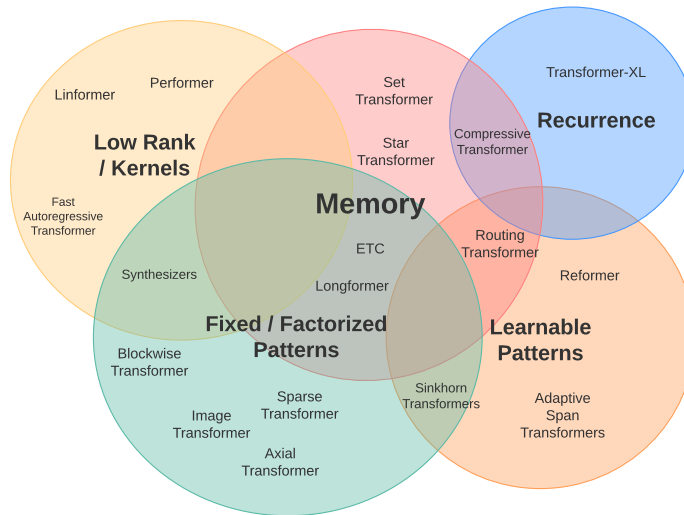


Figure 2.1 – A Venn diagram of different approaches for reducing the computational and memory complexity of attention as illustrated and detailed by [Tay, 2020].

ods, such as the Performer [Choromanski, 2020] and the Nyströmformer [Xiong, 2021b]. The Linformer [Wang, 2020c] reduce the quadratic complexity by projecting the attention matrix into a lower-dimensional space or by using random feature maps.

Moreover, Factorized Transformers such as the Transformer-XL [Dai, 2019] introduce recurrence into the self-attention mechanism, enabling the model to handle longer context than standard Transformers. Methods like Reversible Transformers, including Reformer [Kitaev, 2020], use reversible layers to reduce the memory requirements. This way, they can accommodate longer sequences during training. Finally, work on hybrid models, like the Conformer [Gulati, 2020], combines elements of Transformers with other neural architectures like convolutional neural networks, capturing local and global dependencies more efficiently. Overall, efficient Transformer architectures continue to be an active area of research. The improvements in computational and memory efficiency brought about by these methods open up possibilities for applying Transformers to larger and more complex datasets and tasks. For a more in-depth study of efficient attention variants the reader can refer to the survey by Tay *et al.* [Tay, 2020].

2.1.4 CNNs Augmented with Self-Attention

The success of the self-attention mechanism in various settings encourages researchers in computer vision to augment the popular CNN-based architectures with self-attention. This augmentation aims to capitalize on the strengths of both mechanisms: the strong locality bias of CNNs and the global context awareness of self-attention. Such an approach potentially enables more robust handling of long-range dependencies. An early influential work in this domain, Non-local Neural Networks Wang *et al.* [Wang, 2018] where the authors incorporated a non-local operation into existing CNN architectures. This operation, inspired by the self-attention mechanism of Transformers, computes a weighted sum of all the input feature maps' features at a given position. The non-local block is capable of capturing dependencies regardless of their

relative distance in the input data, significantly enhancing the model’s ability to handle complex spatial relationships. In the same spirit, Squeeze-and-Excitation Networks [Hu, 2018], proposed a channel-wise self-attention mechanism. This approach focuses on interdependencies between the channels of the convolutional layers, allowing for an adaptive recalibration of channel-wise feature responses. The resulting mechanism provides the model with the ability to emphasize informative features selectively.

The work of Ramachandran *et al.* [Ramachandran, 2019] explored using self-attention as a stand-alone layer, replacing all convolutions with a form of self-attention in a ResNet model. BoTNNets by Srinivas *et al.* [Srinivas, 2021] replaces the spatial convolutions of a ResNet by a novel bottleneck Transformer blocks which helps control the computational cost of self-attention. Meanwhile, Bello [Bello, 2021] introduced LambdaNetworks, a model that effectively replaces self-attention with lambda layers which provide a method of capturing long-range interactions without the associated computational costs of self-attention, making it more suitable for larger inputs. Carion *et al.* [Carion, 2020] proposed a novel approach to object detection, treating it as a direct set prediction problem. By leveraging self-attention and Transformers, this model coined DETR bypassed the need for several components traditionally used in object detection, such as non-maximum suppression and anchor boxes. Another interesting application of attention for semantic segmentation is Axial-Attention [Wang, 2020b] which applies self-attention sequentially in the height and width axes of the image. This mechanism effectively handles high-resolution images, demonstrating the advantages of self-attention in semantic segmentation tasks.

2.1.5 Vision Transformers

Vision Transformers (ViT) have become a paradigm-shifting development in computer vision, signifying a move away from traditional convolutional neural networks. Introduced by Dosovitskiy *et al.* [Dosovitskiy, 2021a], ViTs treat images as sequences of patches and apply Transformer-style self-attention to the task of image recognition. Several impactful variants of ViT soon followed. The DeiT work of [Touvron, 2020] focused on improving the data efficiency of ViT, managing to match its performance with fewer training images and setting the standard for optimization and regularization methods for training Vision Transformers (see also [Touvron, 2022b; Touvron, 2022a]). The Swin Transformer [Liu, 2021b], introduced a hierarchical structure with shifted windows to handle images of varying resolution, further bridging the gap between Transformers and CNNs. Pyramid Vision Transformer [Wang, 2021a] incorporated a pyramid structure to extract features at different scales, akin to CNNs, while maintaining the global self-attention mechanism of Transformers.

Subsequent iterations like CaiT [Touvron, 2021c] tackled the problem of making Vision Transformers deeper by proposing LayerScale for stabilizing training for models 48 layers deep. Additionally, Touvron *et al.* proposed a class attention final layer enabling the model to focus on class-related features. Graham *et al.* [Graham, 2021] introduces LeViT which offers an important step towards reducing the computational burden and model size associated with ViTs. LeViT provides favorable trade-offs in terms of throughput even when compared to popular backbones like EfficientNet. MViT [Fan, 2021] introduces an efficient method of spatial reduction which

helps control the computational complexity and enabling MViT to be applied to successfully to video recognition tasks.

The success of Vision Transformers has motivated innovative designs that further blend the advantages of Transformers and CNNs. For example, ConvNext [Liu, 2022b] revisits the design of CNNs by employing key design choices from ViTs like the columnar structure, replacing batch normalization by layer normalization and larger kernel sizes. ConvNext shows that CNNs can be competitive to vision transformers after integrating some of the modern design choices developed for ViTs. In summary, Vision Transformers and their successors have had a profound influence on the field of computer vision, inspiring a wide range of new model architectures that merge the advantages of CNNs and Transformers. This evolving landscape promises a wealth of potential for further exploration and innovation.

2.2 Self-Supervised Learning

Self-supervised learning has recently seen a lot of interest due to its ability to exploit the inherent structure of data to learn useful representations without the dependency on costly and domain-specific annotations. Self-supervised methods come in various shapes and forms, each with unique strengths and applications. In this section, we will explore some of the popular families of self-supervised learning methods including autoencoders, contrastive methods, non-contrastive methods, and clustering methods.

2.2.1 Autoencoders

Autoencoders are neural networks that are trained with the objective of reconstructing their input, typically with some form of constraint in order to learn useful features. They have been applied in a wide range of tasks, including image reconstruction, noise reduction, and representation learning. Denoising Autoencoders [Vincent, 2008] are designed to recover a clean input from a corrupted version, thereby learning robust representations of the data. Contractive Autoencoders [Rifai, 2011] enforce a form of robustness by adding a penalty term to the standard reconstruction error that constrains the derivatives of the encoder functions. The Context Encoder by Pathak *et al.* [Pathak, 2016] is a modification of the traditional autoencoder paradigm to perform inpainting of missing image regions. The autoencoding family of methods have found even more success in the context of Natural Language Processing. BERT [Devlin, 2018] marked a significant advancement by learning contextual relations between words in a text by training a bidirectional Transformer to predict missing (masked) words in a sentence. Subsequently, the highly popular GPT [Radford, 2018] model can be thought of as a Denoising Autoencoder that uses causal masking as the form of noise.

2.2.2 Pretext tasks

Early work in self-supervised learning relied on the idea of exploiting inherent spatial and temporal cues in images and videos. One such approach, Noroozi *et al.* [Noroozi, 2016] proposed

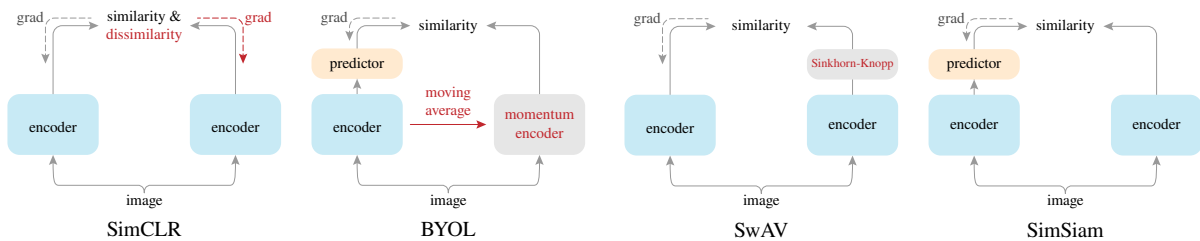


Figure 2.2 – An overview of popular joint embedding methods including contrastive, non-contrastive, and clustering based methods as illustrated by Chen *et al.* [Chen, 2021a]. Figure reproduced by permission of the authors.

the Jigsaw Puzzle, where permuted image patches are rearranged to their correct configuration, serving as a powerful source of supervisory signal for learning rich visual representations. The Context Prediction task by Doersch *et al.* [Doersch, 2015] leverages spatial context as a free and abundant supervisory signal where a CNN is trained to predict the relative positions of random pairs of patches extracted from each image. Meanwhile, Gidaris *et al.* [Gidaris, 2018] developed RotNet, which involves predicting the rotation of an image to learn visual representations by inferring the correct orientation of an object. Larsson *et al.* [Larsson, 2016] introduced a task of automatic colorization, which uses the task of filling in color in grayscale images to learn representations that capture both semantics and texture of the input images. Some other works made use of the inherent temporal cues in video such as [Wang, 2015] which proposed a self-supervised learning method that learns visual representations using videos. This approach exploits the temporal coherence and context in videos to learn robust visual representations.

2.2.3 Joint embeddings methods

A highly popular family of self-supervised approaches are joint embedding methods where an encoder is trained to be invariant to certain types of distortions and transformations. This can be achieved explicitly via contrastive or clustering methods or implicitly through non-contrastive methods that rely on applying regularization to the representations to avoid trivial solutions and collapse. An overview of notable examples of different joint embeddings methods is illustrated in Figure 2.2.

Contrastive Methods. Dosovitskiy *et al.* [Dosovitskiy, 2014] proposed instance discrimination where a model is tasked to classify an image correctly as itself under different transformations via an N -way classification objective where N corresponds to the number of unique images in the dataset. A more efficient the formulation is proposed in CPC [Oord, 2018] in the form of the InfoNCE objective. In the context of self-supervised learning, given a query q and a set of K keys consisting of one positive key k^+ (from the same sample as the query) and $K - 1$ negative keys k^- (from different samples), the InfoNCE loss can be defined as follows:

$$\mathcal{L}_{\text{NCE}} = -\mathbb{E} \left[\log \frac{\exp(f(q, k^+))}{\sum_{k \in k^+, k^-} \exp(f(q, k))} \right] \quad (2.5)$$

The InfoNCE loss encourages the model to assign higher similarity to positive pairs than to negative pairs. As a result, during the learning process, the model learns to map positive pairs closer together and negative pairs further apart in the embedding space. This property makes the InfoNCE loss particularly well-suited for contrastive learning tasks. CPC is primarily proposed for learning representations from audio, but InfoNCE was widely adopted by the majority of the subsequent contrastive approaches for vision tasks. Most notably, Chen *et al.* [Chen, 2020c] presented a simple yet effective method termed SimCLR, which significantly improved the state-of-the-art self-supervised learning performance. The SimCLR framework employs data augmentation as a key ingredient to generate positive pairs as well as identifying that using a large batch size, longer training and MLP head can significantly improve the performance of such methods. He *et al.* [He, 2020] introduced MoCo which constructs the dictionary of negative samples dynamically with a queue and a momentum-updated encoder. This approach allowed training with a large number of negative examples, which improved the quality of the learned representations.

Clustering Methods. Another family of effective self-supervised learning approaches are clustering-based methods. effective. DeepCluster [Caron, 2018] leverages clustering to enforce a pseudo-label based learning objective, thereby learning rich visual features in an unsupervised manner. Subsequently, Asano *et al.* [Asano, 2019b] introduced a method called SeLa to combine the process of clustering and representation learning in a single framework. The most notable of the clustering-based methods is SwAV [Caron, 2020a] which enhances the self-supervised learning process by promoting consistency between cluster assignments produced for different transformations of the same image.

Non-contrastive Methods. Non-contrastive methods have become increasingly popular because they sidestep the need for sampling a high number of negative pairs for every update. BYOL [Grill, 2020] was the first method to show a competitive performance can be achieved without relying on negative pairs. This was achieved by using an online and offline encoders where the weights of the offline encoder are updated via an exponential moving average. Combining this with a new formulation for the model head by adding a predictor, BYOL achieved a surprising result by outperforming many contrastive methods with no negative pairs and without collapse. SimSiam [Chen, 2021a] extended this idea, utilizing two identical networks that predict each other’s outputs, but with an added stop-gradient operation to prevent the collapse of the representations. Barlow Twins [Zbontar, 2021] proposed a method where the outputs of two identical networks are decorrelated, enforcing diversity in the learned representations and preventing collapse. VicReg [Bardes, 2021] introduced an objective that regularizes the representation’s variance along each dimension independently. Finally, Caron *et al.* [Caron, 2021] proposed DINO which is based on a self-distillation framework. The student is encouraged to match the output of a teacher that is updated with a exponential moving average. DINO was also the first method to show that self-supervised learning when used in conjunction with Vision Transformers can lead to outstanding results.

2.3 Vision-Language pre-training

In addition to learning representations through self-supervised learning, as discussed in the previous section, weakly supervised pre-training has been a highly successful paradigm in recent years. It leverages noisy labels, which are easier to obtain, as a substitute for expensive curated annotations. Early work by [Joulin, 2016] made use of image collections with paired textual captions to learn rich semantic representations. The full potential of this idea was realized through aggressive scaling of the image and text collection to millions [Radford, 2021] or even billions [Ilharco, 2021] of pairs. CLIP [Radford, 2021] and ALIGN [Jia, 2021a] use a simple contrastive objective to align matching visual and textual pairs. CLIP models can be very flexible as they learn about a wide range of visual concepts directly from natural language. CoCa [Yu, 2022] incorporates a text decoder and an additional captioning loss, which contributes to its strong performance in Visual Question Answering (VQA) and captioning tasks. Zhai *et al.* [Zhai, 2022] demonstrated that a strong pre-trained vision encoder can be easily augmented with open-set recognition capabilities by training a text encoder to align with the visual features resulting in performance often superior to training both encoders jointly from scratch. Alayrac *et al.* [Alayrac, 2022] explores interweaving images and text utterances by adapting a vision encoder to work jointly with a pre-trained Large Language Model yielding outstanding results in VQA and serving as a general-purpose visual chatbot.

2.4 Lossless Compression

The goal of lossless compression is to encode samples from a discrete probability distribution $x \sim p_d(x)$ $x \in \mathcal{X}$ as bit strings $m = \text{enc}(x) \in \{0, 1\}^*$ of shortest possible length $\ell(m)$ such that x can be decoded without loss of information $x = \text{dec}(m)$. Block codes assign a unique binary code of equal length to each event in \mathcal{X} . The expected length per symbol of such codes is $\ell(m) = \lceil \log_2 |\mathcal{X}| \rceil$ bits. We can improve this result if we consider to give shorter code words to more frequently occurring symbols and longer codes to less frequently occurring ones. In fact, the smallest expected achievable average code length per point using an approximation to the true data distribution $p(x)$, is given by the *cross-entropy*:

$$H(p_d, p) := \mathbb{E}_{x \sim p_d} [-\log p(x)]. \quad (2.6)$$

This quantity is bounded from below (seen via Jensen’s inequality) by the *entropy* of the “true” probability distribution p_d , i.e. when $p = p_d$ [MacKay, 2003; Cover, 1991]. Entropy coders reach this bound up to a small constant ϵ . An entropy codec is a tuple of an inverse pair of functions, enc_{p_X} and dec_{p_X} that achieve near optimal compression rates on sequences of symbols.

$$\begin{aligned} \text{enc}_p &: m, x \mapsto \hat{m} \\ \text{dec}_p &: \hat{m} \mapsto (m, x), \end{aligned} \quad (2.7)$$

such that $\ell(\hat{m}) = \ell(m) + \log_2 p(x) + \epsilon$.

Chapter 3

Cross Covariance Image Transformers

Objectives

Following tremendous success in natural language processing, transformers have recently shown much promise for computer vision. The self-attention operation underlying transformers yields global interactions between all tokens, *i.e.* words or image patches, and enables flexible modelling of image data beyond the local interactions of convolutions. This flexibility, however, comes with a quadratic complexity in time and memory, hindering application to long sequences and high-resolution images. We propose a “transposed” version of self-attention that operates across feature channels rather than tokens, where the interactions are based on the cross-covariance matrix between keys and queries. The resulting cross-covariance attention (XCA) has linear complexity in the number of tokens, and allows efficient processing of high-resolution images. Our cross-covariance image transformer (XCiT) – built upon XCA – combines the accuracy of conventional transformers with the scalability of convolutional architectures. We validate the effectiveness and generality of XCiT by reporting excellent results on multiple vision benchmarks, including (self-supervised) image classification on ImageNet-1k, object detection and instance segmentation on COCO, and semantic segmentation on ADE20k.

Contents

3.1	Introduction	17
3.2	Related work	19
3.3	Method	20
3.3.1	Background	20
3.3.2	Cross-covariance attention	22
3.3.3	Cross-covariance image transformers	24
3.4	Preliminary study on Vision Transformers (ViT)	26
3.4.1	Impact of resolution versus patch size	26
3.4.2	Approximate attention models in ViT with DeiT training	27
3.4.3	Training and testing with varying resolution	27
3.5	Experimental evaluation	28
3.5.1	Image classification	28
3.5.2	Object detection and instance segmentation	32
3.5.3	Semantic segmentation	34
3.5.4	Implementation details	35
3.6	Conclusion and Future Work	35

3.1 Introduction

Transformers architectures [Vaswani, 2017b] have provided quantitative and qualitative breakthroughs in speech and natural language processing. Recently, Dosovitskiy *et al.* [Dosovitskiy, 2021a] established transformers as a viable architecture for learning visual representations, reporting competitive results for image classification while relying on large-scale pre-training. Touvron *et al.* [Touvron, 2020] have shown on par or better accuracy/throughput compared to strong convolutional baselines such as EfficientNets [Tan, 2019] when training transformers on ImageNet-1k using extensive data augmentation and improved training schemes. Promising results have been obtained for other vision tasks, including image retrieval [El-Nouby, 2021b], object detection and semantic segmentation [Liu, 2021b; Wang, 2021a; Zhang, 2021; Zheng, 2020], as well as video understanding [Arnab, 2021; Bertasius, 2021; Fan, 2021].

One major drawback of transformers is the time and memory complexity of the core self-attention operation, that increases quadratically with the number of input tokens, or similarly number of patches in computer vision. For $w \times h$ images, this translates to a complexity of $\mathcal{O}(w^2h^2)$, which is prohibitive for most tasks involving high-resolution images, such as object detection and segmentation. Various strategies have been proposed to alleviate this complexity, for instance using approximate forms of self-attention [Liu, 2021b; Zhang, 2021], or pyramidal architectures which progressively downsample the feature maps [Wang, 2021a]. However, none of the existing solutions are fully satisfactory, as they either trade complexity for accuracy, or their complexity remains excessive for processing very large images.

We replace the self-attention, as originally introduced by Vaswani *et al.* [Vaswani, 2017b], with a “transposed” attention that we denote as “cross-covariance attention” (XCA). Cross-covariance attention substitutes the explicit full pairwise interaction between tokens by self-attention among features, where the attention map is derived from the cross-covariance matrix computed over the key and query projections of the token features. Importantly, XCA has a linear complexity in the number of patches. To construct our Cross-Covariance Image Transformers (XCiT), we combine XCA with local patch interaction modules that rely on efficient depth-wise convolutions and point-wise feedforward networks commonly used in transformers, see Figure 3.1. XCA can be regarded as a form of a dynamic 1×1 convolution, which multiplies all tokens with the same data-dependent weight matrix. We find that the performance of our XCA layer can be further improved by applying it on blocks of channels, rather than directly mixing all channels together. This “block-diagonal” shape of XCA further reduces the computational complexity with a factor linear in the number of blocks.

Given its linear complexity in the number of tokens, XCiT can efficiently process images with more than thousand pixels in each dimension. Notably, our experiments show that XCiT does not compromise the accuracy and achieves similar results to DeiT [Touvron, 2020] and CaiT [Touvron, 2021c] in comparable settings. Moreover, for dense prediction tasks such as object detection and image segmentation, our models outperform popular ResNet [He, 2016] backbones as well as the recent transformer-based models [Liu, 2021b; Wang, 2021a; Zhang, 2021]. Finally, we also successfully apply XCiT to the self-supervised feature learning using

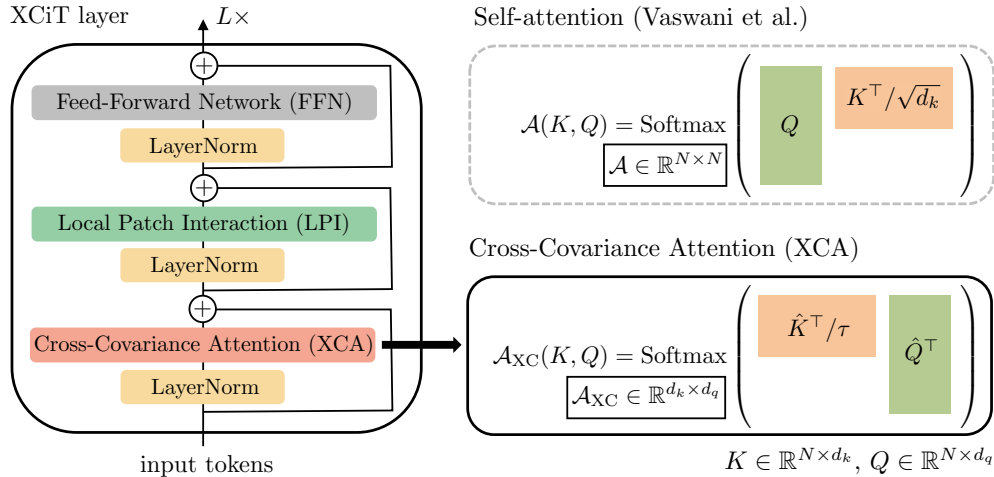


Figure 3.1 – Our XCiT layer consists of three main blocks, each preceded by LayerNorm and followed by a residual connection: (i) the core cross-covariance attention (XCA) operation, (ii) the local patch interaction (LPI) module, and (iii) a feed-forward network (FFN). By transposing the query-key interaction, the computational complexity of XCA is linear in the number of data elements N , rather than quadratic as in conventional self-attention.

DINO [Caron, 2021], and demonstrate improved performance compared to a DeiT-based backbone [Touvron, 2020].

Overall, we summarize our contributions as follows:

- We introduce cross-covariance attention (XCA), which provides a “transposed” alternative to conventional self-attention, attending over channels instead of tokens. Its complexity is linear in the number of tokens, allowing for efficient processing of high-resolution images, see Figure 3.2.
- XCA attends to a fixed number of channels, irrespective of the number of tokens. As a result, our models are significantly more robust to changes in image resolution at test time, and are therefore more amenable to process variable-size images.
- For image classification, we demonstrate that our models are on par with state-of-the-art vision transformers for multiple model sizes using a simple columnar architecture, *i.e.*, in which we keep the resolution constant across layers. In particular, our XCiT-L24 model achieves 86.0% top-1 accuracy on ImageNet, outperforming its CaiT-M24 [Touvron, 2021c] and NFNet-F2 [Brock, 2021] counterparts with comparable numbers of parameters.
- For dense prediction tasks with high-resolution images, our models outperform ResNet and multiple transformer-based backbones. On the COCO benchmark, we achieve a strong performance of 48.5% and 43.7% mAP for object detection and instance segmentation respectively. Moreover, we report 48.4% mIoU for semantic segmentation on the ADE20k benchmark, outperforming the state-of-the-art Swin Transformer [Liu, 2021b] backbones across all comparable model sizes.
- Finally, our XCiT model is highly effective in self-supervised learning setups, achieving 80.9% top-1 accuracy on ImageNet-1k using DINO [Caron, 2021].

3.2 Related work

Deep vision transformers. Training deep vision transformers can be challenging due to instabilities and optimization issues. Touvron *et al.* [Touvron, 2021c] successfully train models with up to 48 layers using LayerScale, which weighs contributions of residual blocks across layers and improves optimization. Additionally, the authors introduce class attention layers which decouple the learning of patch features and the feature aggregation stage for classification.

Spatial structure in vision transformers. Yuan *et al.* [Yuan, 2021b] propose applying a soft split for patch projection with overlapping patches which is applied repeatedly across model layers, reducing the number of patches progressively. Han *et al.* [Han, 2021] introduce a transformer module for intra-patch structure, exploiting pixel-level information and integrating with an inter-patch transformer to attain higher representation power. d’Ascoli *et al.* [dAscoli, 2021] consider the initialization of self-attention blocks as a convolutional operator, and demonstrate that such initialization improves the performance of vision transformers in low-data regimes. Graham *et al.* [Graham, 2021] introduce LeViT, which adopts a multi-stage architecture with progressively reduced feature resolution similar to popular convolutional architectures, allowing for models with high inference speed while retaining a strong performance. Moreover, the authors adopt a convolution-based module for extracting patch descriptors. Yuan *et al.* [Yuan, 2021a] improve both the performance and the convergence speed of vision transformers by replacing the linear patch projection with convolutional layers and max-pooling, as well as modifying the feed-forward networks in each transformer layer to incorporate depth-wise convolutions.

Efficient attention. Numerous methods for efficient self-attention have been proposed in the literature to address the quadratic complexity of self-attention in the number of input tokens. These include restricting the span of the self-attention to local windows [Parmar, 2018; Qiu, 2019], strided patterns [Child, 2019], axial patterns [Ho, 2019], or an adaptive computation across layers [Sukhbaatar, 2019]. Other methods provide an approximation of the self-attention matrix which can be achieved by a projection across the token dimension [Wang, 2020c], or through a factorization of the softmax-attention kernel [Choromanski, 2020; Katharopoulos, 2020; Shen, 2021; Xiong, 2021a], which avoids explicit computation of the attention matrix. While conceptually different, our XCA performs similar computations without being sensitive to the choice of the kernel. Similarly, Lee-Thorp *et al.* [Lee-Thorp, 2021] achieve faster training by substituting self-attention with unparametrized Fourier Transform. Other efficient attention methods rely on local attention and adding a small number of global tokens, thus allowing interaction among all tokens only by hopping through the global tokens [Ainslie, 2020; Beltagy, 2020; Jaegle, 2021; Zaheer, 2020].

Transformers for high-resolution images. Several works adopt visual transformers to high-resolution image tasks beyond image classification, such as object detection and image segmentation. Wang *et al.* [Wang, 2021a] design a model with a pyramidal architecture and address complexity by gradually reducing the spatial resolution of keys and values. Similarly, for

video recognition Fan *et al.* [Fan, 2021] utilize pooling to reduce the resolution across the spatial and temporal dimensions to allow for an efficient computation of the attention matrix. Zhang *et al.* [Zhang, 2021] adopt global tokens and local attention to reduce the model complexity, while Liu *et al.* [Liu, 2021b] provide an efficient method for local attention with shifted windows. In addition, Zheng *et al.* [Zheng, 2020] and Ranftl *et al.* [Ranftl, 2021] study problems like semantic segmentation and monocular depth estimation with the quadratic self-attention operation.

Data-dependent layers. Our XCiT layer can be regarded as a “dynamic” 1×1 convolution, which multiplies all token features with the same data-dependent weight matrix, derived from the key and query cross-covariance matrix. In the context of convolutional networks, Dynamic Filter Networks [Brabandere, 2016] explore a related idea, using a filter generating subnetwork to produce convolutional filters based on features in previous layers. Squeeze-and-Excitation networks [Hu, 2018] use data dependent 1×1 convolutions in convolutional architectures. Spatially average-pooled features are fed to a 2-layer MLP which produces per channel scaling parameters. Closer in spirit to our work, Lambda layers propose a way to ensure global interaction in ResNet models [Bello, 2021]. Their “content-based lambda function” is computing a similar term as our cross-covariance attention, but differing in how the softmax and ℓ_2 normalizations are applied. Moreover, Lambda layers also include specific position-based lambda functions, and LambdaNetworks are based on ResNets while XCiT follows the ViT architecture. Recently *data-independent* analogues of self-attention have also been found to be an effective alternative to convolutional and self-attention layers for vision tasks [Ding, 2021b; Melas-Kyriazi, 2021; Tolstikhin, 2021; Touvron, 2021a]. These methods treat entries in the attention map as learnable parameters, rather than deriving the attention map dynamically from queries and keys, but their complexity remains quadratic in the number of tokens. Zhao *et al.* [Zhao, 2020] consider alternative attention forms in computer vision.

3.3 Method

In this section, we first recall the self-attention mechanism, and the connection between the Gram and covariance matrices, which motivated our work. We then propose our cross-covariance attention operation (XCA) – which operates along the feature dimension instead of token dimension in conventional transformers – and combine it with local patch interaction and feedforward layers to construct our Cross-Covariance Image Transformer (XCiT). See Figure 3.1 for an overview.

3.3.1 Background

Token self-attention. Self-attention, as introduced by Vaswani *et al.* [Vaswani, 2017b], operates on an input matrix $X \in \mathbb{R}^{N \times d}$, where N is the number of tokens, each of dimensionality d . The input X is linearly projected to queries, keys and values, using the weight matrices $W_q \in \mathbb{R}^{d \times d_q}$, $W_k \in \mathbb{R}^{d \times d_k}$ and $W_v \in \mathbb{R}^{d \times d_v}$, such that $Q = XW_q$, $K = XW_k$ and $V = XW_v$, where $d_q = d_k$. Keys and values are used to compute an attention map $\mathcal{A}(K, Q) = \text{Softmax}(QK^\top / \sqrt{d_k})$,

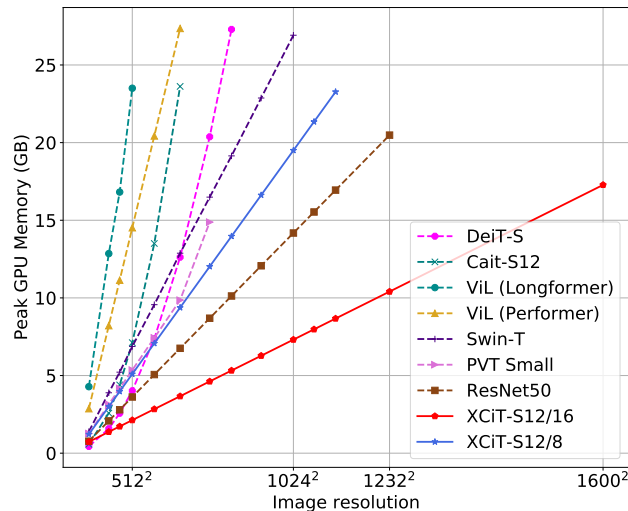


Figure 3.2 – Inference memory usage of vision transformer variants. Our XCiT models scale linearly in the number of tokens, which makes it possible to scale to much larger image sizes, even in comparison to approaches employing approximate self-attention or a pyramidal design. All measurements are performed with a batch size of 64 on a single V100-32GB GPU.

and the output of the self-attention operation is defined as the weighted sum of N token features in V with the weights corresponding to the attention map: $\text{Attention}(Q, K, V) = \mathcal{A}(K, Q)V$. The computational complexity of self-attention scales quadratically in N , due to pairwise interactions between all N elements.

Relationship between Gram and covariance matrices. To motivate our cross-covariance attention operation, we recall the relation between Gram and covariance matrices. The unnormalised $d \times d$ covariance matrix is obtained as $C = X^T X$. The $N \times N$ Gram matrix contains all pairwise inner products: $G = X X^T$. The non-zero part of the eigenspectrum of the Gram and covariance matrix are equivalent, and the eigenvectors of C and G can be computed in terms of each other. If V are the eigenvectors of G , then the eigenvectors of C are given by $U = X V$. To minimise the computational cost, the eigendecomposition of either the Gram or covariance matrix can be obtained in terms of the decomposition of the other, depending on which of the two matrices is the smallest.¹

We draw upon this strong connection between the Gram and covariance matrices to consider if it is possible to avoid the quadratic cost to compute the $N \times N$ attention matrix, which is computed from the analogue of the $N \times N$ Gram matrix $Q K^T = X W_q W_k^T X^T$. Below we consider how we can use the $d_k \times d_q$ cross-covariance matrix, $K^T Q = W_k^T X^T X W_q$, which can be computed in linear time in the number of elements N , to define an attention mechanism.

1. For C to represent the covariance, X should be centered, *i.e.* $X \mathbf{1} = \mathbf{0}$. For the relation between C and G , however, centering is not required.

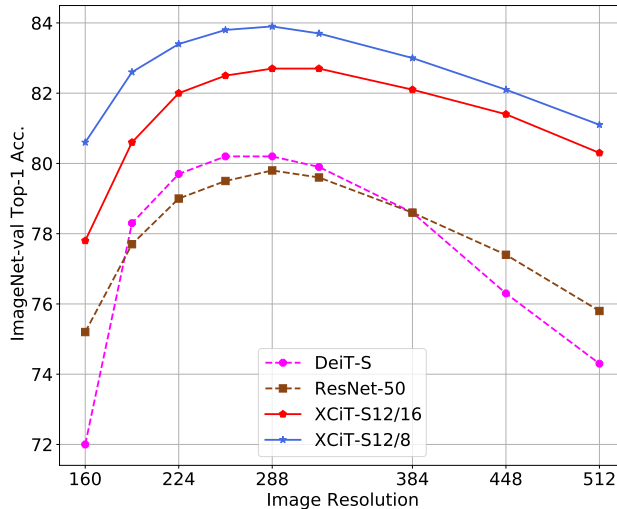


Figure 3.3 – Performance when changing the resolution at test-time for models with a similar number of parameters. All networks were trained at resolution 224, w/o distillation. XCiT is more tolerant to changes of resolution than the Gram-based DeiT and benefit more from the “FixRes” effect Touvron *et al.* [Touvron, 2019] when inference is performed at a larger resolution than at train-time.

3.3.2 Cross-covariance attention

We propose a cross-covariance based self-attention function that operates along the feature dimension, rather than along the token dimension as in token self-attention. Using the definitions of queries, keys and values from above, the cross-covariance attention function is defined as:

$$\text{XC-Attention}(Q, K, V) = V \mathcal{A}_{\text{XC}}(K, Q), \quad \mathcal{A}_{\text{XC}}(K, Q) = \text{Softmax}\left(\hat{K}^\top \hat{Q} / \tau\right), \quad (3.1)$$

where each output token embedding is a convex combination of the d_v features of its corresponding token embedding in V . The attention weights \mathcal{A} are computed based on the cross-covariance matrix.

ℓ_2 -Normalization and temperature scaling. In addition to building our attention operation on the cross-covariance matrix, we make a second modification compared to token self-attention. We restrict the magnitude of the query and key matrices by ℓ_2 -normalising them, such that each column of length N of the normalised matrices \hat{Q} and \hat{K} has unit norm, and every element in $d \times d$ cross-covariance matrix $\hat{K}^\top \hat{Q}$ is in the range $[-1, 1]$. We observed that controlling the norm strongly enhances the stability of training, especially when trained with a variable numbers of tokens. However, restricting the norm reduces the representational power of the operation by removing a degree of freedom. Therefore, we introduce a learnable temperature parameter τ which scales the inner products before the Softmax, allowing for sharper or more uniform distribution of attention weights.

Block-diagonal cross-covariance attention. Instead of allowing all features to interact among each other, we divide them into a h groups, or “heads”, in a similar fashion as multi-head token self-attention. We apply the cross-covariance attention separately per head where for each head, we learn separate weight matrices to project X to queries, keys and values, and collect the corresponding weight matrices in the tensors $W_q \in \mathbb{R}^{h \times d \times d_q}$, $W_k \in \mathbb{R}^{h \times d \times d_k}$ and $W_v \in \mathbb{R}^{h \times d \times d_v}$, where we set $d_k=d_q=d_v=d/h$. Restricting the attention within heads has two advantages: (i) the complexity of aggregating the values with the attention weights is reduced by a factor h ; (ii) more importantly, we empirically observe that the block-diagonal version is easier to optimize, and typically leads to improved results. This observation is in line with observations made for Group Normalization [Wu, 2018a], which normalizes groups of channels separately based on their statistics, and achieves favorable results for computer vision tasks compared to Layer Normalization [Ba, 2016], which combines all channels in a single group. Figure 3.6 shows that each head learns to focus on semantically coherent parts of the image, while being flexible to change what type of features it attends to based on the image content.

Complexity analysis. The usual token self-attention with h heads has a time complexity of $\mathcal{O}(N^2d)$ and memory complexity of $\mathcal{O}(hN^2+Nd)$. Due to the quadratic complexity, it is problematic to scale token self-attention to images with a large number of tokens. Our cross-covariance attention overcomes this drawback as its computational cost of $\mathcal{O}(Nd^2/h)$ scales linearly with the number of tokens, as does the memory complexity of $\mathcal{O}(d^2/h+Nd)$. Therefore, our model scales much better to cases where the number of tokens N is large, and the feature dimension d is relatively small, as is typically the case, in particularly when splitting the features into h heads.

Runtime and Memory Usage We present the peak memory usage as well as the throughput of multiple models including full-attention and efficient vision transformers in Table 3.1. Additionally, in Figure 3.4 we plot the processing speed represented as millisecond per image as a function of image resolution for various models. We can observe that XCiT provides a strong trade-off, possessing the best scalability in terms of peak memory, even when compared to ResNet-50. Additionally, the processing time scales linearly with respect to resolution, with only ResNet-50 providing a better trade-off on that front.

Model	#params ($\times 10^6$)	ImNet Top-1 @224	Image Resolution							
			224 ²		384 ²		512 ²		1024 ²	
			im/sec	mem (MB)	im/sec	mem (MB)	im/sec	mem (MB)	im/sec	mem (MB)
ResNet-50	25	79.0	1171	772	434	2078	245	3618	61	14178
DeiT-S	22	79.9	974	433	263	1580	116	4020	N/A	OOM
CaiT-S12	26	80.8	671	577	108	2581	38	7117	N/A	OOM
PVT-Small	25	79.8	777	1266	256	3142	134	5354	N/A	OOM
Swin-T	29	81.3	704	1386	220	3890	120	6873	29	26915
XCiT-S12/16	26	82.0	781	731	266	1372	151	2128	37	7312

Table 3.1 – **Inference throughput and peak GPU memory usage** for our XCiT small model compared to other models of comparable size that include token self-attention. All models tested using batch size of 64 on a V100 GPU with 32GB memory.

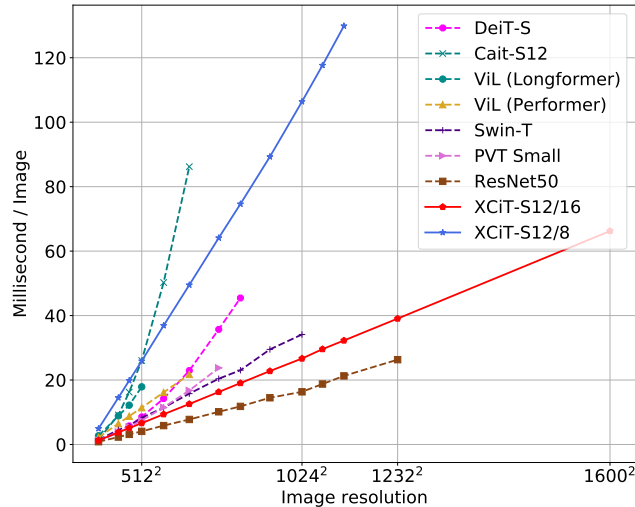


Figure 3.4 – We present the millisecond per image during inference of multiple models. Our XCiT-S12/16 model provides a speed up for images with higher resolution compared to existing vision transformers, especially the ones with quadratic complexity like DeiT and CaiT.

3.3.3 Cross-covariance image transformers

To construct our cross-covariance image transformers (XCiT), we adopt a columnar architecture which maintains the same spatial resolution across layers, similarly to [Dosovitskiy, 2021a; Touvron, 2020; Touvron, 2021c]. We combine our cross-covariance attention (XCA) block with the following additional modules, each one being preceded by a LayerNorm [Ba, 2016]. See Figure 3.1 for an overview. Since in this section we specifically design the model for computer vision tasks, tokens correspond to image patches in this context.

Local patch interaction. In the XCA block communication between patches is only implicit through the shared statistics. To enable explicit communication across patches we add a simple Local Patch Interaction (LPI) block after each XCA block. LPI consists of two depth-wise 3×3 convolutional layers with Batch Normalization and GELU non-linearity in between. Due to its depth-wise structure, the LPI block has a negligible overhead in terms of parameters, as well as a very limited overhead in terms of throughput and memory usage during inference.

Feed-forward network. As is common in transformer models, we add a point-wise feedforward network (FFN), which has a single hidden layer with $4d$ hidden units. While interaction between features is confined within groups in the XCA block, and no feature interaction takes place in the LPI block, the FFN allows for interaction across all features.

Global aggregation with class attention. When training our models for image classification, we utilize the class attention layers as proposed by Touvron *et al.* [Touvron, 2021c]. These layers aggregate the patch embeddings of the last XCiT layer through writing to a CLS token

Model	Depth	d	#heads	#params	GFLOPs		ImageNet-1k-val top-1 acc. (%)		
					@224/16	@384/8	@224/16	@224/16 Υ	@384/8 Υ \uparrow
XCiT-N12	12	128	4	3M	0.5	6.4	69.9	72.2	77.8
XCiT-T12	12	192	4	7M	1.2	14.3	77.1	78.6	82.4
XCiT-T24	24	192	4	12M	2.3	27.3	79.4	80.4	83.7
XCiT-S12	12	384	8	26M	4.8	55.6	82.0	83.3	85.1
XCiT-S24	24	384	8	48M	9.1	106.0	82.6	83.9	85.6
XCiT-M24	24	512	8	84M	16.2	188.0	82.7	84.3	85.8
XCiT-L24	24	768	16	189M	36.1	417.9	82.9	84.9	86.0

Table 3.2 – **XCiT models**. Design choices include model depth, patch embeddings dimensionality d , and the number of heads h used in XCA. By default our models are trained and tested at resolution 224 with patch sizes of 16×16 . We also train with distillation using a convolutional teacher (denoted Υ) as proposed by Touvron *et al.* [Touvron, 2020]. Finally, we report performance of our strongest models obtained with 8×8 patch size, fine-tuned (\uparrow) and tested at resolution 384×384 (column @384/8), using distillation with a teacher that was also fine-tuned @384.

by one-way attention between the CLS tokens and the patch embeddings. The class attention is also applied per head, *i.e.* feature group.

Handling images of varying resolution. In contrast to the attention map involved in token self-attention, in our case the covariance blocks are of fixed size independent of the input image resolution. The softmax always operates over the same number of elements, which may explain why our models behave better when dealing with images of varying resolutions (see Figure 3.3). In XCiT we include additive sinusoidal positional encoding [Vaswani, 2017b] with the input tokens. We generate them in 64 dimensions from the 2d patch coordinates and then linearly project to the transformer working dimension d . This choice is orthogonal to the use of learned positional encoding, as in ViT [Dosovitskiy, 2021a]. However, it is more flexible since there is no need to interpolate or fine-tune the network when changing the image size.

Model configurations. In Table 3.2 we list different variants of our model which we use in our experiments, with different choices for model width and depth. For the patch encoding layer, unless mentioned otherwise, we adopt the alternative used by Graham *et al.* [Graham, 2021] with convolutional patch projection layers. We also experimented with a linear patch projection as described in [Dosovitskiy, 2021a], see our ablation in Table 4.8. Our default patch size is 16×16 , as in other vision transformer models including ViT [Dosovitskiy, 2021a], DeiT [Touvron, 2020] and CaiT [Touvron, 2021c]. We also experiment with smaller 8×8 patches, which has been observed to improve performance [Caron, 2021]. Note that this is efficient with XCiT as its complexity scales linearly with the number of patches, while ViT, DeiT and CaiT scale quadratically.

Pseudo-code We provide a PyTorch-style pseudo code of the Cross-covariance attention operation. The pseudo code resembles the Timm library [Wightman, 2019] implementation of token self-attention. We show that XCA only requires few modifications, namely the ℓ_2 normalization, setting the learnable temperature parameters and a transpose operation of the keys, queries and

values.

Algorithm 3.1: Pseudocode of XCA in a PyTorch-like style.

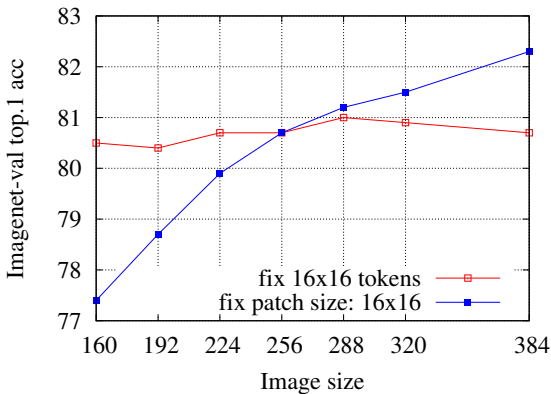
```

# self.qkv: nn.Linear(dim, dim * 3, bias=qkv_bias) # self.temp: nn.Parameter(torch.ones(num_headss, 1, 1))
def forward(self, x): B, N, C = x.shape qkv = self.qkv(x).reshape(B, N, 3, self.num_heads, C // self.
    num_heads) qkv =
    qkv.permute(2, 0, 3, 1, 4) q, k, v = qkv[0], qkv[1], qkv[2] # split into query, key and value
    q = q.transpose(-2, -1) k = k.transpose(-2, -1) # Transpose to shape (B, h, C, N) v = v.transpose(-2,
    -1)
    q = F.normalize(q, dim=-1, p=2) # L2 Normalization across the token dimension k = F.normalize(k, dim
    =-1, p=2)
    attn = (k @ q.transpose(-2, -1)) # Computing the block diagonal cross-covariance matrix attn = attn *
    self.temp # Adjusting the activations scale with temperature parameter attn = attn.softmax(dim=-1) #
    d x d attention map
    x = attn @ v # Apply attention to mix channels per token x = x.permute(0, 3, 1, 2).reshape(B, N, C)
    x = self.proj(x) return x
  
```

3.4 Preliminary study on Vision Transformers (ViT)

In this section, we report the results associated with our preliminary study on high-resolution transformers. Most of the experiments were carried out on the ViT architecture [Dosovitskiy, 2021a] with DeiT training [Touvron, 2020], and intended to analyze different aspects of transformers when considering images with varying resolution or high-resolution images specifically.

3.4.1 Impact of resolution versus patch size



Variable patch size						
Image Size	80	112	160	256	320	384
Patch Size	5	7	10	16	20	24
Top-1	78.2	79.7	80.5	80.7	80.9	80.7

Variable number of tokens size						
Image Size	160	224	256	288	320	384
# of tokens	100	196	256	324	400	576
Top-1	77.4	79.9	80.7	81.2	81.5	82.3

Figure 3.5 – **Impact of input resolution on accuracy for DeiT-S.** We consider different image resolutions, and either (1) **increase the patch size while keeping the number of tokens fixed**; or (2) **keep the patch size fixed and use more tokens**. Larger input images are beneficial if the number of tokens increases. The impact of a change of a resolution for a constant number of patches (of varying size) is almost neutral. As one can observe, the main driver of performance is the number of patches. The patch size has a limited impact on the accuracy, except when considering very small ones. We have observed and confirmed similar trends with XCiT models.

3.4.2 Approximate attention models in ViT with DeiT training

In Table 3.3, we report the results that we obtain by replacing the Multi-headed Self-attention operation with efficient variants [Ho, 2019; Shen, 2021; Wang, 2020c; Wang, 2021a] in the DeiT-S backbone. First, we can notice that for all efficient self-attention choices there is a clear drop in performance compared to the DeiT-S baseline. The spatial reduction attention (SRA) proposed in PVT [Wang, 2021a] has a significantly weaker performance compared to the full-attention with a quadratic complexity that is more efficient than full-attention by only a constant factor R^2 . Linformer [Wang, 2020c] provides a better accuracy compared to SRA, however, it is also clearly weaker than full-attention. Moreover, Linformer does not have the flexibility of processing variable length sequences which limits its application in many computer vision tasks. Efficient attention [Shen, 2021] provides a better trade-off than the aforementioned methods, with improved accuracy and linear complexity. However, it has a 3.6% drop in performance compared to full-attention. Finally, axial attention [Ho, 2019] provides the strongest performance among the efficient attention variants we studied with a 1.5% drop in accuracy compared to the baseline. We observe a saving in memory usage, but a drop in speed due to the separate row and column attention operations. Our observations are consistent with [Dosovitskiy, 2021a].

Model	Complexity	Top-1
DeiT-S [Touvron, 2020]	$\mathcal{O}(N^2)$	79.9
SRA (Average Pool) [Wang, 2021a]	$\mathcal{O}(N^2/R^2)$	73.5
SRA (Convolutional) [Wang, 2021a]	$\mathcal{O}(N^2/R^2)$	74.0
Linformer ($k=\sqrt{n}$) [Wang, 2020c]	$\mathcal{O}(kN)$	75.7
Efficient Transformer [Shen, 2021]	$\mathcal{O}(N)$	76.3
Axial [Ho, 2019]	$\mathcal{O}(N\sqrt{N})$	78.4

Table 3.3 – **ImageNet Top-1 accuracy of efficient self-attention variants** (after 300 epochs of training).

3.4.3 Training and testing with varying resolution

For several tasks, it is important that the network is able to handle images of varying resolutions. This is the case, for instance, for image segmentation, image detection, or image retrieval where the object of interest may have very different sizes. We present an analysis of train/test resolution trade-off in Table 3.4.

Test / Train	160	224	256	288	320	MS
160	77.2	75.9	73.3	68.2	59.6	76.3
224	78.0	79.9	79.9	79.0	77.9	79.6
256	77.3	80.4	80.7	80.2	79.9	80.6
288	76.3	80.4	81.0	81.2	80.8	81.0
320	75.0	80.1	80.9	81.3	81.5	81.3

Table 3.4 – **Trade-off between train and test resolutions for DeiT**. MS refers to multi-scale training, where the models have seen images from different resolutions at training time.

3.5 Experimental evaluation

In this section we demonstrate the effectiveness and versatility of XCiT on multiple computer vision benchmarks, and present ablations providing insight on the importance of its different components.

Model	#params	FLOPs	Res.	ImNet	V2
EfficientNet-B5 RA [Cubuk, 2020]	30M	9.9B	456	83.7	—
RegNetY-4GF [Radosavovic, 2020]	21M	4.0B	224	80.0	72.4
DeiT-SY [Touvron, 2020]	22M	4.6B	224	81.2	68.5
Swin-T [Liu, 2021b]	29M	4.5B	224	81.3	—
CaiT-XS24Y ↑ [Touvron, 2021c]	26M	19.3B	384	84.1	74.1
XCiT-S12/16Y	26M	4.8B	224	83.3	72.5
XCiT-S12/16Y ↑	26M	14.3B	384	84.7	74.1
XCiT-S12/8Y ↑	26M	55.6B	384	85.1	74.8
EfficientNet-B7 RA [Cubuk, 2020]	66M	37.0B	600	84.7	—
NFNet-F0 [Brock, 2021]	72M	12.4B	256	83.6	72.6
RegNetY-8GF [Radosavovic, 2020]	39M	8.0B	224	81.7	72.4
Swin-S [Liu, 2021b]	50M	8.7B	224	83.0	—
CaiT-S24Y ↑ [Touvron, 2021c]	47M	32.2B	384	85.1	75.4
XCiT-S24/16Y	48M	9.1B	224	83.9	73.3
XCiT-S24/16Y ↑	48M	26.9B	384	85.1	74.6
XCiT-S24/8Y ↑	48M	105.9B	384	85.6	75.7
RegNetY-16GF [Radosavovic, 2020]	84M	16.0B	224	82.9	72.4
Swin-B ↑ [Liu, 2021b]	88M	47.0B	384	84.2	—
DeiT-BY ↑ [Touvron, 2020]	87M	55.5B	384	85.2	75.2
CaiT-S48Y ↑ [Touvron, 2021c]	89M	63.8B	384	85.3	76.2
XCiT-M24/16Y	84M	16.2B	224	84.3	73.6
XCiT-M24/16Y ↑	84M	47.7B	384	85.4	75.1
XCiT-M24/8Y ↑	84M	187.9B	384	85.8	76.1
NFNet-F3 [Brock, 2021]	255M	114.8B	416	85.7	75.2
CaiT-M24Y ↑ [Touvron, 2021c]	186M	116.1B	384	85.8	76.1
XCiT-L24/16Y	189M	36.1B	224	84.9	74.6
XCiT-L24/16Y ↑	189M	106.0B	384	85.8	75.8
XCiT-L24/8Y ↑	189M	417.8B	384	86.0	76.6

Table 3.5 – **ImageNet classification.** Number of parameters, FLOPs, image resolution, and top-1 accuracy on ImageNet-1k and ImageNet-V2. Training strategies vary across models, transformer-based models and the reported RegNet mostly follow recipes from DeiT [Touvron, 2020].

3.5.1 Image classification

We use ImageNet-1k [Deng, 2009b] to train and evaluate our models for image classification. It consists of 1.28M training images and 50k validation images, labeled across 1,000 semantic categories. Our training setup follows the DeiT recipe [Touvron, 2020]. We train our model for 400 epochs with the AdamW optimizer [Loshchilov, 2017] using a cosine learning rate decay. In order to enhance the training of larger models, we utilize LayerScale [Touvron, 2021c] and adjust the stochastic depth [Huang, 2016] for each of our models accordingly. Following [Touvron, 2021c], images are cropped with crop ratio of 1.0 for evaluation. In addition to the ImageNet-1k validation set, we report results for ImageNet-V2 [Recht, 2019] which has a distinct test set. Our implementation is based on the Timm library [Wightman, 2019].

Architecture	CIFAR ₁₀	CIFAR ₁₀₀	Flowers102	Cars	iNat ₁₈	iNat ₁₉
EfficientNet-B7 [Tan, 2019]	<u>98.9</u>	91.7	98.8	94.7	—	—
ViT-B/16 [Dosovitskiy, 2021a]	98.1	87.1	89.5	—	—	—
ViT-L/16 [Dosovitskiy, 2021a]	97.9	86.4	89.7	—	—	—
DeiT-B/16 [Touvron, 2020] Υ	99.1	91.3	98.8	92.9	<u>73.7</u>	<u>78.4</u>
XCiT-S24/16 Υ	99.1	91.2	97.4	92.8	68.8	76.1
XCiT-M24/16 Υ	99.1	<u>91.4</u>	98.2	93.4	72.6	78.1
XCiT-L24/16 Υ	99.1	91.3	<u>98.3</u>	<u>93.7</u>	75.6	79.3

Table 3.6 – Evaluation on transfer learning.

Results on ImageNet. We present a family of seven models in Table 3.2 with different operating points in terms of parameters and FLOPs. We observe that the performance of the XCiT models benefits from increased capacity both in depth and width. Additionally, consistent with [Touvron, 2020; Touvron, 2021c] we find that using hard distillation with a convolutional teacher improves the performance. Because of its linear complexity in the number of tokens, it is feasible to train XCiT at 384×384 resolution with small 8×8 patches, *i.e.* 2304 tokens, which provides a strong boost in performance across all configurations.

We compare to the state-of-the-art convolutional and transformer-based architectures [Brock, 2021; Liu, 2021b; Radosavovic, 2020; Tan, 2019; Touvron, 2021c] in Table 3.5. By varying the input image resolution and/or patch size, our models provide competitive or superior performance across model sizes and FLOP budgets. First, the models operating on 224×224 and 16×16 (*e.g.* XCiT-S12/16) enjoy high accuracy at relatively few FLOPs compared to their counterparts with comparable parameter count and FLOPs. Second, our models with 16×16 and 384×384 resolution images (*e.g.* XCiT-S12/16 \uparrow) yield an improved accuracy at the expense of higher FLOPs, and provide superior or on-par performance compared to state-of-the-art models with comparable computational requirements. Finally, XCiT linear complexity allows us to scale to process 384×384 images with 8×8 patch sizes (*e.g.* XCiT-S12/8 \uparrow), achieving the highest accuracy across the board, albeit at a relatively high FLOPs count.

Transfer learning In order to further demonstrate the flexibility and generality of our models, we report transfer learning experiments in Table 3.6 for models that have been pre-trained using ImageNet-1k and finetuned for other datasets including CIFAR-10, CIFAR-100 [Krizhevsky, 2009], Flowers-102 [Nilsback, 2008], Stanford Cars [Krause, 2013] and iNaturalist [Horn, 2017]. We observe that the XCiT models provide competitive performance when compared to strong baselines like ViT-B, ViT-L, DeiT-B and EfficientNet-B7.

Class attention visualization. In Figure 3.6 we show the class attention map obtained in the feature aggregation stage. Each head focuses on different semantically coherent regions in the image (*e.g.* faces or umbrellas). Furthermore, heads tend to focus on similar patterns across images (*e.g.* bird head or human face), but adapts by focusing on other salient regions when such patterns are absent.



Figure 3.6 – Visualization of the attention map between the CLS token and individual patches in the class-attention stage. For each column, each row represents the attention map w.r.t. one head, corresponding to the image in the first row. Each head seems salient to semantically coherent regions. Heads are sensitive to similar features within the same or across images (*e.g.* people or bird faces). They trigger on different concepts when such features are missing (*e.g.*, cockpit for race cars).

Queries and Keys magnitude visualizations Our XCA operation relies on the cross-covariance matrix of the queries \hat{Q} and keys \hat{K} which are ℓ_2 normalized across the patch dimension. Therefore, each element in the $d \times d$ matrix represents a cosine similarity whose value is strongly influenced by the magnitude of each patch. In Figure 3.7 we visualize the magnitude of patch embeddings in the queries and keys matrices. We observe that patch embeddings with higher magnitude corresponds to more salient regions in the image, providing a very cheap visualization and interpretation of which regions in the image contribute more in the cross-covariance attention.

Robustness to resolution changes. In Figure 3.3 we report the accuracy of XCiT-S12, DeiT-S and ResNet-50 trained on 224×224 images and evaluated at different image resolutions. While DeiT outperforms ResNet-50 when train and test resolutions are similar, it suffers from a larger drop in performance as the image resolution deviates farther from the training resolution. XCiT displays a substantially increased accuracy when train and test resolutions are similar, while also being robust to resolution changes, in particular for the model with 8×8 patches.

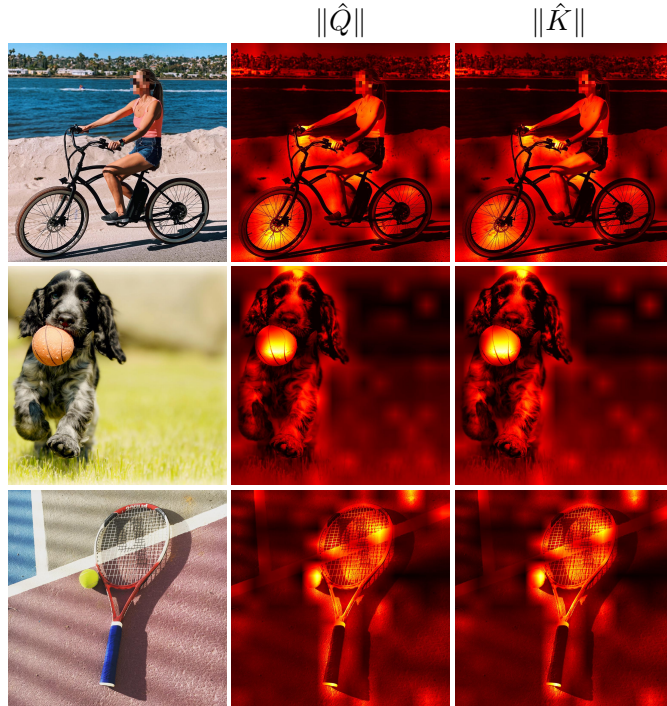


Figure 3.7 – **Visualization of the queries \hat{Q} and keys \hat{K} norm across the feature dimension.** We empirically observe that magnitude of patch embeddings in the queries and keys correlates with the saliency of their corresponding region in the image.

SSL Method	Model	#params	FLOPs	Linear	k -NN
MoBY [Xie, 2021]	Swin-T [Liu, 2021b]	29M	4.5B	75.0	–
DINO [Caron, 2021]	ResNet-50 [He, 2016]	23M	4.1B	74.5	65.6
DINO [Caron, 2021]	ViT-S/16 [Dosovitskiy, 2021a]	22M	4.6B	76.1	72.8
DINO [Caron, 2021]	ViT-S/8 [Dosovitskiy, 2021a]	22M	22.4B	79.2	77.2
DINO [Caron, 2021]	XCiT-S12/16	26M	4.9B	77.8	76.0
DINO [Caron, 2021]	XCiT-S12/8	26M	18.9B	79.2	77.1
DINO [Caron, 2021]	ViT-B/16 [Dosovitskiy, 2021a]	87M	17.5B	78.2	76.1
DINO [Caron, 2021]	ViT-B/8 [Dosovitskiy, 2021a]	87M	78.2B	80.1	77.4
DINO [Caron, 2021]	XCiT-M24/16	84M	16.2B	78.8	76.4
DINO [Caron, 2021]	XCiT-M24/8	84M	64.0B	80.3	77.9
DINO [Caron, 2021]	XCiT-M24/8 \uparrow 384	84M	188.0B	80.9	-

Table 3.7 – **Self-supervised learning.** Top-1 acc. on ImageNet-1k. We report with a crop-ratio 0.875 for consistency with DINO. For the last row it is set to 1.0 (improves from 80.7% to 80.9%). All models are trained for 300 epochs.

Self-supervised learning. We train XCiT in a self-supervised manner using DINO [Caron, 2021] on ImageNet-1k. In Table 3.7 we report performance using the linear and k -NN protocols as in [Caron, 2021]. Across model sizes XCiT obtains excellent accuracy with both protocols, substantially improving DINO with ResNet-50 or ViT architectures, as well as over those reported for Swin-Transformer trained with MoBY [Xie, 2021]. Comparing the larger models to ViT, we also observed improved performance for XCiT achieving a strong 80.3% accuracy. For fair comparison, all reported models have been trained for 300 epochs. Further improved performance of small models is reported by Caron *et al.* [Caron, 2021] when training for 800 epochs,

Model	Ablation	ImNet top-1 acc.
XCiT-S12/16	Baseline	82.0
XCiT-S12/8		83.4
XCiT-S12/16	Linear patch proj.	81.1
XCiT-S12/8		83.1
XCiT-S12/16	w/o LPI layer	80.8
	w/o XCA layer	75.9
XCiT-S12/16	w/o ℓ_2 -normal.	failed
	w/o learned temp. τ	81.8

Table 3.8 – **Ablations** of various architectural design choices on the task of ImageNet-1k classification using the XCiT-S12 model. Our baseline model uses the convolutional projection adopted from LeViT.

which we expect to carryover to XCiT based on the results presented here.

Analysis and ablations. In Table 3.8 we provide ablation experiments to analyse the impact of different design choices for our XCiT-S12 model. First, we observe the positive effect of using the convolutional patch projection as compared to using linear patch projection, for both 8×8 and 16×16 patches. Second, while removing the LPI layer reduces the accuracy by only 1.2% (from 82.0 to 80.8), removing the XCA layer results in a large drop of 6.1%, underlining the effectiveness of XCA. We noticed that the inclusion of two convolutional components – convolutional patch projection and LPI – not only brings improvements in accuracy, but also accelerates training. Third, although we were able to ensure proper convergence without ℓ_2 -normalization of queries and keys by tweaking the hyper-parameters, we found that it provides stability across model size (depth and width) and other hyper-parameters. Finally, while the learnable softmax temperature parameter is not critical, removing it drops accuracy by 0.2%.

3.5.2 Object detection and instance segmentation

Our XCiT models can efficiently process high-resolution images (see Figure 3.2). Additionally, XCiT has a better adaptability to varying image resolutions compared to ViT models (see Figure 3.3). These two properties make XCiT a good fit for dense prediction tasks including detection and segmentation.

We evaluate XCiT for object detection and instance segmentation using the COCO benchmark [Lin, 2014] which consists of 118k training and 5k validation images including bounding boxes and mask labels for 80 categories. We integrate XCiT as backbone in the Mask R-CNN [He, 2017] detector with FPN [Lin, 2017]. Since the XCiT architecture is inherently columnar, we make it FPN-compatible by extracting features from different layers (*e.g.*, [4, 6, 8, 12] for XCiT-S12). All features have a constant stride of 8 or 16 based on the patch size, and the feature resolutions are adjusted to have strides of [4, 8, 16, 32], similar to ResNet-FPN backbones, where the downsampling is achieved by max pooling and the upsampling is obtained using a single transposed convolution layer. The model is trained for 36 epochs (3x schedule) using the AdamW optimizer with learning rate of 10^{-4} , 0.05 weight decay and 16 batch size. We adopt the multiscale training and augmentation strategy of DETR [Carion, 2020]. Our implementation is

Backbone	#params	AP ^b	AP ₅₀ ^b	AP ₇₅ ^b	AP ^m	AP ₅₀ ^m	AP ₇₅ ^m
ResNet18 [He, 2016]	31.2M	36.9	57.1	40.0	33.6	53.9	35.7
PVT-Tiny [Wang, 2021a]	32.9M	39.8	62.2	43.0	37.4	59.3	39.9
ViL-Tiny [Zhang, 2021]	26.9M	41.2	64.0	44.7	37.9	59.8	40.6
XCiT-T12/16	26.1M	42.7	64.3	46.4	38.5	61.2	41.1
XCiT-T12/8	25.8M	44.5	66.4	48.8	40.3	63.5	43.2
ResNet50 [He, 2016]	44.2M	41.0	61.7	44.9	37.1	58.4	40.1
PVT-Small [Wang, 2021a]	44.1M	43.0	65.3	46.9	39.9	62.5	42.8
ViL-Small [Zhang, 2021]	45.0M	43.4	64.9	47.0	39.6	62.1	42.4
Swin-T [Liu, 2021b]	47.8M	46.0	68.1	50.3	41.6	65.1	44.9
XCiT-S12/16	44.3M	45.3	67.0	49.5	40.8	64.0	43.8
XCiT-S12/8	43.1M	47.0	68.9	51.7	42.3	66.0	45.4
ResNet101 [He, 2016]	63.2M	42.8	63.2	47.1	38.5	60.1	41.3
ResNeXt101-32	62.8M	44.0	64.4	48.0	39.2	61.4	41.9
PVT-Medium [Wang, 2021a]	63.9M	44.2	66.0	48.2	40.5	63.1	43.5
ViL-Medium [Zhang, 2021]	60.1M	44.6	66.3	48.5	40.7	63.8	43.7
Swin-S [Liu, 2021b]	69.1M	48.5	70.2	53.5	43.3	67.3	46.6
XCiT-S24/16	65.8M	46.5	68.0	50.9	41.8	65.2	45.0
XCiT-S24/8	64.5M	48.1	69.5	53.0	43.0	66.5	46.1
ResNeXt101-64 [Xie, 2017]	101.9M	44.4	64.9	48.8	39.7	61.9	42.6
PVT-Large [Wang, 2021a]	81.0M	44.5	66.0	48.3	40.7	63.4	43.7
ViL-Large [Zhang, 2021]	76.1M	45.7	67.2	49.9	41.3	64.4	44.5
XCiT-M24/16	101.1M	46.7	68.2	51.1	42.0	65.6	44.9
XCiT-M24/8	98.9M	48.5	70.3	53.4	43.7	67.5	46.9

Table 3.9 – **COCO object detection and instance segmentation** performance on the mini-val set. All backbones are pre-trained on ImageNet-1k, use Mask R-CNN model [He, 2017] and are trained with the same 3x schedule.

based on the mmdetection library [Chen, 2019].

Results on COCO. In Table 3.9 we report object detection and instance segmentation results of four variants of XCiT using 16×16 and 8×8 patches. We compare to ResNets [He, 2016] and concurrent efficient vision transformers [Liu, 2021b; Wang, 2021a; Zhang, 2021]. All models are trained using the 3x schedule after ImageNet-1k pre-training. Note that other results with higher absolute numbers have been achieved when pre-training on larger datasets [Liu, 2021b] or with longer schedules [Bello, 2021], and are therefore not directly comparable to the reported results. First, across all model sizes XCiT outperforms the convolutional ResNet [He, 2016] and ResNeXt [Xie, 2017] by a large margin with either patch size. Second, we observe a similar increase in accuracy compared to PVT [Wang, 2021a] and ViL [Zhang, 2021] backbones. Finally, XCiT provides a competitive performance with Swin [Liu, 2021b]². For relatively small models, XCiT-S12/8 outperforms its Swin-T counterpart with a decent margin. On the other hand, Swin-S provides slightly stronger results compared to XCiT-S24/8. Utilizing smaller 8×8 patches leads to a consistent gain across all models.

². We use report the results provided by the authors in their open-sourced code <https://github.com/SwinTransformer/Swin-Transformer-Object-Detection>

Backbone	Semantic FPN		UperNet	
	#params	mIoU	#params	mIoU
ResNet18 [He, 2016]	15.5M	32.9	-	-
PVT-Tiny [Wang, 2021a]	17.0M	35.7M	-	-
XCiT-T12/16	8.4M	38.1	33.7M	41.5
XCiT-T12/8	8.4M	39.9	33.7	43.5
ResNet50 [He, 2016]	28.5M	36.7	66.5M	42.0
PVT-Small [Wang, 2021a]	28.2M	39.8	-	-
Swin-T [Liu, 2021b]	-	-	59.9M	44.5
XCiT-S12/16	30.4M	43.9	52.4M	45.9
XCiT-S12/8	30.4M	44.2	52.3M	46.6
ResNet101 [He, 2016]	47.5M	38.8	85.5M	43.8
ResNeXt101-32 [Xie, 2017]	47.1M	39.7	-	-
PVT-Medium [Wang, 2021a]	48.0M	41.6	-	-
Swin-S [Liu, 2021b]	-	-	81.0M	47.6
XCiT-S24/16	51.8M	44.6	73.8M	46.9
XCiT-S24/8	51.8M	47.1	73.8M	48.1
ResNeXt101-64 [Xie, 2017]	86.4M	40.2	-	-
PVT-Large [Wang, 2021a]	65.1M	42.1	-	-
Swin-B [Liu, 2021b]	-	-	121.0M	48.1
XCiT-M24/16	90.8M	45.9	109.0M	47.6
XCiT-M24/8	90.8M	46.9	108.9M	48.4

Table 3.10 – **ADE20k semantic segmentation** performance using Semantic FPN [Kirillov, 2019] and UperNet [Xiao, 2018] (in comparable settings). We do not include comparisons with other state-of-the-art models that are pre-trained on larger datasets [Liu, 2021b; Ranftl, 2021; Zheng, 2020].

3.5.3 Semantic segmentation

We further show transferability of our models with semantic segmentation experiments on the ADE20k dataset [Zhou, 2017], which consists of 20k training and 5k validation images with labels over 150 semantic categories. We integrate our backbones in two segmentation methods: Semantic FPN [Kirillov, 2019] and UperNet [Xiao, 2018]. We train for 80k and 160k iterations for Semantic FPN and UperNet respectively. Following [Liu, 2021b], the models are trained using batch size 16 and an AdamW optimizer with learning rate of 6×10^{-5} and 0.01 weight decay. We apply the same method of extracting FPN features as explained in Section 3.5.2. We report the performance using the standard single scale protocol (without multi-scale and flipping). Our implementation is based on the mmsegmentation library [Contributors, 2020].

Results on ADE20k. We present the semantic segmentation performance using XCiT backbones in Table 3.10. First, for Semantic FPN [Kirillov, 2019], XCiT provides a superior performance compared to ResNet, ResNeXt and PVT backbones using either option of patch size. Second, compared to Swin Transformers using the same UperNet decoder [Xiao, 2018], XCiT with 8×8 patches consistently achieves a higher mIoU for different models. XCiT with 16×16 patches provides a strong performance especially for smaller models where XCiT-S12/16 outperforms Swin-T.

3.5.4 Implementation details

Sinusoidal Positional Encoding We adopt a sinusoidal positional encoding as proposed by Vaswani *et al.* [Vaswani, 2017b] and adapted to the 2D case by Carion *et al.* [Carion, 2020]. However we depart from this method in that we first produce this encoding in an intermediate 64-d space before projecting it to the working space of the transformers. More precisely, in our implementation each of the x and y coordinates is encoded using 32 dimensions corresponding to cosine and sine functions with different frequencies (16 frequency for each function). The encoding of both coordinates are eventually concatenated to obtain a 64 dimension 2D positional encoding. Finally, the 64 dimension positional encoding is linearly projected to the working dimension of the model d .

Obtaining Feature Pyramid for Dense Prediction For state-of-the-art detection and segmentation models, FPN is an important component which provides features of multiple scales. We adapt XCiT to be compatible with FPN detection and segmentation methods through a simple re-scaling of the features extracted from different layers. In particular, for models with 12 layers, we extract features from the 4th, 6th, 8th and 12th layers respectively. As for models with 24 layers, we extract features from 8th, 12th, 16th and 24th layers. Concerning the re-scaling of the features, the 4 feature levels are downsized by a ratio of 4, 8, 16 and 32 compared to the input image size. Feature downsizing is performed with max pooling and upsampling is achieved using a single layer of transposed convolutions with kernel size $k = 2$ and stride $s = 2$.

3.6 Conclusion and Future Work

Contributions. We present an alternative to token self-attention which operates on the feature dimension, eliminating the need for expensive computation of quadratic attention maps. We build our XCiT models with the cross-covariance attention as its core component and demonstrate the effectiveness and generality of our models on various computer vision tasks. In particular, it exhibits a strong image classification performance on par with state-of-the-art transformer models while similarly robust to changing image resolutions as convnets. XCiT is effective as a backbone for dense prediction tasks, providing excellent performance on object detection, instance and semantic segmentation. Finally, we showed that XCiT can be a strong backbone for self-supervised learning, matching the state-of-the-art results with less compute. XCiT is a generic architecture that can readily be deployed in other research domains where self-attention has shown success.

Limitations. Our models enable training with smaller patches and on higher-resolution images, which leads to clear performance gains. However, for tasks like image classification this gain comes at a cost of relatively high number of FLOPs. In order to address this issue, other components, like FFN, could also be re-examined. Another point is that XCiT models seem to overfit more than their CaiT counterparts, see Table 3.5. They are more similar to some convnets in that respect.

Chapter 4

Sample Efficient Self-Supervised Pre-training with Vision Transformers

Objectives

Pre-training models on large-scale datasets, like ImageNet, is a standard practice in computer vision. This paradigm is especially effective for tasks with small training sets, for which high-capacity models tend to overfit. In this work, we consider a self-supervised pre-training scenario that only leverages the target task data. We consider datasets, like Stanford Cars, Sketch or COCO, which are order(s) of magnitude smaller than Imagenet. Our study shows that denoising autoencoders, such as BEiT or a variant that we introduce in this chapter, are more robust to the type and size of the pre-training data than popular joint embedding self-supervised methods. We obtain competitive performance compared to ImageNet pre-training on a variety of classification datasets, from different domains. On COCO, when pre-training solely using COCO images, the detection and instance segmentation performance surpasses the supervised ImageNet pre-training in a comparable setting.

Contents

4.1	Introduction	37
4.2	Related Work	39
4.3	Analysis	40
4.3.1	Sample Efficiency	41
4.3.2	Learning using non object-centric images	42
4.3.3	Tokenizers	42
4.4	Methodology	44
4.4.1	SplitMask	44
4.4.2	Encoder-Decoder Architecture	44
4.4.3	Global Contrastive Loss	45
4.5	Experiments	45
4.5.1	Datasets	45
4.5.2	Dense Prediction	47
4.5.3	Image Classification	47
4.5.4	Pre-training using ImageNet	49
4.5.5	Ablations	49
4.5.6	Implementation Details	51
4.6	Conclusion	52

4.1 Introduction

Modern computer vision neural networks are heavily parametrized: they routinely have tens or hundreds of millions of parameters [He, 2016; Dosovitskiy, 2021a; Liu, 2021b; Radosavovic, 2020]. This has been the key to their success for leveraging large-scale image collections such as ImageNet. However these high-capacity models tend to overfit on small, or even medium-sized datasets consisting of hundreds of thousands of images. This problem was pointed out by Oquab et al. [Oquab, 2014] in 2014:

“Learning CNNs [...] amounts to estimating millions of parameters and requires a very large number of annotated image samples. This property currently prevents the application of CNNs to problems with limited training data.”

The authors describe a learning setting [Oquab, 2014; Yosinski, 2014] that is nowadays the dominant learning paradigm for data-starving problems:

(1) pre-train a model on a large dataset like Imagenet [Deng, 2009b], and in turn (2) finetune the weights of the models on the target task for which we have a limited amount of data. The second training stage typically adopts a shorter optimization procedure than the one employed when training from scratch (*i.e.*, from randomly generated weights).

This simple approach has led to impressive results, which are state-of-the-art in many tasks such as detection [He, 2017; Carion, 2020], segmentation [Chen, 2014] and action recognition [Carreira, 2017]. Despite this success, we point out that it is difficult to disentangle the benefits offered by such a large-scale curated label dataset from the limitations of this pre-training paradigm. Putting aside the discussion on the collection effort (cost, requiring in-domain expertise, etc), we point out that pre-training a model on a dataset and fine-tuning it on another can introduce two sort of discrepancies.

First, this setting introduces a domain shift between the images used to pre-train the model and those targeted by the fine-tuning stage. Imagenet images may be sufficiently representative of natural images (despite the collecting bias). To date, most researchers consider that the benefit of having a large amount of images vastly compensates the domain discrepancy on benchmarks involving natural images, such as the fine-grained iNaturalist datasets [Van Horn, 2018; Horn, 2017] or even out-of-domain distributions such as sketches, paintings or clipart.

The second question, discussed by Doersch *et al.* [Doersch, 2020], is the so-called *supervision collapse*. This phenomenon is inherent to pre-training with a fixed set of labels: the network learns to focus on the mapping between images and the labels of the pre-training stage, but can discard information that is relevant to other downstream tasks. In other terms, pre-training on large-scale classification datasets does not necessarily align with the goal of learning general-purpose features, as it uses only a subset of the available information controlled by the given dataset categorization bias [Rosch, 1973].

These limitations have motivated the development of self-supervised pre-training methods which learn directly from data, without relying on annotations. Most notably, the contrastive and joint embedding approaches [He, 2020; Caron, 2020a; Caron, 2021; Chen, 2020c; Grill, 2020] can serve as effective pre-training strategies. While obtaining a strong performance on numerous tasks, such methods have a strong bias towards ImageNet data since the transformations have

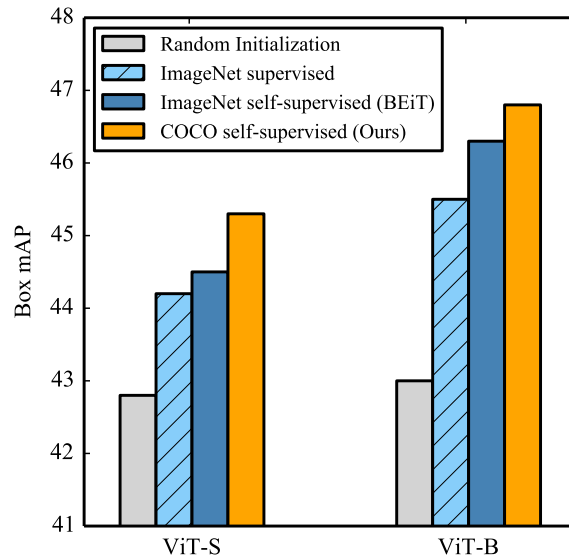


Figure 4.1 – We demonstrate that self-supervised pre-training using denoising autoencoders like BEiT and our variant SplitMask are more robust to the type and/or size of pre-training data used. For example, the object detection performance of such models, when pre-trained only using COCO images and a Mask R-CNN pipeline, outperforms both supervised and BEiT self-supervised baselines pre-trained on ImageNet, as well as a randomly initialized baseline trained for a long schedule.

been hand-designed to perform well on the ImageNet benchmark. Some of the most effective transformations, like cropping, rely on the images being object-centric [Purushwalkam, 2020]. When applied on uncurated data, these methods degrade significantly and require larger datasets to obtain similar performance [Goyal, 2021].

This is in contrast with natural language processing, where nowadays, most applications use large models which were pre-trained on uncurated data. In particular, the (masked) language modeling loss has been applied to transformer networks, leading to the BERT model [Devlin, 2018], which is now the foundation of most NLP models. Inspired by this success, Bao *et al.* [Bao, 2021] have shown the potential of the *Masked Image Modeling (MIM)* task to pre-train vision transformers. Such a model can be thought of as a denoising autoencoder [Vincent, 2008] where the noise corresponds to the patch masking operation. This technique has been successfully applied to ImageNet, but research questions remain:

- (1) How much does this pre-training technique rely on the number of pre-training samples, and in particular, does it require millions of images to be useful?
- (2) Is this technique robust to different distributions of training images? In particular, is it an effective paradigm to learn with non object-centric or uncurated images?

If the answer to both questions is positive, it will enable pre-training using a larger variety of datasets, including the training sets of many tasks that are smaller or belong to a different domain than ImageNet.

Overall, we make the following contributions:

- First, we demonstrate that denoising autoencoders are more sample efficient than joint embedding techniques, enabling pre-training without relying on large-scale datasets (e.g.

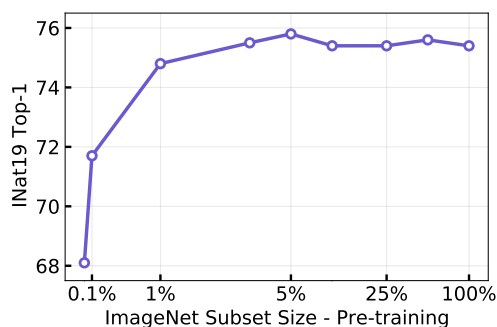


Figure 4.2 – Pre-training using different ImageNet subsets. Transfer performance does not improve beyond using a subset as small as 5% when trained for the same number of iterations.

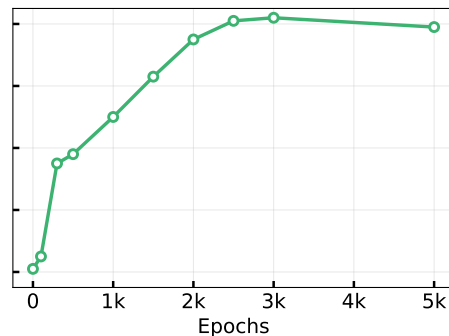


Figure 4.3 – Varying the number of pre-training epochs for the 10% subset. The performance first increases with longer training. Then we observe a plateau and a slight overfitting.

ImageNet);

- Second, as a consequence of the better sample efficiency, we show on multiple datasets that it is possible to pre-train directly on the target task data and obtain a competitive performance, even with datasets that are orders of magnitude smaller than ImageNet;
- Third, we demonstrate that denoising autoencoders can be successfully applied to non object-centric images such as COCO, achieving performance similar to the one obtained when pre-training with ImageNet, unlike joint embedding techniques which seem to suffer a drop in performance.

4.2 Related Work

In this section, we briefly review some previous work on self-supervised learning, including autoencoders and instance discrimination methods.

Pre-training with autoencoders has a long history in deep learning, where it was initially used as a greedy layer-wise method to improve optimization [Hinton, 2006; Bengio, 2007; Ranzato, 2007; Vincent, 2010; Vincent, 2008]. In the context of unsupervised feature learning for image classification, different tasks related to denoising autoencoders have been considered, such as in-painting [Pathak, 2016], colorization [Zhang, 2016] or de-shuffling of image patches [Noroozi, 2016]. In natural language processing, denoising autoencoders have been applied by masking or randomly replacing some tokens of the input, and reconstructing the original sequence, leading to the BERT model [Devlin, 2018]. Similar methods have been proposed to pre-train sequence-to-sequence models, by considering additional kind of noises such as word shuffling or deleting [Raffel, 2019; Lewis, 2019].

There has been efforts to adopt such successful ideas in NLP to computer vision, but with limited success. Chen *et al.* [Chen, 2020b] proposed iGPT, a transformer-based autoregressive model that operates over image pixels, while Atito *et al.* [Atito, 2021] trained a ViT model on denoising of images where the noise is applied at pixel level. More recently, Bao *et al.* [Bao,

2021] introduced the Masked Image Modeling loss in computer vision, where image patches are masked, and the goal is to predict the discretized label of the missing patches corresponding to their visual words as defined by a pre-trained discrete VAE [Ramesh, 2021b].

Instance discrimination is a set of self-supervised techniques which consider that each image corresponds to its own class [Dosovitskiy, 2014; Dosovitskiy, 2015]. A set of data augmentations (or transformations) is then applied to each image to generate multiple examples for each class. The global image representations are trained in a contrastive framework, typically using the InfoNCE loss [Wu, 2018b], to have high similarity for instances transformed from the same source image and low similarity with all other images. As the performance of these methods depends on the number of negatives, it either requires large batches or memory banks to work well [Wu, 2018b; He, 2020; Chen, 2020c]. It was later shown that when using a momentum encoder [He, 2020], simpler loss functions that did not directly discriminate against other images could be used [Grill, 2020; Caron, 2021; Zbontar, 2021; Bardes, 2021; Chen, 2021a]. Finally, a related line of work is to use clustering techniques to pre-train deep neural networks [Xie, 2016; Yang, 2016; Caron, 2018; Asano, 2019b; Caron, 2020b].

Transformer networks were originally introduced in the context of machine translation, replacing recurrent neural networks by an attention-based mechanism [Vaswani, 2017b]. Transformers were later applied to image recognition, by splitting images into patches, embedding these independently, and then processing the obtained representations as a sequence [Dosovitskiy, 2021a]. Initially, only vision transformers pre-trained on very large collections obtained good performance, but smaller models trained on ImageNet with heavy augmentation can also yield competitive tradeoffs [Touvron, 2020].

Pre-training data is an important ingredient of self-supervised learning, and multiple works have studied its impact on the transfer performance of models. While it is possible to learn high quality features from non-curated (eg. YFCC or IG) data using instance discrimination, this usually requires an order of magnitude more data than ImageNet [Caron, 2019; Goyal, 2021]. Similarly, one can perform supervised pre-training using weakly supervised data, such as using hashtags as labels, but this strategy also requires a large amount of data to work well [Joulin, 2016; Mahajan, 2018; Dosovitskiy, 2021a]. On the other hand, it was shown that for many natural language processing tasks, increasing the size of the pre-training dataset did not lead to strong improvement when using denoising autoencoders [Raffel, 2019]. Finally, some work studied how much could be learned from a single pre-training image [Asano, 2019a] or from synthetic data [Kataoka, 2020; Krishna, 2021].

4.3 Analysis

In this section, we study the impact of the pre-training data on the performance of denoising autoencoder, and how they compare to those of joint embedding methods. More precisely, we

Method	IMNet 1% <i>epochs: 30k</i>	IMNet 10% <i>epochs: 3k</i>	IMNet Full <i>epochs: 300</i>	COCO <i>epochs: 3k</i>
Supervised	71.6	75.0	75.8	—
DINO [Caron, 2021]	70.1	73.1	78.4	71.9
BEiT [Bao, 2021]	74.1	74.5	75.2	74.4
SplitMask	74.8	75.4	75.4	76.3

Table 4.1 – Analysis of different self-supervision methods transfer performance to the iNaturalist-2019 dataset when varying the size of the ImageNet subset used in the pre-training stage, in addition to using non object-centric dataset like COCO for pre-training. We observe that denoising autoencoders have a more robust behaviour w.r.t. pre-training data size or nature compared to joint embedding methods like DINO as well as supervised pre-training.

investigate how the number of images, and their nature, influence the quality of self-supervised models. In this preliminary analysis, we consider the recent method BEiT and SplitMask, our variant as detailed in Section 4.4, as representatives of denoising autoencoders, and DINO [Caron, 2021] of a joint embedding method, respectively.

4.3.1 Sample Efficiency

Denoising autoencoders vs Supervised/DINO First, we start by studying the impact of the pre-training dataset size, by varying the number of ImageNet examples we use to train models. We consider subsets of ImageNet containing 10% and 1% of the total number of examples, and use the balanced (in terms of classes) subsets from [Assran, 2021]. To decouple the effect of using smaller datasets and the effect of doing less training updates, we adapt the number of epochs to keep the number of iterations constant. This means that we perform 3k and 30k epochs on ImageNet 10% and 1% respectively. We report results in Table 4.1. Observe how pre-training with an autoencoder loss such as masked image modeling is robust to the reduction in dataset size. In contrast, like for supervised pre-training, the performance of models pre-trained with DINO self-supervision degrades when training with smaller datasets.

Pre-training number of samples We plot the iNaturalist-2019 transfer performance as a function of ImageNet subset size used during pre-training using SplitMask in Figure 4.2. We observe that the peak performance is achieved using only 5% of the ImageNet samples and adding more samples does not provide additional boost, given the number of updates are kept constant. We also observe that using only a single image per class, which corresponds to the 0.1% subset containing 1000 samples, leads to a non-trivial boost (+4 points) over training from scratch. This is a strong indication that denoising autoencoders are highly sample efficient unsupervised learning methods.

Pre-training schedule length Furthermore, we plot the transfer performance as a function of number of pre-training epochs in Figure 4.3 using the 10% ImageNet subset. It can be observed that training for long schedules of nearly 3k epochs, matching the total number of updates for that of full ImageNet with 300 epochs, is crucial to achieve such strong performance

	DALL-E	Rand. Proj.	Rand. Patches	K-Means
iNat19	75.2	75.2	75.3	75.0

Table 4.2 – Ablation study on the effect of different tokenization methods. We compare the DALL-E tokenizer originally used in BEiT with patch level techniques: random projection, random patches and k-means clustering. We observe that the DALL-E tokenizer can be effectively replaced by simpler methods that do not require training on a large dataset.

for smaller subsets. However, we observe slight overfitting for very long schedules. This problem is more predominant for pre-training using very small datasets like Stanford-Cars as illustrated in Figure 4.6.

4.3.2 Learning using non object-centric images

We now study the impact of changing the nature of the pre-training data. In particular we use images that are not object-centric, like in Imagenet. To this end, instead of pre-training using Imagenet, we pre-train with images from the COCO dataset only. As COCO contains roughly 118k images, this dataset is approximately equivalent in terms of size to the ImageNet 10% subset. Again, to disentangle the effect of training with a different number of iterations, we adapt the number of epochs: we use 3k epochs on COCO.

We report the results of this experiments in Table 4.1. When pre-trained on COCO, DINO drops significantly compared to full ImageNet pre-training (-8.3). Interestingly, the drop is higher than using 10% ImageNet even though the numbers of samples is roughly the same. We hypothesize this is because COCO images are not biased to be object-centric, while this joint embedding method was designed and developed using ImageNet as benchmark. In contrast, BEiT’s performance only decreases slightly while SplitMask attains +0.7 improvement over full ImageNet pre-training. This is an interesting property which makes such models prime candidates for learning effectively from uncurated images in the wild.

4.3.3 Tokenizers

The BEiT method, as proposed by Bao *et al.* [Bao, 2021], relies on the discrete VAE tokenizer from DALL-E, which has been pretrained on a large weakly supervised dataset. Since we want to study whether it is possible to pre-train models solely on small datasets, or non object-centric ones, we replace the DALL-E tokenizer by a simple alternative. To this end, we consider different simple alternatives to discretize images at the patch level without any pre-training. Each of these techniques is applied on each patch independently, making them relatively lightweight and more efficient than the original tokenizer considered in BEiT.

Given a vocabulary of size V , each element of the vocabulary is represented by a unit vector $\mathbf{e}_i \in \mathbb{R}^d$, where $i \in \{1, \dots, V\}$ and d is the dimension of patches (in the case of 8x8 patches, $d = 192$). Then, to tokenize an image, we associate each patch to the element of the vocabulary which has the highest cosine similarity with the patch in the pixel space. Hence, for a patch \mathbf{x} ,

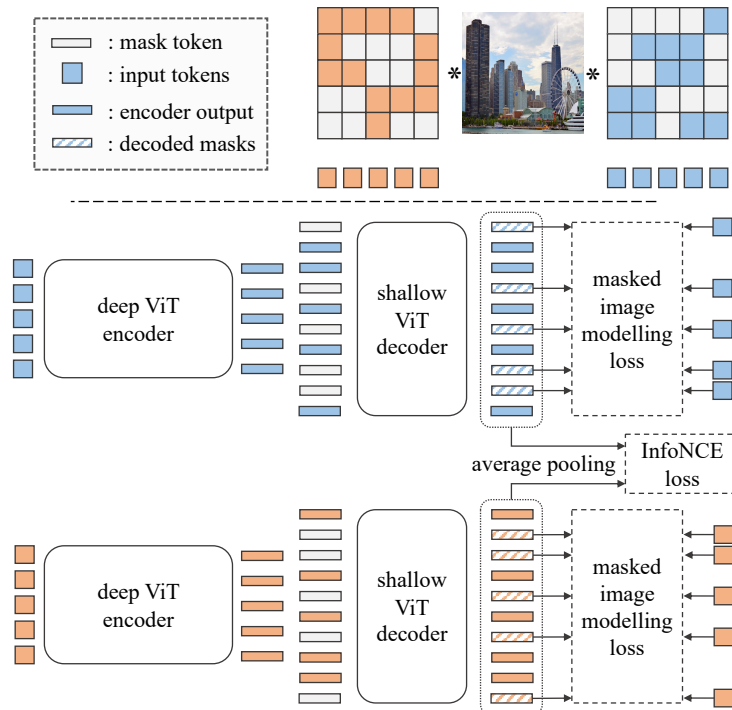


Figure 4.4 – SplitMask consists of three steps. First, the input image patches are split into two disjoint subsets. Second, a shared deep ViT encoder processes each subset separately. The encoder outputs on each branch are augmented with a set of special mask tokens, representing the positions of the missing patches, and fed to a shallow ViT decoder. The decoder output corresponding to the mask tokens is used to solve a MIM task similar to BEiT. Finally, a global image descriptor is extracted from the decoder outputs of each branch by means of average pooling. The descriptors are trained to have high similarity using a contrastive loss (InfoNCE).

its corresponding token t is obtained as

$$t = \operatorname{argmax}_{i \in \{1, \dots, V\}} \mathbf{x}^\top \mathbf{e}_i. \quad (4.1)$$

We now discuss three simple ways to obtain the elements of the vocabulary \mathbf{e}_i . First, we can sample random vectors with uniform element-wise distribution, and call the corresponding tokenizer *random projection*. Second, we can sample V random patches uniformly in the set of all patches of images from the training set, and refer to the tokenizer as *random patches*. Finally, we can perform k-means clustering on the patches of images from the training set, and use the centroids as elements of the vocabulary. We refer to this last tokenizer, which was once widely employed in computer vision for bag-of-words representations, as *k-means*.

We train a ViT-base model on the ImageNet dataset, using these three tokenizers, as well as the DALL-E tokenizer originally considered by BEiT. We report results in Table 4.2. We observe that replacing the DALL-E tokenizer by simpler choices does not lead to any significant degradation in accuracy. This also provides a 26% relative runtime improvement for base models over its counterpart using the DALL-E tokenizer on 16 GPUs with a batch size of 1024.

4.4 Methodology

In this section, we introduce SplitMask, a variant of denoising autoencoders based on vision transformers. An overview of our method is illustrated in Figure 5.2.

4.4.1 SplitMask

Our approach is based on three steps, which we refer to as *split*, *inpaint* and *match*. As in standard vision transformers, an image is first broken down into patches of 16×16 pixels. Then, we *split* the patches into two disjoint subsets \mathcal{A} and \mathcal{B} , which are processed independently by our deep ViT encoder. Next, using the patch representations of the subset \mathcal{A} and a shallow decoder (e.g. 2 layers), we *inpaint*¹ the patches of the subset \mathcal{B} , by solving a MIM task, and vice versa. Finally, we obtain a global image descriptor by average pooling of the patch representations from the decoder output corresponding to each branch.

The feature aggregation is over both observed and hallucinated patches. We try to *match* the global descriptors of the image obtained from subset \mathcal{A} to that obtained from subset \mathcal{B} . In other words, we use the masking operation of the mask image modeling loss as a data augmentation for a contrastive learning loss similar to NPID or SimCLR. Note, SplitMask does not add any significant computational cost over MIM methods like BEiT to produce this global contrastive training signal.

4.4.2 Encoder-Decoder Architecture

We now discuss in more details the architecture of the model that we use to implement the SplitMask pipeline described in the previous subsection. Our method relies on an encoder-decoder architecture. The encoder of our model is a standard vision transformer, with absolute positional embeddings. In contrast to BEiT method, our encoder does not process representations of the *masked* tokens, but only of the observed ones². Hence, an image is divided into patches, which are linearly embedded, and positional embeddings are added to these representations. These representations are split into two subsets \mathcal{A} and \mathcal{B} , which are processed independently by standard transformer layers. Before feeding the output representations to the decoder, we insert mask embeddings that includes the position information of the missing patches in the sequences \mathcal{A} and \mathcal{B} . Finally, using the decoded representations of the masked patches, we predict their corresponding visual words using a cross entropy loss function.

Thus, if an image contains n patches, the encoder processes two sequences of size $n/2$, while the decoder processes two sequences of size n . Since in practice we use decoder which is much more lightweight than standard vision transformers, the computational complexity of our models is similar to a standard ViT. One advantage of our approach compared to BEiT is that at each iteration, the encoder processes all the patches of the image. The loss function is also computed over all the patches of the image, instead of only on a subset.

1. Inpainting in this context is implemented by solving a Masked Image Modeling task rather than the typical inpainting by reconstruction of pixels.

2. Concurrent to this work, He *et al.* [He, 2021b] proposed MAE. This is an encoder-decoder architecture where the encoder processing the observed patches only, similar to what we do in our SplitMask variant.

4.4.3 Global Contrastive Loss

In addition to the MIM loss, which is computed at the patch level, our approach also uses a contrastive loss at the image level. To this end, we apply an average pooling operation over all the output representations of the decoder (including representations of the masked patches). For each image, we obtain two representations \mathbf{x}_a and \mathbf{x}_b , corresponding to the subsets \mathcal{A} and \mathcal{B} of observed patches. We then apply the InfoNCE loss [Oord, 2018] over these representations:

$$\ell(\mathbf{x}_a) = \frac{\exp(\mathbf{x}_a^\top \mathbf{x}_b / \tau)}{\sum_{\mathbf{y} \in \{\mathbf{x}_b\} \cup \mathcal{N}} \exp(\mathbf{x}_a^\top \mathbf{y} / \tau)}, \quad (4.2)$$

where τ is a temperature hyper-parameter and \mathcal{N} is a set of negatives, corresponding to the representations of the other images in the batch. Following previous work [Chen, 2020c], we symmetrize the contrastive loss, and apply it similarly on the representation \mathbf{x}_b from the subset \mathcal{B} . The motivation for adding this contrastive loss is to encourage the model to produce globally coherent features that are consistent across different choices of observed subsets without relying on any hand-designed transformations. Using our design of SplitMask, we attain such signal with almost no overhead.

4.5 Experiments

In this section, we perform empirical evaluations of denoising autoencoders, and the impact of the pre-training data on downstream task performance. In particular, we study how well pre-training performs when only the target task data is used instead of relying on a large-scale dataset such as ImageNet. We perform experiments on different tasks, such as classification, detection and instance segmentation. We consider datasets of varying size, including some significantly smaller than ImageNet. We also compare our variant SplitMask method to BEiT, either pre-trained on target task data or ImageNet, in addition to the supervised pre-training baselines. Finally, we perform an ablation study on our method to investigate the impact of its different components on finetuning and linear evaluation.

4.5.1 Datasets

We study the pre-training and finetuning of computer vision models on a variety of datasets, see Table 4.3 for details. For image classification, we consider the iNaturalist 2018 and 2019 [Van Horn, 2018], Stanford Cars [Krause, 2013] and Food101 [Bossard, 2014] datasets, which all contain fine-grained categories. We also consider three subsets from the DomainNet dataset [Peng, 2019], *clipart*, *painting* and *sketch*, which are not natural images and hence from different domains than ImageNet. For object detection and instance segmentation, we use the COCO dataset [Lin, 2014]. Finally, we also use the ADE20k dataset [Zhou, 2017] for semantic segmentation. The training set sizes of these different datasets vary from 8k to 437k images, thus all being significantly smaller than ImageNet, some more than two order of magnitude smaller. This allows to investigate under different data regimes how feasible it is to pre-train directly on

Dataset	#Train	#Test	#Classes	Epochs
ImageNet [Deng, 2009b]	1,281,167	50,000	1000	300
iNaturalist 2018 [Van Horn, 2018]	437,513	24,426	8,142	800
iNaturalist 2019 [Horn, 2017]	265,240	3,003	1,010	1,400
Food 101 [Bossard, 2014]	75,750	25,250	101	5,000
Stanford Cars [Krause, 2013]	8,144	8,041	196	5,000
Clipart [Peng, 2019]	34,019	14,818	345	5,000
Painting [Peng, 2019]	52,867	22,892	345	5,000
Sketch [Peng, 2019]	49,115	21,271	345	5,000
ADE20k [Zhou, 2017]	20,210	2,000	150	21,000
COCO [Lin, 2014]	118,287	5,000	80	3,000

Table 4.3 – Data size, number of classes and number of pre-training epochs details for all datasets used for pre-training.

Method	Backbone	Pre-training			AP ^b	AP ₅₀ ^b	AP ₇₅ ^b	AP ^m	AP ₅₀ ^m	AP ₇₅ ^m	
		Supervised	IMNet	COCO							
Random Initialization	ViT-S	x	x	x	38.3	60.1	41.4	35.6	57.1	37.7	
Random Initialization†		x	x	x	42.8	64.5	45.6	39.1	61.5	41.7	
DeiT [Touvron, 2020]		✓	✓	x	44.2	66.6	47.9	40.1	63.2	42.7	
BEiT [Bao, 2021]		x	✓	x	44.5	66.2	48.8	40.3	63.2	43.1	
DINO [Caron, 2021]		x	x	✓	43.7	65.5	47.7	39.6	62.3	42.3	
BEiT		x	x	✓	44.7	66.3	48.8	40.2	63.1	43.2	
SplitMask		x	x	✓	45.3	66.9	49.4	40.6	63.6	43.5	
Random Initialization		ViT-B	x	x	x	40.7	62.7	44.2	37.1	59.1	39.4
Random Initialization†			x	x	x	43.0	64.2	46.9	38.8	61.3	41.6
DeiT [Touvron, 2020]			✓	✓	x	45.5	67.9	49.2	41.0	64.6	43.8
BEiT [Bao, 2021]	x		✓	x	46.3	67.6	50.6	41.6	64.5	44.9	
DINO [Caron, 2021]	x		x	✓	43.1	64.4	46.9	38.9	61.4	41.4	
BEiT	x		x	✓	46.7	67.7	51.2	41.8	65.0	44.6	
SplitMask	x		x	✓	46.8	67.9	51.5	42.1	65.3	45.1	

Table 4.4 – COCO detection and instance segmentation performance, using a Mask R-CNN pipeline, for models with different pre-training recipes. We see that BEiT and SplitMask pre-training using COCO outperform supervised ImageNet pre-training of DeiT as well as self-supervised ImageNet pre-training using BEiT. †: Method uses a longer 6x schedule instead of the default 3x following He *et al.* [He, 2018].

the target task data, alleviating the need for a large scale curated dataset as ImageNet.

As previously mentioned, we want to perform a constant number of updates during pre-training, and we thus adapt the number of epochs when training on target task data to match the number of updates corresponding to 300 epochs on ImageNet. For smaller classification datasets, we limit the number of pre-training epochs to 5000 since we observed pre-training for longer generally does not result in further improvement in terms of downstream performance. For very small datasets, like Stanford Cars, we observed an overfitting behaviour with training for very long schedules (e.g. more than 5k epochs, see Figure 4.6). Note that the adjusted number of pre-training epochs is provided in Table 4.3.

4.5.2 Dense Prediction

4.5.2.1 Object detection and Instance Segmentation

First, we evaluate our approach on the COCO object detection and instance segmentation dataset using the Mask R-CNN pipeline [He, 2017] and report our results in Table 4.4. We compare models pre-trained on the COCO dataset alone with their equivalent counterparts that were pre-trained on ImageNet, either in a supervised or self-supervised fashion. First, we observe that BEiT models which were pre-trained on the COCO dataset alone obtain better downstream task performance than the same models pre-trained on ImageNet. For example, when using a ViT-base backbone, pre-training on COCO instead of ImageNet leads to a boost of +0.4 in box AP.

Additionally, we observe that a similar pre-training of DINO using COCO images provides a relatively weak performance, only outperforming random initialization. This indicates that strong pre-training on COCO is a unique property of denoising autoencoders and it does not extend to other self-supervised learning methods.

Finally, we observe that SplitMask leads to a consistent improvement compared to the BEiT baseline, such as +0.6 box AP when using a ViT-small and +0.3 mask AP for ViT-base backbones. All put together, in a comparable setting, we obtain a +1.1 box AP increase while not using ImageNet. Since COCO contains one order of magnitude less images than ImageNet, this suggests that large scale datasets are not necessary for pre-training.

4.5.2.2 Semantic Segmentation

For semantic segmentation, we compare our denoising autoencoder models, pre-trained solely using ADE20k images, to their counterparts pre-trained on ImageNet. The results are reported in Table 4.5. All models use an UperNet pipeline [Xiao, 2018]. We observe that denoising autoencoders can provide a very competitive performance on such a challenging task even when pre-trained using a relatively small sample size of 20k images. The performance matches that of BEiT self-supervised pre-training using ImageNet and only marginally lower than supervised ImageNet pre-training.

We have found that adapting the random cropping strategy is a crucial implementation detail that helps improve the denoising autoencoders pre-training performance on such dataset. In particular, we reduce the maximal size of the crop from 100% to 25% of the raw image size.

4.5.3 Image Classification

We perform empirical evaluation on a number classification datasets and report our results in Table 4.6. Overall, we find that BEiT or SplitMask pre-training, using solely the target datasets images, consistently obtains either the strongest or, at worst, the second strongest performance when compared to different options of self-supervised and supervised pre-training using ImageNet as well as training from scratch [Liu, 2021a].

Method	Pre-training			mIoU
	Supervised	IMNet	ADE20k	
Random Init.	✗	✗	✗	25.4
DeiT [Touvron, 2020]	✓	✗	✗	46.1
BEiT [Bao, 2021]	✗	✓	✗	45.6
BEiT	✗	✗	✓	45.6
SplitMask	✗	✗	✓	45.7

Table 4.5 – Semantic segmentation performance for different pre-trained models on ADE20k using an UperNet pipeline [Xiao, 2018]. All models reported use a ViT-B architecture. In spite of the small size of the ADE20k dataset, performance of our models provides a performance competitive to those pre-trained using ImageNet.

Method	Backbone	Supervised pre-training	Data Used		iNat-18	iNat-19	Food 101	Cars	Clipart	Painting	Sketch
			IMNet	Target	437k	265k	75k	8k	34k	52k	49k
Liu <i>et al.</i> [Liu, 2021a] [‡]	CVT-13	✗	✗	✓	—	—	—	—	60.6	55.2	57.6
	ResNet-50	✗	✗	✓	—	—	—	—	63.9	53.5	59.6
Random Init.	ViT-S	✗	✗	✓	59.6	67.5	84.7	35.3	41.0	38.4	37.2
DeiT [Touvron, 2020]		✓	✓	✓	<u>69.9</u>	75.8	91.5	92.2	79.6	74.2	72.5
BEiT [Bao, 2021]		✗	✓	✓	68.1	75.2	90.5	92.4	75.3	68.7	68.5
BEiT		✗	✗	✓	68.8	<u>76.1</u>	90.7	<u>92.7</u>	—	69.0	—
SplitMask		✗	✗	✓	70.1	76.3	91.5	92.8	<u>78.3</u>	<u>69.2</u>	<u>70.7</u>
Random Init.		ViT-B	✗	✗	✓	59.6	68.1	83.3	36.9	41.9	37.6
DeiT [Touvron, 2020]	✓		✓	✓	<u>73.2</u>	77.7	91.9	92.1	80.0	73.8	72.6
BEiT [Bao, 2021]	✗		✓	✓	71.6	78.6	91.0	93.9	78.0	71.5	71.4
BEiT	✗		✗	✓	72.4	<u>79.3</u>	<u>91.7</u>	92.7	—	70.7	—
SplitMask	✗		✗	✓	74.6	80.4	91.2	<u>93.1</u>	<u>79.3</u>	<u>72.0</u>	<u>72.1</u>
Random Init.	CVT-13		✗	✗	✓	59.6	68.1	83.3	36.9	41.9	37.6

Table 4.6 – Comparison between finetuning performance on the target datasets of different sizes and domains when pre-trained using the target datasets themselves, ImageNet pre-training (both supervised and self-supervised), and training from scratch. Both denoising autoencoders (BEiT and SplitMask) obtain competitive performance when solely using the target data. ‡: Liu *et al.* [Liu, 2021a] use a different pre-training setup and backbones.

Method	Backbone	Epochs	Top-1
MocoV3 [Chen, 2021b]	ViT-S	300	81.4
DINO [Caron, 2021]		300	81.5
BEiT [Bao, 2021]		300	81.3
SplitMask		300	81.5
MocoV3 [Chen, 2021b]	ViT-B	300	83.2
DINO [Caron, 2021]		400	83.6
BEiT [Bao, 2021]		300	82.8
BEiT [Bao, 2021]		800	83.2
SplitMask		300	83.6

Table 4.7 – Finetuning performance on ImageNet. Here, epochs refer to the number of pre-training epochs on ImageNet.

BEiT pre-training: ImageNet vs Target First, we compare ImageNet pre-training to the target data pre-training with BEiT and observe that for many cases, pre-training on the target data alone leads to better results. This is true for the ViT-small backbone across all the datasets including Stanford cars (+1.1% acc), which consists of only 8k images. When using a ViT-base backbone, pre-training on the target task data outperforms BEiT self-supervised

ImageNet pre-training for datasets as small as Food101 (+0.7 acc), which is more than 10x smaller than ImageNet. Second, we observe that SplitMask leads to further improvement in performances for multiple datasets: for example, on the iNaturalist 2018 dataset, we see +3.0 in accuracy with a ViT-base model.

Supervised ImageNet pre-training As it was already observed in previous work [Chen, 2020c; Chen, 2021b; Caron, 2021], we also see in many cases that self-supervised training outperforms supervised pre-training on ImageNet. For example, on the iNaturalist datasets, training with the target task data alone (including a pre-training step) gives better results than pre-training on ImageNet with labels: with a ViT-base model and the SplitMask method, we see an improvement of +2.7% in top-1 accuracy. As for the *clipart*, *painting* and *sketch* datasets, we see that SplitMask provides a competitive performance, outperforming an ImageNet pre-trained BEiT across all datasets for ViT-S. However, for the aforementioned datasets, supervised pre-training achieves the best performance for both ViT-S and ViT-B.

We note that when pre-training using the *clipart* and *sketch* datasets with the BEiT method, we experienced numerical instability that prevented the model from converging with long schedules (e.g. 5000 epochs). However, the instability problem was not observed for SplitMask models. Nevertheless, more investigation might be needed to fully understand how to optimize pre-training of such models.

4.5.4 Pre-training using ImageNet

In addition to our main study concerning the robustness of denoising autoencoders w.r.t the size and type of pre-training data, we study SplitMask in the more commonly used setting of pre-training and finetuning using ImageNet.

In Table 4.7 we show the performance of our SplitMask method using the ViT-S and ViT-B backbones and 300 epochs pre-training compared to other recent transformer-based self-supervised learning methods. It can be observed that SplitMask provides a strong performance, outperforming both BEiT and MocoV3 for all backbones. Additionally, SplitMask achieves a performance on par with DINO while being significantly cheaper and simpler to train. Note that while SplitMask and BEiT attain a strong finetuning performance, denoising autoencoding methods typically fall behind in terms of linear probing compared to instance discrimination methods like DINO.

4.5.5 Ablations

4.5.5.1 SplitMask vs BEiT

We ablate our proposed components in SplitMask compared to a BEiT baseline in Table 4.8. All models use a ViT-B backbone and pre-trained for 300 epochs. First, we observe that the ImageNet finetuning performance improves with a margin (+0.5) by simply adopting the encoder-decoder architecture and processing two disjoint subsets per iteration. Second, the global contrastive loss on its own, without the MIM objective, provides a very weak performance.

This is expected since there is no training signal for the local patch representations, and a global matching objective with 50% masking of patches may be too hard, providing a noisy training signal and hindering the model’s ability to learn informative features.

Our full SplitMask model that uses both the [MIM](#) and contrastive objectives obtains the best performance and outperforms BEiT by a large margin of +0.8. The Linear probing performance of SplitMask is stronger than BEiT. However, both models provide a relatively weak performance on this benchmark compared to instance discrimination methods, whose final layers are more aligned to the classification task. Note, SplitMask adds a negligible computing overhead compared to the BEiT baseline: its wall-clock training time is marginally higher as detailed in [Table 4.8](#). All models are trained using 16 GPUs and batch size of 2048.

Method	Split	Inpaint	Match	Finetune	Lin.	Hours
BEiT [Bao, 2021]	✗	✓	✗	82.8	41.0	32.5
SplitMask	✓	✓	✗	83.3	46.4	31.0
	✓	✗	✓	79.3	4.0	32.5
	✓	✓	✓	83.6	46.5	34.0

Table 4.8 – Ablations of different components in our SplitMask model in comparison with a BEiT baseline. All models including the baseline have been trained for 300 epochs using a ViT-B backbone.

4.5.5.2 Encoder-Decoder vs BEiT

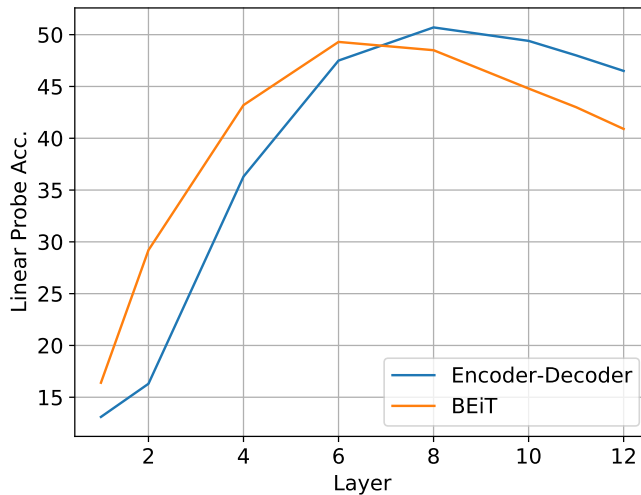


Figure 4.5 – Linear probing accuracy on ImageNet for SplitMask and BEiT using features extracted from different layers.

An advantage of the encoder-decoder design we propose in [4.4.2](#) is that it encourages decoupling of general-purpose encoding of image features, which is required for the downstream tasks, and features specific to solving the pretext task of [MIM](#). In particular, compared to BEiT the encoder is not capable of solving the pretext task on its own since it does not have access to the mask token. Therefore, it can only help solve the task by providing informative represen-

tation to the decoder which is the component responsible of solving the pretext task. We can see in Figure 4.5 that this property improves the transferability of later layers representation to downstream tasks compared to BEiT which has a stronger drop in linear probing performance in later layers.

4.5.5.3 Overfitting during pre-training

We observed that for pre-training of very small datasets (e.g. Stanford-Cars), longer pre-training schedules can be counterproductive. For example, if we follow the assumption we need to pre-training for the same number of updates of ImageNet pre-training for 300 epochs, the Stanford-Cars equivalent schedule would be 45k epochs. However, as we see in Figure 4.6, pre-training longer than 5k epochs leads to a severe drop in finetuning performance.

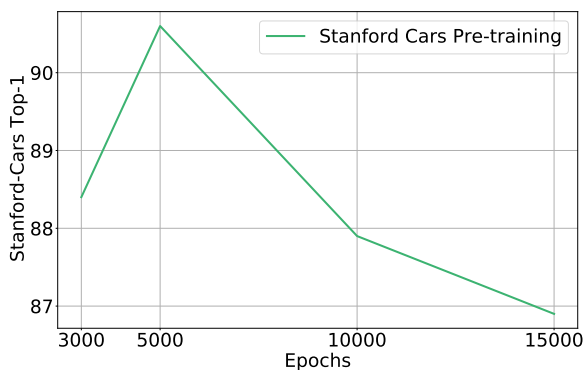


Figure 4.6 – Finetuning performance for the Stanford Cars datasets as a function of number of pre-training epochs using the same datasets images.

4.5.6 Implementation Details

Tokenizers. Similarly to the tokenizer used in [Bao, 2021], all tokenizers presented in Table 4.2 have a vocabulary of size 8192. For the random tokenizer, we sample 8192 vectors with uniform component-wise distribution. For the random patches tokenizer we sample 8192 patches from different images. For the K-means tokenizer, the 8192 elements of the vocabulary are obtained by applying the K-means algorithm to 3 millions patches sampled from the dataset.

Pre-training. We use the original ViT formulation as proposed by Dosovitskiy *et al.* [Dosovitskiy, 2021a] and we follow the pre-training hyperparameters of Bao *et al.* [Bao, 2021]. All baselines reported use the same backbone implementation and trained in similar settings. For SplitMask, by default, we use random block masking [Bao, 2021] of 50% masking ratio to obtain a mask and its complement to extract the two subsets. The maximum and minimum number of patches per block is 75 and 16 respectively. We use the standard random cropping and horizontal flipping as data augmentations. We use 2 transformer layers for the decoder with embedding dimensions matching that of the encoder.

However, for the smallest datasets (i.e. Stanford-Cars, ClipArt, Sketch and Paintings), we found that stronger data augmentation and more aggressive masking prevents early overfitting. In particular, we use a uniform masking of 75% (like in the work by He et al. [He, 2021b]), as well as using random greyscale, solarization, Gaussian blur and color jittering as additional forms of data augmentation.

The BEiT baselines pre-trained on ImageNet and reported in Table 4.4 and 4.6 use the DALL-E tokenizer. Other BEiT and SplitMask models have been pre-trained using our random projection tokenizer. For the InfoNCE loss we use $\tau = 0.2$ following Chen *et al.* [Chen, 2021b].

Object detection and Instance segmentation. We use the Mask R-CNN detection method [He, 2017] with ViT backbone as our detection method. In order to obtain features compatible with the Feature Pyramid Network (FPN) design [Lin, 2017], we use max pooling and transposed convolution operations similar to El-Nouby *et al.* [El-Nouby, 2021d]. To accommodate for the variable resolution we replace the absolute positional encoding for our models and the baselines with sinusoidal positional encoding [Vaswani, 2017b]. All models are trained using the 3x schedule (36 epochs) unless mentioned otherwise. We use the training hyper-parameters used by Liu *et al.* [Liu, 2021b].

4.6 Conclusion

In this chapter, we have raised the question of how to pre-train models with self-supervised learning, wondering in particular on whether large scale datasets such as Imagenet are necessary for pre-training. Our study on ImageNet shows that taking a smaller pre-training dataset does not lead to a big performance drop for denoising autoencoders, as opposed to instance discrimination self-supervised techniques or supervised pre-training. Similarly, training on non object-centric images does not impact the downstream task performance significantly.

Building upon these observations, we have pre-trained models directly on the target task data, instead of ImageNet, and performed evaluations on datasets of various sizes. We have shown that it is possible to pre-train on datasets 10x smaller than ImageNet, for example obtaining +0.5 box AP gains by solely using COCO images. *We believe that this is strong evidence that large-scale datasets, such as ImageNet, are not necessary for self-supervised pre-training when using denoising autoencoders.*

Chapter 5

Image Compression with Product Quantized Masked Image Modeling

Objectives

Recent neural compression methods have been based on the popular hyperprior framework. It relies on Scalar Quantization and offers a very strong compression performance. This contrasts from recent advances in image generation and representation learning, where Vector Quantization is more commonly employed.

In this chapter, we attempt to bring these lines of research closer by revisiting vector quantization for image compression. We build upon the VQ-VAE framework and introduce several modifications. First, we replace the vanilla vector quantizer by a product quantizer which implicitly defines high-quality quantizers that would otherwise require intractably large codebooks. Second, inspired by the success of Masked Image Modeling in the context of self-supervised learning, we propose a novel conditional entropy model which improves entropy coding by modeling the co-dependencies of the quantized latent codes. The resulting PQ-MIM model is surprisingly effective. It outperforms HiFiC in terms of FID and KID when optimized with perceptual losses. Finally, since PQ-MIM is compatible with image generation frameworks, we show qualitatively that it can operate under a hybrid mode between compression and generation. As a result, we explore the extreme compression regime where an image is compressed into 200 bytes, i.e., less than a tweet.

Contents

5.1	Introduction	54
5.2	Related work	56
5.3	Product Quantized Masked Image Modeling	57
5.3.1	High-level architecture : PQ-VAE	58
5.3.2	Image entropy model	59
5.3.3	Training the PQ-MIM	61
5.4	Experiments	62
5.4.1	Experimental setup	62
5.4.2	Main experimental results	64
5.4.3	Analysis and Ablations	65
5.4.4	Limitations	67
5.5	Conclusion	68

5.1 Introduction

Image compression played a crucial role in accelerating multiple events in history: it was a major component of the Voyager mission [Ludwig, 2016]. Efficient image codecs have accelerated the rapid growth of the internet by enabling the transmission of images in a few dozen of kilobytes, thanks to the emergence of effective lossy methods. This democratization was accompanied by standardization efforts to facilitate interoperability, which led to the emergence of standards such as the Joint Photographic Experts Groups (JPEG). Subsequent formats have leveraged scientific advances on all components of source coding, ranging from transforms [Antonini, 1992], and quantization [Gray, 1998], to entropy coding [Witten, 1987; Taubman, 2000], eventually leading to modern video compression codecs enabling streaming and video-conferencing applications.

Neural methods have recently become increasingly popular for image compression as well as other image processing tasks, such as denoising [Tian, 2020], super-resolution [Bruna, 2016; Dong, 2015; Ledig, 2017; Wang, 2021b] or image reconstruction [Wang, 2020a; Knoll, 2020]. In typical scenarios, neural image compression is not necessarily mature enough to take over standard techniques like the BPG format inherited from the High-Efficiency Video Coding standard [Sullivan, 2012]. This is because they do not offer a significant quantitative advantage over prior works that would justify the higher complexity, which depends on the context and operational constraints. A key advantage of neural compression methods is their enhanced qualitative reconstruction when incorporating an adversarial loss or likewise psycho-visual objectives favoring visually appealing reconstructions [Agustsson, 2019; Mentzer, 2020]. From this perspective, neural compression is related to image generation. The two subfields, however, are currently dominated by different approaches, noticeably they employ different discretization procedures. Indeed, while earlier neural compression methods utilized vector quantization [Agustsson, 2017, VQ], recent methods mostly employ scalar quantization (SQ). In contrast, the recent literature on image generation [Chang, 2022; Yu, 2021; Esser, 2021; Rombach, 2021] relies on Vector Quantization jointly with a distortion criterion akin to those used in compression.

In this work we aim to reduce the methodological gap and to make a step towards the unification of neural image compression and image generation, and allowing image compression to more directly benefit from the rapid advances in image generation methods. Patch-based masking methods for self-supervised learning [Bao, 2022; He, 2021a; El-Nouby, 2021a] have recently demonstrated their potential for image generation [Chang, 2022]. Inspired by this work we propose a compression approach built upon Vector Quantized Variational Auto-Encoders [Oord, 2017; Razavi, 2019]. In this context, we focus on two intertwined questions: (1) How to define a vector quantizer offering a range of rate-distortion operating points? (2) How to define an entropy model minimizing the cost of storing the quantization indexes, while avoiding the prohibitive complexity of an auto-regressive model?

To address the challenges above, we revisit vector quantization in image compression, and investigate product quantization [Jégou, 2010] (PQ) in a compression system derived from VQ-VAE [Oord, 2016b]. We show that PQ offers a strong and scalable rate-distortion trade-off. We

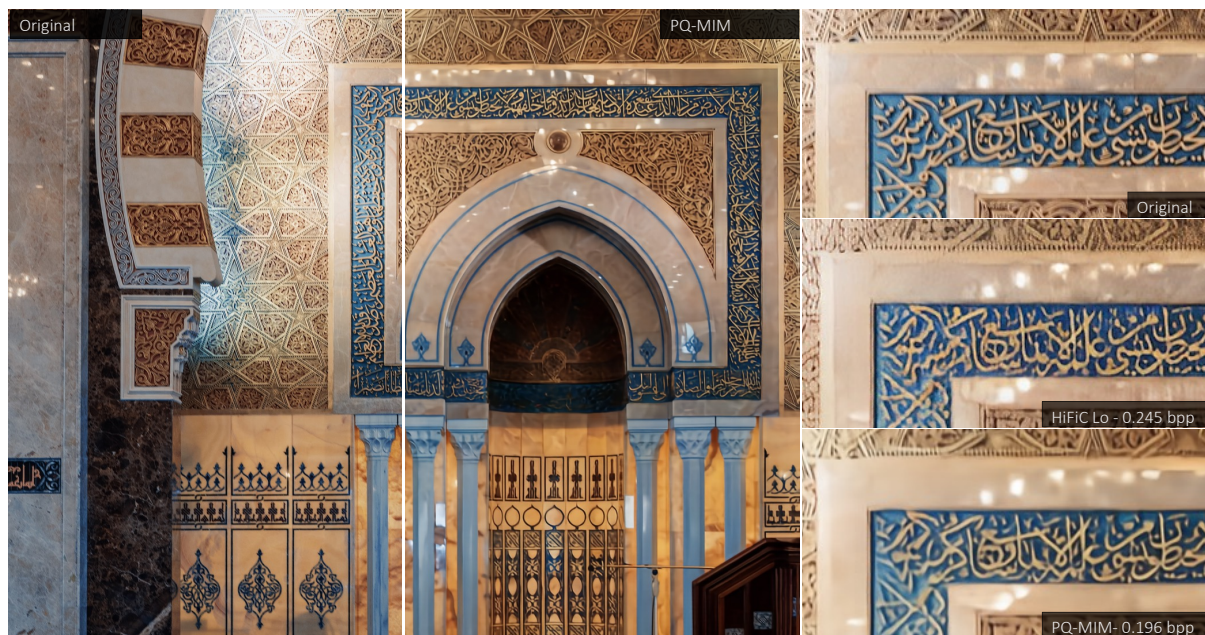


Figure 5.1 – **Qualitative example of PQ-MIM compression.** PQ-MIM provides a strong compression performance. We retain many of the details present original image with minimal blurring effect even with compression rate as low as 0.196 bpp. Moreover, compared to HiFiC we achieve a lower rate for the same image. PQ-MIM provides colors that are more faithful to the original image, while HiFiC has a darkening effect and some high-frequency artifacts. On the other hand, PQ-MIM can have some smoothing effect that can cause loss of detail for particular regions (e.g. some of the Arabic letters in the example above). More qualitative examples are provided in Figure 5.11.

then we focus on the spatial entropy modeling and coding of the quantization indexes in the VQ or PQ latent layer, hence we name our method as (Vector/Product)-Quantized Masked Image Modeling (VQ-MIM and PQ-MIM). To this end, we introduce a multi-stage vector-quantized image model: we gradually reduce the conditional entropy of the patch latent codes by increasing the number of observed patches we condition on for each stage. The conditional distribution over patches is estimated by a transformer model and provided to an entropy coder, symmetrically on the emitter and receiver sides.

In summary, we make the following contributions:

- We introduce a novel Masked Image Modeling conditional entropy model that significantly reduces the rates by leveraging the spatial inter-dependencies between latent codes.
- We introduce product quantization for VQ-VAE. This simple PQ-VAE variant offers a strong and scalable rate-distortion trade-off.
- When trained with adversarial and perpetual losses, PQ-MIM exhibits a strong performance in terms of perceptual metrics like FID and KID, outperforming HiFiC [Mentzer, 2020].
- We qualitatively show that PQ-MIM is capable of operating in a hybrid mode, between generative and compression, without requiring further training and finetuning. This allows for higher resilience to corrupted or missing signal where our model can fill-in the missing information.

5.2 Related work

Neural image compression Early approaches to neural image compression reach back to the late 1980s [Sonehara, 1989; Sicuranza, 1990; Bottou, 1998]. Recent rapid advances in explicit and implicit density modelling [Goodfellow, 2014; Kingma, 2014; Larochelle, 2011; Oord, 2016a; Rezende, 2015; Salimans, 2017] have renewed interest in posing image compression as a learning problem. Due to the connection to variational learning [Gregor, 2016; Frey, 1998; Alemi, 2018], variational auto-encoders have been the primary choice for lossy image compression. In contrast to standard variational models, the evaluation of neural compression models focuses on achievable bitrates, multi-scale applicability and computational complexity. The initial works in the field used fully convolutional architectures for encoding/decoding [Ballé, 2017; Theis, 2017; Mentzer, 2018]. The resulting encoded latent image representations are quantized and compressed via an entropy coder with learned explicit density model or “entropy bottleneck”. Initial variational approaches directly modeled a single level of code densities. Ballé extended these models by introducing a second “hyperprior” that yielded improved performance [Ballé, 2018]. Hyperprior models have been the basis for several subsequent advances with further improvements for density modeling, such as joint autoregressive models [Minnen, 2018], Gaussian mixture/attention [Cheng, 2020], and channel-wise auto-regressive models [Minnen, 2020]. Another line of work has proposed to use vector quantization with histogram-based probabilities for image compression [Agustsson, 2017; Lu, 2019]. Contrary to VQ-VAE models, these models typically optimize the rate (or a surrogate of the rate) directly and may include a spatial component for the quantized vectors. Yang et al. [Yang, 2020] showed multiple ways to improve the encoding process, including the fine-tuning of the discretization process and employing bits-back coding in the entropy bottleneck. Finally, [Santurkar, 2018; Mentzer, 2020; Rippel, 2017; Agustsson, 2019] showed that altering the distortion metric to include an additional adversarial loss can make a large difference for compression rate. Another interesting line of work considers image compression by training image-specific networks, or network adapters, that map image coordinates to RGB values, and compressing the image-specific parameters [Dupont, 2021; Dupont, 2022; Strüpler, 2022].

VQ-models for image generation. There has been significant interest in generative models based in discrete image representations, as introduced by VQ-VAE [Oord, 2017; Razavi, 2019]. A discrete representation of reduced spatial resolution is learned by means of an autoencoder which quantizes the latent representations. This discrete representation is coupled with a strong prior, for example implemented as an autoregressive pixel-CNN model. VQ-GAN [Esser, 2021] replaces the prior architecture with a transformer model [Vaswani, 2017a], and introduces an adversarial loss term to learn an autoencoder with more visually pleasing reconstructions and improved sample quality. This approach has been extended to text-based generative image models by extending the prior to model a longer sequence that combines the discrete image representation with a prefix that encoding the conditioning text. This has yielded impressive results by scaling the model capacity and training data to tens or hundreds of million text-image pairs [Ding, 2021a; Gafni, 2022; Ramesh, 2021a]. A fundamental limitation of autoregressive generative

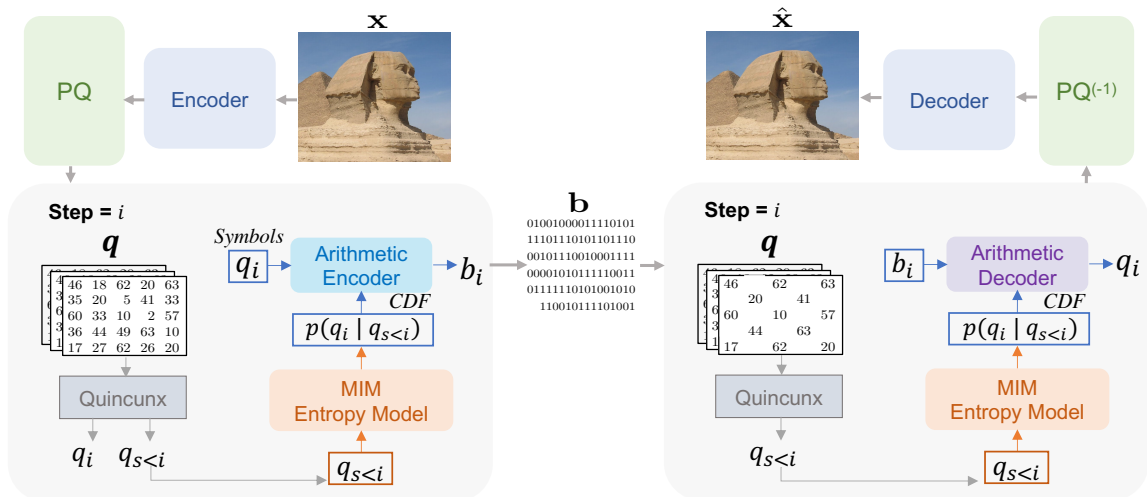


Figure 5.2 – **PQ-MIM overview**. Our model consists of (i) a transformer based encoder and decoder, (ii) a masked image model (MIM) for conditional entropy modeling, and (iii) an entropy coder, *e.g.* an arithmetic coder (AE/AD). The input image \mathbf{x} is projected to a set of latent features, followed by product quantization to yield quantization indices \mathbf{q} . The arithmetic coder encodes (and decodes) \mathbf{q} into a bitstream \mathbf{b} in a lossless manner. The elements in \mathbf{q} are spatially split into groups, as detailed in Figure 5.4. **Conditional Entropy Modeling**. Our model estimates the conditional probabilities of the discrete indices in S steps. Every step, a subset of the tokens q_i is selected using the quincunx pattern. Our MIM transformer estimates $p(q_i|q_{s<i})$ and passes it to the Arithmetic encoder as a CDF, effectively reducing the lossless compression cost.

models is that they sample data sequentially, requiring separate non-parallel evaluation of the predictive (transformer) model to sample each token. To alleviate this, several data items can be sampled independently in parallel, conditioning on all previously sampled tokens. This has been leveraged to speed-up pixel-CNNs for small images and video by two to three orders of magnitude [Reed, 2017]. More recently, MaskGIT [Chang, 2022] and follow-up work [Lezama, 2022] explored this for generative models of VQ-VAE representations, and find that similar or better sample quality is obtained by parallel sampling of image patch subsets in few steps, reducing the generation time significantly. Despite their success for image synthesis we are not aware of the use of such models for image compression in earlier work.

5.3 Product Quantized Masked Image Modeling

This work takes a step towards to closing the gap between neural image compression and image generation methodologies. We revisit vector quantization for image compression and propose an entropy model inspired by masked image modelling. Our compression pipeline, depicted in Figure 5.2, relies on three neural networks:

1. **The Encoder network** $\hat{E} : \mathcal{X} \rightarrow \mathcal{Z}$ maps input images $\mathbf{x} \in \mathcal{X}$ to a quantized representation $\mathbf{z} \in \mathcal{Z}$.
2. **The Mask Image Model (MIM)** compresses the quantized representations without loss of information. This network is involved both on the compression and decompression side.

3. **The Decoder network** $G : \mathcal{Z} \rightarrow \mathcal{X}$ produces an estimate $\hat{\mathbf{x}} = G(\hat{E}(\mathbf{x}))$ of the original image \mathbf{x} .

We now detail the architecture, in particular our PQ proposal, of the image model that we employ in the statistical lossless coding, as well as the training scheme.

5.3.1 High-level architecture: PQ-VAE

High-level architecture. We follow recent work on discrete generative image models for the design of image encoder and decoder [Chang, 2022; Esser, 2021; Oord, 2017; Razavi, 2019; Yu, 2021]. The encoder $E : \mathcal{X} \rightarrow \mathbb{R}^{w \times h \times d}$ takes an RGB image \mathbf{x} of resolution $W \times H$ as input and maps it to a latent representation $E(\mathbf{x})$ with d feature channels and a reduced spatial resolution $w \times h$, downsampling the input resolution by a factor $f = W/w = H/h$. The $T = w \times h$ elements of the latent representation $E(\mathbf{x})$ are quantized with a vector quantizer $Q(\cdot)$ to produce the quantized latent representation $\mathbf{z} = Q(E(\mathbf{x})) = \hat{E}(\mathbf{x})$, where each element in $E(\mathbf{x})$ is replaced with its nearest cluster center. The decoder G uses the quantized latents \mathbf{z} to reconstruct the image.

Product Quantization. In VQ-VAE, the quantizer Q is simply an online k-means quantizer that produces quantization indices from real-valued vectors. We denote by $\mathbf{q} \in \{1, \dots, V\}^T$ the map of quantization indices indicating for each element of \mathbf{z} which of the V centroids is selected. The higher V the more precise is the approximation \mathbf{z} , leading to higher bit rates. For instance, assuming that indices are coded with a naive coding scheme (see next section), the bit rate is doubled when moving to $V = 256$ to $V = 65536$ centroids.

However, scaling the number of centroids K is possible up to thousands of centroids, but beyond that it is computationally prohibitive. Additionally, it is challenging to train large codebooks where each centroid has a very low probability of being updated. To address this problem, we replace the online k-means quantizer by a product quantizer [Jégou, 2010] (PQ): the latent vector \mathbf{z} is split into M subvectors as $\mathbf{z} = [\mathbf{z}^1, \dots, \mathbf{z}^j, \dots, \mathbf{z}^M]$ of dimension M/d . Each subvector is quantized by a distinct quantizer having V_s quantization value. The set of quantizers implicitly defines a vector quantizer in the latent space with $V = V_s^M$ distinct centroids. Hence, we can easily define very large codebooks without the computational and optimization problems mentioned above, because both the assignment and learning are marginalized over the different subspaces. Empirically, we observe in Figure 5.3 that PQ provides a better scaling behaviour for higher rates compared to VQ whose codebook size needs to grow exponentially to achieve the same rates.

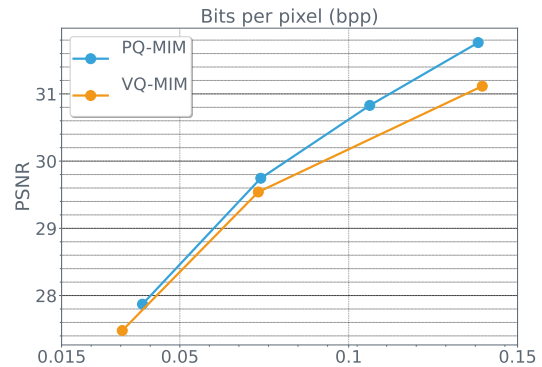


Figure 5.3 – **PQ and VQ comparison.** While VQ provides a comparable performance to PQ for extremely low rates where the codebook size is small, PQ exhibits better scaling behaviour for higher rates.

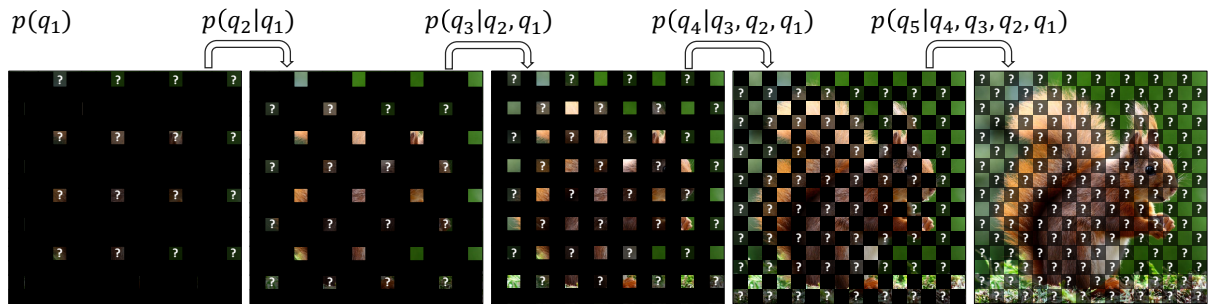


Figure 5.4 – **Illustration of PQ-MIM with a quincunx pattern.** We employ the quincunx pattern both on the encoder and decoder side. Each panel represents one of five stages in which we en/de-code a set of tokens in parallel using a probability model p_s implemented by a transformer with parameters θ_s . The transformer predicts the tokens in \mathbf{q}_s , displayed in grayscale and marked by “?”, and takes as input the preceding groups of tokens $\mathbf{q}_1, \dots, \mathbf{q}_{s-1}$ that are displayed in color. The distribution provided by this neural network is fed to an arithmetic en/de-coder.

Neural network. Without loss of generality, we choose all neural network models to be identical. This is not a requirement but this offers the property that the encoder and decoder have identical complexities, and that the memory and compute peaks are identical. More specifically we choose a cross-variance transformer [El-Nouby, 2021c] (XCiT), whose complexity is linear with respect to image resolution. In contrast, standard vision transformers [Dosovitskiy, 2021b] (ViT) are quadratic in the image surface, which is prohibitive for high resolution images that can typically require strong compression. We point out that recent work has shown that Swin-Transformers [Zhu, 2021] could be a compelling choice as well in the context of image compression. Formally they also have a quadratic complexity, but this is amortized by the hierarchical structure of this architecture.

5.3.2 Image entropy model

In this section we present PQ-MIM. The objective is to compress the discrete representations \mathbf{q} without loss of information, producing a bitstream of the compressed representation that can be transmitted or stored. During the decoding stage, we invert the aforementioned lossless compression. This model can be regarded as the VQ/PQ-VAE counterpart of adaptive contextual arithmetic coders, like EBCOT [Taubman, 2000] or CABAC [Richardson, 2004], proposed in early compression standards, in that it couples a conditional probabilistic model with an arithmetic coder.

Lossless compression. A naive manner for lossless compression of the discrete image codes $\mathbf{q} = \{q_t\}_{t=1}^T$ is to use fixed-length codes. In that case, each code word is assigned to a unique binary representation of equal length, resulting in $\lceil \log_2 V \rceil$ bits per element q_t . This approach is computationally very efficient as fixed-length codes are not model-based, and as such do not require likelihood estimation, and because all codes are of equal length by construction, the computation is perfectly parallelizable. However, theoretically this coding scheme could be Shannon optimal only if codewords are uniformly and independently distributed. Those assumptions are not met in practice due to the architecture choices we have made previously:

k-means does not produce uniformly distributed indices except in singular cases [Gray, 1998].

Entropy model. To improve the bitrate, we hence rely on an entropy coder, which provides an inverse pair of functions, enc_p and dec_p , achieving near optimal compression rates on sequences of symbols for any distribution p . The better p matches the (unknown) underlying data distribution, the better the compression rate [Cover, 1991]. Generally, more powerful generative models will ensure better compression performance.

Fully autoregressive generative models $p(\mathbf{q}) = \prod_{t=1}^T p(q_t|\mathbf{q}_{<t})$ are powerful [Ding, 2021a; Esser, 2021; Gafni, 2022; Ramesh, 2021a; Yu, 2021], however, they are inconvenient in that the likelihood estimation for this type of model is not trivially parallelizable: each patch index must be processed sequentially as it is used to condition subsequent patch indices. Thus, similar to prior works [Chang, 2022; Reed, 2017] we propose a masked image model, which we use to predict the image patch indices in several stages. Specifically, we partition \mathbf{q} into S subsets $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_S$ of patch indices, that we refer to as *tokens* by analogy to language modelling:

$$\mathbf{q} = \bigcup_{s=1}^S \mathbf{q}_s. \quad (5.1)$$

We model the elements in each subset conditionally independent given all preceding groups:

$$p(\mathbf{q}) = \prod_{s=1}^S p(\mathbf{q}_s|\mathbf{q}_{<s}; \theta_s), \quad (5.2)$$

$$p(\mathbf{q}_s|\mathbf{q}_{<s}; \theta_s) = \prod_{q_t \in \mathbf{q}_s} p(q_t|\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_{s-1}; \theta_s). \quad (5.3)$$

Since $p(\mathbf{q}_1)$ is not conditioned on any previous elements, it fully factorizes over $q_t \in \mathbf{q}_1$, and we model it as the marginal distribution over the vocabulary observed on the training data. The non-trivial conditional distributions $p(\mathbf{q}_s|\mathbf{q}_{<s}; \theta_s)$ for $s \geq 2$ are modeled using transformer networks, which have few inductive biases and have been successful across many tasks, including image generation. An overview of the MIM entropy model is illustrated in Figure 5.2.

Amortized encoding and decoding. From a computational point of view, our proposal allows for the compression (resp. decompression) to proceed in S stages. In each stage s we encode/decode the set \mathbf{q}_s of tokens conditioned on the groups $\mathbf{q}_1, \dots, \mathbf{q}_{s-1}$ encoded/decoded in preceding stages, but independently among the tokens in the set \mathbf{q}_s . This allows for parallelization among the elements in each subset \mathbf{q}_s , and requires strictly S forward passes through the model independent of the image size, rather than T sequential forwards passes for fully autoregressive models.

Quincunx partitioning. In practice, we typically use $S = 5$ stages. We have explored different patterns to partition the T tokens over the S stages. In particular we consider the “quincunx” regular grid pattern, where in each stage we double the number of tokens to predict, see Figure 5.4 for an illustration. This multi-level refinement was previously explored for image compression in the context of lifting schemes designed with oriented wavelets [Chappelier, 2006]. In our experiments we contrast this partitioning with alternative ones with other patterns and

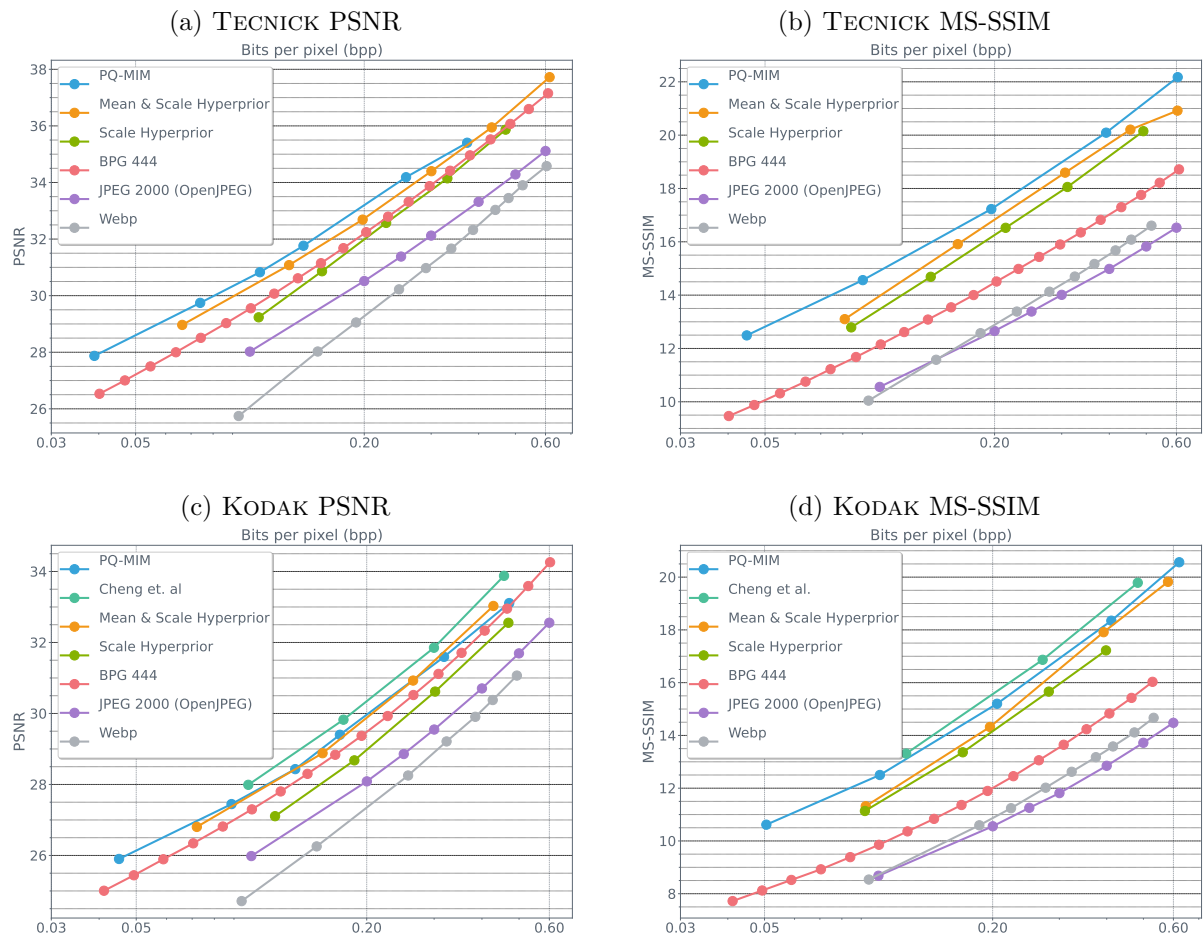


Figure 5.5 – **Rate-Distortion performance for Tecnick and Kodak datasets.** We report PQ-MIM PSNR and MS-SSIM performance for various operating points. PQ-MIM provides a competitive performance, particularly for MS-SSIM, compared to standard codecs such as JPEG 2000 [Taubman, 2012] and BPG [Bellard,] as well as recent neural methods [Cheng, 2020; Minnen, 2018; Ballé, 2018].

subset cardinalities (Figure 5.9).

5.3.3 Training the PQ-MIM

Reconstruction objective and training. The goal of lossy image compression is to match the image and its reconstruction as closely as possible according to some distortion metric. In this chapter, we train our model to reduce the image distortion and the quantization loss using the following objective:

$$L = L_{\text{Rec}} + \eta \cdot L_{\text{PQ}} \quad (5.4)$$

Following VQ-VAE, our quantization objective L_{PQ} consists of embedding and commitment losses, averaged over the M different PQ sub-vectors. For the distortion loss L_{Rec} , we present two setups where we use different types of distortion measures:

- **MSE & MS-SSIM.** Typical distortion measures used in the majority of the neural compression literature such as mean squared error (MSE) or multi-scale structural similarity [Wang,

2003] (MS-SSIM). For this setup, the model is trained solely using one distortion measure at a time.

$$L_{\text{Rec}}(\mathbf{x}, \hat{E}, G) = L_{\text{MSE/MS-SSIM}}(\mathbf{x}, \hat{\mathbf{x}}) \quad (5.5)$$

— **Perceptual measures.** Alternatively, we report a setup where we utilize perceptual objectives such as LPIPS [Zhang, 2018] and adversarial training [Goodfellow, 2014] to enhance psycho-visual image quality. The distortion loss is defined as:

$$L_{\text{Rec}}(\mathbf{x}, \hat{E}, G) = L_{\text{MSE}}(\mathbf{x}, \hat{\mathbf{x}}) + \alpha \cdot L_{\text{Perc}}(\mathbf{x}, \hat{\mathbf{x}}) + \gamma \cdot L_{\text{Adv}}(\hat{E}, G, D), \quad (5.6)$$

where α and γ are weighing coefficients and the adversarial loss L_{Adv} is defined as:

$$L_{\text{Adv}}(\hat{E}, G, D) = \mathbb{E}_{\mathbf{x}}[\ln D(\mathbf{x})] + \mathbb{E}_{\hat{\mathbf{x}}}[\ln(1 - D(\hat{\mathbf{x}}))], \quad (5.7)$$

where $D(\cdot)$ is the discriminator and $\mathbb{E}_{\mathbf{x}}$ denotes the expectation over \mathbf{x} sampled uniformly from the training set. Similarly $\mathbb{E}_{\hat{\mathbf{x}}}$ denotes the expectation over reconstructed training images. Note that, unlike the MSE and perceptual losses, the adversarial loss does not compare individual images and their reconstructions, but aims to match the *distributions* of original images and their reconstructions.

Training the entropy model. Our MIM module is an XCiT transformer that accepts T tokens as input representing the image patches. During training, we randomly mask a set of tokens by sampling from a uniform distribution $U(0, 1)$. The masked tokens are replaced with a mask embedding vector, while the observed token indices are mapped to their corresponding continuous representation using an embedding look-up table. The MIM module outputs a context vector for every masked token which in turn is passed to M linear heads, representing the different PQ sub-vector indices, followed by a softmax to yield a distribution $p(\mathbf{q}_s | \mathbf{q}_{<s})$. The module is trained using a standard cross-entropy objective. While we train the autoencoder and the MIM modules simultaneously, we do not backpropagate gradients from the MIM module to the encoder E or the quantization parameters, so the two components can be trained separately in sequence.

5.4 Experiments

We first present our experimental setup in Section 5.4.1 and then present our results in Section 5.4.2. We provide ablation studies in Section 5.4.3. We will share our code and models.

5.4.1 Experimental setup

Rate-distortion control. For all our experiments we fix the codebook size $V = 256$ and only vary the number of sub-vectors $M \in \{2, 4, 6\}$ for two different down-sampling factors $f \in \{8, 16\}$.

PQ-VAE implementation details. Our PQ-VAE training uses the straight-through estimator [Bengio, 2013] to propagate gradients through the quantization bottleneck. As for the quantization, the elements of the latent representation \mathbf{z} are first linearly projected to a low

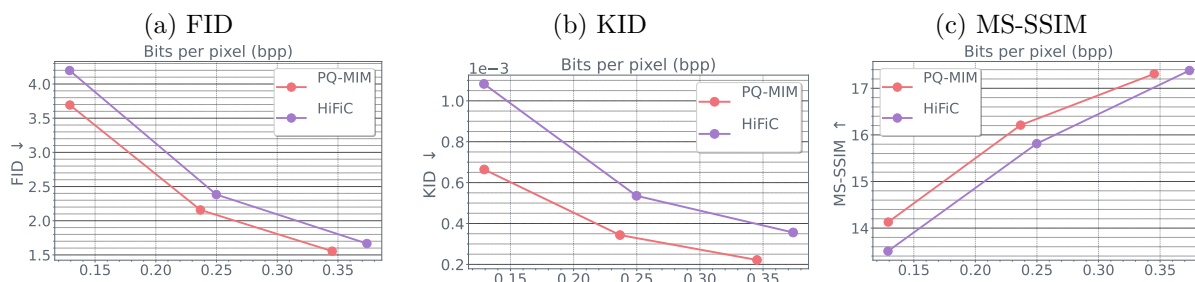


Figure 5.6 – **Perceptual training and evaluation using CLIC 2020 test-set.** Performance of our adversarially trained PQ-MIM w.r.t perceptual metrics compared to HiFiC [Mentzer, 2020]. PQ-MIM provides a stronger performance on FID, KID and MS-SSIM across all reported operating points.

dimensional look-up vector (dim=8 per sub-vector) followed by ℓ_2 normalization, following Yu *et al.* [Yu, 2021]. We train our model using ImageNet [Deng, 2009a] for 50 epochs with a batch size of 256. We use an AdamW [Loshchilov, 2017] optimizer with a peak learning rate of 1.10^{-3} , weight decay of 0.02 and $\beta_2 = 0.95$. We apply a linear warmup for the first 5 epochs of training followed by a cosine decay schedule for the remaining 45 epochs to a minimum learning rate of 5.10^{-5} . Unless mentioned otherwise, for all experiments, the encoder and decoder use an XCiT-L6 with 6 layers and a hidden dimension of 768. We use sinusoidal positional embedding [Vaswani, 2017a] such that our model can flexibly operate on variable-sized images.

For the results reported in Figure 5.5, the models are trained using solely MSE ($\eta = 0.5$) or MS-SSIM ($\eta = 10.0$) distortion losses for their corresponding plots. As for the models trained with perceptual objectives (Figure 5.6 and Table 5.1), they are trained with a weighted sum of MSE, LPIPS ($\alpha = 1$) and adversarial loss ($\gamma = 0.1$). We use a ProjectedGAN Discriminator [Sauer, 2021] architecture. The perceptual training is initialized with an MSE only trained checkpoint and trained for 50 epochs using ImageNet with a learning rate of 10^{-4} and weight decay of 5.10^{-5} . Similar to HiFiC [Mentzer, 2020], we freeze the encoder during the perceptual training. Additionally, we find that clipping the gradient norm to a maximum value of 4.0 improves the training stability. Our discriminator takes only the decoded image as input and does not rely on any other conditional signal.

MIM implementation details. Our MIM module is an XCiT-L12 with 12 layers and embedding dimension of 768. MIM is trained simultaneously with the PQ-VAE, but the gradients are not backpropagated to the encoder or the quantizer parameters. The PQ indices q are split into inputs and targets for the MIM model as defined by the quincunx partitioning pattern. By default we use $S = 5$ stages. All stages are processed with the same MIM model. The masked patches are replaced by a learnable “mask” token embedding. The loss is computed only for the masked patches. Since every token is assigned M PQ indices, the output of the MIM transformer is passed to M separate linear heads to predict M softmax normalized distributions over their corresponding codebooks. The marginal distribution of the codebook is computed as a normalized histogram over the ImageNet training set. For entropy coding, we use the implementation of the `torchac`¹ arithmetic coder.

1. <https://github.com/fab-jul/torchac>

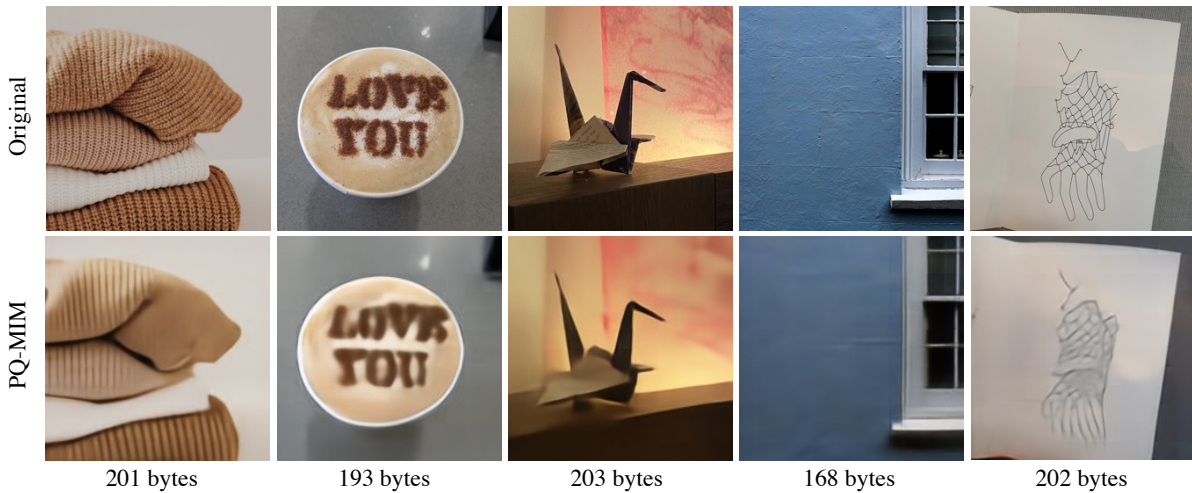


Figure 5.7 – **Extreme Image Compression.** PQ-MIM exhibits non-trivial compression performance at the extreme compression regime (e.g. 0.03 bpp), leading to compressed image codes that can fit in a short tweet (280 characters).

Datasets. We train our models using ImageNet [Deng, 2009a]. For data augmentation, we apply random resized cropping to 256×256 images and horizontal flipping. For evaluation and comparison to prior work, we use KODAK [Kodak, 1993] and TECNICK [Asuni, 2014] datasets for PSNR and MS-SSIM. Moreover, we compute the perceptual metrics (FID [Heusel, 2017], KID [Bińkowski, 2018]) for perceptually trained models using the CLIC 2020 test-set [Toderici, 2020] (428 images) using the same patch cropping scheme detailed by [Mentzer, 2020].

Baselines. We compare to several existing neural compression baselines: the scale hyperprior model [Ballé, 2018], mean & scale hyperprior [Minnen, 2018], GMM hyperprior [Cheng, 2020], and HiFiC [Mentzer, 2020]. Among the non-neural codecs, we compare to the popular BPG [Bellard,], WebP, and JPEG2000.

5.4.2 Main experimental results

Comparison to existing (neural) codecs. We compare PQ-MIM to other approaches across a wide range of bitrates in Figure 5.5². Note that in our evaluation we consider bitrates that are an order of magnitude lower than what is typically studied in the literature: most previous studies were limited to 0.1 bpp and above, see *e.g.* [Ballé, 2018; Cheng, 2020; Minnen, 2018]. The extremely low bitrates we consider make it possible to transmit a 256×256 image in an SMS or a tweet (280 characters)³ as shown in Figure 5.7. PQ-MIM achieves a strong and competitive performance for both KODAK and TECNICK datasets, outperforming all prior neural and standard codecs with the exception of GMM hyperprior [Cheng, 2020]. We observe that PQ-MIM is particularly strong for low rates, making it a good fit for extreme compression scenarios. Moreover, PQ-MIM exhibits a particularly strong MS-SSIM performance which was designed to model the human visual contrast perception [Wang, 2004; Wang, 2003].

2. All results for the baselines are reported using the authors’ official repositories

3. For example, a bitrate of 0.03 yields $256 \times 256 \times 0.03/8 \approx 246$ bytes for a 256×256 image.

Table 5.1 – **Discriminator Architecture.** We investigate multiple discriminators including StyleGAN [Karras, 2019], ProjectedGAN [Sauer, 2021] and UNet [Schönfeld, 2020].

Discriminator	FID ↓	KID ↓	MS-SSIM ↑
None	26.1	1.2×10^{-2}	15.3
StyleGAN	3.57	4.8×10^{-4}	13.3
ProjectedGAN	3.69	6.6×10^{-4}	14.1
UNet	3.87	6.7×10^{-4}	13.9

Table 5.2 – **Different MIM masking policies.** PQ-MIM with quincunx pattern with 5 steps reduces the bpp significantly (27%). Additionally, PQ-MIM is orders of magnitude cheaper in terms of FLOPs compared to an autoregressive raster order masking pattern.

Masking policy	#steps	bpp	MACs/Pixel (M)
Marginal baseline	1	0.512	0.69
Raster order	T	00M	24.4×10^3
Quincunx	5	0.373	6.66

Perceptual metrics comparison. In Figure 5.6, we compare PQ-MIM to HiFiC [Mentzer, 2020] in perceptual quality measures like FID and KID as well as MS-SSIM. HiFiC is based on the mean & scale hyperprior model [Minnen, 2018], but adds an adversarially trained discriminator model to improve the perceptual quality of the image reconstructions. PQ-MIM, with perceptual training, achieves a strong performance for all reported metrics, outperforming HiFiC for all operating points.

5.4.3 Analysis and Ablations

Model size and architecture. In Figure 5.8, we analyze the effect of using XCiT of different capacities for the autoencoder with respect to rate-distortion trade-off. We observe that the performance improves with higher capacity autoencoders, but there is a diminishing return with a further increase in capacity. For all our experiments we use an XCiT-L6 since it achieves the best performance.

Masking patterns. In Table 5.2, we compare to predicting tokens one-by-one autoregressively in a raster-scan order, the same pattern used in VQ-VAE based generative image models such as DALL-E [Ramesh, 2021a] and VQ-GAN [Esser, 2021]. In contrast to PQ-MIM, raster-scan models require causal attention, which makes XCiT not a good fit. We use a standard ViT model instead. However, due to the quadratic complexity of ViT and the high resolution of images typically used for evaluation of compression method (e.g. TECNICK), our autoregressive variant consistently exceeded the memory limits, even when using A100 GPUs with 40GB memory. Moreover, raster scan fully-autoregressive models results in extremely expensive FLOP count since it needs T separate forward passes per image. On the other hand, our stage-wise MIM with quincunx pattern requires 4 evaluations, and does not scale with the image resolution as is the case for raster, making it a more practical solution.

Number of prediction stages. We compare the quincunx masking pattern with masking based on the confidence score following MaskGiT [Chang, 2022] patch selection procedure for image generation. In the latter case, at a given step, we pick the patches to transmit dynamically based on their confidence score. The confidence is defined as the maximum across the probabilities assigned over the vocabulary by the model.⁴ For the quincunx and confidence-based

4. Note that for decoding the same confidence score can be used to identify the group of tokens to decode.

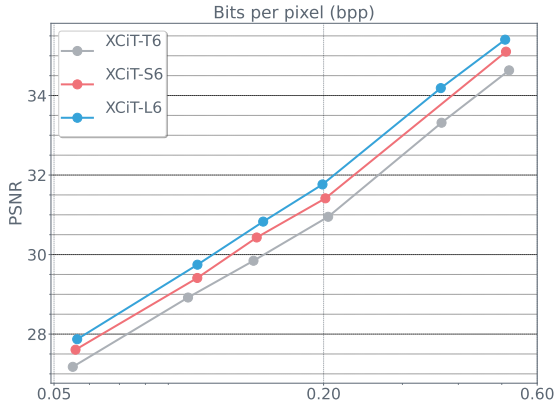


Figure 5.8 – **Autoencoder capacity.** The RD performance of different encoder/decoder capacities. We use the same model trunk for (i) the encoder; (ii) the PQ-MIM entropy model ; and (iii) the decoder. Increasing the model size from an XCiT-T6 model (3.5M params) to XCiT-L6 (47M params) increases the performance by typically +0.8dB.

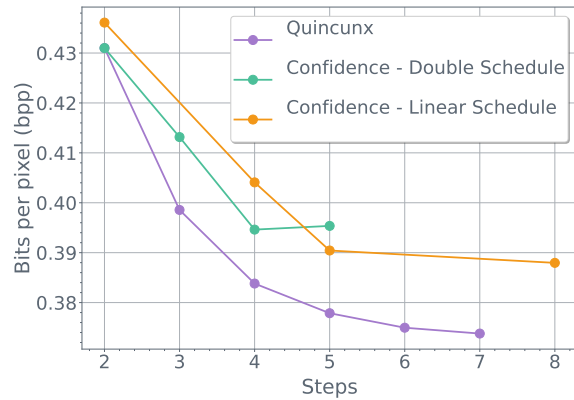


Figure 5.9 – **MIM number of steps.** In addition to quincunx, we explore using the prediction confidence as the patch masking policy following MaskGiT [Chang, 2022]. We test a linear and doubling schedules. Quincunx provides a higher rate saving, even compared to a confidence policy with a longer schedule.

masking policy we use the same 5-stage scheme, in which the number of tokens in subsequent groups doubles in size. We ablate the number of prediction stages S for the quincunx and confidence-based sampling. For the latter we consider two options: (i) a linear scheme where each group of tokens contains (approximately) the same number of tokens; and (ii) a doubling scheme where each subsequent group of tokens is double the size of the previous group, as is also used for the quincunx pattern. Every point on the curves in Figure 5.9 corresponds to the bitrate when encoding/decoding with a given number of steps. For example, $S = 2$ steps means we have only two steps each encoding/decoding 50% of the patches. For the 3-steps doubling schedule the groups have sizes of 1/4, 1/4, 1/2, and so on. For all three patterns the bitrate monotonically decreases with the number of steps, as more tokens can be predicted from larger contexts.

On the one hand, using confidence-based masking patterns, doubling and linear schemes lead to mostly comparable bitrates for the same number of steps, with further improvement for linear with more steps however with diminishing return. On the other hand, quincunx provides a stronger reduction ratio with strictly 5 steps, even when compared with confidence-based masking with a higher number of steps.

Predicting missing patches. When encoding/decoding the discrete image representation \mathbf{q} , we reduce the bitrate with our models by their ability to predict the remaining tokens given those of preceding stages. To illustrate the predictive ability of our model, we consider an experiment where we remove a subset of the tokens (sampled randomly), and use our model to fill in the missing patches by conditioning on the observed ones. We then use the PQ-VAE decoder to decode the discrete latent codes including the filled-in ones. In Figure 5.10 we show the results obtained when removing 10%, 20%, 30% and 50% of the patches. MIM can model

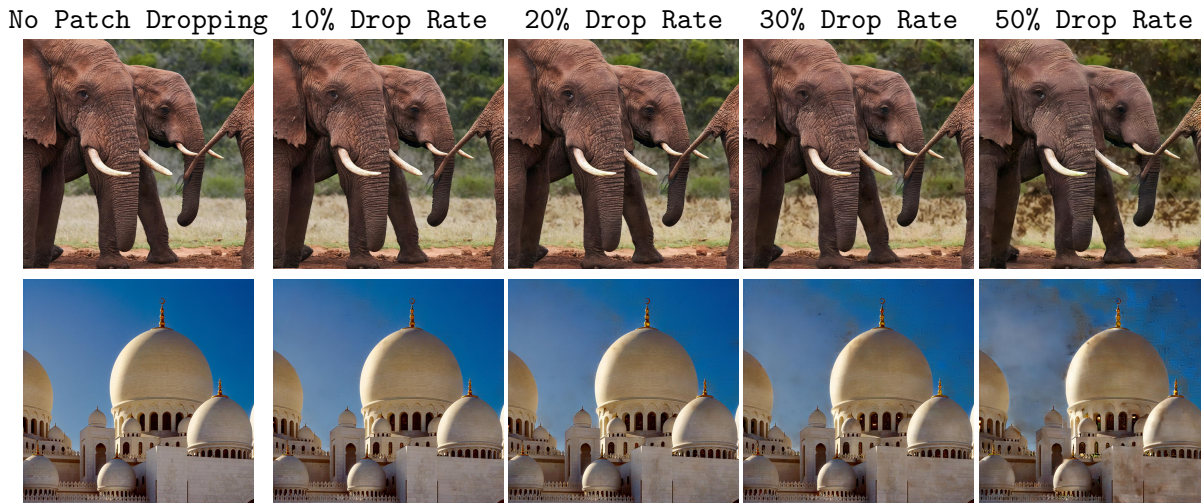


Figure 5.10 – **PQ-MIM can operate on a partial set of transmitted tokens.** Since PQ-MIM is compatible with generation, the conditional entropy model can be repurposed to predict the PQ codes for missing parts in an image. We show results for different dropping rates of transmitted tokens. PQ-MIM exhibits strong inpainting abilities. Even for the extreme case where half of the image patches are dropped, PQ-MIM can still retain a large percentage of the original image structure and details.

redundancies among the patches, which corroborates our findings of its compression ability.

5.4.4 Limitations

Image compression, and compression of visual data in general, is an important technology to scale the distribution of visual data. This is ever more important to cope with the growing quantity of visual data that is streamed in the form of video and for augmented and virtual reality applications. Compression is also critical to allow users with low-bandwidth connections to benefit from applications relying on sharing of image, video, or virtual reality data.

Caveats of learned neural compression models: biases and performance. As with any machine learning model, potential biases in the training data may be transferred into the model via training. In our setting, this can affect the autoencoding reconstruction abilities for content under-represented in the training data, as well as the compression abilities of the model for such data. Such biases should be assessed before the deployment of the model. Beyond rate-distortion trade-offs, important evaluation dimensions include the energy and latency performance of compression models. Current neural compression methods, including ours, need to be further optimized to be competitive with existing codecs on these criteria.

Specific limitations of our approach. In our work, we specifically focus on a high compression regime, with compression rate lower than 0.6 bit per pixel. This is a favorable case for our approach: it benefits from the generative model capability inherited from the VQ-VAE. Compared to VQ, PQ allows for higher rates, but it becomes comparatively less effective when increasing the number of subquantizers, as one can deduce from the slopes of the rate-distortion curves in Fig. 5.5. It is also computationally more expensive.

5.5 Conclusion

In this chapter, we have revisited vector quantization for neural image compression. We introduced a product-quantization variants of VQ-VAE and shown that it has better scaling properties in terms of bit-rate. Additionally, we introduced a novel conditional entropy model based on masked image modeling. We have shown that combined with the quincunx partitioning pattern, PQ-MIM provides a strong reduction in bit-rate. PQ-MIM exhibits a competitive performance for PSNR and MS-SSIM metric compared to strong neural compression baselines. Furthermore, when we train PQ-MIM using perceptual losses, it provides a strong performance on multiple metrics (e.g. FID, KID and MS-SSIM) compared the strong baseline of HiFiC. Finally, we have shown that PQ-MIM can operate in a hybrid compression/generation mode where it can fill the gaps for non-transmitted patches.



(a) Original



(c) PQ-MIM (bpp=0.165)



(b) Original



(d) PQ-MIM (bpp=0.124)

Figure 5.11 – **Qualitative examples** More qualitative examples for image compression of high resolution images with PQ-MIM.

Chapter 6

Learning a shared embedding space for six modalities with ImageBind

Objectives

We present IMAGEBIND, an approach to learn a joint embedding across six different modalities - images, text, audio, depth, thermal, and IMU data. We show that all combinations of paired data are not necessary to train such a joint embedding, and only image-paired data is sufficient to bind the modalities together. IMAGEBIND can leverage recent large scale vision-language models, and extends their zero-shot capabilities to new modalities just by using their natural pairing with images. It enables novel emergent applications ‘out-of-the-box’ including cross-modal retrieval, composing modalities with arithmetic, cross-modal detection and generation. The emergent capabilities improve with the strength of the image encoder and we set a new state-of-the-art on emergent zero-shot recognition tasks across modalities, outperforming specialist supervised models. Finally, we show strong few-shot recognition results outperforming prior work, and that IMAGEBIND serves as a new way to evaluate vision models for visual and non-visual tasks.

Contents

6.1	Introduction	71
6.2	Related Work	72
6.3	Method	73
6.3.1	Preliminaries	73
6.3.2	Binding modalities with images	74
6.3.3	Implementation Details	75
6.4	Experiments	76
6.4.1	Emergent zero-shot classification	77
6.4.2	Comparison to prior work	78
6.4.3	Few-shot classification	79
6.4.4	Analysis and Applications	80
6.4.5	Pretraining details	82
6.4.6	Datasets and Metrics	82
6.5	Ablation Study	84
6.5.1	Scaling the Image Encoder	85
6.5.2	Training Loss and Architecture	85
6.6	Discussion and Limitations	88

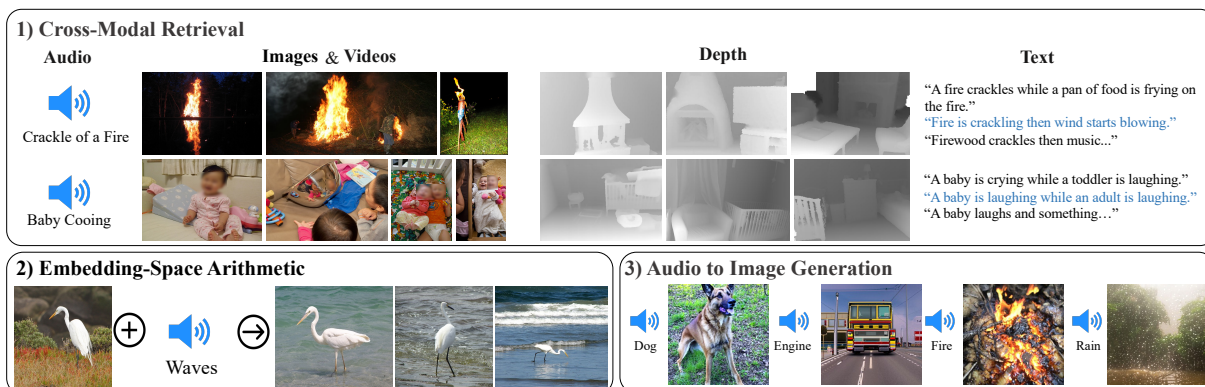


Figure 6.1 – **ImageBind’s joint embedding space enables novel multimodal capabilities.** By aligning six modalities’ embedding into a common space, IMAGEBIND enables: **1)** Cross-Modal Retrieval, which shows *emergent* alignment of modalities such as audio, depth or text, that aren’t observed together. **2)** Adding embeddings from different modalities naturally composes their semantics. And **3)** Audio-to-Image generation, by using our audio embeddings with a pre-trained DALLE-2 [Ramesh, 2022b] decoder designed to work with CLIP text embeddings.

6.1 Introduction

A single image can bind together many experiences – an image of a beach can remind us of the sound of waves, the texture of the sand, a breeze, or even inspire a poem. This ‘binding’ property of images offers many sources of supervision to learn visual features, by aligning them with any of the sensory experiences associated with images. Ideally, for a single joint embedding space, visual features should be learned by aligning to all of these sensors. However, this requires acquiring all types and combinations of paired data with the same set of images, which is infeasible.

Recently, many methods learn image features aligned with text [Schuhmann, 2021; Radford, 2021; Mahajan, 2018; Yu, 2022; Yuan, 2021c; Jia, 2021a; Alayrac, 2022], audio [Morgado, 2021; Patrick, 2021; Owens, 2018; Bachmann, 2022; Tian, 2019; Arandjelovic, 2017] *etc.* These methods use a single pair of modalities or, at best, a few visual modalities. However, the final embeddings are limited to the pairs of modalities used for training. Thus, video-audio embeddings cannot directly be used for image-text tasks and vice versa. A major obstacle in learning true joint embedding is the absence of large quantities of multimodal data where all modalities are present together.

In this chapter, we present IMAGEBIND, which learns a single shared representation space by leveraging multiple types of image-paired data. It does not need datasets where all modalities co-occur with each other. Instead, we leverage the binding property of images and we show that just aligning each modality’s embedding to image embeddings leads to an emergent alignment across all of the modalities. In practice, IMAGEBIND leverages web-scale (image, text) paired data and combines it with naturally occurring paired data such as (video, audio), (image, depth) *etc.* to learn a single joint embedding space. This allows IMAGEBIND to implicitly align the text embeddings to other modalities such as audio, depth *etc.*, enabling zero-shot recognition capabilities on that modality without explicit semantic or textual pairing. Moreover, we show

that it can be initialized with large-scale vision-language models such as CLIP [Radford, 2021], thereby leveraging the rich image and text representations of these models. Thus, IMAGEBIND can be applied to a variety of different modalities and tasks with little training.

We use large-scale image-text paired data along with naturally paired ‘self-supervised’ data across four new modalities - audio, depth, thermal, and Inertial Measurement Unit (IMU) readings – and show strong emergent zero-shot classification and retrieval performance on tasks for each of these modalities. These emergent properties improve as the underlying image representation is made stronger. On the audio classification and retrieval benchmarks, IMAGEBIND’s emergent zero-shot classification matches or outperforms specialist models trained with direct audio-text supervision on benchmarks like ESC, Clotho, AudioCaps. IMAGEBIND representations also outperform specialist supervised models on few-shot evaluation benchmarks. Finally, we show that IMAGEBIND’s joint embeddings can be used for a wide variety of compositional tasks as illustrated in fig. 6.1, including cross-modal retrieval, combining embeddings via arithmetic, detecting audio sources in images, and generating images given audio input.

6.2 Related Work

IMAGEBIND builds upon several advances in vision-language, multimodal, and self-supervised research.

Language Image Pre-training. Training images jointly with linguistic signals like words or sentences has been shown to be an effective method for zero-shot, open-vocabulary recognition and text to image retrieval [Frome, 2013; Faghri, 2017; Socher, 2014; Kiros, 2014]. Language as supervision can further be used for learning strong video representations [Alayrac, 2020; Miech, 2019b; Miech, 2020]. Joulin et al. [Joulin, 2016] show that using large-scale image dataset with noisy captions yields strong visual features. Recently, CLIP [Radford, 2021], ALIGN [Jia, 2021a] and Florence [Yuan, 2021c] collect large collections of image and text pairs and train models to embed image and language inputs in a joint space using contrastive learning, exhibiting impressive zero-shot performance. CoCa [Yu, 2022] adds an image captioning objective on top of the contrastive loss for improved performance. Flamingo [Alayrac, 2022] handles arbitrarily interleaved images and texts, and achieves state of the art on many few-shot learning benchmarks. LiT [Zhai, 2022] adopts contrastive training for fine-tuning and observes freezing image encoders works the best. This prior line of works mostly considers image and text, while our work enables zero-shot recognition on multiple modalities.

Multi-Modal Learning. Our work binds multiple modality representations in a joint embedding space. Prior works explored joint training of multiple modalities in a supervised [Girdhar, 2022b; Likhoshesterov, 2021] or self-supervised contexts [Tian, 2019; Girdhar, 2022a; Wang, 2022; Morgado, 2021; Arandjelovic, 2017]. The success of image and language pre-training methods such as CLIP has inspired approaches that revisits learning deep semantic representations through matching other modalities with linguistic inputs. Various methods adapt CLIP to extract semantically strong video representations [Xue, 2022; Luo, 2021; Fang, 2021; Lin, 2022]. Most related to our method, Nagrani et al. [Nagrani, 2022] create a weakly-labeled dataset

for paired video-audio and captions that allows for training multi-modal video-audio encoder to match textual features resulting in strong audio and video retrieval and captioning performance. AudioCLIP [Guzhov, 2021] adds audio as an additional modality into a CLIP framework, enabling zero-shot audio classification. In contrast, IMAGEBIND does not require explicit paired data between all modalities and instead leverages image as a natural weak supervision for unifying modalities.

Feature Alignment Pre-trained CLIP models have been utilized as teachers to supervise other models due to the strength of its visual representations [Wei, 2022; Peng, 2022; Liu, 2022a]. Moreover, CLIP joint image and text embedding space has also been leveraged for a variety of zero-shot tasks like detection [Zhou, 2022; Gu, 2021], segmentation [Li, 2022], mesh animation [Youwang, 2022] *etc.* showing the power of joint embedding spaces. PointCLIP [Zhang, 2022] finds a pre-trained CLIP encoder can be used for 3D recognition by projecting a point cloud to a number of 2D depth map views, which in turn are encoded using CLIP visual encoder. In multilingual neural machine translation, a similar phenomenon to the emergence behavior of IMAGEBIND is commonly observed and utilized: if languages are trained in the same latent space through learned implicit bridging, translation can be done between language pairs on which no paired data is provided [Johnson, 2017; Lample, 2017].

6.3 Method

Our goal is to learn a single joint embedding space for all modalities by using images to bind them together. We align each modality’s embedding to image embeddings, such as text to image using web data and IMU to video using video data captured from egocentric cameras with IMU. We show that the resulting embedding space has a powerful emergent zero-shot behavior that automatically associates pairs of modalities without seeing any training data for that specific pair. We illustrate our approach in fig. 6.2.

6.3.1 Preliminaries

Aligning specific pairs of modalities. Contrastive learning [Hadsell, 2006] is a general technique for learning an embedding space by using pairs of related examples (positives) and unrelated examples (negatives). Using pairs of aligned observations, contrastive learning can align pairs of modalities such as (image, text) [Radford, 2021], (audio, text) [Guzhov, 2021], (image, depth) [Tian, 2019], (video, audio) [Morgado, 2021] *etc.* However, in each case, the joint embeddings are trained and evaluated using the same pairs of modalities. Thus, (video, audio) embeddings are not directly applicable for text-based tasks while (image, text) embeddings cannot be applied for audio tasks.

Zero-shot image classification using text prompts. CLIP [Radford, 2021] popularized a ‘zero-shot’ classification task based on an aligned (image, text) embedding space. This involves constructing a list of text descriptions that describe the classes in a dataset. An input image is classified based on its similarity to the text descriptions in the embedding space. Unlocking such zero-shot classification for other modalities requires specifically training using paired text data,

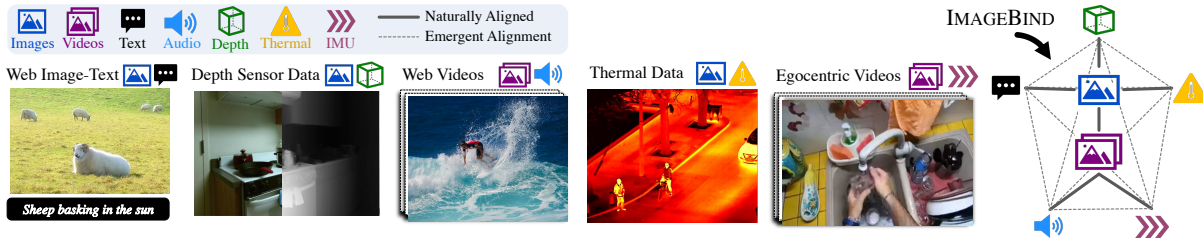


Figure 6.2 – **ImageBind overview**. Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, *etc.* IMAGEBIND links all these modalities in a common embedding space, enabling new emergent alignments and capabilities.

e.g., (audio, text) [Guzhov, 2021] or (point-clouds, text) [Zhang, 2022]. In contrast, IMAGEBIND unlocks zero-shot classification for modalities *without* paired text data.

6.3.2 Binding modalities with images

IMAGEBIND uses pairs of modalities $(\mathcal{I}, \mathcal{M})$, where \mathcal{I} represents images and \mathcal{M} is another modality, to learn a single joint embedding. We use large-scale web datasets with (image, text) pairings that span a wide range of semantic concepts. Additionally, we use the natural, self-supervised pairing of other modalities – audio, depth, thermal, and Inertial Measurement Unit (IMU) – with images.

Consider the pair of modalities $(\mathcal{I}, \mathcal{M})$ with aligned observations. Given an image \mathbf{I}_i and its corresponding observation in the other modality \mathbf{M}_i , we encode them into normalized embeddings: $\mathbf{q}_i = f(\mathbf{I}_i)$ and $\mathbf{k}_i = g(\mathbf{M}_i)$ where f, g are deep networks. The embeddings and the encoders are optimized using an InfoNCE [Oord, 2018] loss:

$$L_{\mathcal{I}, \mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau)}{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau) + \sum_{j \neq i} \exp(\mathbf{q}_i^\top \mathbf{k}_j / \tau)}, \quad (6.1)$$

where τ is a scalar temperature that controls the smoothness of the softmax distribution and j denotes unrelated observations, also called ‘negatives’. We follow [Wu, 2018b] and consider every example $j \neq i$ in the mini-batch to be a negative. The loss makes the embeddings \mathbf{q}_i and \mathbf{k}_i closer in the joint embedding space, and thus aligns \mathcal{I} and \mathcal{M} . In practice, we use a symmetric loss $L_{\mathcal{I}, \mathcal{M}} + L_{\mathcal{M}, \mathcal{I}}$.

Emergent alignment of unseen pairs of modalities. IMAGEBIND uses modalities paired with images, *i.e.*, pairs of the form $(\mathcal{I}, \mathcal{M})$ to align each the embeddings from each modality \mathcal{M} to those from images. We observe an emergent behavior in the embedding space that aligns two pairs of modalities $(\mathcal{M}_1, \mathcal{M}_2)$ even though we only train using the pairs $(\mathcal{I}, \mathcal{M}_1)$ and $(\mathcal{I}, \mathcal{M}_2)$. This behavior allows us to perform a wide variety of zero-shot and cross-modal retrieval tasks without training for them. We achieve state-of-the-art zero-shot text-audio classification results without observing a single sample of paired (audio, text).

6.3.3 Implementation Details

IMAGEBIND is conceptually simple and can be implemented in many different ways. We deliberately choose a vanilla implementation that is flexible and allows for an effective study and easy adoption. In [section 6.5](#), we present design decisions that are critical for good emergent ‘binding’.

Encoding modalities. We use a Transformer architecture [[Vaswani, 2017b](#)] for all the modality encoders. We use the Vision Transformer (ViT) [[Dosovitskiy, 2021a](#)] for images. Following [[Girdhar, 2022a](#)], we use the same encoder for images and videos. We temporally inflate [[Carreira, 2017](#)] the patch projection layer of the ViT and use 2 frame video clips sampled from 2 seconds. We follow [[Gong, 2021](#)] for encoding audio and convert a 2 second audio sampled at 16kHz into spectrograms using 128 mel-spectrogram bins. As the spectrogram is also a 2D signal like an image, we use a ViT with a patch size of 16 and stride 10. We treat thermal images and depth images as one-channel images and also use a ViT to encode them. We follow [[Girdhar, 2022b](#)] to convert depth into disparity maps for scale invariance. We extract the IMU signal consisting of accelerometer and gyroscope measurements across the X , Y , and Z axes. We use 5 second clips resulting in 2K time step IMU readings which are projected using a 1D convolution with a kernel size of 8. The resulting sequence is encoded using a Transformer. Finally, we follow the text encoder design from CLIP [[Radford, 2021](#)].

We use separate encoders for images, text, audio, thermal images, depth images, and IMU. We add a modality-specific linear projection head on each encoder to obtain a fixed size d dimensional embedding, that is normalized and used in the InfoNCE loss from [eq. \(6.1\)](#). In addition to ease of learning, this setup allows us to also initialize a subset of the encoders using pretrained models, *e.g.*, the image and text encoder using CLIP [[Radford, 2021](#)] or OpenCLIP [[Ilharco, 2021](#)].

6.3.3.1 Inference implementation details

Audio/Video: For both these temporal modalities (whether operated upon together during pre-training or separately during inference), we sample fixed length clips to operate on. During training, we randomly sample a clip, typically 2s in length. At inference time, we uniformly sample multiple clips to cover the full length of the input sample. For instance, for 5s ESC videos, we would sample $\lceil \frac{5}{2} \rceil = 3$ clips. For video clips, we sample a fixed number of frames from each clip. For audio, we process each raw audio waveform by sampling it at 16KHz followed by extracting a log mel spectrogram with 128 frequency bins using a 25ms Hamming window with hop length of 10ms. Hence, for a t second audio we get a $128 \times 100t$ dimensional input.

IMU: For IMU, we sample fixed length clips of 5 seconds, centered around time-stamps that are aligned with narrations. For each clip, we get a 6×2000 dimensional input and we measure the zero-shot performance for scenario classification using each clip as an independent testing sample.

Dataset	Task	#cls	Metric	#test
Audioset Audio-only (AS-A) [Gemmeke, 2017]	Audio cls.	527	mAP	19048
ESC 5-folds (ESC) [Piczak, 2015]	Audio cls.	50	Acc	400
Clotho (Clotho) [Font, 2013]	Retrieval	-	Recall	1045
AudioCaps (AudioCaps) [Kim, 2019]	Retrieval	-	Recall	796
VGGSound (VGGs) [Chen, 2020a]	Audio cls.	309	Acc	14073
SUN Depth-only (SUN-D) [Song, 2015]	Scene cls.	19	Acc	4660
NYU-v2 Depth-only (NYU-D) [Silberman, 2012]	Scene cls.	10	Acc	653
LLVIP (LLVIP) [Jia, 2021b]	Person cls.	2	Acc	15809
Ego4D (Ego4D) [Grauman, 2022]	Scenario cls.	108	Acc	68865

Table 6.1 – **Emergent zero-shot classification datasets** for **audio**, **depth**, **thermal**, and **Inertial Measurement Unit (IMU)** modalities. We evaluate IMAGEBIND without training for any of these tasks and without training on paired text data for these modalities. For each dataset, we report the task (classification or retrieval), number of classes (#cls), metric for evaluation (Accuracy or mean Average Precision), and the number of test samples (#test).







											
	IN1K	P365	K400	MSR-VTT	NYU-D	SUN-D	AS-A	VGGS	ESC	LLVIP	Ego4D
Random	0.1	0.27	0.25	0.1	10.0	5.26	0.62	0.32	2.75	50.0	0.9
IMAGEBIND	77.7	45.4	50.0	36.1	54.0	35.1	17.6	27.8	66.9	63.4	25.0
Text Paired	-	-	-	-	41.9*	25.4*	28.4 [†]	-	68.6 [†]	-	-
Absolute SOTA	91.0	60.7	89.9	57.7	76.7	64.9	49.6	52.5	97.0	-	-

Table 6.2 – **Emergent zero-shot classification** of IMAGEBIND using text prompts highlighted in blue. IMAGEBIND aligns images with text, depth, audio, thermal and IMU modalities. The resulting embedding space can associate text embeddings with the non-image modalities, and leads to strong emergent zero-shot classification. We show strong performance even on non-visual modalities such as audio and IMU. We compare to ‘Text Paired’ baselines wherever possible, which trains with paired text data for that modality. *We use the OpenCLIP ViT-H [Ilharco, 2021] on depth rendered as grayscale images. [†][Guzhov, 2021] that uses AS class names as supervision during training, and hence is not “zero-shot”. Overall, IMAGEBIND shows strong emergent zero-shot performance, even compared to such upper bounds. We also report the absolute state-of-the-art (SOTA) on each dataset for reference, which typically uses additional supervision, model ensembles *etc.* We report the top-1 classification accuracy for all datasets except MSR-VTT (Recall@1) and Audioset Audio-only (mAP).

6.4 Experiments

Naturally paired modalities and datasets. We use IMAGEBIND on six modalities - image/video, text, audio, depth, thermal images, and IMU. As described in section 6.3.3, we treat videos as 2 frame images and process them the same as images. For the naturally available paired data, we use the (video, audio) pairs from the Audioset dataset [Gemmeke, 2017], (image, depth) pairs from the SUN RGB-D dataset [Song, 2015], (image, thermal) pairs from the LLVIP dataset [Jia, 2021b] and (video, IMU) pairs from the Ego4D dataset [Grauman, 2022]. For these pairs of modalities, we do not use any extra supervision like class labels, text *etc.* Since SUN RGB-D and LLVIP are relatively small, we follow [Girdhar, 2022b] and replicate them 50× for training.

Large scale image-text pairs. We leverage image-text supervision from large-scale web data [Radford, 2021]. For ease of experimentation, we use pretrained models that are trained on billions of (image, text) pairs. Specifically, we use the pretrained vision (ViT-H 630M params) and text encoders (302M params) from OpenCLIP [Ilharco, 2021] in our experiments.

Encoders for each modality. We convert audio into 2D mel-spectrograms [Gong, 2021], and thermal and depth modalities into 1 channel images and use ViT-B, ViT-S encoders respectively. The image and text encoders are kept frozen during the IMAGEBIND training and the audio, depth, thermal, and IMU encoders are updated.

Emergent zero-shot vs. zero-shot. Methods such as CLIP [Radford, 2021], AudioCLIP [Guzhov, 2021] *etc.* train with modality pairs, (image, text) and (audio, text), to demonstrate zero-shot classification using text-prompts for the same modality. In contrast, IMAGEBIND binds modalities together using only image-paired data. Thus, just by training on (image, text) and (image, audio), IMAGEBIND can perform zero-shot classification of audio using text prompts. As we do not directly train for this ability, we term it *emergent* zero-shot classification to distinguish it from methods that specifically train using paired text-supervision for all modalities.

Evaluation on downstream tasks. We comprehensively evaluate IMAGEBIND on a many different downstream tasks using different protocols. We summarize the main datasets used for evaluation in table. 6.1.

6.4.1 Emergent zero-shot classification

We evaluate IMAGEBIND on emergent zero-shot classification and use the text prompt templates from [Radford, 2021]. We report the results in table. 6.2. Each task measures IMAGEBIND’s ability to associate text embeddings to the other modalities without observing them together during training. Given the novelty of our problem setting, there are no “fair” baselines to compare IMAGEBIND with. Nevertheless, we compare to prior work that uses text paired with certain modalities (*e.g.* audio [Nagrani, 2022; Guzhov, 2021]), and for certain “visual-like” modalities such as depth and thermal, we use the CLIP model directly. We also report the best reported supervised upper bound per benchmark.

IMAGEBIND achieves a high emergent zero-shot classification performance. On each benchmark, IMAGEBIND achieves strong gains and even compares favorably to supervised specialist models trained for the specific modality and task. These results demonstrate that IMAGEBIND aligns the modalities and implicitly transfers the text supervision associated with images to other modalities like audio. In particular, IMAGEBIND shows strong alignment for non-visual modalities like audio and IMU suggesting that their naturally available pairing with images is a powerful source of supervision. For completeness, we also report the standard zero-shot image (ImageNet [Russakovsky, 2015] - IN1K, Places-365 [Zhou, 2014] - P365) and video (Kinetics400 [Kay, 2017] - K400, MSR-VTT 1k-A [Xu, 2016] - MSR-VTT) tasks. As the image & text encoders are initialized (and frozen) using OpenCLIP, these results match those of OpenCLIP.

6.4.1.1 Zero-shot evaluation details

Query Templates. For all evaluations, we use the default set of templates from CLIP [Radford, 2021].¹ Note that we use the same templates for non visual modalities like audio and depth as

1. https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb

	Emergent	Clotho		AudioCaps		ESC
		R@1	R@10	R@1	R@10	Top-1
<i>Uses audio and text supervision</i>						
AudioCLIP [Guzhov, 2021]	✗	—	—	—	—	68.6
<i>Uses audio and text loss</i>						
AVFIC [Nagrani, 2022]	✗	3.0	17.5	8.7	37.7	—
<i>No audio and text supervision</i>						
IMAGEBIND	✓	6.0	28.4	9.3	42.3	66.9
<i>Supervised</i>						
AVFIC finetuned [Nagrani, 2022]	✗	8.4	38.6	—	—	—
ARNLQ [Oncescu, 2021]	✗	12.6	45.4	24.3	72.1	—

Table 6.3 – **Emergent zero-shot audio retrieval and classification.** We compare IMAGEBIND to prior work on zero-shot audio retrieval and audio classification. Without using audio-specific supervision, IMAGEBIND outperforms prior methods on zero-shot retrieval and has comparable performance on the classification task. IMAGEBIND’s emergent zero-shot performance approaches those of specialist supervised models.

	Modality	Emergent	MSR-VTT		
			R@1	R@5	R@10
MIL-NCE [Miech, 2019a]	V	✗	8.6	16.9	25.8
SupportSet [Patrick, 2020]	V	✗	10.4	22.2	30.0
FIT [Bain, 2021]	V	✗	15.4	33.6	44.1
AVFIC [Nagrani, 2022]	A+V	✗	19.4	39.5	50.3
IMAGEBIND	A	✓	6.8	18.5	27.2
IMAGEBIND	A+V	✗	36.8	61.8	70.0

Table 6.4 – **Zero-shot text based retrieval** on MSR-VTT 1K-A. We compare IMAGEBIND’s emergent retrieval performance using audio and observe that it performs favorably to methods that use the stronger video modality for retrieval.

well since we only use semantic/textual supervision associated with images.

6.4.2 Comparison to prior work

We now compare IMAGEBIND against prior work in zero-shot retrieval and classification tasks.

Zero-shot text to audio retrieval and classification. Unlike IMAGEBIND, prior work trains using paired data for that modality, *e.g.*, AudioCLIP [Guzhov, 2021] uses (audio, text) supervision and AVFIC [Nagrani, 2021] uses automatically mined (audio, text) pairs. We compare their zero-shot text to audio retrieval and classification performance to IMAGEBIND’s emergent retrieval and classification in table. 6.3.

IMAGEBIND significantly outperforms prior work on the audio text retrieval benchmarks. On the Clotho dataset, IMAGEBIND has double the performance of AVFIC despite not using any text pairing for audio during training. Compared to the supervised AudioCLIP model, IMAGEBIND achieves comparable audio classification performance on ESC. Note that AudioCLIP uses class names from AudioSet as text targets for audio-text training, hence is referred to as ‘supervised’.

IMAGEBIND’s strong performance on all three benchmarks validates its ability to align the audio and text modalities using images as a bridge.

Text to audio and video retrieval. We use the MSR-VTT 1k-A benchmark to evaluate the text to audio and video retrieval performance in [table. 6.4](#). Only using audio, IMAGEBIND achieves strong emergent retrieval performance compared to the video retrieval performance of prior work like MIL-NCE. The text to video performance for our model is strong (36.1% R@1 in [table. 6.2](#)) as it uses OpenCLIP’s vision and text encoders and outperforms many prior methods. However, combining the audio and video modalities further boosts performance showing the utility of IMAGEBIND’s features over an already strong retrieval model.

6.4.3 Few-shot classification

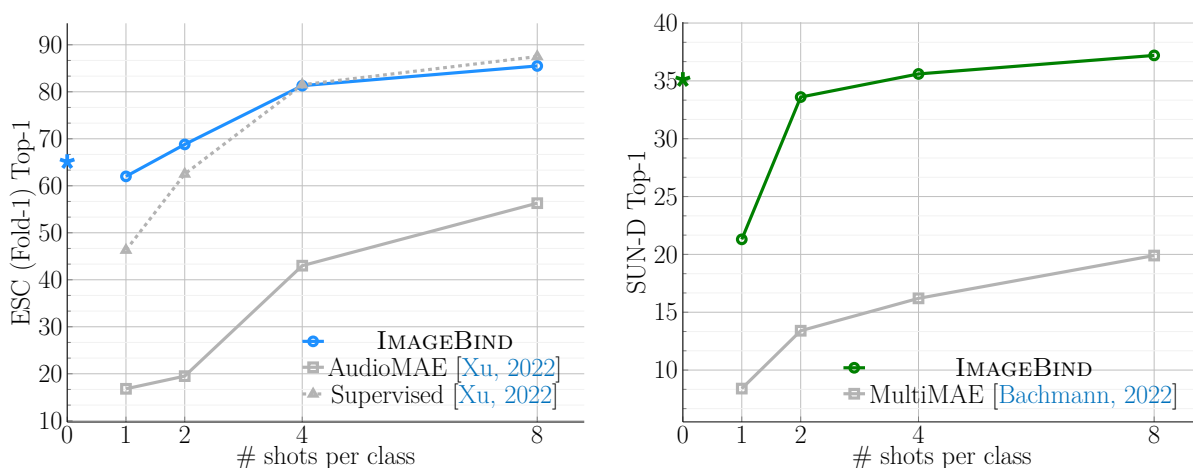


Figure 6.3 – **Few-shot classification on audio and depth.** We report the emergent zero-shot classification performance on each benchmark (denoted by \star). We train linear classifiers on fixed features for the ≥ 1 -shot case. **(Left)** In all settings, IMAGEBIND outperforms the self-supervised AudioMAE model. IMAGEBIND even outperforms a supervised AudioMAE model upto 4 shot learning showing its strong generalization. **(Right)** We compare with the MultiMAE model trained with images, depth, and semantic segmentation masks. IMAGEBIND outperforms MultiMAE across all few-shot settings on few-shot depth classification.

We now evaluate the label-efficiency of IMAGEBIND by evaluating on few-shot classification. We use the audio and depth encoders from IMAGEBIND and evaluate them on audio and depth classification respectively in [fig. 6.3](#). For ≥ 1 -shot results, we follow [[Radford, 2021](#); [Morgado, 2021](#)] and train linear classifiers on fixed features.

On few-shot audio classification ([fig. 6.3](#) left), we compare with (1) self-supervised AudioMAE model trained on audio from Audioset and (2) a supervised AudioMAE model finetuned on audio classification. Both baselines use the same capacity ViT-B audio encoder as IMAGEBIND. IMAGEBIND significantly outperforms the AudioMAE model on all settings with gains of $\sim 40\%$ accuracy in top-1 accuracy on ≤ 4 -shot classification. IMAGEBIND also matches or outperforms the supervised model on ≥ 1 -shot classification. IMAGEBIND’s emergent zero-shot performance surpasses the supervised ≤ 2 -shot performance.

For few-shot depth classification, we compare with the multimodal MultiMAE [[Bachmann,](#)

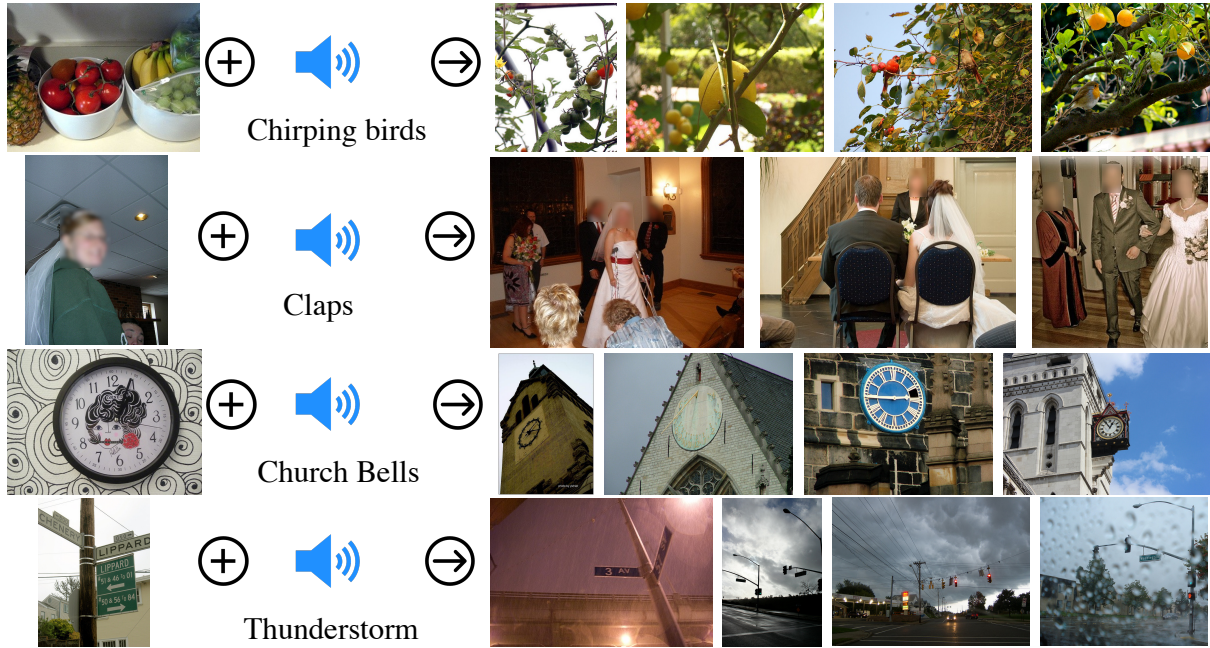


Figure 6.4 – **Embedding space arithmetic** where we add image and audio embeddings, and use them for image retrieval. The composed embeddings naturally capture semantics from different modalities. Embeddings from an image of fruits + the sound of birds retrieves images of birds surrounded by fruits.

2022] ViT-B/16 model trained on images, depth, and semantic segmentation data. IMAGEBIND significantly outperforms MultiMAE across all the few-shot settings. Altogether, these results show the strong generalization of IMAGEBIND audio and depth features trained with image alignment.

6.4.3.1 Few-shot evaluation details

For the few-shot results in Figures 6.3 using the ESC and SUN datasets, we sampled k training samples per class, where $k \in \{1, 2, 4, 8\}$. We fix the k samples such that our model and the baselines use exactly the same samples during training. For all few-shot evaluations, including the baselines, we freeze the encoder parameters and only train a linear classifier.

Audio: For audio few-shot training with ESC, our model and the baselines are trained using AdamW with a learning rate of 1.6×10^{-3} and weight decay of 0.05 for 50 epochs.

Depth: For depth few-shot training with SUN, our model and the baselines are trained using AdamW with a learning rate of 10^{-2} and no weight decay for 60 epochs.

6.4.4 Analysis and Applications

Multimodal embedding space arithmetic. We study whether IMAGEBIND’s embeddings can be used to compose information across modalities. In fig. 6.4, we show image retrievals obtained by adding together image and audio embeddings. The joint embedding space allows for us to compose two embeddings: *e.g.*, image of fruits on a table + sound of chirping birds and retrieve an image that contains both these concepts, *i.e.*, fruits on trees with birds. Such

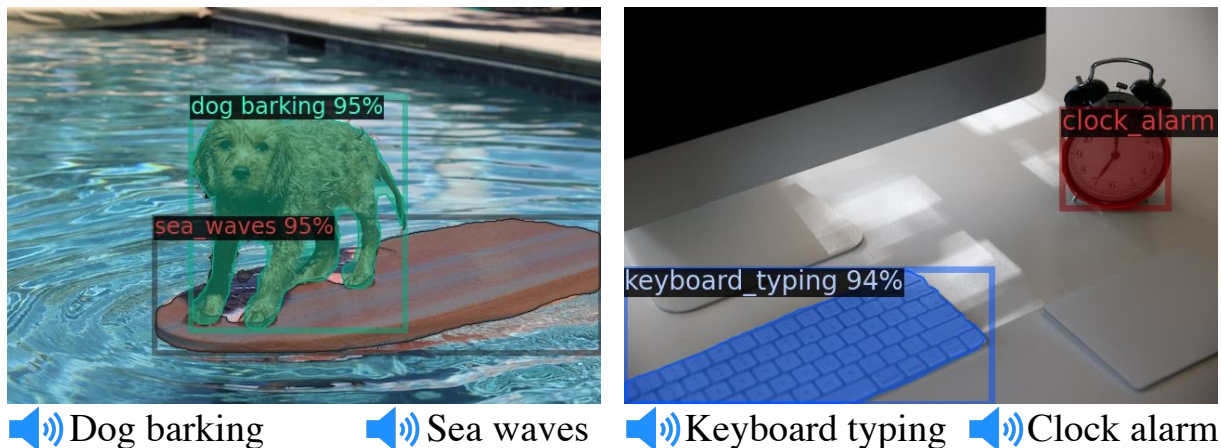


Figure 6.5 – **Object detection with audio queries.** Simply replacing Detic [Zhou, 2022]’s CLIP-based ‘class’ embeddings with our audio embeddings leads to an object detector promptable with audio. This requires no re-training of any model.

emergent compositionality whereby semantic content from different modalities can be composed will likely enable a rich variety of compositional tasks.

Without re-training, we can ‘upgrade’ existing vision models that use CLIP embeddings to use IMAGEBIND embeddings from other modalities such as audio.

Upgrading text-based detectors to audio-based. We use a pretrained text-based detection model, Detic [Zhou, 2022], and simply replace its CLIP-based ‘class’ (text) embeddings with IMAGEBIND’s audio embeddings. Without training, this creates an ‘audio’-based detector that can detect and segment objects based on audio prompts. As shown in fig. 6.5, we can prompt the detector with the barking sound of a dog to localize a dog.

Upgrading text-based diffusion models to audio-based. We use a pretrained DALLE-2 [Ramesh, 2022b] diffusion model (private reimplementation) and replace its prompt embeddings by our audio embeddings. In fig. 6.1, we observe that we can repurpose the diffusion model to generate plausible images using different types of sounds.

6.4.4.1 Qualitative evaluation details

Cross-modal nearest neighbors. We perform the retrieval on the embedding feature after temperature scaling. The nearest neighbors are computed using cosine distance. In fig. 6.1, we show retrievals for audio from ESC, image retrievals from IN1K and COCO, depth from SUN-D, and text from AudioCaps.

Embedding arithmetic. For arithmetic, we again use the embedding features after temperature scaling. We ℓ_2 normalize the features and sum the embeddings after scaling them by 0.5. We use the combined feature to perform nearest neighbor retrieval using cosine distance, as described above. In fig. 6.1, we show combination of images and audio from IN1K and ESC, and show retrievals from IN1K.

Audio→Image Generation. For generating images from audio clips, we rely on an in-house reproduced implementation of DALLE-2 [Ramesh, 2022b]. In DALLE-2, to produce images from

Config	AS	SUN	LLVIP	Ego4D
Vision encoder	ViT-Huge			
embedding dim.	768	384	768	512
number of heads	12	8	12	8
number of layers	12	12	12	6
Optimizer	AdamW			
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.95$			
Peak learning rate	1.6e-3	1.6e-3	5e-4	5e-4
Weight decay	0.2	0.2	0.05	0.5
Batch size	2048	512	512	512
Gradient clipping	1.0	1.0	5.0	1.0
Warmup epochs	2			
Sample replication	1.25	50	25	1.0
Total epochs	64	64	64	8
Stoch. Depth [Huang, 2016]	0.1	0.0	0.0	0.7
Temperature	0.05	0.2	0.1	0.2
Augmentations:				
RandomResizedCrop				
size	—	224px		—
interpolation	—	Bilinear	Bilinear	—
RandomHorizontalFlip	—	$p = 0.5$	$p = 0.5$	—
RandomErase	—	$p = 0.25$	$p = 0.25$	—
RandAugment	—	9/0.5	9/0.5	—
Color Jitter	—	0.4	0.4	—
Frequency masking	12	—	—	—

Table 6.5 – Pretraining hyperparameters

text prompts, the image generation model relies on text embeddings produced by the pre-trained CLIP-L/14 text encoder. Since IMAGEBIND naturally aligns CLIP’s-embedding space to that of other modalities proposed in the chapter, we can upgrade the DALLE-2 model to generate images by prompting it with these new unseen modalities. We achieve zero-shot audio to image generation with DALLE-2 by simply using the temperature-scaled audio embeddings generated by IMAGEBIND’s audio encoder as a proxy for the CLIP’s text embeddings in the DALLE-2’s image generation model.

Detecting objects using audio. We extract all audio descriptors from the validation set of ESC using an IMAGEBIND ViT-B/32 encoder, yielding 400 descriptors in total. We use an off-the-shelf CLIP-based Detic [Zhou, 2022] model and use the audio descriptors as the classifier for Detic in place of CLIP text-based ‘class’ embeddings. We use a score threshold of 0.9 for the qualitative results in Figure 6.5.

6.4.5 Pretraining details

In Table 6.5 we detail the hyperparameters used to pre-train each of the models reported in Table 6.2.

6.4.6 Datasets and Metrics

Audioset (AS) [Gemmeke, 2017]. This dataset is used for both training and evaluation. It contains 10s videos from YouTube annotated into 527 classes. It consists of 3 pre-defined splits,

the balanced split with about 20K videos, test split with 18K videos, and an unbalanced training split with about 2M videos. For **training**, we use the 2M unbalanced set without any labels, and only use it for audio-video matching. For **zero-shot evaluation** in table 6.2, we use the test set and compute logits for each class using the textual class names along with the templates as described later in section 6.4.1.1. The metric used is top-1 accuracy.

ESC-50 (ESC) [Piczak, 2015]. We use this dataset for evaluating the learned representations in a zero-shot manner. The task here is “Environmental Sound Classification” (ESC). It consists of 2000 5s audio clips classified into 50 classes. It has pre-defined 5 fold evaluation, each consisting of 400 test audio clips. In this work, we compute 0-shot predictions on the evaluation set for each fold and report the 5-fold average performance. For ablations we use only the first fold for computational ease. The metric used is top-1 accuracy.

Clotho (Clotho) [Font, 2013]. This is a dataset of audio from the Freesound platform with textual descriptions. It consists of a dev and test set of 2893 and 1045 audio clips respectively, with each clip associated with 5 descriptions. We consider the text→audio retrieval task, and consider each of the 5 associated captions as a separate test query and retrieve from the set of audio clips. The metric used is recall@ K , where a given test query is assumed to be correctly solved if the ground truth audio is retrieved within the top- K retrieved audio clips.

AudioCaps (AudioCaps) [Kim, 2019]. This is a dataset of audiovisual clips from YouTube accompanied by textual descriptions. It consists of clips from the Audioset dataset as described earlier. We use the splits as provided in [Oncescu, 2021],² which removes clips that overlap with the VGGSound dataset. We end up with 48198 training, 418 validation and 796 test clips. We only use the test set for zero-shot evaluation of our model. The task is text→audio retrieval, and evaluation is performed using recall@ K .

VGGSound (VGGS) [Chen, 2020a]. This dataset contains about 200K video clips of 10s length, annotated with 309 sound classes consisting of human actions, sound-emitting objects and human-object interactions. We only use the audio from the test set (with 14073 clips) for 0-shot classification. The evaluation is done using the top-1 accuracy metric.

SUN RGB-D (SUN). We use the registered RGB and Depth maps provided in the SUN RGB-D [Song, 2015] dataset **train** set (\sim 5K pairs) for training our model. We follow [Girdhar, 2022b] to post process the depth maps in two steps - 1) we use in-filled depth values and 2) convert them to disparity for scale normalization. This dataset is only used in training, so we do not use any metadata or class labels.

SUN Depth-only (SUN-D). We use only the \sim 5K depth maps from the **val** split of the SUN RGB-D [Song, 2015] dataset and denote them as SUN Depth-only. This dataset is only used for evaluation and we do not use the RGB images. We process the depth maps similar to SUN RGB-D (in-filled depth, converted to disparity). We use the 19 scene classes in the dataset and use their class names for constructing the zero-shot classification templates.

NYU-v2 Depth-only (NYU-D). We use the 794 **val** set depth maps from the NYU-v2 Depth-only [Silberman, 2012] dataset for evaluation only. We post-process the depth similar to

2. https://www.robots.ox.ac.uk/~vgg/research/audio-retrieval/resources/benchmark-files/AudioCaps_retrieval_dataset.tar.gz

SUN Depth-only. We use the 10 scene class names in the dataset. The 10th scene class, called ‘other’, correspond to 18 different semantic classes – [‘basement’, ‘cafe’, ‘computer lab’, ‘conference room’, ‘dINETTE’, ‘exercise room’, ‘foyer’, ‘furniture store’, ‘home storage’, ‘indoor balcony’, ‘laundry room’, ‘office kitchen’, ‘playroom’, ‘printer room’, ‘reception room’, ‘student lounge’, ‘study’, ‘study room’]. For zero-shot evaluation, we compute the cosine similarity of the 10th class as the maximum cosine similarity among these 18 classnames.

LLVIP (LLVIP). The LLVIP dataset [Jia, 2021b] consists of RGB image and Thermal (infrared low-light) image pairs. The dataset was collected in an outdoor setting using fixed cameras observing street scenes and contains RGB images taken in a low-light paired with infrared images (8~14um frequency). The RGB thermal pairs are registered in the dataset release. For training, we use the `train` set with 12025 RGB image and thermal pairs. For evaluation, we use the `val` set with 3463 pairs of RGB and thermal images. Since the original dataset is designed for detection, we post process it for a binary classification task. We crop out pedestrian bounding boxes and random bounding boxes (same aspect ratio and size as pedestrian) to create a balanced set of 15809 total boxes (7931 ‘person’ boxes). For zero-shot classification, we use the following class names for the ‘person’ class - [‘person’, ‘man’, ‘woman’, ‘people’], and [‘street’, ‘road’, ‘car’, ‘light’, ‘tree’] for the background class.

Ego4D (Ego4D) [Grauman, 2022]. For the Ego4D dataset, we consider the task of scenario classification. There are 108 unique scenarios present in the 9,645 videos of the Ego4D dataset. We filter out all videos annotated with more than one scenario which yields 7,485 videos with a single scenario assigned. For each video, We select all time-stamps that contains a synchronized IMU signal as well as aligned narrations. We sample 5 second clips around each time-stamp. The dataset is split randomly such that we have 510,142 clips for training, and 68,865 clips for testing. During training we only use the video frames and their corresponding IMU signal. We use the test split to measure zero-shot scenario classification performance, where each clip of IMU signal is assigned the video-level scenario label as its ground-truth.

6.5 Ablation Study

We investigate various design choices for learning a joint embedding space for different modalities. Since the ablation experimental setup is similar to [section 6.4](#), we only note the main differences (full details in [section 6.4.5](#)). We report results on the ESC fold-1 for the ablation study. We use a ViT-B encoder for the image, audio, depth, and thermal modalities by default and train them for 16 epochs (*vs.* 32 epochs in [section 6.4](#)). For IMU we use a lightweight 6 layer encoder with 512 dimensional width and 8 heads, and train it for 8 epochs. The text encoder follows [Radford, 2021] and is a twelve layer Transformer with a width of 512 dimensions. We initialize the image and text encoder from the CLIP model [Radford, 2021].

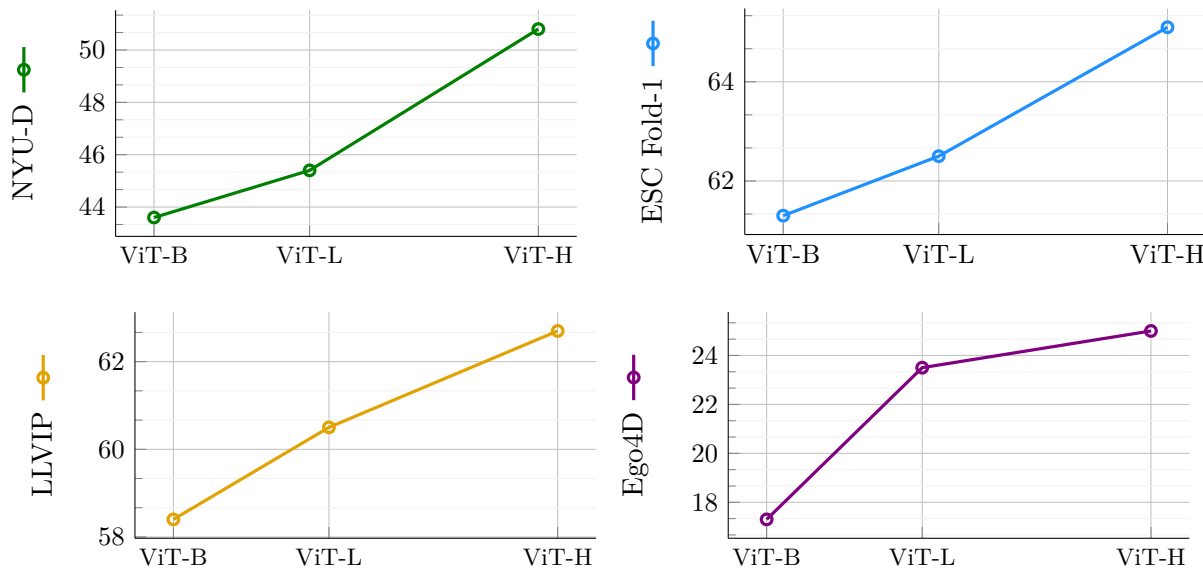


Figure 6.6 – **Scaling the image encoder** size while keeping the other modality encoders’ size fixed. We measure the performance on the emergent zero-shot classification of depth, audio, thermal, and IMU modalities. Scaling the image encoder significantly improves the zero-shot classification results suggesting that a stronger visual representation improves the ‘binding’ of modalities.

6.5.1 Scaling the Image Encoder

The central idea in IMAGEBIND is aligning the embeddings of all modalities to image embeddings. Thus, the image embeddings plays a central role in the emergent alignment of unseen modalities and we study their effect on the emergent zero-shot performance. We vary the size of the image encoder and train an encoder for the depth, audio *etc.* modalities to match the image representation. To isolate the effect of the image representation, we fix the size of the other modality encoders. We use the pretrained CLIP (ViT-B and ViT-L) and OpenCLIP (ViT-H) image and text encoders for this experiment. Our results in fig. 6.6 show that IMAGEBIND’s emergent zero-shot performance on all modalities improves with better visual features. For depth and audio classification, the stronger ViT-H *vs.* the ViT-B image encoder, provides a gain of 7% and 4% respectively. Thus, stronger visual features can improve recognition performance even on non-visual modalities.

6.5.2 Training Loss and Architecture

We study the effect of the training design choices on the emergent zero-shot classification. We focus on two modalities with different characteristics - depth which is visual and spatial, and audio which is non-visual and has a temporal component. We found that studying these diverse modalities led to robust and transferable design decisions.

Contrastive loss temperature. We study the effect of the temperature τ (eq. (6.1)) in table. 6.6a. We experiment with a learnable temperature initialized to 0.07 (parametrized in the log-scale) following [Radford, 2021] *vs.* various values of fixed temperatures. Unlike [Radford, 2021], we observe that a fixed temperature is best for depth, audio and IMU classification. Ad-

Temp →	Learn 0.05 0.07 0.2 1.0	Proj head →	Linear MLP	Batch size →	512 1k 2k 4k
SUN-D	24.1 27.0 27.3 26.7 28.0	SUN-D	26.7 26.5	NYU-D	47.3 46.5 43.0 39.9
ESC	54.8 56.7 52.4 45.4 24.3	ESC	56.7 51.0	ESC	39.4 53.9 56.7 53.9
(a) Temperature for loss.		(b) Projection Head.		(c) Batch size.	
Epochs →	16 32 64	Data aug →	Basic Strong	Spatial align →	None Aligned
SUN-D	26.7 27.9 29.9	SUN-D	25.4 26.7	SUN-D	16.0 26.7
ESC	56.7 61.3 62.9	ESC	56.7 22.6		
(d) Training epochs.		(e) Data aug for image.		(f) Spatial alignment of depth.	
Data aug →	None RandErase	Temporal align →	None Aligned	Data aug →	Basic +Freq mask
SUN-D	24.2 26.7	ESC	55.7 56.7	ESC	56.5 56.7
(g) Depth data aug.		(h) Temporal alignment of audio.		(i) Audio data aug.	

Table 6.6 – **Training loss and architecture** design decisions and their impact on emergent zero-shot classification. Settings for results in section 6.4 highlighted in gray. **(a)** A fixed temperature in the contrastive loss outperforms a learnable one for all modalities. **(b)** A linear projection head for computing the depth or audio embedding works better than an MLP head. **(c)** A smaller batch size works better for depth classification presumably because of the smaller size of (image, depth) datasets. **(d)** Longer training improves the zero-shot classification performance for both modalities. **(e)** Stronger image augmentation improves depth classification while basic augmentation significantly improves audio classification. **(f, g)** Using spatially aligned image and depth crops when training IMAGEBIND significantly improves performance. Similarly, RandErase augmentation is critical to good zero-shot classification on depth. **(h, i)** Temporally aligned audio and video matching gives improved performance and using frequency augmentation for audio gives a slight improvement.

ditionally, we see that a higher temperature is better for training the depth, thermal, and IMU encoders, whereas a lower temperature works best for the audio modality.

Projection head. We vary the projection head used for each encoder from a linear layer to an MLP with 768 hidden dimensions. The results in table. 6.6b show that a linear projection performs better for both modalities. This is in contrast to standard self-supervised methods like SimCLR [Chen, 2020c] whose performance improves with MLP projection heads.

Training epochs. We vary the number training epochs and report the classification performance in table. 6.6d. Longer training consistently improves the emergent zero-shot performance for both modalities across all datasets.

Data augmentation for paired images. During IMAGEBIND training, we augment images either using basic augmentation (cropping, color jitter) or strong augmentation that additionally applies RandAugment [Cubuk, 2020] and RandErase [Zhong, 2020]. We specify the augmentation parameters in section 6.4.5. Stronger augmentation helps depth classification when training on the small number of (image, depth) pairs from the SUN RGB-D dataset. However, for audio, strongly augmenting the video makes the task too challenging, leading to a significant drop of 34% on ESC.

Depth specific design choices. We vary the type of spatial crops used for training in table. 6.6f. Following CMC [Tian, 2019], we use two unaligned random crops from the corresponding image and depth pair *vs.* our default choice of using spatially aligned random crops. Contrary to CMC, we observe that random cropping severely degrades performance: more than 10% on

Image Encoder	Audio Encoder (ESC)		Depth Encoder (SUN)	
	ViT-S	ViT-B	ViT-S	ViT-B
ViT-B	52.8	56.7	30.7	26.7
ViT-H	54.8	60.3	33.3	29.5

Table 6.7 – **Capacity of the audio and depth encoders** and their impact on performance. A stronger image encoder improves performance for both audio and depth tasks. As the number of (image, depth) pairs is small, a smaller encoder improves performance for depth. For audio classification, a larger encoder is better.

	IN1K	ESC	SUN-D
Joint training	14.1	7.1	12.4
Update image only w/ text	35.4	11.8	18.7

Table 6.8 – **Different training procedures** for IMAGEBIND where we train *all encoders from scratch*. While joint training leads to worse performance, only updating the image encoder using the (image, text) loss, and jointly training the models is better.

SUN-D. Unlike vanilla self-supervised learning, our image representations learned from image-text pairs are more semantic and thus spatially misaligned crops hurt performance. In [table 6.6g](#), we observe that RandomErase [[Zhong, 2020](#)] boosts performance on depth classification.

Audio specific design choices. We train for video-audio alignment using temporally aligned samples or unaligned samples and measure the final performance in [table 6.6h](#). Similar to the depth classification observation, temporally aligned samples lead to better performance. [table 6.6i](#) shows that using frequency masking augmentation for audio also provides a small boost in performance.

Capacity of the audio and depth encoders and their impact of the classification performance is reported in [table 6.7](#). A smaller encoder for depth improves performance presumably because of the relatively small size of the (image, depth) dataset. Conversely, we observe that larger audio encoder improves the performance, particularly when paired with a high capacity image encoder.

Training procedure. We study the different choices in training procedure on images, text, depth, and audio. We jointly train all the encoders and do *not* initialize the image and text encoders. We use a subset of the image-text pairs from the LAION dataset [[Schuhmann, 2021](#)], and use the same datasets for depth and audio modalities as in [section 6.5.2](#). Full training details are in [section 6.4.5](#). We experiment with two settings in [table 6.8](#): (1) jointly training all encoders; and (2) stopping the gradient to the image encoder when aligning with audio and depth. Training the image representation solely from the text supervision (second setting) yields the best results not only for zero-shot image classification but also for the emergent zero-shot audio and depth classification.

ImageBind to evaluate pretrained vision models in [table 6.9](#). We initialize the vision encoder using a pretrained model and keep it fixed. We use image-paired data to align and train text, audio, and depth encoders. Compared to the supervised DeiT model, the self-supervised DINO model is better at emergent zero-shot classification on both depth and audio modalities. Moreover, the emergent zero-shot performance is not correlated with the pure vision performance

	IN1K	VGGS	ESC	SUN-D	NYU-D
DINO [Caron, 2021]	64.4	17.2	44.7	26.8	48.8
DeiT [Touvron, 2021b]	74.4 [†]	9.6	25.0	25.2	48.0

Table 6.9 – **ImageBind as an evaluation tool.** We initialize (and fix) the image encoder with different methods and align other modalities. IMAGEBIND measures the impact of visual features on multimodal tasks. [†] trained with IN1K supervision.

on ImageNet suggesting that these tasks measure different properties. IMAGEBIND can serve as a valuable tool to measure vision models’ strength on multimodal applications.

6.6 Discussion and Limitations

IMAGEBIND is a simple and practical way to train a joint embedding space using only image alignment. Our method leads to emergent alignment across all modalities which can be measured using cross-modal retrieval and text-based zero-shot tasks. We enable a rich set of compositional multimodal tasks across different modalities, show a way to evaluate pretrained vision models for non-vision tasks and ‘upgrade’ models like Detic and DALLE-2 to use audio. There are multiple ways to further improve IMAGEBIND. Our image alignment loss can be enriched by using other alignment data. Our embeddings are trained without a specific downstream task, and thus lag the performance of specialist models. More research into adapting general purpose embeddings for each task, including structured prediction tasks such as detection will be beneficial. Finally, new benchmarks, *e.g.* our emergent zero-shot task, to measure emergent abilities of multimodal models are essential to create new and exciting applications.

Chapter 7

Conclusion

Transformers are highly expressive and versatile, making them effective for various modalities. Initially, Transformers were primarily used for machine translation and NLP tasks, and it took a while for the computer vision research community to adopt them. However, the introduction of Vision Transformers by Dosovitskiy *et al.* [Dosovitskiy, 2021a] ignited a surge of research aimed at adapting these architectures to different vision tasks and addressing their limitations. In this manuscript, we delved into Vision Transformers and tackled a number of research questions relating to their design and application. First, we addressed their quadratic complexity limitation. Afterward, we explored the synergy between Masked Image Modeling and Transformers for self-supervised pre-training and image compression. Finally, we leveraged the universality of Transformers architecture to enable the learning of general-purpose representations for modalities with limited data. The methods we have proposed in this thesis have pushed the limit for state-of-the-art performance in a number of vision tasks and unlocked novel use cases by utilizing the power of Transformers. Below, we first re-discuss the challenges faced when first experimenting with Transformers, and how this has motivated some of our contributions. Then we will discuss some perspectives related to the topic of this thesis.

7.1 Summary of Contributions

7.1.1 Reducing the quadratic complexity of Vision Transformers

The quadratic complexity of Transformers limits their applicability to vision tasks that require high image resolutions. We have experienced this first-hand with our effort to improve image retrieval with Transformers [El-Nouby, 2021b]. We have observed that off-the-shelf Vision Transformers can provide superior performance to their CNN counterparts for retrieval applications with low-resolution images. However, for particular object retrieval, where images are typically high in resolution, using Transformers to extract descriptors was extremely challenging. This motivated us to tackle this shortcoming by proposing XCiT [El-Nouby, 2021c]. In Chapter 3, we introduced cross-covariance attention, a novel transposed view of self-attention that has linear complexity in terms of image resolution but quadratic complexity with respect to the model’s hidden dimension, which we can control. We then presented XCiT, a model architecture that relies on cross-covariance attention as its main building block. XCiT has a

simple columnar structure similar to a standard ViT model, with cross-covariance attention used as a drop-in replacement for self-attention. XCiT has several advantages over ViT and other efficient attention variants. It achieves the strongest reduction in memory consumption and latency while retaining performance on par or stronger than its counterparts. At the time it was published, XCiT is the first columnar attention-based architecture that achieved strong performance for dense tasks such as object detection and semantic segmentation. With XCiT, we can develop more efficient and accurate models for a wide range of vision tasks, addressing the limitations of traditional self-attention-based architectures.

7.1.2 Masked Image Modeling as a powerful tool

Masked language modeling has become a popular paradigm for learning strong representations for NLP tasks following the success of BERT [Devlin, 2018]. This has motivated works like BEiT [Bao, 2021] to adapt masked prediction to images. Therefore, Masked image modeling initially arose as a tool for learning visual representations with efforts like BEiT [Bao, 2021], MAE [He, 2021a], and SplitMask [El-Nouby, 2021a]. It has also been found as a useful tool for other tasks like image generation [Chang, 2022] and image compression [El-Nouby, 2023].

Self-supervised pre-training. In Chapter 4 we investigated different self-supervised pre-training paradigms such as contrastive and joint embedding methods. We delved deeper into denoising autoencoding, in particular with masking as the form of noise, using Transformers. Specifically, we proposed a novel self-supervised denoising objective called SplitMask, which splits each image into disjoint sets and operates in a more sample-efficient manner compared to other competing methods like BEiT. We conducted an in-depth analysis of the different properties of SplitMask and popular joint embedding methods, such as DINO, to demonstrate the advantages of our proposed method. First, we showed that masked image modeling approaches, in general, and SplitMask, in particular, have much higher sample efficiency during pre-training compared to DINO and supervised approaches. Second, we observed that SplitMask can effectively learn using more diverse types of visual data compared to joint embedding methods, which in turn have a strong bias towards object-centric images. Our findings indicate that novel use cases can be enabled by methods such as SplitMask. For example, companies or individuals can train strong visual models relying only on their private collection of data, eliminating the dependence on any third-party large-scale datasets that are typically required for successful pre-training. Our results show the utility of masked image modeling as a sample efficient objective. However, the flip perspective of our observations indicate that this family of methods does not exhibit a strong scaling behavior with respect dataset size, which stand in contrast with language models like GPT [Brown, 2020] where immensely scaling the data was a key element of their success.

Entropy Modeling. In Chapter 5, we presented an interesting application of masked image modeling in the context of neural image compression. Initially, we revisited vector quantization, demonstrating that while it can deliver satisfactory outcomes for extremely low bit-rates, the

vocabulary sizes exhibit exponential growth as we increase the permitted bit-rates. This renders vector quantization impractical except within a limited range of low bit-rates. In response to this challenge, we proposed product quantization as a viable alternative. Product quantization partitions the latent into multiple sub-vectors, each of which is quantized independently using a moderately-sized vocabulary. This approach significantly simplifies the optimization problem and yields an improved scaling behavior in relation to bit-rate. Subsequently, we turned our attention to the role of the entropy model in influencing the compression rate. We introduced PQ-MIM, a Transformer pre-trained for masked patch prediction, with an objective closely mirroring those discussed in Chapter 4. Given a specific masked patch, PQ-MIM can serve as a robust entropy model, estimating the assignment probability of each symbol in the vocabulary. This estimation progressively improves with increasing the number of observed patches used for conditioning. Hence, we introduced the Quincunx pattern where an image is segmented into five subsets, following a generalized checkerboard pattern, effectively doubling the number of patches included in each subset. We demonstrated that our final model which combines product quantization and our MIM module with a Quincunx pattern, achieves robust rate-distortion performance. This paves the way for extreme compression use cases, enabling the compression of an image into an ASCII code equivalent to the size of a tweet.

7.1.3 Learning representations for data-scarce modalities

Finally, Chapter 6 introduced ImageBind, a novel approach for achieving strong zero-shot and open-set recognition capabilities in multimodal models. Inspired by the success of contrastive vision and language models like CLIP, we investigate how to extend these abilities to other modalities, such as audio, thermal, depth, and IMU. We leverage existing semantic alignment between the visual and textual modalities and propose a novel approach for aligning new modalities to the visual modality. We utilize pre-trained vision-language models, such as CLIP, which have been trained on large-scale collections of image-text pairs covering a wide range of semantic concepts, to provide a supervisory signal to the other modalities. Since it is more feasible to collect aligned visual data, we align the representations of the new modalities to that of the vision-language model using a simple contrastive objective, such as InfoNCE. This approach allows us to demonstrate that the resulting features for the new modalities are highly semantic, leading to strong zero-shot performance. Additionally, we show that aligning all modalities in a single shared embedding space enables interesting applications for manipulating representations across modalities. For example, we showcase that descriptors from different modalities can be aggregated via simple vector arithmetic, enabling advanced retrieval applications. Furthermore, we demonstrate that ImageBind multimodal representations can power image generation using audio input as well as object detection with audio queries. Overall, ImageBind provides a powerful framework for achieving strong zero-shot and open-set recognition capabilities in multimodal models, opening up new avenues for research in this exciting field.

7.2 Open Problems and Future Directions

In this thesis, we have taken steps towards adapting and pushing the limits of Transformer architectures for various computer vision tasks as well as multimodality. While the community has done significant progress, many use cases and applications require further research to be unlocked. For example, we need more efficient attention for very long sequences as is the case for videos. Additionally, to further improve the visual representations, we require self-supervised approaches that are more scalable and general-purpose. Finally, taking multimodal learning to the next level by supporting fully multimodal generative models.

7.2.1 Efficient Attention for Video Data

Despite XCiT’s compelling performance and reduced complexity for high-resolution images, as discussed in Chapter 3, the model’s quadratic complexity in terms of hidden dimension poses challenges for scaling to larger models. Moreover, more efficient attention mechanisms is required for long-range video understanding and generation, where the sequence length can grow significantly compared to the high-resolution image scenario we focused on with XCiT.

Advancing Representation Learning in Videos. As evidenced by the success of joint embedding and masking-based methods such as SplitMask, self-supervised representation learning for images has seen significant advancements. The corresponding progress for video data has been slower. Videos intuitively contain a wealth of information about the world, including physics and inter-object relationships, yet unsupervised systems pre-trained on videos are yet to outperform those trained on static images. A possible reason is the limited context (16 or 32 frames) typically used to train video models due to computational limitations. To maximally exploit the informational richness of videos, we might need to consider longer contexts, which necessitates a significantly more efficient attention formulation and implementation.

Enhancing Video Generation. The field has witnessed a revolution in image generation [Ramesh, 2022a], which has been followed by promising models for video generation [Singer, 2022]. These video models, however, are not yet on par with their image counterparts. They heavily rely on interpolation techniques and separable spatial and temporal attention to circumvent the computational complexity of attending over long sequences of spatiotemporal patches. We believe for video generation, even more than representation learning, unlocking the expressive power of attention in exploiting long-range dependencies can be invaluable and it has the potential to significantly enhance the quality of generations.

7.2.2 Improving the scalability of Self-supervised Learning

In Chapter 4, we presented the efficacy of self-supervised methods such as SplitMask for the sample-efficient and robust pre-training of vision models. However, the denoising family of methods, including SplitMask, MAE [He, 2021b], and BEiT [Peng, 2022], often exhibit weak

linear probing performance, making finetuning generally necessary to reap the gains. This contrasts with joint embedding methods, which typically deliver strong off-the-shelf performance without finetuning, but they also present challenges in terms of low sample efficiency, bias towards object-centric images, and difficulty in scaling model parameters. A key research question is to identify a family of models that can enjoy the benefits of both methods.

Scaling in terms of Data. Denoising autoencoding methods, particularly SplitMask and MAE, show promising scalability in terms of model parameters due to their strategy of dropping masked patches during pre-training. However, as discussed in Chapter 4 and confirmed by Singh *et al.* [Singh, 2023], these methods do not demonstrate a clear benefit from additional data, which is a significant limitation. An open research question is how to modify the training objectives of these models to improve their scalability with respect to data, while retaining their sample efficiency and other favorable properties.

Enhancing Off-the-shelf Performance. Another limitation of denoising autoencoder methods is their relatively weak off-the-shelf performance, likely due to the generative nature of the training objective, which does not incentivize last-layer features to be highly discriminative, as is the case for joint embedding methods. With the rise of general-purpose large language models and interest in augmenting them with visual reasoning abilities, there is a growing need for unbiased visual features that are not specifically tuned for a fixed set of categories, yet are highly discriminative. Developing pre-training methods that overcome the drawbacks of both joint embedding and denoising methods is a challenging and open research question that warrants significant investigation.

Chapter 8

Résumé Substantiel

L'apprentissage profond a eu un impact considérable dans le domaine de la vision par ordinateur. Cet impact a révolutionné notre manière de percevoir, de comprendre et d'interagir avec le monde numérique. Les progrès rapides ont conduit au développement de nouveaux algorithmes et d'architectures qui ont continuellement fait avancer l'état de l'art dans les tâches de vision par ordinateur. Les avancées en matière d'apprentissage profond ont surpassé les méthodes classiques de vision par ordinateur et ont permis une large gamme d'applications telles que la conduite autonome, l'analyse de l'environnement et la gestion de l'information. Par exemple, la conduite autonome [Chen, 2015], l'imagerie et la vidéo [Ramesh, 2022a ; Singer, 2022], ainsi que l'analyse d'images médicales [Shin, 2016]. L'un des principaux facteurs qui ont contribué au succès de l'apprentissage profond dans le domaine de la vision par ordinateur a été l'innovation dans les architectures de réseau [Shin, 2016 ; He, 2016 ; Szegedy, 2015 ; Hu, 2018 ; Wang, 2018]. Ces innovations ont abouti à des modèles qui peuvent mieux apprendre et représenter des modèles complexes à partir de données à grande échelle. De plus, elles ont permis le développement de modèles capables d'apprendre des représentations hiérarchiques, qui sont cruciales pour comprendre la structure des données visuelles.

Les réseaux neuronaux convolutifs (CNN) [Fukushima, 1980 ; LeCun, 1989] sont depuis longtemps considérés comme l'architecture de référence pour les tâches de vision par ordinateur. Ils sont considérés comme l'architecture la mieux adaptée à ces tâches. Les CNN se caractérisent par leur biais inductif et leur efficacité d'échantillonnage, ce qui leur permet d'apprendre et de généraliser efficacement à partir d'un nombre limité d'exemples d'apprentissage. Leur capacité à apprendre les caractéristiques locales et à représenter de manière robuste les hiérarchies spatiales en fait l'architecture dominante pour les tâches de vision. Le succès des CNN peut être attribué à leur capacité à exploiter la structure spatiale inhérente aux images, ce qui leur permet d'apprendre efficacement des représentations significatives. Malgré leur efficacité, les mêmes biais inductifs et restrictions de conception qui rendent les CNN efficaces peuvent potentiellement limiter leur généralisation à mesure que la quantité de données d'apprentissage augmente. Par exemple, les CNN sont connus pour avoir un fort biais en faveur de la texture, ce qui peut limiter leur capacité à se généraliser à de nouveaux ensembles de données d'images diversifiées.

Le paysage de l'apprentissage profond a commencé à évoluer avec l'introduction de l'architecture Transformer [Vaswani, 2017b]. À l'origine conçue pour la traduction automatique et

les tâches de traitement du langage naturel (NLP), l'architecture Transformer s'est avérée très efficace pour capturer les dépendances à long terme et les modèles complexes dans les données séquentielles. L'innovation clé de cette architecture est le mécanisme d'auto-attention, qui permet au modèle d'évaluer dynamiquement l'importance des différents éléments d'entrée dans le contexte de la séquence entière. Le succès des Transformers dans les tâches de NLP a incité les chercheurs à explorer leur potentiel dans le domaine de la vision par ordinateur. Dans un premier temps, les mécanismes d'auto-attention ont été intégrés aux architectures CNN existantes, donnant ainsi naissance à des modèles hybrides [Wang, 2018] qui combinaient les avantages des deux approches. L'étape logique suivante de cette évolution a été le développement des transformateurs de vision (ViT) [Dosovitskiy, 2021a], qui s'éloignent considérablement des approches traditionnelles basées sur les CNN. Les modèles ViT sont entièrement composés de modules d'auto-attention et de perceptrons multicouches (MLP), éliminant complètement les couches convolutives. Malgré ce changement radical d'architecture, les modèles ViT ont obtenu des résultats de pointe dans une large gamme de tâches de vision par ordinateur, surpassant les performances de leurs homologues basés sur les CNN. De plus, les transformateurs de vision présentent des propriétés différentes de celles des CNN, notamment une plus grande robustesse face aux occlusions et aux perturbations, ainsi qu'un biais moins prononcé en faveur de la texture [Naseer, 2021]. Ces propriétés font des ViT des candidats prometteurs pour améliorer les performances de généralisation dans de nombreuses tâches visuelles.

Les transformateurs ont ouvert de vastes possibilités de recherche dans le domaine de la vision par ordinateur. L'expressivité et la généralité des modèles Transformer les rendent très adaptables à de nombreuses tâches et domaines. De plus, leur capacité à modéliser des dépendances à longue portée leur permet de capturer de manière plus efficace le contexte global et les relations complexes au sein des données visuelles, par rapport aux CNN. Dans cette thèse, nous plongeons plus profondément dans le monde des Transformers, explorant leur potentiel et leurs applications dans diverses tâches de vision par ordinateur. Nous examinerons les facteurs qui ont contribué à leur succès et chercherons des moyens d'améliorer leur adaptation aux tâches de vision par ordinateur. De plus, nous explorerons comment l'unification des architectures à travers les modalités peut conduire à des approches novatrices pour fournir une supervision aux modalités disposant de peu de ressources, en s'appuyant sur des modèles de vision solides.

8.1 Défis

Les transformateurs ouvrent effectivement de nouvelles possibilités d'apprentissage pour obtenir des représentations visuelles plus puissantes grâce à l'apprentissage supervisé et non supervisé. Ils permettent également de développer des approches plus homogènes pour l'apprentissage multimodal, c'est-à-dire l'apprentissage à partir de différentes modalités (par exemple, images et textes) de manière cohérente. Dans ce manuscrit, nous nous attaquons à plusieurs défis liés à ces sujets passionnants.

8.1.1 Complexité computationnelle

L'un des principaux défis posés par les transformateurs, tels qu'ils ont été introduits par VASWANI et al. [Vaswani, 2017b], est la complexité quadratique en fonction de la longueur de la séquence. Le mécanisme d'auto-attention utilisé par les transformateurs nécessite le calcul des interactions entre chaque paire d'éléments de la séquence d'entrée, ce qui se traduit par une complexité de l'ordre de $\mathcal{O}(N^2)$, où N est la longueur de la séquence. Dans le domaine de la vision par ordinateur, cette complexité pose un défi pour l'application directe des transformateurs aux images haute résolution, qui sont souvent nécessaires pour des tâches telles que la détection d'objets, la segmentation sémantique et la compression d'images. Pour surmonter cette limitation, il est nécessaire d'innover au niveau architectural afin de réduire le coût de calcul des transformateurs pour les tâches liées aux images haute résolution.

8.1.2 Apprentissage de la représentation à grande échelle et efficace en termes d'échantillons

Les transformateurs excellent dans la capture des dépendances à longue portée et peuvent apprendre efficacement les représentations hiérarchiques. À l'instar de leur impact sur l'apprentissage des représentations avec des approches telles que BERT [Devlin, 2018] et GPT [Radford, 2018], les transformateurs ont le potentiel de révolutionner le pré-entraînement auto-supervisé pour les tâches de vision. Cependant, pour réaliser pleinement ce potentiel, il est essentiel de développer des méthodes plus efficaces en termes d'échantillonnage qui peuvent être mises à l'échelle. Des méthodes plus efficaces en termes d'échantillons et qui peuvent s'adapter plus facilement en termes de données et de calcul par rapport aux méthodes conjointes existantes. Le pouvoir d'expression de l'architecture Transformer, lorsqu'elle est utilisée en conjonction avec des méthodes plus génériques et moins biaisées, est un atout majeur. Avec des objectifs de débruitage plus génériques et moins biaisés, elle a le potentiel de fournir l'efficacité et l'évolutivité nécessaires pour améliorer considérablement le pré-entraînement auto-supervisé des modèles de vision.

8.1.3 Apprentissage multimodal

En tant qu'architecture universelle, le modèle Transformer a été appliqué avec succès à diverses modalités. Cette polyvalence ouvre la voie à des principes de conception et à des composants communs qui peuvent faciliter l'apprentissage multimodal. En utilisant les transformateurs comme base pour l'apprentissage multimodal, nous pouvons développer des méthodes qui intègrent efficacement les informations provenant de diverses modalités et faciliter le transfert transparent de connaissances entre différentes tâches et domaines. Cependant, alors qu'il peut y avoir une abondance de données pour certaines modalités, d'autres peuvent souffrir d'une grave pénurie de données. Dans de tels cas, le transfert de connaissances rendu possible par l'utilisation d'un cadre commun tel que Transformers peut avoir un impact significatif sur la qualité des données. En tirant parti des différentes modalités et en transférant les connaissances acquises d'un domaine à l'autre, nous pouvons améliorer la performance des modèles dans des domaines

où la rareté des données est un problème majeur.

8.2 Présentation et contributions

Le manuscrit commence par une discussion sur le contexte et la littérature pertinente dans le chapitre 2. Ensuite, nous approfondissons les détails de chacune de nos quatre principales contributions dans les chapitres suivants, comme indiqué ci-dessous.

8.2.1 Maîtriser la complexité quadratique des transformateurs de vision

Dans le chapitre 3, nous étudions la complexité informatique des transformateurs de vision [Dosovitskiy, 2021a]. Plus précisément, nous constatons que l’opération d’auto-attention, qui est au cœur de l’architecture des transformateurs Transformer [Vaswani, 2017b], présente une croissance quadratique en fonction de la résolution de l’image. Cette complexité peut rendre les transformateurs de vision moins efficaces, en particulier lorsqu’il s’agit de traiter des images à haute résolution, comme c’est souvent le cas dans les applications de vision par ordinateur, telles que la segmentation sémantique et la détection d’objets. Pour relever ce défi, nous proposons une nouvelle formulation alternative de l’opération d’auto-attention, que nous appelons *Cross-Covariance Attention* (XCA) [El-Nouby, 2021c]. L’attention à covariance croisée présente une complexité linéaire par rapport à la taille de l’entrée et peut être intégrée de manière transparente en remplacement de l’auto-attention dans les transformateurs de vision. Nous présentons XCiT, une nouvelle architecture pour la vision par ordinateur qui utilise XCA comme élément central. Nous démontrons que XCiT offre des améliorations significatives en termes de mémoire et de débit tout en conservant les excellentes performances des transformateurs de vision.

8.2.2 Pré-entraînement auto-supervisé efficace avec Transformer

L’architecture Transformer a été un catalyseur pour l’innovation en matière de pré-entraînement auto-supervisé dans le domaine du traitement du langage naturel. Elle a conduit au développement de modèles influents tels que BERT [Devlin, 2018] et GPT [Radford, 2018]. Dans le domaine de la vision par ordinateur, des méthodes telles que [Chen, 2020c ; He, 2020 ; Caron, 2021] se sont imposées grâce à leur grande adaptabilité et à leurs performances exceptionnelles. Elles ont démontré leur compétitivité par rapport aux méthodes supervisées pour le fine-tuning des images. Cependant, les méthodes d’intégration conjointe présentent certaines limitations, notamment leur dépendance à l’égard de techniques d’augmentation de données conçues manuellement et adaptées à ImageNet [Deng, 2009b]. De plus, ces méthodes peuvent souffrir de problèmes de taille de modèle, comme l’a souligné [Chen, 2021b]. Motivés par le succès de BERT dans le domaine du NLP, dans le chapitre 4, nous étudions l’efficacité des méthodes de débruitage et d’auto-codage lorsqu’elles sont utilisées en conjonction avec des transformateurs de vision pour le pré-entraînement auto-supervisé. Tout d’abord, nous proposons SplitMask [El-Nouby, 2021a], une nouvelle méthode de pré-entraînement auto-supervisée basée sur la modélisation d’images masquées. De plus, nous constatons que SplitMask offre une meilleure efficacité d’échantillonnage et peut être utilisée pour apprendre des représentations robustes en utilisant des ensembles

de données de plusieurs ordres de grandeur plus petits que ceux requis par les méthodes d'intégration conjointe. En outre, nous observons que SplitMask est bien adapté à l'apprentissage avec une plus large gamme de données visuelles car il n'est pas biaisé par une distribution spécifique d'images, contrairement aux méthodes d'intégration conjointe qui sont généralement centrées sur l'objet d'ImageNet.

8.2.3 Modélisation d'images masquées pour une meilleure compression d'images

La compression neuronale des images s'est révélée être une alternative prometteuse aux codecs traditionnels en raison de sa qualité d'image supérieure sur le plan perceptif et psychovisuel. En général, les systèmes de compression d'images neuronales se composent de trois éléments principaux : un encodeur et un décodeur neuronaux profonds, qui apprennent à mapper l'image et sa représentation latente inverse ; un quantificateur, qui fait correspondre les représentations latentes continues à un ensemble de symboles discrets ; et un modèle d'entropie, qui exploite les redondances dans l'ensemble de symboles discrets pour réduire la longueur finale du flux binaire. Dans le chapitre 5, nous proposons plusieurs contributions. Tout d'abord, nous adoptons XCiT pour concevoir le codeur et le décodeur. Ensuite, nous utilisons la quantification par produit au lieu de la quantification vectorielle. Enfin, nous proposons un nouveau modèle d'entropie pour la compression d'images neuronales basé sur la modélisation d'images masquées. Notre méthode, PQ-MIM [El-Nouby, 2023], permet d'obtenir une forte réduction des débits par rapport aux méthodes basées sur la fréquence, tout en offrant des avantages supplémentaires par rapport aux méthodes autorégressives qui peuvent être prohibitives pour les images à haute résolution. Ainsi, PQ-MIM permet une compression extrême des images à la taille d'un tweet ou d'un SMS.

8.2.4 ImageBind pour l'apprentissage d'un espace d'intégration partagé pour six modalités

Les transformateurs ont émergé comme une architecture polyvalente offrant des performances exceptionnelles dans différentes modalités, telles que le texte [Devlin, 2018 ; Radford, 2018], les images [Dosovitskiy, 2021a ; Touvron, 2020], la vidéo [Gedas Bertasius, 2021 ; Tong, 2022], l'audio [Xu, 2022], et les graphes [Yun, 2019], entre autres. Cela ouvre des perspectives passionnantes pour le développement de systèmes multimodaux avec des composants partagés et des conceptions similaires. Cependant, l'approche prédominante pour former ces modalités repose sur l'apprentissage supervisé, qui consiste à apprendre à partir d'entrées sensorielles un ensemble d'étiquettes catégoriques. Malheureusement, l'apprentissage supervisé est limité par la difficulté de collecter efficacement et de manière évolutive des annotations. Pour surmonter ce défi, l'apprentissage faiblement supervisé est apparu comme un paradigme alternatif qui se base sur la collecte de grandes quantités de données avec des étiquettes bruitées, qui peuvent être plus faciles à acquérir. Ce paradigme a connu un immense succès dans l'apprentissage de représentations visuelles, alimenté par des collections massives de paires d'images et de textes provenant du web ouvert [Joulin, 2016 ; Radford, 2021 ; Zhai, 2022]. Cependant, la génération de collections similaires pour d'autres modalités telles que l'audio, les images thermiques et les

images de profondeur est beaucoup plus difficile. Dans le chapitre 6, nous abordons ce problème en introduisant ImageBind, une nouvelle méthode d'entraînement qui permet d'encoder six modalités différentes dans un espace latent partagé, en tirant parti des performances élevées des modèles de vision et de langage existants.

Le manuscrit se conclut par le chapitre 7, qui résume nos principales conclusions et idées, tout en abordant les limitations de nos approches et en proposant des pistes pour de futures recherches.

Appendix

A Additional results for XCiT

A.1 More XCiT models

We present additional results for our XCiT models in Table A.1. We include performance of 384×384 images using a 16×16 patch size as well as results for images with 224×224 resolution using patch size of 8×8 .

Models	Depth	d	#Blocks	#params	$P = (16 \times 16)$				$P = (8 \times 8)$			
					GFLOPs	@224	@224 Υ	@384 \uparrow	GFLOPs	@224	@224 Υ	@384 \uparrow
XCiT-N12	12	128	4	3M	0.5	69.9	72.2	75.4	2.1	73.8	76.3	77.8
XCiT-T12	12	192	4	7M	1.2	77.1	78.6	80.9	4.8	79.7	81.2	82.4
XCiT-T24	24	192	4	12M	2.3	79.4	80.4	82.6	9.2	81.9	82.6	83.7
XCiT-S12	12	384	8	26M	4.8	82.0	83.3	84.7	18.9	83.4	84.2	85.1
XCiT-S-24	24	384	8	48M	9.1	82.6	83.9	85.1	36.0	83.9	84.9	85.6
XCiT-M24	24	512	8	84M	16.2	82.7	84.3	85.4	63.9	83.7	85.1	85.8
XCiT-L24	24	768	16	189M	36.1	82.9	84.9	85.8	142.2	84.4	85.4	86.0

TABLE A.1 – ImageNet-1k top-1 accuracy of XCiT for additional combinations of image and patch sizes.

A.2 Image Retrieval

Context of this study. Vision-based retrieval tasks such as landmark or particular object retrieval have been dominated in the last years by methods extracting features from high-resolution images. Traditionally, the image description was obtained as the aggregation of local descriptors, like in VLAD [Jégou, 2012]. Most of the modern methods now rely on convolutional neural networks [Berman, 2019; Gordo, 2017; Tolias, 2016b]. In a recent paper, El-Nouby *et al.* [El-Nouby, 2021b] show promising results with vision transformers, however they also underline the inherent scalability limitation associated with the fact that ViT models do not scale well with image resolution. Therefore, it cannot compete with convolutional neural networks whose performance readily improve with higher resolution images. Our XCiT models do not suffer from this limitation : our models scale linearly with the number of pixels, like convnets, and therefore makes it possible to use off-the-shelf methods initially developed for retrieval with high-resolution images.

Datasets and evaluation measure In each benchmark, a set of query images is searched in a database of images and the performance is measured as the mean average precision.

The Holidays [Jégou, 2008] dataset contains images of 500 different objects or scenes. We use the version of the dataset where the orientation of images (portrait or landscape) has been corrected. Oxford [Philbin, 2007] is a dataset of building images, which corresponds to famous landmark in Oxford. A similar dataset has been produced for famous monuments in Paris and referred to as Paris6k [Chum, 2007].

Dataset	number of images		nb of instances
	database	queries	
Holidays	1491	500	500
R-Oxford	4993	70	26

TABLE A.2 – **The basic statistics on the image retrieval datasets.**

We use the revisited version of the Oxford benchmark [Radenović, 2018a], which breaks down the evaluation into easy, medium and hard categories. We report results on the "medium" and "hard" settings, as we observed that the ordering of techniques does not change under the easy measures.

Image representation : global and local description with XCiT We consider three existing methods to extract an image vector representations from the pre-trained XCiT models. Note that to the best of our knowledge, for the first time we extract local features from the output layer of a transformer layer, and treat them as patches fed to traditional state-of-the-art methods based on matching local descriptors or CNN.

CLS token. Similar to EL-NOUBY et al. [El-Nouby, 2021b] with ViT, we use the final vector as the image descriptor. In this context, the introduction of class-attention layers can be regarded as a way to learn the aggregation method.

VLAD. We treat the patches before the class-attention layers as individual local descriptors, and aggregate them into a higher-dimensional vector by employing the Vector of locally aggregated Descriptors [Jégou, 2012].

AMSK. We also apply the aggregated selective match kernel from Tolias *et al.* [Tolias, 2016a]. This method was originally introduced for local descriptors, but got adapted to convolutional networks. To the best of our knowledge this is the state of the art on several benchmarks [Tolias, 2020].

For all these methods, we use the models presented in our main paper, starting from the version fine-tuned at resolution 384×384 . By default the resolution is 768. This is comparable to the choice adopted in the literature for ResNet (e.g., 800 in the work by Berman et al. [Berman, 2019]).

Experimental setting : Image retrieval with models pretrained on Imagenet1k only

We only consider models pre-trained on Imagenet-1k. Note that the literature reports significant improvement when learning or fine-tuning networks [Radenović, 2018b; Tolias, 2020] on specialized datasets (e.g., of buildings for Oxford5k and Paris6k). We consider only XCiT-S12 models, since they have a number of parameters comparable to that of ResNet-50. We report the results in Table A.3.

Scaling resolution. As expected increasing the resolution with XCiT improves the performance steadily up to resolution 768. This shows that our models are very tolerant to resolution changes considering that they have been fine-tuned at resolution 384. The performance starts to saturates at resolution 1024, which led us to keep 784 as the pivot resolution.

Self-supervision. The networks XCiT pre-trained with self-supervision achieve a comparatively better performance than their supervised counterpart on Holidays, however, we have the opposite observation for \mathcal{R} Oxford.

Impact of Image description. We adopt the class-token as the descriptor, and in our experiments we verified that this aggregation method is better than average and GeM pooling [Boureau, 2010; Radenović, 2018b]. In Table A.3 one can see there is a large benefit in employing a patch based method along with our XCiT transformers : XCiT-VLAD performs significantly better than the CLS token, likely thanks to the higher dimensionality. This is further magnified with AMSK, where we obtain results approaching the absolute state of the art on Holidays, despite a sub-optimal training setting for image retrieval. This is interesting since our method has not been fine-tuned for retrieval tasks and we have not been adapted in any significant way beyond applying off-the-shelf this aggregation technique. A direct comparison with ResNet-50 shows that our XCiT method obtains competitive results in this comparable setting, slightly below the ResNet-50 on \mathcal{R} Oxford but significantly better on Holidays.

Base model	parameters	ROxford5k (mAP)		Holidays (mAP)
		Medium	Hard	
XCiT- class token				
XCiT-S12/16		30.1	8.7	86.0
XCiT-S12/8		33.2	12.1	86.4
XCiT-S12/16	resolution 224	12.7	2.4	71.5
XCiT-S12/16	resolution 384	20.1	4.6	83.4
XCiT-S12/16	resolution 512	26.6	5.8	84.6
XCiT-S12/16	resolution 768	30.1	8.7	86.0
XCiT-S12/16	resolution 1024	30.3	11.2	86.3
XCiT-S12/16	self-supervised DINO	35.1	11.9	87.3
XCiT-S12/8	self-supervised DINO	30.9	7.9	88.3
XCiT- VLAD				
XCiT-S12/16	k=256	36.6	11.6	89.9
XCiT-S12/16	k=1024	40.0	13.0	90.7
XCiT- ASMK				
XCiT-S12/8	k=1024	36.5	9.4	90.4
XCiT-S12/8	k=65536	42.0	12.9	92.3
XCiT-S12/16	k=1024	35.2	11.5	90.4
XCiT-S12/16	k=65536	40.0	15.0	92.0
ResNet-50 - ASMK				
Resnet50	k=1024	41.6	14.6	86.0
Resnet50	k=65536	41.9	14.5	87.9
Multigrain-resnet50	k=1024	32.9	9.4	87.9

TABLE A.3 – **Instance retrieval experiments.** The default resolution is 768. The default class token size is 128 dimensions. The "local descriptor" representation extracted from the activations is in 128 dimensions. To our knowledge the state of the art with ResNet-50 on Holidays with Imagenet pre-training only is the Multigrain method [Berman, 2019], which achieves mAP=92.5%. Here we compare against this method under the same training setting, i.e., off-the-shelf network pre-trained on Imagenet1k only and with the same training procedure and resolution. We refer the reader to Tolias *et al.* [Tolias, 2020] for the state of the art on ROxford, which involves some training on the target domain with images depicted building and fine-tuning at the target resolution.

References

- [Agustsson, 2017] E. AGUSTSSON, F. MENTZER, M. TSCHANNEN, L. CAVIGELLI, R. TIMOFTE, L. BENINI et L. v. GOOL, « Soft-to-hard vector quantization for end-to-end learning compressible representations », *Advances in Neural Information Processing Systems*, 2017 (cf. p. 54, 56).
- [Agustsson, 2019] E. AGUSTSSON, M. TSCHANNEN, F. MENTZER, R. TIMOFTE et L. van GOOL, « Generative Adversarial Networks for Extreme Learned Image Compression », *International Conference on Computer Vision*, 2019 (cf. p. 54, 56).
- [Ainslie, 2020] J. AINSLIE, S. ONTANON, C. ALBERTI, V. CVICEK, Z. FISHER, P. PHAM, A. RAVULA, S. SANGHAI, Q. WANG et L. YANG, « ETC : Encoding Long and Structured Inputs in Transformers », *Conference on Empirical Methods in Natural Language Processing*, 2020 (cf. p. 19).
- [Alayrac, 2022] J.-B. ALAYRAC, J. DONAHUE, P. LUC, A. MIECH, I. BARR, Y. HASSON, K. LENC, A. MENSCH, K. MILLICAN, M. REYNOLDS et al., « Flamingo : a visual language model for few-shot learning », *arXiv preprint arXiv :2204.14198*, 2022 (cf. p. 15, 71, 72).
- [Alayrac, 2020] J.-B. ALAYRAC, A. RECASENS, R. SCHNEIDER, R. ARANDJELOVIC, J. RAMAPURAM, J. DE FAUW, L. SMAIRA, S. DIELEMAN et A. ZISSERMAN, « Self-Supervised MultiModal Versatile Networks. », *Advances in Neural Information Processing Systems*, 2020 (cf. p. 72).
- [Alemi, 2018] A. ALEMI, B. POOLE, I. FISCHER, J. DILLON, R. A. SAUROUS et K. MURPHY, « Fixing a broken ELBO », *International Conference on Learning Representations*, 2018 (cf. p. 56).
- [Antonini, 1992] M. ANTONINI, M. BARLAUD, P. MATHIEU et I. DAUBECHIES, « Image coding using wavelet transform », t. 1, n° 2, p. 205-220, 1992 (cf. p. 54).
- [Arandjelovic, 2017] R. ARANDJELOVIC et A. ZISSERMAN, « Look, listen and learn », *International Conference on Computer Vision*, 2017 (cf. p. 71, 72).
- [Arnab, 2021] A. ARNAB, M. DEGHANI, G. HEIGOLD, C. SUN, M. LUČIĆ et C. SCHMID, « Vivit : A video vision transformer », *arXiv preprint arXiv :2103.15691*, 2021 (cf. p. 17).
- [Asano, 2019a] Y. M. ASANO, C. RUPPRECHT et A. VEDALDI, « A critical analysis of self-supervision, or what we can learn from a single image », *arXiv preprint arXiv :1904.13132*, 2019 (cf. p. 40).
- [Asano, 2019b] Y. M. ASANO, C. RUPPRECHT et A. VEDALDI, « Self-labelling via simultaneous clustering and representation learning », *arXiv preprint arXiv :1911.05371*, 2019 (cf. p. 14, 40).
- [Assran, 2021] M. ASSRAN, M. CARON, I. MISRA, P. BOJANOWSKI, A. JOULIN, N. BALLAS et M. RABBAT, « Semi-Supervised Learning of Visual Features by Non-Parametrically Predicting View Assignments with Support Samples », *arXiv preprint arXiv :2104.13963*, 2021 (cf. p. 41).
- [Asuni, 2014] N. ASUNI et A. GIACHETTI, « TESTIMAGES : A large-scale archive for testing visual devices and basic image processing algorithms (SAMPLING 1200 RGB set) », *STAG : Smart Tools and Apps for Graphics*, 2014 (cf. p. 64).
- [Atito, 2021] S. ATITO, M. AWAIS et J. KITTLER, « Sit : Self-supervised vision transformer », *arXiv preprint arXiv :2104.03602*, 2021 (cf. p. 39).
- [Ba, 2016] J. L. BA, J. R. KIROS et G. E. HINTON, « Layer normalization », *arXiv preprint arXiv :1607.06450*, 2016 (cf. p. 23, 24).
- [Bachmann, 2022] R. BACHMANN, D. MIZRAHI, A. ATANOV et A. ZAMIR, « MultiMAE : Multi-modal Multi-task Masked Autoencoders », *arXiv preprint arXiv :2204.01678*, 2022 (cf. p. 71, 79).
- [Bahdanau, 2014] D. BAHDANAU, K. CHO et Y. BENGIO, « Neural machine translation by jointly learning to align and translate », *arXiv preprint arXiv :1409.0473*, 2014 (cf. p. 8).

- [Bain, 2021] M. BAIN, A. NAGRANI, G. VAROL et A. ZISSERMAN, « Frozen in time : A joint video and image encoder for end-to-end retrieval », *arXiv preprint arXiv :2104.00650*, 2021 (cf. p. 78).
- [Ballé, 2017] J. BALLÉ, V. LAPARRA et E. P. SIMONCELLI, « End-to-end Optimized Image Compression », *International Conference on Learning Representations*, 2017 (cf. p. 56).
- [Ballé, 2018] J. BALLÉ, D. MINNEN, S. SINGH, S. J. HWANG et N. JOHNSTON, « Variational image compression with a scale hyperprior », *International Conference on Learning Representations*, 2018 (cf. p. 56, 61, 64).
- [Bao, 2021] H. BAO, L. DONG et F. WEI, « BEiT : BERT Pre-Training of Image Transformers », *arXiv preprint arXiv :2106.08254*, 2021 (cf. p. 38, 39, 41, 42, 46, 48, 50, 51, 90).
- [Bao, 2022] H. BAO, L. DONG et F. WEI, « BEiT : BERT Pre-Training of Image Transformers », *International Conference on Learning Representations*, 2022 (cf. p. 54).
- [Bardes, 2021] A. BARDES, J. PONCE et Y. LECUN, « Vicreg : Variance-invariance-covariance regularization for self-supervised learning », *arXiv preprint arXiv :2105.04906*, 2021 (cf. p. 14, 40).
- [Bellard,] F. BELLARD, *BPG Image format* (cf. p. 61, 64).
- [Bello, 2021] I. BELLO, « LambdaNetworks : Modeling long-range interactions without attention », *arXiv preprint arXiv :2102.08602*, 2021 (cf. p. 11, 20, 33).
- [Beltagy, 2020] I. BELTAGY, M. E. PETERS et A. COHAN, « Longformer : The long-document transformer », *arXiv preprint arXiv :2004.05150*, 2020 (cf. p. 9, 19).
- [Bengio, 2007] Y. BENGIO, P. LAMBLIN, D. POPOVICI et H. LAROCHELLE, « Greedy layer-wise training of deep networks », *Advances in neural information processing systems*, 2007, p. 153-160 (cf. p. 39).
- [Bengio, 2013] Y. BENGIO, N. LÉONARD et A. COURVILLE, « Estimating or propagating gradients through stochastic neurons for conditional computation », t. arXiv :1308.3432, 2013 (cf. p. 62).
- [Berman, 2019] M. BERMAN, H. JÉGOU, A. VEDALDI, I. KOKKINOS et M. DOUZE, « MultiGrain : a unified image embedding for classes and instances », *arXiv preprint arXiv :1902.05509*, 2019 (cf. p. 100, 101, 103).
- [Bertasius, 2021] G. BERTASIUS, H. WANG et L. TORRESANI, « Is Space-Time Attention All You Need for Video Understanding ? », *arXiv preprint arXiv :2102.05095*, 2021 (cf. p. 17).
- [Bińkowski, 2018] M. BIŃKOWSKI, D. J. SUTHERLAND, M. ARBEL et A. GRETTON, « Demystifying MMD GANs », *International Conference on Learning Representations*, 2018 (cf. p. 64).
- [Bossard, 2014] L. BOSSARD, M. GUILLAUMIN et L. VAN GOOL, « Food-101 – Mining Discriminative Components with Random Forests », *European Conference on Computer Vision*, 2014 (cf. p. 45, 46).
- [Bottou, 1998] L. BOTTOU, P. HAFFNER, P. G. HOWARD, P. SIMARD, Y. BENGIO et Y. LE CUN, « High quality document image compression with “DjVu” », *Journal of Electronic Imaging*, t. 7, n° 3, p. 410-425, 1998 (cf. p. 56).
- [Boureau, 2010] Y.-L. BOUREAU, J. PONCE et Y. LECUN, « A Theoretical Analysis of Feature Pooling in Visual Recognition », *International Conference on Machine Learning*, 2010 (cf. p. 102).
- [Brabandere, 2016] B. D. BRABANDERE, X. JIA, T. TUYTELAARS et L. V. GOOL, « Dynamic Filter Networks », *Advances in Neural Information Processing Systems*, 2016 (cf. p. 20).
- [Brock, 2021] A. BROCK, S. DE, S. L. SMITH et K. SIMONYAN, « High-Performance Large-Scale Image Recognition Without Normalization », *arXiv preprint arXiv :2102.06171*, 2021 (cf. p. 8, 18, 28, 29).
- [Brown, 2020] T. B. BROWN, B. MANN, N. RYDER, M. SUBBIAH, J. KAPLAN, P. DHARIWAL, A. NEELAKANTAN, P. SHYAM, G. SASTRY, A. ASKELL et al., « Language models are few-shot learners », *preprint arXiv :2005.14165*, 2020 (cf. p. 90).
- [Bruna, 2016] J. BRUNA, P. SPRECHMANN et Y. LECUN, « Super-resolution with deep convolutional sufficient statistics », *International Conference on Learning Representations*, 2016 (cf. p. 54).
- [Carion, 2020] N. CARION, F. MASSA, G. SYNNAEVE, N. USUNIER, A. KIRILLOV et S. ZAGORUYKO, « End-to-End Object Detection with Transformers », *European Conference on Computer Vision*, 2020 (cf. p. 11, 32, 35, 37).
- [Caron, 2018] M. CARON, P. BOJANOWSKI, A. JOULIN et M. DOUZE, « Deep clustering for unsupervised learning of visual features », *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, p. 132-149 (cf. p. 14, 40).
- [Caron, 2019] M. CARON, P. BOJANOWSKI, J. MAIRAL et A. JOULIN, « Unsupervised pre-training of image features on non-curated data », *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, p. 2959-2968 (cf. p. 40).

References

- [Caron, 2020a] M. CARON, I. MISRA, J. MAIRAL, P. GOYAL, P. BOJANOWSKI et A. JOULIN, « Unsupervised learning of visual features by contrasting cluster assignments », *arXiv preprint arXiv :2006.09882*, 2020 (cf. p. 14, 37).
- [Caron, 2020b] M. CARON, I. MISRA, J. MAIRAL, P. GOYAL, P. BOJANOWSKI et A. JOULIN, « Unsupervised learning of visual features by contrasting cluster assignments », *arXiv preprint arXiv :2006.09882*, 2020 (cf. p. 40).
- [Caron, 2021] M. CARON, H. TOUVRON, I. MISRA, H. JÉGOU, J. MAIRAL, P. BOJANOWSKI et A. JOULIN, « Emerging properties in self-supervised vision transformers », *arXiv preprint arXiv :2104.14294*, 2021 (cf. p. 4, 14, 18, 25, 31, 37, 40, 41, 46, 48, 49, 88, 97).
- [Carreira, 2017] J. CARREIRA et A. ZISSERMAN, « Quo vadis, action recognition ? a new model and the kinetics dataset », *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 6299-6308 (cf. p. 37, 75).
- [Chang, 2022] H. CHANG, H. ZHANG, L. JIANG, C. LIU et W. T. FREEMAN, « MaskGIT : Masked Generative Image Transformer », *Computer Vision and Pattern Recognition*, 2022 (cf. p. 54, 57, 58, 60, 65, 66, 90).
- [Chappelier, 2006] V. CHAPPELIER et C. GUILLEMOT, « Oriented wavelet transform for image compression and denoising », t. 15, n° 10, p. 2892-2903, 2006 (cf. p. 60).
- [Chen, 2015] C. CHEN, A. SEFF, A. KORNHAUSER et J. XIAO, « Deepdriving : Learning affordance for direct perception in autonomous driving », *Proceedings of the IEEE international conference on computer vision*, 2015, p. 2722-2730 (cf. p. 1, 94).
- [Chen, 2020a] H. CHEN, W. XIE, A. VEDALDI et A. ZISSERMAN, « Vggsound : A large-scale audio-visual dataset », *ICASSP*, 2020 (cf. p. 76, 83).
- [Chen, 2019] K. CHEN, J. WANG, J. PANG, Y. CAO, Y. XIONG, X. LI, S. SUN, W. FENG, Z. LIU, J. XU, Z. ZHANG, D. CHENG, C. ZHU, T. CHENG, Q. ZHAO, B. LI, X. LU, R. ZHU, Y. WU, J. DAI, J. WANG, J. SHI, W. OUYANG, C. C. LOY et D. LIN, « MMDetection : Open MMLab Detection Toolbox and Benchmark », *arXiv preprint arXiv :1906.07155*, 2019 (cf. p. 33).
- [Chen, 2014] L.-C. CHEN, G. PAPANDREOU, I. KOKKINOS, K. MURPHY et A. L. YUILLE, « Semantic image segmentation with deep convolutional nets and fully connected crfs », *arXiv preprint arXiv :1412.7062*, 2014 (cf. p. 37).
- [Chen, 2020b] M. CHEN, A. RADFORD, R. CHILD, J. WU, H. JUN, D. LUAN et I. SUTSKEVER, « Generative pretraining from pixels », *International Conference on Machine Learning*, 2020 (cf. p. 39).
- [Chen, 2020c] T. CHEN, S. KORNBLITH, M. NOROUZI et G. HINTON, « A simple framework for contrastive learning of visual representations », *arXiv preprint arXiv :2002.05709*, 2020 (cf. p. 4, 14, 37, 40, 45, 49, 86, 97).
- [Chen, 2021a] X. CHEN et K. HE, « Exploring simple siamese representation learning », *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, p. 15 750-15 758 (cf. p. 13, 14, 40).
- [Chen, 2021b] X. CHEN, S. XIE et K. HE, « An empirical study of training self-supervised vision transformers », *arXiv preprint arXiv :2104.02057*, 2021 (cf. p. 4, 48, 49, 52, 97).
- [Cheng, 2020] Z. CHENG, H. SUN, M. TAKEUCHI et J. KATTO, « Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules », *Computer Vision and Pattern Recognition*, 2020 (cf. p. 56, 61, 64).
- [Child, 2019] R. CHILD, S. GRAY, A. RADFORD et I. SUTSKEVER, « Generating long sequences with sparse transformers », *arXiv preprint arXiv :1904.10509*, 2019 (cf. p. 9, 19).
- [Choromanski, 2020] K. CHOROMANSKI, V. LIKHOSHERSTOV, D. DOHAN, X. SONG, A. GANE, T. SARLOS, P. HAWKINS, J. DAVIS, A. MOHIUDDIN, L. KAISER et al., « Rethinking attention with performers », *arXiv preprint arXiv :2009.14794*, 2020 (cf. p. 10, 19).
- [Chum, 2007] O. CHUM, J. PHILBIN, J. SIVIC, M. ISARD et A. ZISSERMAN, « Total recall : Automatic query expansion with a generative feature model for object retrieval », *International Conference on Computer Vision*, 2007 (cf. p. 101).
- [Contributors, 2020] M. CONTRIBUTORS, *MMSegmentation : OpenMMLab Semantic Segmentation Toolbox and Benchmark*, <https://github.com/open-mmlab/mms Segmentation>, 2020 (cf. p. 34).
- [Cover, 1991] T. M. COVER, *Elements of information theory*. John Wiley & Sons, 1991 (cf. p. 15, 60).
- [Cubuk, 2020] E. D. CUBUK, B. ZOPH, J. SHLENS et Q. V. LE, « Randaugment : Practical automated data augmentation with a reduced search space », *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020 (cf. p. 28, 86).

- [dAscoli, 2021] S. D’ASCOLI, H. TOUVRON, M. LEAVITT, A. MORCOS, G. BIROLI et L. SAGUN, « ConViT : Improving Vision Transformers with Soft Convolutional Inductive Biases », *arXiv preprint arXiv :2103.10697*, 2021 (cf. p. 19).
- [Dai, 2019] Z. DAI, Z. YANG, Y. YANG, J. CARBONELL, Q. V. LE et R. SALAKHUTDINOV, « Transformer-xl : Attentive language models beyond a fixed-length context », *arXiv preprint arXiv :1901.02860*, 2019 (cf. p. 10).
- [Deng, 2009a] J. DENG, W. DONG, R. SOCHER, L.-J. LI, K. LI et L. FEI-FEI, « ImageNet : A Large-Scale Hierarchical Image Database », *Computer Vision and Pattern Recognition*, 2009 (cf. p. 63, 64).
- [Deng, 2009b] J. DENG, W. DONG, R. SOCHER, L.-J. LI, K. LI et L. FEI-FEI, « Imagenet : A large-scale hierarchical image database », *Computer Vision and Pattern Recognition*, 2009 (cf. p. 4, 28, 37, 46, 97).
- [Devlin, 2018] J. DEVLIN, M.-W. CHANG, K. LEE et K. TOUTANOVA, « Bert : Pre-training of deep bidirectional transformers for language understanding », *arXiv preprint arXiv :1810.04805*, 2018 (cf. p. 3-5, 9, 12, 38, 39, 90, 96-98).
- [Ding, 2021a] M. DING, Z. YANG, W. HONG, W. ZHENG, C. ZHOU, D. YIN, J. LIN, X. ZOU, Z. SHAO, H. YANG et J. TANG, « CogView : Mastering Text-to-Image Generation via Transformers », *Advances in Neural Information Processing Systems*, 2021 (cf. p. 56, 60).
- [Ding, 2021b] X. DING, X. ZHANG, J. HAN et G. DING, « RepMLP : Re-parameterizing Convolutions into Fully-connected Layers for Image Recognition », *arXiv preprint arXiv :2105.01883*, 2021 (cf. p. 20).
- [Doersch, 2015] C. DOERSCH, A. GUPTA et A. A. EFROS, « Unsupervised visual representation learning by context prediction », *International Conference on Computer Vision*, 2015 (cf. p. 13).
- [Doersch, 2020] C. DOERSCH, A. GUPTA et A. ZISSERMAN, « Crosstransformers : spatially-aware few-shot transfer », *arXiv preprint arXiv :2007.11498*, 2020 (cf. p. 37).
- [Dong, 2015] C. DONG, C. C. LOY, K. HE et X. TANG, « Image super-resolution using deep convolutional networks », t. 38, n° 2, p. 295-307, 2015 (cf. p. 54).
- [Dosovitskiy, 2021a] A. DOSOVITSKIY, L. BEYER, A. KOLESNIKOV, D. WEISSENBORN, X. ZHAI, T. UNTERTHINER, M. DEGHANI, M. MINDERER, G. HEIGOLD, S. GELLY et al., « An image is worth 16x16 words : Transformers for image recognition at scale », *International Conference on Learning Representations*, 2021 (cf. p. 2, 3, 5, 11, 17, 24-27, 29, 31, 37, 40, 51, 75, 89, 95, 97, 98).
- [Dosovitskiy, 2021b] A. DOSOVITSKIY, L. BEYER, A. KOLESNIKOV, D. WEISSENBORN, X. ZHAI, T. UNTERTHINER, M. DEGHANI, M. MINDERER, G. HEIGOLD, S. GELLY, J. USZKOREIT et N. HOULSBY, « An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale », *International Conference on Learning Representations*, 2021 (cf. p. 59).
- [Dosovitskiy, 2015] A. DOSOVITSKIY, P. FISCHER, J. T. SPRINGENBERG, M. RIEDMILLER et T. BROX, « Discriminative unsupervised feature learning with exemplar convolutional neural networks », *IEEE transactions on pattern analysis and machine intelligence*, t. 38, n° 9, p. 1734-1747, 2015 (cf. p. 40).
- [Dosovitskiy, 2014] A. DOSOVITSKIY, J. T. SPRINGENBERG, M. RIEDMILLER et T. BROX, « Discriminative unsupervised feature learning with convolutional neural networks », *Advances in neural information processing systems*, t. 27, p. 766-774, 2014 (cf. p. 13, 40).
- [Dupont, 2021] E. DUPONT, A. GOLIŃSKI, M. ALIZADEH, Y. W. TEH et A. DOUCET, « COIN : COMpression with Implicit Neural representations », *ICLR Neural Compression Workshop*, 2021 (cf. p. 56).
- [Dupont, 2022] E. DUPONT, H. LOYA, M. ALIZADEH, A. GOLIŃKI, Y. W. TEH et A. DOUCET, « COIN++ : Data Agnostic Neural Compression », t. arXiv :2201.12904, 2022 (cf. p. 56).
- [Esser, 2021] P. ESSER, R. ROMBACH et B. OMMER, « Taming Transformers for High-Resolution Image Synthesis », *Computer Vision and Pattern Recognition*, 2021 (cf. p. 54, 56, 58, 60, 65).
- [Faghri, 2017] F. FAGHRI, D. J. FLEET, J. R. KIROS et S. FIDLER, « Vse++ : Improving visual-semantic embeddings with hard negatives », *arXiv preprint arXiv :1707.05612*, 2017 (cf. p. 72).
- [Fan, 2021] H. FAN, B. XIONG, K. MANGALAM, Y. LI, Z. YAN, J. MALIK et C. FEICHTENHOFER, « Multiscale Vision Transformers », *arXiv preprint arXiv :2104.11227*, 2021 (cf. p. 11, 17, 20).
- [Fang, 2021] H. FANG, P. XIONG, L. XU et Y. CHEN, « Clip2video : Mastering video-text retrieval via image clip », *arXiv preprint arXiv :2106.11097*, 2021 (cf. p. 72).
- [Font, 2013] F. FONT, G. ROMA et X. SERRA, « Freesound technical demo », *ACM international conference on Multimedia*, 2013 (cf. p. 76, 83).
- [Frey, 1998] B. J. FREY, « Bayesian networks for pattern classification, data compression, and channel coding. », thèse de doct., University of Toronto, 1998 (cf. p. 56).

References

- [Frome, 2013] A. FROME, G. S. CORRADO, J. SHLENS, S. BENGIO, J. DEAN, M. RANZATO et T. MIKOLOV, « Devise : A deep visual-semantic embedding model », *Advances in neural information processing systems*, t. 26, 2013 (cf. p. 72).
- [Fukushima, 1980] K. FUKUSHIMA, « Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position », *Biological cybernetics*, t. 36, n° 4, p. 193-202, 1980 (cf. p. 1, 94).
- [Gafni, 2022] O. GAFNI, A. POLYAK, O. ASHUAL, S. SHEYNIN, D. PARIKH et Y. TAIGMAN, « Make-A-Scene : Scene-Based Text-to-Image Generation with Human Priors », *European Conference on Computer Vision*, 2022 (cf. p. 56, 60).
- [Gedas Bertasius, 2021] L. T. GEDAS BERTASIOUS Heng Wang, « Is Space-Time Attention All You Need for Video Understanding ? », 2021 (cf. p. 5, 98).
- [Geirhos, 2018] R. GEIRHOS, P. RUBISCH, C. MICHAELIS, M. BETHGE, F. A. WICHMANN et W. BRENDEL, « ImageNet-trained CNNs are biased towards texture ; increasing shape bias improves accuracy and robustness », *arXiv preprint arXiv :1811.12231*, 2018 (cf. p. 1).
- [Gemmeke, 2017] J. F. GEMMEKE, D. P. W. ELLIS, D. FREEDMAN, A. JANSEN, W. LAWRENCE, R. C. MOORE, M. PLAKAL et M. RITTER, « Audio Set : An ontology and human-labeled dataset for audio events », *International Conference on Acoustics, Speech, and Signal Processing*, 2017 (cf. p. 76, 82).
- [Gidaris, 2018] S. GIDARIS, P. SINGH et N. KOMODAKIS, « Unsupervised representation learning by predicting image rotations », *arXiv preprint arXiv :1803.07728*, 2018 (cf. p. 13).
- [Girdhar, 2022a] R. GIRDHAR, A. EL-NOUBY, M. SINGH, K. V. ALWALA, A. JOULIN et I. MISRA, « OmniMAE : Single Model Masked Pretraining on Images and Videos », *arXiv preprint arXiv :2206.08356*, 2022 (cf. p. 72, 75).
- [Girdhar, 2022b] R. GIRDHAR, M. SINGH, N. RAVI, L. van der MAATEN, A. JOULIN et I. MISRA, « Omnivore : A Single Model for Many Visual Modalities », *Computer Vision and Pattern Recognition*, 2022 (cf. p. 72, 75, 76, 83).
- [Gong, 2021] Y. GONG, Y.-A. CHUNG et J. GLASS, « AST : Audio Spectrogram Transformer », *Interspeech*, 2021 (cf. p. 75, 77).
- [Goodfellow, 2014] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE et Y. BENGIO, « Generative Adversarial Nets », *Advances in Neural Information Processing Systems*, 2014 (cf. p. 56, 62).
- [Gordo, 2017] A. GORDO, J. ALMAZÁN, J. REVAUD et D. LARLUS, « End-to-End Learning of Deep Visual Representations for Image Retrieval », *International journal of Computer Vision*, t. 124, 2017 (cf. p. 100).
- [Goyal, 2021] P. GOYAL, M. CARON, B. LEFAUDEX, M. XU, P. WANG, V. PAI, M. SINGH, V. LIPTCHINSKY, I. MISRA, A. JOULIN et al., « Self-supervised pretraining of visual features in the wild », *arXiv preprint arXiv :2103.01988*, 2021 (cf. p. 38, 40).
- [Graham, 2021] B. GRAHAM, A. EL-NOUBY, H. TOUVRON, P. STOCK, A. JOULIN, H. JÉGOU et M. DOUZE, « LeViT : a Vision Transformer in ConvNet’s Clothing for Faster Inference », *arXiv preprint arXiv :2104.01136*, 2021 (cf. p. 11, 19, 25).
- [Grauman, 2022] K. GRAUMAN, A. WESTBURY, E. BYRNE, Z. CHAVIS, A. FURNARI, R. GIRDHAR, J. HAMBURGER, H. JIANG, M. LIU, X. LIU et al., « Ego4d : Around the world in 3,000 hours of egocentric video », *Computer Vision and Pattern Recognition*, 2022 (cf. p. 76, 84).
- [Gray, 1998] R. M. GRAY et D. L. NEUHOFF, « Quantization », t. 44, n° 6, p. 2325-2383, 1998 (cf. p. 54, 60).
- [Gregor, 2016] K. GREGOR, F. BESSE, D. JIMENEZ REZENDE, I. DANIHELKA et D. WIERSTRA, « Towards conceptual compression », *Advances in Neural Information Processing Systems*, 2016 (cf. p. 56).
- [Grill, 2020] J.-B. GRILL, F. STRUB, F. ALTCHÉ, C. TALLEC, P. H. RICHEMOND, E. BUCHATSKAYA, C. DOERSCH, B. A. PIRES, Z. D. GUO, M. G. AZAR et al., « Bootstrap your own latent : A new approach to self-supervised learning », *arXiv preprint arXiv :2006.07733*, 2020 (cf. p. 14, 37, 40).
- [Gu, 2021] X. GU, T.-Y. LIN, W. KUO et Y. CUI, « Open-vocabulary object detection via vision and language knowledge distillation », *arXiv preprint arXiv :2104.13921*, 2021 (cf. p. 73).
- [Gulati, 2020] A. GULATI, J. QIN, C.-C. CHIU, N. PARMAR, Y. ZHANG, J. YU, W. HAN, S. WANG, Z. ZHANG, Y. WU et al., « Conformer : Convolution-augmented transformer for speech recognition », *arXiv preprint arXiv :2005.08100*, 2020 (cf. p. 9, 10).

- [Guzhov, 2021] A. GUZHOV, F. RAUE, J. HEES et A. DENGEL, *AudioCLIP : Extending CLIP to Image, Text and Audio*, 2021. arXiv : [2106.13043 \[cs.SD\]](#) (cf. p. [73](#), [74](#), [76-78](#)).
- [Hadsell, 2006] R. HADSELL, S. CHOPRA et Y. LECUN, « Dimensionality reduction by learning an invariant mapping », *Computer Vision and Pattern Recognition*, 2006 (cf. p. [73](#)).
- [Han, 2021] K. HAN, A. XIAO, E. WU, J. GUO, C. XU et Y. WANG, « Transformer in transformer », *arXiv preprint arXiv :2103.00112*, 2021 (cf. p. [19](#)).
- [He, 2021a] K. HE, X. CHEN, S. XIE, Y. LI, P. DOLLAR et R. GIRSHICK, « Masked Autoencoders Are Scalable Vision Learners », *Computer Vision and Pattern Recognition*, 2021 (cf. p. [54](#), [90](#)).
- [He, 2021b] K. HE, X. CHEN, S. XIE, Y. LI, P. DOLLAR et R. GIRSHICK, « Masked Autoencoders Are Scalable Vision Learners », *arXiv preprint arXiv :2111.06377*, 2021 (cf. p. [44](#), [52](#), [92](#)).
- [He, 2020] K. HE, H. FAN, Y. WU, S. XIE et R. GIRSHICK, « Momentum contrast for unsupervised visual representation learning », *Computer Vision and Pattern Recognition*, 2020 (cf. p. [4](#), [14](#), [37](#), [40](#), [97](#)).
- [He, 2018] K. HE, R. GIRSHICK et P. DOLLAR, « Rethinking im-agenet pre-training », *arXiv preprint arXiv :1811.08883*, 2018 (cf. p. [46](#)).
- [He, 2017] K. HE, G. GKIOXARI, P. DOLLAR et R. GIRSHICK, « Mask r-cnn », *International Conference on Computer Vision*, 2017 (cf. p. [32](#), [33](#), [37](#), [47](#), [52](#)).
- [He, 2016] K. HE, X. ZHANG, S. REN et J. SUN, « Deep residual learning for image recognition », *Computer Vision and Pattern Recognition*, 2016 (cf. p. [1](#), [7](#), [17](#), [31](#), [33](#), [34](#), [37](#), [94](#)).
- [Heusel, 2017] M. HEUSEL, H. RAMSAUER, T. UNTERTHINER, B. NESSLER et S. HOCHREITER, « GANs trained by a two time-scale update rule converge to a local Nash equilibrium », *Advances in Neural Information Processing Systems*, 2017 (cf. p. [64](#)).
- [Hinton, 2006] G. E. HINTON et R. R. SALAKHUTDINOV, « Reducing the dimensionality of data with neural networks », *science*, t. 313, n° 5786, p. 504-507, 2006 (cf. p. [39](#)).
- [Ho, 2019] J. HO, N. KALCHBRENNER, D. WEISSENBORN et T. SALIMANS, « Axial attention in multidimensional transformers », *arXiv preprint arXiv :1912.12180*, 2019 (cf. p. [19](#), [27](#)).
- [Horn, 2017] G. V. HORN, O. MAC AODHA, Y. SONG, A. SHEPARD, H. ADAM, P. PERONA et S. J. BELONGIE, « The iNaturalist Species Classification and Detection Dataset », *arXiv preprint arXiv :1707.06642*, 2017 (cf. p. [29](#), [37](#), [46](#)).
- [Howard, 2017] A. G. HOWARD, M. ZHU, B. CHEN, D. KALENICHENKO, W. WANG, T. WEYAND, M. ANDREETTO et H. ADAM, « Mobilenets : Efficient convolutional neural networks for mobile vision applications », *arXiv preprint arXiv :1704.04861*, 2017 (cf. p. [8](#)).
- [Hu, 2018] J. HU, L. SHEN et G. SUN, « Squeeze-and-excitation networks », *Computer Vision and Pattern Recognition*, 2018 (cf. p. [1](#), [11](#), [20](#), [94](#)).
- [Huang, 2016] G. HUANG, Y. SUN, Z. LIU, D. SEDRA et K. Q. WEINBERGER, « Deep networks with stochastic depth », *European Conference on Computer Vision*, 2016 (cf. p. [28](#), [82](#)).
- [Ilharco, 2021] G. ILHARCO, M. WORTSMAN, R. WIGHTMAN, C. GORDON, N. CARLINI, R. TAORI, A. DAVE, V. SHANKAR, H. NAMKOONG, J. MILLER, H. HAJISHIRZI, A. FARHADI et L. SCHMIDT, *OpenCLIP*, 2021 (cf. p. [15](#), [75](#), [76](#)).
- [Jaegle, 2021] A. JAEGLER, F. GIMENO, A. BROCK, A. ZISSERMAN, O. VINYALS et J. CARREIRA, « Perceiver : General Perception with Iterative Attention », *arXiv preprint arXiv :2103.03206*, 2021 (cf. p. [19](#)).
- [Jégou, 2010] H. JÉGOU, C. SCHMID, H. HARZALLAH et J. VERBEEK, « Accurate image search using the contextual dissimilarity measure », t. 32, n° 1, p. 2-11, 2010 (cf. p. [54](#), [58](#)).
- [Jégou, 2008] H. JÉGOU, M. DOUZE et C. SCHMID, « Hamming embedding and weak geometric consistency for large scale image search », *European Conference on Computer Vision*, 2008 (cf. p. [101](#)).
- [Jégou, 2012] H. JÉGOU, F. PERRONNIN, M. DOUZE, J. SÁNCHEZ, P. PEREZ et C. SCHMID, « Aggregating local image descriptors into compact codes », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, t. 34, n° 9, 2012 (cf. p. [100](#), [101](#)).
- [Jia, 2021a] C. JIA, Y. YANG, Y. XIA, Y.-T. CHEN, Z. PAREKH, H. PHAM, Q. LE, Y.-H. SUNG, Z. LI et T. DUERIG, « Scaling up visual and vision-language representation learning with noisy text supervision », *International Conference on Machine Learning*, 2021 (cf. p. [15](#), [71](#), [72](#)).
- [Jia, 2021b] X. JIA, C. ZHU, M. LI, W. TANG et W. ZHOU, « LLVIP : A visible-infrared paired dataset for low-light vision », *International Conference on Computer Vision*, 2021 (cf. p. [76](#), [84](#)).

References

- [Johnson, 2017] M. JOHNSON, M. SCHUSTER, Q. V. LE, M. KRİKUN, Y. WU, Z. CHEN, N. THORAT, F. VIÉGAS, M. WATTENBERG, G. CORRADO et al., « Google’s multilingual neural machine translation system : Enabling zero-shot translation », *Transactions of the Association for Computational Linguistics*, t. 5, p. 339-351, 2017 (cf. p. 73).
- [Joulin, 2016] A. JOULIN, L. VAN DER MAATEN, A. JABRI et N. VASILACHE, « Learning visual features from large weakly supervised data », *European Conference on Computer Vision*, Springer, 2016, p. 67-84 (cf. p. 5, 15, 40, 72, 98).
- [Karras, 2019] T. KARRAS, S. LAINE et T. AILA, « A Style-Based Generator Architecture for Generative Adversarial Networks », *Computer Vision and Pattern Recognition*, 2019 (cf. p. 65).
- [Kataoka, 2020] H. KATAOKA, K. OKAYASU, A. MATSUMOTO, E. YAMAGATA, R. YAMADA, N. INOUE, A. NAKAMURA et Y. SATOH, « Pre-training without natural images », *Proceedings of the Asian Conference on Computer Vision*, 2020 (cf. p. 40).
- [Katharopoulos, 2020] A. KATHAROPOULOS, A. VYAS, N. PAPPAS et F. FLEURET, « Transformers are RNNs : Fast autoregressive transformers with linear attention », *International Conference on Machine Learning*, 2020 (cf. p. 19).
- [Kay, 2017] W. KAY, J. CARREIRA, K. SIMONYAN, B. ZHANG, C. HILLIER, S. VIJAYANARASIMHAN, F. VIOLA, T. GREEN, T. BACK, P. NATSEV, A. SULEYMAN et A. ZISSERMAN, « The kinetics human action video dataset », *arXiv preprint arXiv :1705.06950*, 2017 (cf. p. 77).
- [Kim, 2019] C. D. KIM, B. KIM, H. LEE et G. KIM, « Audiocaps : Generating captions for audios in the wild », *NAACL*, 2019 (cf. p. 76, 83).
- [Kingma, 2014] D. KINGMA et M. WELLING, « Auto-Encoding Variational Bayes », *International Conference on Learning Representations*, 2014 (cf. p. 56).
- [Kirillov, 2019] A. KIRILLOV, R. GIRSHICK, K. HE et P. DOLLÁR, « Panoptic feature pyramid networks », *Computer Vision and Pattern Recognition*, 2019 (cf. p. 34).
- [Kiros, 2014] R. KIROS, R. SALAKHUTDINOV et R. S. ZEMEL, « Unifying visual-semantic embeddings with multimodal neural language models », *arXiv preprint arXiv :1411.2539*, 2014 (cf. p. 72).
- [Kitaev, 2020] N. KITAEV, Ł. KAISER et A. LEVSKAYA, « Reformer : The efficient transformer », *arXiv preprint arXiv :2001.04451*, 2020 (cf. p. 10).
- [Knoll, 2020] F. KNOLL, K. HAMMERNIK, C. ZHANG, S. MOELLER, T. POCK, D. K. SODICKSON et M. AKCAKAYA, « Deep-learning methods for parallel magnetic resonance imaging reconstruction : A survey of the current approaches, trends, and issues », *IEEE Signal Processing Magazine*, t. 37, n° 1, p. 128-140, 2020 (cf. p. 54).
- [Kodak, 1993] E. KODAK, *Kodak lossless true color image suite (PhotoCD PCD0992)*, 1993 (cf. p. 64).
- [Krause, 2013] J. KRAUSE, M. STARK, J. DENG et L. FEI-FEI, « 3D Object Representations for Fine-Grained Categorization », *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, 2013 (cf. p. 29, 45, 46).
- [Krishna, 2021] K. KRISHNA, J. BIGHAM et Z. C. LIPTON, « Does Pretraining for Summarization Require Knowledge Transfer ? », *arXiv preprint arXiv :2109.04953*, 2021 (cf. p. 40).
- [Krizhevsky, 2009] A. KRIZHEVSKY, « Learning Multiple Layers of Features from Tiny Images », CIFAR, rapp. tech., 2009 (cf. p. 29).
- [Krizhevsky, 2012] A. KRIZHEVSKY, I. SUTSKEVER et G. E. HINTON, « Imagenet classification with deep convolutional neural networks », *Advances in Neural Information Processing Systems*, 2012 (cf. p. 7).
- [Lample, 2017] G. LAMPLE, A. CONNEAU, L. DENOYER et M. RANZATO, « Unsupervised machine translation using monolingual corpora only », *arXiv preprint arXiv :1711.00043*, 2017 (cf. p. 73).
- [Larochelle, 2011] H. LAROCHELLE et I. MURRAY, « The Neural Autoregressive Distribution Estimator », 2011 (cf. p. 56).
- [Larsson, 2016] G. LARSSON, M. MAIRE et G. SHAKHAROVICH, « Learning representations for automatic colorization », *Computer Vision—ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, 2016, p. 577-593 (cf. p. 13).
- [LeCun, 1989] Y. LECUN, B. BOSER, J. S. DENKER, D. HENDERSON, R. E. HOWARD, W. HUBBARD et L. D. JACKEL, « Backpropagation applied to handwritten zip code recognition », *Neural computation*, t. 1, n° 4, p. 541-551, 1989 (cf. p. 1, 7, 94).
- [LeCun, 1998] Y. LECUN, L. BOTTOU, Y. BENGIO et P. HAFFNER, « Gradient-based learning applied to document recognition », *Proceedings of the IEEE*, 1998 (cf. p. 7).

- [Ledig, 2017] C. LEDIG, L. THEIS, F. HUSZAR, J. CABALLERO, A. CUNNINGHAM, A. ACOSTA, A. AITKEN, A. TEJANI, J. TOTZ, Z. WANG et W. SHI, « Photo-realistic single image super-resolution using a generative adversarial network », *Computer Vision and Pattern Recognition*, 2017 (cf. p. 54).
- [Lee-Thorp, 2021] J. LEE-THORP, J. AINSLIE, I. ECKSTEIN et S. ONTANON, « FNet : Mixing Tokens with Fourier Transforms », *arXiv preprint arXiv :2105.03824*, 2021 (cf. p. 19).
- [Lewis, 2019] M. LEWIS, Y. LIU, N. GOYAL, M. GHAZVININEJAD, A. MOHAMED, O. LEVY, V. STOYANOV et L. ZETTLEMOYER, « Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension », *arXiv preprint arXiv :1910.13461*, 2019 (cf. p. 39).
- [Lezama, 2022] J. LEZAMA, H. CHANG, L. JIANG et I. ESSA, « Improved Masked Image Generation with Token-Critic », *European Conference on Computer Vision*, 2022 (cf. p. 57).
- [Li, 2022] B. LI, K. Q. WEINBERGER, S. BELONGIE, V. KOLTUN et R. RANFTL, « Language-driven semantic segmentation », *arXiv preprint arXiv :2201.03546*, 2022 (cf. p. 73).
- [Likhoshesterov, 2021] V. LIKHOSHERSTOV, A. ARNAB, K. CHOROMANSKI, M. LUCIC, Y. TAY, A. WELLER et M. DEGHANI, « PolyViT : Co-training Vision Transformers on Images, Videos and Audio », *arXiv preprint arXiv :2111.12993*, 2021 (cf. p. 72).
- [Lin, 2017] T.-Y. LIN, P. DOLLÁR, R. GIRSHICK, K. HE, B. HARIHARAN et S. BELONGIE, « Feature pyramid networks for object detection », *Computer Vision and Pattern Recognition*, 2017 (cf. p. 32, 52).
- [Lin, 2014] T.-Y. LIN, M. MAIRE, S. BELONGIE, J. HAYS, P. PERONA, D. RAMANAN, P. DOLLÁR et C. L. ZITNICK, « Microsoft coco : Common objects in context », *European Conference on Computer Vision*, 2014 (cf. p. 32, 45, 46).
- [Lin, 2022] Z. LIN, S. GENG, R. ZHANG, P. GAO, G. de MELO, X. WANG, J. DAI, Y. QIAO et H. LI, « Frozen CLIP Models are Efficient Video Learners », *European Conference on Computer Vision*, 2022 (cf. p. 72).
- [Liu, 2022a] X. LIU, J. ZHOU, T. KONG, X. LIN et R. JI, « Exploring target representations for masked autoencoders », *arXiv preprint arXiv :2209.03917*, 2022 (cf. p. 73).
- [Liu, 2021a] Y. LIU, E. SANGINETO, W. BI, N. SEBE, B. LEPRI et M. NADAI, « Efficient Training of Visual Transformers with Small Datasets », *Advances in Neural Information Processing Systems*, 2021 (cf. p. 47, 48).
- [Liu, 2021b] Z. LIU, Y. LIN, Y. CAO, H. HU, Y. WEI, Z. ZHANG, S. LIN et B. GUO, « Swin transformer : Hierarchical vision transformer using shifted windows », *arXiv preprint arXiv :2103.14030*, 2021 (cf. p. 11, 17, 18, 20, 28, 29, 31, 33, 34, 37, 52).
- [Liu, 2022b] Z. LIU, H. MAO, C.-Y. WU, C. FEICHTENHOFER, T. DARRELL et S. XIE, « A convnet for the 2020s », *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 11 976-11 986 (cf. p. 12).
- [Loshchilov, 2017] I. LOSHCILOV et F. HUTTER, « Decoupled weight decay regularization », *arXiv preprint arXiv :1711.05101*, 2017 (cf. p. 28, 63).
- [Lu, 2019] X. LU, H. WANG, W. DONG, F. WU, Z. ZHENG et G. SHI, « Learning a deep vector quantization network for image compression », *IEEE Access*, t. 7, p. 118 815-118 825, 2019 (cf. p. 56).
- [Ludwig, 2016] R. LUDWIG et J. TAYLOR, *Voyager telecommunications*. John Wiley et Sons, Inc, 2016 (cf. p. 54).
- [Luo, 2021] H. LUO, L. JI, M. ZHONG, Y. CHEN, W. LEI, N. DUAN et T. LI, « CLIP4Clip : An Empirical Study of CLIP for End to End Video Clip Retrieval. CoRR abs/2104.08860 (2021) », *arXiv preprint arXiv :2104.08860*, 2021 (cf. p. 72).
- [MacKay, 2003] D. J. C. MACKAY, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003 (cf. p. 15).
- [Mahajan, 2018] D. MAHAJAN, R. GIRSHICK, V. RAMANATHAN, K. HE, M. PALURI, Y. LI, A. BHARAMBE et L. VAN DER MAATEN, « Exploring the limits of weakly supervised pretraining », *Proceedings of the European conference on computer vision (ECCV)*, 2018, p. 181-196 (cf. p. 40, 71).
- [Melas-Kyriazi, 2021] L. MELAS-KYRIAZI, « Do You Even Need Attention? A Stack of Feed-Forward Layers Does Surprisingly Well on ImageNet », *arXiv preprint arXiv :2105.02723*, 2021 (cf. p. 20).
- [Mentzer, 2018] F. MENTZER, E. AGUSTSSON, M. TSCHANNEN, R. TIMOFTE et L. van GOOL, « Conditional probability models for deep image compression », *Computer Vision and Pattern Recognition*, 2018 (cf. p. 56).
- [Mentzer, 2020] F. MENTZER, G. D. TODERICI, M. TSCHANNEN et E. AGUSTSSON, « High-Fidelity Generative Image Compression », *Advances in Neural Information Processing Systems*, 2020 (cf. p. 54-56, 63-65).

References

- [Miech, 2020] A. MIECH, J.-B. ALAYRAC, L. SMAIRA, I. LAPTEV, J. SIVIC et A. ZISSERMAN, « End-to-end learning of visual representations from uncurated instructional videos », *Computer Vision and Pattern Recognition*, 2020 (cf. p. 72).
- [Miech, 2019a] A. MIECH, D. ZHUKOV, J.-B. ALAYRAC, M. TAPASWI, I. LAPTEV et J. SIVIC, « HowTo100M : Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips », *International Conference on Computer Vision*, 2019 (cf. p. 78).
- [Miech, 2019b] A. MIECH, D. ZHUKOV, J.-B. ALAYRAC, M. TAPASWI, I. LAPTEV et J. SIVIC, « Howto100m : Learning a text-video embedding by watching hundred million narrated video clips », *International Conference on Computer Vision*, 2019 (cf. p. 72).
- [Minnen, 2018] D. MINNEN, J. BALLÉ et G. D. TODERICI, « Joint Autoregressive and Hierarchical Priors for Learned Image Compression », *Advances in Neural Information Processing Systems*, 2018 (cf. p. 56, 61, 64, 65).
- [Minnen, 2020] D. MINNEN et S. SINGH, « Channel-wise autoregressive entropy models for learned image compression », 2020 (cf. p. 56).
- [Morgado, 2021] P. MORGADO, N. VASCONCELOS et I. MISRA, « Audio-Visual Instance Discrimination with Cross-Modal Agreement », *Computer Vision and Pattern Recognition*, 2021 (cf. p. 71-73, 79).
- [Nagrani, 2022] A. NAGRANI, P. H. SEO, B. SEYBOLD, A. HAUTH, S. MANEN, C. SUN et C. SCHMID, « Learning Audio-Video Modalities from Image Captions », *European Conference on Computer Vision*, 2022 (cf. p. 72, 77, 78).
- [Nagrani, 2021] A. NAGRANI, S. YANG, A. ARNAB, A. JANSEN, C. SCHMID et C. SUN, « Attention bottlenecks for multimodal fusion », *Advances in Neural Information Processing Systems*, 2021 (cf. p. 78).
- [Naseer, 2021] M. M. NASEER, K. RANASINGHE, S. H. KHAN, M. HAYAT, F. SHAHBAZ KHAN et M.-H. YANG, « Intriguing properties of vision transformers », *Advances in Neural Information Processing Systems*, t. 34, p. 23296-23308, 2021 (cf. p. 2, 95).
- [Nilsback, 2008] M.-E. NILSBACK et A. ZISSERMAN, « Automated Flower Classification over a Large Number of Classes », *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008 (cf. p. 29).
- [Noroozi, 2016] M. NOROOZI et P. FAVARO, « Unsupervised learning of visual representations by solving jigsaw puzzles », *European conference on computer vision*, Springer, 2016, p. 69-84 (cf. p. 12, 39).
- [El-Nouby, 2021a] A. EL-NOUBY, G. IZACARD, H. TOUVRON, I. LAPTEV, H. JEGOU et E. GRAVE, « Are Large-scale Datasets Necessary for Self-Supervised Pre-training ? », *arXiv preprint arXiv :2112.10740*, 2021 (cf. p. 4, 54, 90, 97).
- [El-Nouby, 2023] A. EL-NOUBY, M. J. MUCKLEY, K. ULLRICH, I. LAPTEV, J. VERBEEK et H. JEGOU, « Image Compression with Product Quantized Masked Image Modeling », *Transactions on Machine Learning Research*, 2023 (cf. p. 4, 90, 98).
- [El-Nouby, 2021b] A. EL-NOUBY, N. NEVEROVA, I. LAPTEV et H. JÉGOU, « Training vision transformers for image retrieval », *arXiv preprint arXiv :2102.05644*, 2021 (cf. p. 17, 89, 100, 101).
- [El-Nouby, 2021c] A. EL-NOUBY, H. TOUVRON, M. CARON, P. BOJANOWSKI, M. DOUZE, A. JOULIN, I. LAPTEV, N. NEVEROVA, G. SYNNAEVE, J. VERBEEK et al., « XCiT : Cross-Covariance Image Transformers », *Advances in Neural Information Processing Systems*, 2021 (cf. p. 3, 59, 89, 97).
- [El-Nouby, 2021d] A. EL-NOUBY, H. TOUVRON, M. CARON, P. BOJANOWSKI, M. DOUZE, A. JOULIN, I. LAPTEV, N. NEVEROVA, G. SYNNAEVE, J. VERBEEK et al., « XCiT : Cross-Covariance Image Transformers », *arXiv preprint arXiv :2106.09681*, 2021 (cf. p. 52).
- [Oncescu, 2021] A.-M. ONCESCU, A. KOEPKE, J. F. HENRIQUES, Z. AKATA et S. ALBANIE, « Audio retrieval with natural language queries », *arXiv preprint arXiv :2105.02192*, 2021 (cf. p. 78, 83).
- [Oord, 2016a] A. v. d. OORD, S. DIELEMAN, H. ZEN, K. SIMONYAN, O. VINYALS, A. GRAVES, N. KALCHBRENNER, A. SENIOR et K. KAVUKCUOGLU, « WaveNet : a generative model for raw audio », *ISCA Speech Synthesis Workshop*, 2016 (cf. p. 56).
- [Oord, 2016b] A. v. d. OORD, N. KALCHBRENNER, O. VINYALS, L. ESPEHOLT, A. GRAVES et K. KAVUKCUOGLU, « Conditional Image Generation with PixelCNN Decoders », *Advances in Neural Information Processing Systems*, 2016 (cf. p. 54).
- [Oord, 2017] A. v. d. OORD, O. VINYALS et K. KAVUKCUOGLU, « Neural Discrete Representation Learning », *Advances in Neural Information Processing Systems*, 2017 (cf. p. 54, 56, 58).

- [Oord, 2018] A. v. d. OORD, Y. LI et O. VINALS, « Representation learning with contrastive predictive coding », *arXiv preprint arXiv :1807.03748*, 2018 (cf. p. 13, 45, 74).
- [Oquab, 2014] M. OQUAB, L. BOTTOU, I. LAPTEV et J. SIVIC, « Learning and transferring mid-level image representations using convolutional neural networks », *Computer Vision and Pattern Recognition*, 2014 (cf. p. 37).
- [Owens, 2018] A. OWENS et A. A. EFROS, « Audio-visual scene analysis with self-supervised multisensory features », *European Conference on Computer Vision*, 2018 (cf. p. 71).
- [Parmar, 2018] N. PARMAR, A. VASWANI, J. USZKOREIT, L. KAISER, N. SHAZEER, A. KU et D. TRAN, « Image transformer », *International Conference on Machine Learning*, 2018 (cf. p. 19).
- [Pathak, 2016] D. PATHAK, P. KRAHENBUHL, J. DONAHUE, T. DARRELL et A. A. EFROS, « Context encoders : Feature learning by inpainting », *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 2536-2544 (cf. p. 12, 39).
- [Patrick, 2021] M. PATRICK, Y. M. ASANO, R. FONG, J. F. HENRIQUES, G. ZWEIG et A. VEDALDI, « Multi-modal self-supervision from generalized data transformations », *International Conference on Computer Vision*, 2021 (cf. p. 71).
- [Patrick, 2020] M. PATRICK, P.-Y. HUANG, Y. ASANO, F. METZE, A. HAUPTMANN, J. HENRIQUES et A. VEDALDI, « Support-set bottlenecks for video-text representation learning », *arXiv preprint arXiv :2010.02824*, 2020 (cf. p. 78).
- [Peng, 2019] X. PENG, Q. BAI, X. XIA, Z. HUANG, K. SAENKO et B. WANG, « Moment matching for multi-source domain adaptation », *International Conference on Computer Vision*, 2019 (cf. p. 45, 46).
- [Peng, 2022] Z. PENG, L. DONG, H. BAO, Q. YE et F. WEI, « Beit v2 : Masked image modeling with vector-quantized visual tokenizers », *arXiv preprint arXiv :2208.06366*, 2022 (cf. p. 73, 92).
- [Philbin, 2007] J. PHILBIN, O. CHUM, M. ISARD, J. SIVIC et A. ZISSERMAN, « Object Retrieval with Large Vocabularies and Fast Spatial Matching », *Computer Vision and Pattern Recognition*, 2007 (cf. p. 101).
- [Piczak, 2015] K. J. PICZAK, « ESC : Dataset for environmental sound classification », 2015 (cf. p. 76, 83).
- [Purushwalkam, 2020] S. PURUSHWALKAM et A. GUPTA, « Demystifying contrastive self-supervised learning : Invariances, augmentations and dataset biases », *arXiv preprint arXiv :2007.13916*, 2020 (cf. p. 38).
- [Qiu, 2019] J. QIU, H. MA, O. LEVY, S. W.-t. YIH, S. WANG et J. TANG, « Blockwise self-attention for long document understanding », *arXiv preprint arXiv :1911.02972*, 2019 (cf. p. 19).
- [Radenović, 2018a] F. RADENOVIĆ, A. ISCEN, G. TOLIAS, Y. AVRITHIS et O. CHUM, « Revisiting Oxford and Paris : Large-scale image retrieval benchmarking », *Computer Vision and Pattern Recognition*, 2018 (cf. p. 101).
- [Radenović, 2018b] F. RADENOVIĆ, G. TOLIAS et O. CHUM, « Fine-tuning CNN image retrieval with no human annotation », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018 (cf. p. 102).
- [Radford, 2021] A. RADFORD, J. W. KIM, C. HALLACY, A. RAMESH, G. GOH, S. AGARWAL, G. SASTRY, A. ASKELL, P. MISHKIN, J. CLARK et al., « Learning transferable visual models from natural language supervision », *International Conference on Machine Learning*, 2021 (cf. p. 5, 15, 71-73, 75-77, 79, 84, 85, 98).
- [Radford, 2018] A. RADFORD, K. NARASIMHAN, T. SALIMANS et I. SUTSKEVER, « Improving language understanding with unsupervised learning », *Technical report, OpenAI*, 2018 (cf. p. 3-5, 9, 12, 96-98).
- [Radosavovic, 2020] I. RADOSAVOVIC, R. P. KOSARAJU, R. GIRSHICK, K. HE et P. DOLLÁR, « Designing network design spaces », *Computer Vision and Pattern Recognition*, 2020 (cf. p. 8, 28, 29, 37).
- [Raffel, 2019] C. RAFFEL, N. SHAZEER, A. ROBERTS, K. LEE, S. NARANG, M. MATENA, Y. ZHOU, W. LI et P. J. LIU, « Exploring the limits of transfer learning with a unified text-to-text transformer », *arXiv preprint arXiv :1910.10683*, 2019 (cf. p. 39, 40).
- [Ramachandran, 2019] P. RAMACHANDRAN, N. PARMAR, A. VASWANI, I. BELLO, A. LEVSKAYA et J. SHLENS, « Stand-alone self-attention in vision models », *arXiv preprint arXiv :1906.05909*, 2019 (cf. p. 11).
- [Ramesh, 2022a] A. RAMESH, P. DHARIWAL, A. NICHOL, C. CHU et M. CHEN, « Hierarchical Text-Conditional Image Generation with CLIP Latents », 2022 (cf. p. 1, 92, 94).
- [Ramesh, 2022b] A. RAMESH, P. DHARIWAL, A. NICHOL, C. CHU et M. CHEN, « Hierarchical text-conditional image generation with clip latents », *arXiv preprint arXiv :2204.06125*, 2022 (cf. p. 71, 81).
- [Ramesh, 2021a] A. RAMESH, M. PAVLOV, G. GOH, S. GRAY, C. VOSS, A. RADFORD, M. CHEN et I. SUTSKEVER, « Zero-Shot Text-to-Image Generation », *International Conference on Machine Learning*, 2021 (cf. p. 56, 60, 65).

References

- [Ramesh, 2021b] A. RAMESH, M. PAVLOV, G. GOH, S. GRAY, C. VOSS, A. RADFORD, M. CHEN et I. SUTSKEVER, « Zero-shot text-to-image generation », *arXiv preprint arXiv :2102.12092*, 2021 (cf. p. 40).
- [Ranftl, 2021] R. RANFTL, A. BOCHKOVSKIY et V. KOLTUN, « Vision Transformers for Dense Prediction », *arXiv preprint arXiv :2103.13413*, 2021 (cf. p. 20, 34).
- [Ranzato, 2007] M. RANZATO, C. POULTNEY, S. CHOPRA, Y. LECUN et al., « Efficient learning of sparse representations with an energy-based model », *Advances in neural information processing systems*, t. 19, p. 1137, 2007 (cf. p. 39).
- [Razavi, 2019] A. RAZAVI, A. van den OORD et O. VINYALS, « Generating Diverse High-Fidelity Images with VQ-VAE-2 », *Advances in Neural Information Processing Systems*, 2019 (cf. p. 54, 56, 58).
- [Recht, 2019] B. RECHT, R. ROELOFS, L. SCHMIDT et V. SHANKAR, « Do imagenet classifiers generalize to imagenet ? », *International Conference on Machine Learning*, 2019 (cf. p. 28).
- [Reed, 2017] S. REED, A. van den OORD, N. KALCHBRENNER, S. G. COLMENAREJO, Z. WANG, D. BELOV et N. de FREITAS, « Parallel Multiscale Autoregressive Density Estimation », *International Conference on Machine Learning*, 2017 (cf. p. 57, 60).
- [Rezende, 2015] D. REZENDE et S. MOHAMED, « Variational Inference with Normalizing Flows », *International Conference on Machine Learning*, 2015 (cf. p. 56).
- [Richardson, 2004] I. E. RICHARDSON, *H. 264 and MPEG-4 video compression : video coding for next-generation multimedia*. John Wiley & Sons, 2004 (cf. p. 59).
- [Rifai, 2011] S. RIFAI, P. VINCENT, X. MULLER, X. GLOROT et Y. BENGIO, « Contractive auto-encoders : Explicit invariance during feature extraction », *Proceedings of the 28th international conference on international conference on machine learning*, 2011, p. 833-840 (cf. p. 12).
- [Rippel, 2017] O. RIPPEL et L. BOURDEV, « Real-Time Adaptive Image Compression », *International Conference on Machine Learning*, 2017 (cf. p. 56).
- [Rombach, 2021] R. ROMBACH, A. BLATTMANN, D. LORENZ, P. ESSER et B. OMMER, *High-Resolution Image Synthesis with Latent Diffusion Models*, 2021. arXiv : 2112.10752 [cs.CV] (cf. p. 54).
- [Rosch, 1973] E. H. ROSCH, « Natural categories », *Cognitive Psychology*, 1973 (cf. p. 37).
- [Russakovsky, 2015] O. RUSSAKOVSKY, J. DENG, H. SU, J. KRAUSE, S. SATHEESH, S. MA, Z. HUANG, A. KARPATY, A. KHOSLA, M. BERNSTEIN, A. C. BERG et L. FEI-FEI, « ImageNet Large Scale Visual Recognition Challenge », *IJCV*, 2015 (cf. p. 77).
- [Salimans, 2017] T. SALIMANS, A. KARPATY, X. CHEN et D. KINGMA, « PixelCNN++ : Improving the PixelCNN with discretized logistic mixture likelihood and other modifications », *International Conference on Learning Representations*, 2017 (cf. p. 56).
- [Santurkar, 2018] S. SANTURKAR, D. BUDDEN et N. SHAVIT, « Generative compression », *2018 Picture Coding Symposium (PCS)*, IEEE, 2018, p. 258-262 (cf. p. 56).
- [Sauer, 2021] A. SAUER, K. CHITTA, J. MÜLLER et A. GEIGER, « Projected GANs Converge Faster », *Advances in Neural Information Processing Systems*, 2021 (cf. p. 63, 65).
- [Schönfeld, 2020] E. SCHÖNFELD, B. SCHIELE et A. KHOREVA, « A U-Net based discriminator for generative adversarial networks », *Computer Vision and Pattern Recognition*, 2020 (cf. p. 65).
- [Schuhmann, 2021] C. SCHUHMAN, R. VENCU, R. BEAUMONT, R. KACZMARCZYK, C. MULLIS, A. KATTA, T. COOMBES, J. JITSEV et A. KOMATSUZAKI, « Laion-400m : Open dataset of clip-filtered 400 million image-text pairs », *arXiv preprint arXiv :2111.02114*, 2021 (cf. p. 71, 87).
- [Shen, 2021] Z. SHEN, M. ZHANG, H. ZHAO, S. YI et H. LI, « Efficient attention : Attention with linear complexities », *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021 (cf. p. 19, 27).
- [Shin, 2016] H.-C. SHIN, H. R. ROTH, M. GAO, L. LU, Z. XU, I. NOGUES, J. YAO, D. MOLLURA et R. M. SUMMERS, « Deep convolutional neural networks for computer-aided detection : CNN architectures, dataset characteristics and transfer learning », *IEEE transactions on medical imaging*, t. 35, n° 5, p. 1285-1298, 2016 (cf. p. 1, 94).
- [Sicuranza, 1990] G. SICURANZA, G. ROMPONI et S. MARSÌ, « Artificial neural network for image compression », *Electronics letters*, t. 26, n° 7, p. 477-479, 1990 (cf. p. 56).
- [Silberman, 2012] N. SILBERMAN, D. HOIEM, P. KOHLI et R. FERGUS, « Indoor segmentation and support inference from rgbd images », *European Conference on Computer Vision*, 2012 (cf. p. 76, 83).

- [Simonyan, 2014] K. SIMONYAN et A. ZISSERMAN, « Very deep convolutional networks for large-scale image recognition », *arXiv preprint arXiv :1409.1556*, 2014 (cf. p. 7).
- [Singer, 2022] U. SINGER, A. POLYAK, T. HAYES, X. YIN, J. AN, S. ZHANG, Q. HU, H. YANG, O. ASHUAL, O. GAFNI et al., « Make-a-video : Text-to-video generation without text-video data », *arXiv preprint arXiv :2209.14792*, 2022 (cf. p. 1, 92, 94).
- [Singh, 2023] M. SINGH, Q. DUVAL, K. V. ALWALA, H. FAN, V. AGGARWAL, A. ADCOCK, A. JOULIN, P. DOLLÁR, C. FEICHTENHOFER, R. GIRSHICK et al., « The effectiveness of MAE pre-pretraining for billion-scale pretraining », *arXiv preprint arXiv :2303.13496*, 2023 (cf. p. 93).
- [Socher, 2014] R. SOCHER, A. KARPATY, Q. V. LE, C. D. MANNING et A. Y. NG, « Grounded compositional semantics for finding and describing images with sentences », *Transactions of the Association for Computational Linguistics*, 2014 (cf. p. 72).
- [Sonehara, 1989] N. SONEHARA, « Image data compression using a neural network model », *Proc. Int. Joint Conf. on Neural Networks*, 1989, p. II-35 (cf. p. 56).
- [Song, 2015] S. SONG, S. P. LICHTENBERG et J. XIAO, « Sun rgb-d : A rgb-d scene understanding benchmark suite », *Computer Vision and Pattern Recognition*, 2015 (cf. p. 76, 83).
- [Srinivas, 2021] A. SRINIVAS, T.-Y. LIN, N. PARMAR, J. SHLENS, P. ABBEEL et A. VASWANI, « Bottleneck transformers for visual recognition », *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, p. 16 519-16 529 (cf. p. 11).
- [Strüpler, 2022] Y. STRÜPLER, J. POSTELS, R. YANG, L. van GOOL et F. TOMBARI, « Implicit Neural Representations for Image Compression », *European Conference on Computer Vision*, 2022 (cf. p. 56).
- [Sukhbaatar, 2019] S. SUKHAATAR, E. GRAVE, P. BOJANOWSKI et A. JOULIN, « Adaptive attention span in transformers », *arXiv preprint arXiv :1905.07799*, 2019 (cf. p. 19).
- [Sullivan, 2012] G. J. SULLIVAN, J.-R. OHM, W.-J. HAN et T. WIEGAND, « Overview of the high efficiency video coding (HEVC) standard », *IEEE Transactions on circuits and systems for video technology*, t. 22, n° 12, p. 1649-1668, 2012 (cf. p. 54).
- [Szegedy, 2015] C. SZEGEDY, W. LIU, Y. JIA, P. SERMANET, S. REED, D. ANGUELOV, D. ERHAN, V. VANHOUCKE et A. RABINOVICH, « Going deeper with convolutions », *Computer Vision and Pattern Recognition*, 2015 (cf. p. 1, 7, 94).
- [Tan, 2019] M. TAN et Q. LE, « Efficientnet : Rethinking model scaling for convolutional neural networks », *International Conference on Machine Learning*, PMLR, 2019 (cf. p. 8, 17, 29).
- [Taubman, 2000] D. TAUBMAN, « High performance scalable image compression with EBCOT », t. 9, n° 7, p. 1158-1170, 2000 (cf. p. 54, 59).
- [Taubman, 2012] D. TAUBMAN et M. MARCELLIN, « JPEG2000 Image Compression Fundamentals, Standards and Practice : Image Compression Fundamentals, Standards and Practice », *Springer Science and Business Media*, 2012 (cf. p. 61).
- [Tay, 2020] Y. TAY, M. DEGHANI, D. BAHRI et D. METZLER, « Efficient transformers : A survey.(2020) », *arXiv preprint cs.LG/2009.06732*, 2020 (cf. p. 10).
- [Theis, 2017] L. THEIS, W. SHI, A. CUNNINGHAM et F. HUSZÁR, « Lossy image compression with compressive autoencoders », *International Conference on Learning Representations*, 2017 (cf. p. 56).
- [Thomee, 2016] B. THOMEE, D. A. SHAMMA, G. FRIEDLAND, B. ELIZALDE, K. NI, D. POLAND, D. BORTH et L.-J. LI, « YFCC100M : The new data in multimedia research », *Communications of the ACM*, t. 59, n° 2, p. 64-73, 2016.
- [Tian, 2020] C. TIAN, L. FEI, W. ZHENG, Y. XU, W. ZUO et C.-W. LIN, « Deep learning on image denoising : An overview », t. 131, p. 251-275, 2020 (cf. p. 54).
- [Tian, 2019] Y. TIAN, D. KRISHNAN et P. ISOLA, « Contrastive Multiview Coding », *arXiv preprint arXiv :1906.05849*, 2019 (cf. p. 71-73, 86).
- [Toderici, 2020] G. TODERICI, W. SHI, R. TIMOFTE, J. B. LUCAS THEIS, E. AGUSTSSON, N. JOHNSTON et F. MENTZER, *Workshop and Challenge on Learned Image Compression (CLIC2020)*, CVPR, 2020 (cf. p. 64).
- [Tolias, 2016a] G. TOLIAS, Y. AVRITHIS et H. JÉGOU, « Image search with selective match kernels : aggregation across single and multiple images », *International journal of Computer Vision*, t. 116, n° 3, 2016 (cf. p. 101).
- [Tolias, 2020] G. TOLIAS, T. JENICEK et O. CHUM, « Learning and aggregating deep local descriptors for instance-level recognition », *European Conference on Computer Vision*, 2020 (cf. p. 101-103).

References

- [Tolias, 2016b] G. TOLIAS, R. SICRE et H. JÉGOU, « Particular Object Retrieval With Integral Max-Pooling of CNN Activations », *International Conference on Learning Representations*, 2016 (cf. p. 100).
- [Tolstikhin, 2021] I. TOLSTIKHIN, N. HOULSBY, A. KOLESNIKOV, L. BEYER, X. ZHAI, T. UNTERTHINER, J. YUNG, A. STEINER, D. KEYSERS, J. USZKOREIT, M. LUCIC et A. DOSOVITSKIY, « MLP-Mixer : An all-MLP Architecture for Vision », *arXiv preprint arXiv :2105.01601*, 2021 (cf. p. 20).
- [Tong, 2022] Z. TONG, Y. SONG, J. WANG et L. WANG, « VideoMAE : Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training », *Advances in Neural Information Processing Systems*, 2022 (cf. p. 5, 98).
- [Touvron, 2019] H. TOUVRON, A. VEDALDI, M. DOUZE et H. JÉGOU, « Fixing the train-test resolution discrepancy », *Advances in Neural Information Processing Systems*, 2019 (cf. p. 22).
- [Touvron, 2021a] H. TOUVRON, P. BOJANOWSKI, M. CARON, M. CORD, A. EL-NOUBY, E. GRAVE, A. JOULIN, G. SYNNAEVE, J. VERBEEK et H. JÉGOU, « ResMLP : Feedforward networks for image classification with data-efficient training », *arXiv preprint arXiv :2105.03404*, 2021 (cf. p. 20).
- [Touvron, 2020] H. TOUVRON, M. CORD, M. DOUZE, F. MASSA, A. SABLAYROLLES et H. JÉGOU, « Training data-efficient image transformers and distillation through attention », *arXiv preprint arXiv :2012.12877*, 2020 (cf. p. 5, 11, 17, 18, 24-29, 40, 46, 48, 98).
- [Touvron, 2021b] H. TOUVRON, M. CORD, M. DOUZE, F. MASSA, A. SABLAYROLLES et H. JÉGOU, « Training data-efficient image transformers & distillation through attention », *International Conference on Machine Learning*, 2021 (cf. p. 88).
- [Touvron, 2022a] H. TOUVRON, M. CORD et H. JÉGOU, « Deit iii : Revenge of the vit », *Computer Vision–ECCV 2022 : 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, Springer, 2022, p. 516-533 (cf. p. 11).
- [Touvron, 2022b] H. TOUVRON, M. CORD, A. EL-NOUBY, J. VERBEEK et H. JÉGOU, « Three things everyone should know about vision transformers », *Computer Vision–ECCV 2022 : 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, Springer, 2022, p. 497-515 (cf. p. 11).
- [Touvron, 2021c] H. TOUVRON, M. CORD, A. SABLAYROLLES, G. SYNNAEVE et H. JÉGOU, « Going deeper with Image Transformers », *arXiv preprint arXiv :2103.17239*, 2021 (cf. p. 11, 17-19, 24, 25, 28, 29).
- [Van Horn, 2018] G. VAN HORN, O. MAC AODHA, Y. SONG, Y. CUI, C. SUN, A. SHEPARD, H. ADAM, P. PERONA et S. BELONGIE, « The inaturalist species classification and detection dataset », *Computer Vision and Pattern Recognition*, 2018 (cf. p. 37, 45, 46).
- [Vaswani, 2017a] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. GOMEZ, L. KAISER et I. POLOSUKHIN, « Attention Is All You Need », *Advances in Neural Information Processing Systems*, 2017 (cf. p. 56, 63).
- [Vaswani, 2017b] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER et I. POLOSUKHIN, « Attention is all you need », *Advances in Neural Information Processing Systems*, 2017 (cf. p. 1-3, 8, 9, 17, 20, 25, 35, 40, 52, 75, 94, 96, 97).
- [Vincent, 2008] P. VINCENT, H. LAROCHELLE, Y. BENGIO et P.-A. MANZAGOL, « Extracting and composing robust features with denoising autoencoders », *Proceedings of the 25th international conference on Machine learning*, 2008, p. 1096-1103 (cf. p. 12, 38, 39).
- [Vincent, 2010] P. VINCENT, H. LAROCHELLE, I. LAJOIE, Y. BENGIO, P.-A. MANZAGOL et L. BOTTOU, « Stacked denoising autoencoders : Learning useful representations in a deep network with a local denoising criterion. », *Journal of machine learning research*, t. 11, n° 12, 2010 (cf. p. 39).
- [Wang, 2020a] G. WANG, J. C. YE et B. DE MAN, « Deep learning for tomographic image reconstruction », *Nature Machine Intelligence*, t. 2, n° 12, p. 737-748, 2020 (cf. p. 54).
- [Wang, 2020b] H. WANG, Y. ZHU, B. GREEN, H. ADAM, A. YUILLE et L.-C. CHEN, « Axial-deeplab : Stand-alone axial-attention for panoptic segmentation », *Computer Vision–ECCV 2020 : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, Springer, 2020, p. 108-126 (cf. p. 11).
- [Wang, 2022] R. WANG, D. CHEN, Z. WU, Y. CHEN, X. DAI, M. LIU, Y.-G. JIANG, L. ZHOU et L. YUAN, « BEVT : Bert pretraining of video transformers », *Computer Vision and Pattern Recognition*, 2022 (cf. p. 72).
- [Wang, 2020c] S. WANG, B. LI, M. KHABSA, H. FANG et H. MA, « Linformer : Self-attention with linear complexity », *arXiv preprint arXiv :2006.04768*, 2020 (cf. p. 10, 19, 27).
- [Wang, 2021a] W. WANG, E. XIE, X. LI, D.-P. FAN, K. SONG, D. LIANG, T. LU, P. LUO et L. SHAO, « Pyramid vision transformer : A versatile backbone for dense prediction without convolutions », *arXiv preprint arXiv :2102.12122*, 2021 (cf. p. 11, 17, 19, 27, 33, 34).

- [Wang, 2018] X. WANG, R. GIRSHICK, A. GUPTA et K. HE, « Non-local neural networks », *Computer Vision and Pattern Recognition*, 2018 (cf. p. 1, 2, 10, 94, 95).
- [Wang, 2015] X. WANG et A. GUPTA, « Unsupervised learning of visual representations using videos », *Proceedings of the IEEE international conference on computer vision*, 2015, p. 2794-2802 (cf. p. 13).
- [Wang, 2021b] Z. WANG, J. CHEN et S. C. H. HOI, « Deep Learning for Image Super-Resolution : A Survey », t. 43, n° 10, p. 3365-3387, 2021 (cf. p. 54).
- [Wang, 2004] Z. WANG, A. C. BOVIK, H. R. SHEIKH et E. P. SIMONCELLI, « Image quality assessment : from error visibility to structural similarity », t. 13, n° 4, p. 600-612, 2004 (cf. p. 64).
- [Wang, 2003] Z. WANG, E. P. SIMONCELLI et A. C. BOVIK, « Multiscale structural similarity for image quality assessment », *Asilomar Conference on Signals, Systems & Computers*, 2003 (cf. p. 61, 64).
- [Wei, 2022] Y. WEI, H. HU, Z. XIE, Z. ZHANG, Y. CAO, J. BAO, D. CHEN et B. GUO, « Contrastive Learning Rivals Masked Image Modeling in Fine-tuning via Feature Distillation », *arXiv preprint arXiv :2205.14141*, 2022 (cf. p. 73).
- [Wightman, 2019] R. WIGHTMAN, *PyTorch Image Models*, <https://github.com/rwightman/pytorch-image-models>, 2019 (cf. p. 25, 28).
- [Witten, 1987] I. H. WITTEN, R. M. NEAL et J. G. CLEARY, « Arithmetic coding for data compression », *Communications of the ACM*, t. 30, n° 6, p. 520-540, 1987 (cf. p. 54).
- [Wu, 2018a] Y. WU et K. HE, « Group normalization », *European Conference on Computer Vision*, 2018 (cf. p. 23).
- [Wu, 2018b] Z. WU, Y. XIONG, S. X. YU et D. LIN, « Unsupervised feature learning via non-parametric instance discrimination », *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, p. 3733-3742 (cf. p. 40, 74).
- [Xiao, 2018] T. XIAO, Y. LIU, B. ZHOU, Y. JIANG et J. SUN, « Unified perceptual parsing for scene understanding », *European Conference on Computer Vision*, 2018 (cf. p. 34, 47, 48).
- [Xie, 2016] J. XIE, R. GIRSHICK et A. FARHADI, « Unsupervised deep embedding for clustering analysis », *International conference on machine learning*, PMLR, 2016, p. 478-487 (cf. p. 40).
- [Xie, 2017] S. XIE, R. GIRSHICK, P. DOLLÁR, Z. TU et K. HE, « Aggregated residual transformations for deep neural networks », *Computer Vision and Pattern Recognition*, 2017 (cf. p. 33, 34).
- [Xie, 2021] Z. XIE, Y. LIN, Z. YAO, Z. ZHANG, Q. DAI, Y. CAO et H. HU, « Self-Supervised Learning with Swin Transformers », *arXiv preprint arXiv :2105.04553*, 2021 (cf. p. 31).
- [Xiong, 2021a] Y. XIONG, Z. ZENG, R. CHAKRABORTY, M. TAN, G. FUNG, Y. LI et V. SINGH, « Nyströmformer : A Nyström-Based Algorithm for Approximating Self-Attention », *arXiv preprint arXiv :2102.03902*, 2021 (cf. p. 19).
- [Xiong, 2021b] Y. XIONG, Z. ZENG, R. CHAKRABORTY, M. TAN, G. FUNG, Y. LI et V. SINGH, « Nyströmformer : A nyström-based algorithm for approximating self-attention », *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021 (cf. p. 10).
- [Xu, 2022] H. XU, J. LI, A. BAEVSKI, M. AULI, W. GALUBA, F. METZE et C. FEICHTENHOFER, « Masked autoencoders that listen », *Advances in Neural Information Processing Systems*, 2022 (cf. p. 5, 79, 98).
- [Xu, 2016] J. XU, T. MEI, T. YAO et Y. RUI, « Msr-vtt : A large video description dataset for bridging video and language », *Computer Vision and Pattern Recognition*, 2016 (cf. p. 77).
- [Xue, 2022] H. XUE, Y. SUN, B. LIU, J. FU, R. SONG, H. LI et J. LUO, « CLIP-ViP : Adapting Pre-trained Image-Text Model to Video-Language Representation Alignment », *arXiv preprint arXiv :2209.06430*, 2022 (cf. p. 72).
- [Yang, 2016] J. YANG, D. PARIKH et D. BATRA, « Joint unsupervised learning of deep representations and image clusters », *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016 (cf. p. 40).
- [Yang, 2020] Y. YANG, R. BAMLER et S. MANDT, « Improving inference for neural image compression », *Advances in Neural Information Processing Systems*, 2020 (cf. p. 56).
- [Yosinski, 2014] J. YOSINSKI, J. CLUNE, Y. BENGIO et H. LIPSON, « How transferable are features in deep neural networks? », *arXiv preprint arXiv :1411.1792*, 2014 (cf. p. 37).
- [Youwang, 2022] K. YOUWANG, K. JI-YEON et T.-H. OH, « CLIP-Actor : Text-Driven Recommendation and Stylization for Animating Human Meshes », *ECCV*, 2022 (cf. p. 73).

References

- [Yu, 2021] J. YU, X. LI, J. Y. KOH, H. ZHANG, R. PANG, J. QIN, A. KU, Y. XU et Y. W. JASON BALDRIDGE, « Vector-quantized Image Modeling with Improved VQGAN », *International Conference on Learning Representations*, 2021 (cf. p. 54, 58, 60, 63).
- [Yu, 2022] J. YU, Z. WANG, V. VASUDEVAN, L. YEUNG, M. SEYEDHOSSEINI et Y. WU, « Coca : Contrastive captioners are image-text foundation models », *arXiv preprint arXiv :2205.01917*, 2022 (cf. p. 15, 71, 72).
- [Yuan, 2021a] K. YUAN, S. GUO, Z. LIU, A. ZHOU, F. YU et W. WU, « Incorporating Convolution Designs into Visual Transformers », *arXiv preprint arXiv :2103.11816*, 2021 (cf. p. 19).
- [Yuan, 2021b] L. YUAN, Y. CHEN, T. WANG, W. YU, Y. SHI, Z. JIANG, F. E. TAY, J. FENG et S. YAN, « Tokens-to-token ViT : Training vision transformers from scratch on ImageNet », *arXiv preprint arXiv :2101.11986*, 2021 (cf. p. 19).
- [Yuan, 2021c] L. YUAN, D. CHEN, Y.-L. CHEN, N. CODELLA, X. DAI, J. GAO, H. HU, X. HUANG, B. LI, C. LI et al., « Florence : A new foundation model for computer vision », *arXiv preprint arXiv :2111.11432*, 2021 (cf. p. 71, 72).
- [Yun, 2019] S. YUN, M. JEONG, R. KIM, J. KANG et H. J. KIM, « Graph transformer networks », *Advances in neural information processing systems*, t. 32, 2019 (cf. p. 5, 98).
- [Zaheer, 2020] M. ZAHEER, G. GURUGANESH, A. DUBEY, J. AINSLIE, C. ALBERTI, S. ONTANON, P. PHAM, A. RAVULA, Q. WANG, L. YANG et al., « Big bird : Transformers for longer sequences », *arXiv preprint arXiv :2007.14062*, 2020 (cf. p. 9, 19).
- [Zbontar, 2021] J. ZBONTAR, L. JING, I. MISRA, Y. LECUN et S. DENY, « Barlow twins : Self-supervised learning via redundancy reduction », *arXiv preprint arXiv :2103.03230*, 2021 (cf. p. 14, 40).
- [Zhai, 2022] X. ZHAI, X. WANG, B. MUSTAFA, A. STEINER, D. KEYSERS, A. KOLESNIKOV et L. BEYER, « Lit : Zero-shot transfer with locked-image text tuning », *Computer Vision and Pattern Recognition*, 2022 (cf. p. 5, 15, 72, 98).
- [Zhang, 2021] P. ZHANG, X. DAI, J. YANG, B. XIAO, L. YUAN, L. ZHANG et J. GAO, « Multi-Scale Vision Longformer : A New Vision Transformer for High-Resolution Image Encoding », *arXiv preprint arXiv :2103.15358*, 2021 (cf. p. 17, 20, 33).
- [Zhang, 2018] R. ZHANG, P. ISOLA, A. EFROS, E. SHECHTMAN et O. WANG, « The Unreasonable Effectiveness of Deep Features as a Perceptual Metric », *Computer Vision and Pattern Recognition*, 2018 (cf. p. 62).
- [Zhang, 2022] R. ZHANG, Z. GUO, W. ZHANG, K. LI, X. MIAO, B. CUI, Y. QIAO, P. GAO et H. LI, « Pointclip : Point cloud understanding by clip », *Computer Vision and Pattern Recognition*, 2022 (cf. p. 73, 74).
- [Zhang, 2016] R. ZHANG, P. ISOLA et A. A. EFROS, « Colorful image colorization », *European conference on computer vision*, Springer, 2016, p. 649-666 (cf. p. 39).
- [Zhao, 2020] H. ZHAO, J. JIA et V. KOLTUN, « Exploring Self-Attention for Image Recognition », *Computer Vision and Pattern Recognition*, 2020 (cf. p. 20).
- [Zheng, 2020] S. ZHENG, J. LU, H. ZHAO, X. ZHU, Z. LUO, Y. WANG, Y. FU, J. FENG, T. XIANG, P. H. TORR et al., « Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers », *arXiv preprint arXiv :2012.15840*, 2020 (cf. p. 17, 20, 34).
- [Zhong, 2020] Z. ZHONG, L. ZHENG, G. KANG, S. LI et Y. YANG, « Random erasing data augmentation », *Conference on Artificial Intelligence*, 2020 (cf. p. 86, 87).
- [Zhou, 2014] B. ZHOU, A. LAPEDRIZA, J. XIAO, A. TORRALBA et A. OLIVA, « Learning Deep Features for Scene Recognition using Places Database », *Advances in Neural Information Processing Systems*, 2014 (cf. p. 77).
- [Zhou, 2017] B. ZHOU, H. ZHAO, X. PUIG, S. FIDLER, A. BARRIUSO et A. TORRALBA, « Scene parsing through ade20k dataset », *Computer Vision and Pattern Recognition*, 2017 (cf. p. 34, 45, 46).
- [Zhou, 2022] X. ZHOU, R. GIRDHAR, A. JOULIN, P. KRÄHENBÜHL et I. MISRA, « Detecting Twenty-thousand Classes using Image-level Supervision », *European Conference on Computer Vision*, 2022 (cf. p. 73, 81, 82).
- [Zhu, 2021] Y. ZHU, Y. YANG et T. COHEN, « Transformer-based Transform Coding », *International Conference on Learning Representations*, 2021 (cf. p. 59).

RÉSUMÉ

Les transformateurs ont révolutionné l'apprentissage de la représentation dans de nombreuses modalités, obtenant des résultats de pointe dans le traitement du langage naturel, la vision par ordinateur, la parole et bien d'autres domaines. Cette thèse explore le potentiel des modèles de transformateurs pour la vision par ordinateur. Nous proposons des innovations architecturales pour surmonter certaines de leurs limites. Nous développons des méthodes de pré-entraînement auto-supervisé efficaces en termes d'échantillons, et considérons l'utilisation de ces transformateurs dans un contexte d'apprentissage multimodal. Dans un premier temps, nous proposons l'attention à covariance croisée pour réduire la complexité quadratique de l'attention d'origine et obtenir des performances similaires avec une empreinte mémoire et un coût de calcul moindres, ce qui permet d'appliquer les transformateurs de vision à des images à plus haute résolution. Nous étudions ensuite le pré-entraînement auto-supervisé pour les transformateurs de vision. Nous proposons SplitMask, une méthode de débruitage automatique basée sur la modélisation d'images masquées. Contrairement aux méthodes de plongements conjointes, SplitMask ne nécessite pas d'ensembles de données de pré-entraînement à grande échelle et peut être appliqué à diverses données visuelles. SplitMask est aussi performant que les méthodes de plongements conjoints lorsqu'il est entraîné sur des ensembles de données deux fois plus petits, ce qui met en évidence l'amélioration de l'efficacité d'apprentissage avec peu de données. En outre, nous appliquons la modélisation d'images masquée à la compression d'images neuronales sous la forme d'un modèle entropique amélioré. Cela permet d'obtenir de bonnes performances en matière de débit-distorsion dans les régimes où la compression d'image est extrême, tels la taille d'un SMS ou d'un tweet. Enfin, nous proposons ImageBind, une méthode d'apprentissage d'un espace de plongement partagé entre six modalités. En résumé, cette thèse démontre le potentiel des transformateurs pour la vision par ordinateur grâce à des innovations architecturales, de nouveaux objectifs auto-supervisés et un transfert de connaissances multimodal. Les méthodes proposées dans cette thèse repoussent les limites des transformateurs en vision en améliorant leur passage à l'échelle et leur généralité, en permettant un apprentissage de la représentation plus efficace en termes d'échantillons, et en facilitant le transfert entre les modalités.

MOTS CLÉS

transformateurs de vision, apprentissage auto-supervisé, apprentissage faiblement supervisé, apprentissage multimodal, compression d'image.

ABSTRACT

Transformers have revolutionized representation learning across modalities, achieving state-of-the-art results in natural language processing, computer vision, speech, and beyond. This thesis explores the potential of Transformer models for computer vision. We propose architectural innovations to overcome their limitations, developing sample-efficient self-supervised pre-training methods, and advancing multimodal learning with Transformers. First, we propose Cross-Covariance Attention to reduce the quadratic complexity of self-attention achieving similar performance as vision transformers with lower memory footprint and computational cost, enabling the application of vision transformers to higher-resolution images. We then investigate self-supervised pre-training for vision transformers. We propose SplitMask, a denoising autoencoding method based on masked image modeling. Unlike joint embedding methods, SplitMask does not require large-scale pre-training datasets and can be applied to diverse visual data. SplitMask matches the performance of joint embedding methods when pre-trained on datasets two orders of magnitude smaller, highlighting its improved sample efficiency. Moreover, we apply masked image modeling to neural image compression in the form of an improved entropy model yielding a strong rate-distortion performance and enabling the compression of images to the size of a short SMS or tweet. Finally, we propose ImageBind, a method for learning a shared embedding space across six modalities. ImageBind leverages the abundance of images and text on the web to enable transfer to modalities with scarce annotations like depth, thermal, audio, and IMU. In summary, this thesis demonstrates the potential of Transformers for computer vision through architectural innovations, new self-supervised objectives, and multimodal knowledge transfer. The methods proposed in this thesis push the boundaries of transformers in vision by enhancing their scalability and generality, enabling more sample-efficient representation learning, and facilitating transfer across modalities.

KEYWORDS

vision transformers, self-supervised learning, weakly supervised learning, multimodal learning, image compression