Transformer-Based Zero-Shot Detection via Contrastive Learning

Wei Liu, Hui Chen, Yongqiang Ma, Jianji Wang, Nanning Zheng

Transformer-based Zero-Shot Detection via Contrastive Learning

Wei Liu[1][0000−0002−0899−6442], Hui Chen[1][0000−0002−7108−4067], Yongqiang Ma[1][0000−0002−6063−5601], Jianji Wang[1][0000−0002−4284−3933], and Nanning Zheng[1,2][0000−0003−1608−8257]

[1] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
[2] Correspondence: nnzheng@xjtu.edu.cn
{xianjiaodaliuwei,chenhui0622}@stu.xjtu.edu.cn
musayq@xjtu.edu.cn,wangjianji@mail.xjtu.edu.cn

Abstract. Zero-Shot Detection (ZSD) is a challenging computer vision problem that enables simultaneous classification and localization of previously unseen objects via auxiliary information. Most of the existing methods learn a biased visual-semantic mapping function, which prefers predicting seen classes during testing, and they only focus on region of interest and ignore contextual information in an image. To tackle these problems, we propose a novel framework for ZSD named Transformer-based Zero-Shot Detection via Contrastive Learning (TZSDC). The proposed TZSDC contains four components: transformer-based backbone, Foreground-Background (FB) separation module, Instance-Instance Contrastive Learning (IICL) module, and Knowledge-Transfer (KT) module. The transformer backbone encodes long-range contextual information with less inductive bias. The FB module separates foreground and background by scoring objectness from images. The IICL module optimizes the visual structure in embedding space to make it more discriminative and the KT module transfers knowledge from seen classes to unseen classes via category similarity. Benefiting from these modules, the accurate alignment between the contextual visual features and semantic features can be achieved. Experiments on MSCOCO well validate the effectiveness of the proposed method for ZSD and generalized ZSD.

Keywords: Zero-Shot Detection, Transformer, Contrastive Learning

1 Introduction

In recent years, deep learning has made great progress in object detection [1,8]. However, these methods strongly rely on large-scale annotated data. When lacking sufficient annotated data, the performance of these methods drops rapidly [12,13]. In reality, it is difficult for detectors to generalize to new target domains where annotated data is scarce or absent. However, it's easier for humans to recognize a new class by analogy with similar objects they know.

In order to solve the above problems, Zero-Shot Object detection (ZSD) [7,9, 14,17,18] is proposed to classify and locate unseen classes with only seen classes contained during training. Most ZSD models [9, 14, 17] usually learn a visual-semantic mapping function using visual data and related semantic information of seen classes. At the testing stage, they use the learned model to map visual features into an embedding space and perform the nearest neighbor search to predict unseen classes. Several studies [7,26] use a generative model to synthesize features of unseen classes, and then retrain a classifier of unseen classes, turning zero-shot learning into supervised learning.

These methods [9,17] learn the model on seen classes while ignoring semantic information available for unseen classes, making the model significantly biased towards the seen classes when testing, which will greatly degrade the performance of ZSD and generalized ZSD (GZSD). Besides, current zero-shot detection networks that are based on one-stage or two-stage detection methods for secondary design, only focus on local information near an object's region of interest and do not explicitly encode long-range dependencies between objects, which are crucial to detect multiple objects in an image.

In this paper, we develop a novel framework for ZSD called Transformer-based Zero-Shot Detection via Contrastive Learning (TZSDC), which consists of four modules: transformer-based detector named Deformable DETR [27], Foreground-Background (FB) separation module, Instance-Instance Contrastive Learning (IICL) module, and Knowledge-Transfer (KT) module. We use the Deformable DETR to encode the input images for contextual features. To alleviate the confusion between unseen classes and backgrounds, the FB module makes full use of the existing visual background to compute an objectness score for the output query embeddings. Meanwhile, in order to make visual features in the embedding space more discriminative, the IICL module performs contrastive learning between instances to optimize the visual manifold structure, so that the intra-class spacing is more compact and the inter-class distances are far away from each other. To alleviate the bias problem that the learned model prefers seen classes, the KT module realizes the knowledge transfer from seen classes to unseen classes via category similarity.

The main contributions of the paper can be summarized as (i) We propose a novel framework TZSDC that integrates the transformer and contrastive learning into zero-shot detection, achieving an accurate visual-semantic alignment. (ii) We design a FB module to alleviate the confusion of unseen classes and the background, and a KT module to realize the knowledge transfer from seen classes to unseen classes through category similarity. (iii) Experiments on MSCOCO verify that the proposed method can effectively improve the performance on ZSD and GZSD tasks.

2 Related work

Object Detection In the past few years, object detection has received huge attention and developed rapidly. For traditional object detection frameworks,

there are mainly two types, one-stage methods such as SSD [11], YOLO [19], FCOS [22], and two-stage methods such as Faster R-CNN [20], R-FCN [3]. Their general methods are to generate bounding boxes, determine which box contains objects, and then classify high-confidence boxes. However, due to the design of convolution, they only focus on local information near the region of interest. In recent years, Transformer [2, 27] is developing rapidly in the field of computer vision, DETR [2] applies the transformer to the field of target detection, and Deformable DETR [27] adopts the idea of deformable convolution [4], which integrates multi-scale information and accelerates the convergence speed of DETR. DETR [2] can encode long-range dependencies at multi-scales to enrich contextual information. In this work, we choose Deformable DETR as our basic detection framework.

Zero-shot Learning (ZSL) ZSL is a classic task in computer vision. It aims to use seen examples to train networks and reason about unseen classes with the help of semantic information. Zero-shot learning can be divided into embedding models and generative models. The embedding models [5, 21] mainly learn a mapping function to convert visual features and semantics into an embedding space and then classify by searching the nearest semantic descriptor in the embedding space. In our work, we adopt a basic visual-semantic embedding model, take the latent space as the embedding space, and exploit the similarity between seen and unseen classes to explicitly transfer knowledge from the source class to the target class, promoting better visual semantic alignment.

Zero-Shot Object Detection (ZSD) ZSD is a recently proposed task that can identify and localize unseen objects. Most of them focus on learning embedding functions from visual space to semantic space. MS-Zero [6] designed an asymmetric mapping method to reduce the impact of new noise on the classifier, which first maps visual features to semantic space respectively, and then maps semantic features to visual space. Polarity Loss [16] was proposed to find a more suitable alignment of visual and semantic information, which is an improvement on the basis of Focal Loss to solve the problem of imbalance between positive and negative samples. BLRPN-ZSD [25] designed a background perceptron to use external annotations to solve the confusion of unseen classes and backgrounds. In our work, we introduce the foreground objectness branch to learn from existing visual background data to better separate unseen classes from the background. At the same time, we introduce a contrastive network in the classification branch to explicitly transfer knowledge from the source class to the target class, which is helpful to alleviate the domain transfer problem and the visual-semantic gap.

3 Method

Problem Settings In ZSD, we are given S seen classes in $\mathcal{Y}^s$ and U seen classes in $\mathcal{Y}^u$, where seen classes and unseen classes are disjoint. We can denote

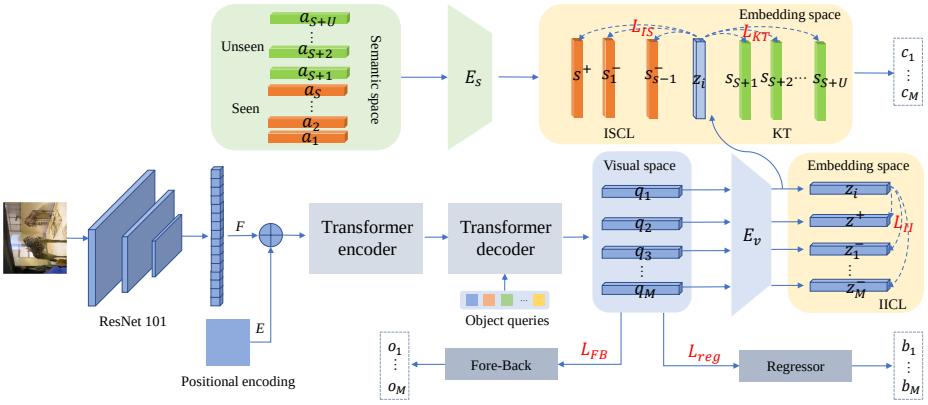

Fig. 1. Illustration of TZSDC. First, TZSDC extracts multi-scale features F from the input image with ResNet101. Next $E + F$ are fed into the transformer encoder and decoder, where E represents the position encoding. After the decoder, a set of query embeddings $\mathcal{Q} = \{\boldsymbol{q}_i\}_{i=1}^{M}, \boldsymbol{q}_i \in R^D$ are obtained from M learnable object quries and FB module seperates the foreground and background. Next, E_v and E_S are used to map query embeddings $Q = \{\boldsymbol{q}_i\}_{i=1}^{M}$ and semantic embeddings $A = \{\boldsymbol{a}_c\}_{c=1}^{S+U}$ into a common embedding space. IICL and ISCL are performed for better visual-semantic alignment through instance-instance and instance-semantic contrastive learning.

that $\mathcal{Y}^s \cap \mathcal{Y}^u = \emptyset$, $\mathcal{Y}^s \cup \mathcal{Y}^u = \mathcal{Y}$. We use $\mathcal{Y}^s = \{Y_1, Y_2, \cdots, Y_S\}$ to represent the seen classes and $\mathcal{Y}^u = \{Y_{S+1}, Y_{S+2}, \cdots, Y_{S+U}\}$ to represent unseen classes. Let $\mathcal{D}^{tr} = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}^s\}_{i=1}^{N}$ be the training dataset containing N images. During training and testing, semantic word-vectors $A = \{\boldsymbol{a}_c\}_{c=1}^{S+U}$ are provided for each class $c \in \mathcal{Y}^s \cup \mathcal{Y}^u$ to conduct a knowledge transfer. The task of ZSD is to learn a detector to recognize and localize unseen classes during testing.

3.1 Overall Architecture

The overall framework for zero-shot detection is shown in Fig. 1. It adopts the standard Deformable DETR [27] as a backbone for ZSD by introducing (i) a Foreground-Background (FB) separation module to reduce confusion of unseen classes and backgrounds; (ii) an Instance-Instacne/Instance-Semantic Contrastive Learning (IICL/ISCL) module to optimize the visual manifold structure in the embedding space; (iii) a Knowledge Transfer (KT) module to transfer knowledge from seen to unseen classes via category similarity.

Given an input image $\boldsymbol{x} \in \mathcal{X}$, ResNet 101 extracts multi-scale features F with a sine-cosine position encoding E added to preserve the position information. Then multi-scale features with position encoding are fed into the transformer encoder and decoder which contain deformable convolution [4]. Driven by cross-attention and self-attention mechanism, the decoder converts a set of M learnable object queries into a set of M query embeddings $\mathcal{Q} = \{\boldsymbol{q}_i\}_{i=1}^{M}, \boldsymbol{q}_i \in R^D$ which

contain the relative positional relationship between objects. And then the query embeddings $\mathcal{Q}$ are fed into the regressor, FB module, IICL module, and KT module. The FB module computes an objectness score for the output query embeddings, using the existing visual background data to achieve foreground and background separation by binary cross-entropy loss. The mapping functions E_v and E_S are used to map query embeddings $Q = \{\boldsymbol{q}_i\}_{i=1}^{M}$ and semantic embeddings $A = \{\boldsymbol{a}_c\}_{c=1}^{S+U}$ into a common embedding space. Then in IICL module, query embeddings with the same label are regarded as positive samples $\boldsymbol{z}^+$, the others are regarded as negative samples $\boldsymbol{z}^-$, and instance-instance contrastive learning is performed to optimize the visual manifold structure. Besides, the ISCL module selects semantic feature $\boldsymbol{s}_i$ corresponding to the class of $\boldsymbol{z}_i$ as the only positive sample, the remaining $S-1$ semantic features as negative samples to perform instance-semantic contrastive learning. Meanwhile, unseen semantic embeddings are used in the KT module to enable knowledge transfer via category similarity between seen and unseen embeddings.

3.2 FB Module

The decoder outputs $\mathcal{Q} = \{\boldsymbol{q}_i\}_{i=1}^{M}, \boldsymbol{q}_i \in R^D$ from M learnable object queries, each of which has a corresponding bounding box and category. And then the classifier recognizes the query embeddings into $S+U+1$ classes: S seen classes, U unseen classes, and background. However, most query embeddings will be predicted as background due to a lack of supervision from visual images of unseen classes. In order to alleviate the confusion between unseen classes and backgrounds, we introduce $FB : R^D \rightarrow [0, 1]$ to separate the foreground and background.

Considering M is generally larger than the number of categories $S + U$, for those queries without actual categories, we regard them as backgrounds. FB module computes an objectness score o_i for query embeddings q_i. The objective of the FB module is to assign higher confidence to query embeddings corresponding to foreground objects than to those corresponding to the backgrounds. Therefore, the foreground and background separation loss function is defined as follows:

$$\mathcal{L}_{FB} = -\sum_{i=1}^{M} m_j \log o_i + (1 - m_i) \log (1 - o_i) \tag{1}$$

where:

$$m_{ij} = \begin{cases} 1, & y_i \; is \; the \; foreground \\ 0, & y_i \; is \; the \; background \end{cases} \tag{2}$$

3.3 IICL/ISCL Module and KT Module

Due to the gap between visual and semantic features, using visual space or semantic space as a common embedding space is not ideal. In order to align the two spaces, we use two functions E_v, E_s to map query embeddings $Q = \{\boldsymbol{q}_i\}_{i=1}^{M}$

and semantic embeddings $A = \{\boldsymbol{a}_c\}_{c=1}^{S+U}$ into a common embedding space to optimize the manifold of visual and semantic features.

$$\boldsymbol{z}_i = E_v\left(\boldsymbol{q}_i\right) \tag{3}$$

$$\boldsymbol{s}_j = E_s\left(\boldsymbol{a}_j\right) \tag{4}$$

where $\boldsymbol{q}_i$ represents the query feature of the i-th class, $\boldsymbol{a}_j$ represents the semantic feature of the j-th class.

Instance-Instance Contrastive Learning (IICL) Instances with similar semantic attributes are usually close together in the embedding space, thus leading to misclassifications. To reduce such misclassifications, we need to optimize the manifold structure of visual features in the embedding space. Inspired by the alignment (closeness of features from positive pairs) and uniformity of the feature distribution to contrastive loss [23], we utilize contrastive loss to learn discriminative features. Given an input image, query embeddings with the same label are regarded as positive samples $\boldsymbol{z}^+$, and the others are regarded as negative samples $\boldsymbol{z}^-$. We assume that there are P_i positive samples and N_i negative samples for i-th object in the input image. The IICL loss $\mathcal{L}_{II}$ is as follows:

$$\mathcal{L}_{II} = E\left[-\log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}^+\right)/\tau_v)}{\sum_{k=1}^{P_i} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_k^+/\tau_v\right) + \sum_{j=1}^{N_i} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_j^-/\tau_v\right)}\right] \tag{5}$$

where τ_v is the temperature parameter of $\mathcal{L}_{II}$.

Instance-Semantic Contrastive Learning (ISCL) The above loss drives positive pairs between visual features compact. Besides, in order to achieve an accurate visual-semantic alignment, we use the semantic information of the source class for supervision and select semantic feature $\boldsymbol{s}_i$ corresponding to the class of $\boldsymbol{z}_i$ as the only positive sample, the remaining $S-1$ semantic features as negative samples. The instance-semantic contrastive loss $\mathcal{L}_{IS}$ can be calculated as follows:

$$\mathcal{L}_{IS} = E\left[-\log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{s}^+\right)/\tau_s)}{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{s}^+/\tau_s\right) + \sum_{j=1}^{S-1} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{s}_j^-/\tau_s\right)}\right] \tag{6}$$

where τ_s is the temperature parameter of $\mathcal{L}_{IS}$ and S is the number of seen classes.

Knowledge Transfer (KT) The task of ZSD is to recognize and locate unseen classes. If only visual information and semantic embeddings of seen classes are used during training, it is easy to bias the model to seen classes. In order to alleviate the bias problem, we use category similarity between seen classes and unseen classes to transfer knowledge from the source classes to the target classes.

We assume that the semantic attribute of the unseen class can be obtained by the linear combination of the attributes of seen classes, which is also widely

adopted in ZSL [24]. For example, "zebra" has a shape like "horse", and the color is black and white like "panda". Inspired by this observation, we use least square regression (LSR) to obtain the reconstruction coefficient of each seen class semantic attribute. The reconstruction coefficient is the category similarity, which is calculated as follows:

$$\boldsymbol{d}_u = \arg\min_{\boldsymbol{d}_u} \left\| \boldsymbol{a}_u - \sum_{k=1}^{S} \boldsymbol{a}_k d_{uk} \right\|_2^2 + \beta \left\| \boldsymbol{d}_u \right\|_2 \tag{7}$$

Where d_{uk} is the category similarity between the u-th unseen class and k-th seen class, and $a_u \in \{\boldsymbol{a}_c\}_{c=S+1}^{S+U}$, $a_k \in \{\boldsymbol{a}_c\}_{c=1}^{S}$, β is the regularization coefficient. After we get the similarity $\boldsymbol{d}_u$ between unseen classes and seen classes, we can use the images of seen classes to learn the similar unseen classes. The knowledge transfer loss $\mathcal{L}_{\mathrm{KT}}$ is defined as:

$$\mathcal{L}_{KT} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=S+1}^{S+U} d_{jy_i} \log \widetilde{\zeta_{ij}} + (1 - d_{jy_i}) \log \left(1 - \widetilde{\zeta_{ij}}\right) \tag{8}$$

Where $\zeta_{ij} = \boldsymbol{z}_i \cdot \boldsymbol{s}_j$, $\boldsymbol{z}_i$, $\boldsymbol{s}_j$ are calculated by Eq.(3) and Eq.(4). $\widetilde{\zeta_{ij}}$ is the normalization of ζ_{ij}.

Regression Loss As with Deformable DETR [27], we use a linear combination of the L1 loss and the IOU loss as our regression loss $\mathcal{L}_{reg}$:

$$\mathcal{L}_{reg} = \lambda_{iou} \mathcal{L}_{iou} \left(b_i, \hat{b}_i\right) + \lambda_{L1} \left\| \left(b_i - \hat{b}_i\right) \right\|_1 \tag{9}$$

3.4 Training and Inference

Training The proposed method includes FB loss $\mathcal{L}_{FB}$, regression loss $\mathcal{L}_{reg}$, IICL loss $\mathcal{L}_{II}$, ISCL loss $\mathcal{L}_{IS}$, KT loss $\mathcal{L}_{KT}$. The total loss function is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{reg} + \mathcal{L}_{IS} + \alpha \mathcal{L}_{KT} + \gamma \mathcal{L}_{II} + \lambda \mathcal{L}_{FB}. \tag{10}$$

Where α, γ, λ are hyper-parameters to balance each loss term. We train our model using a two-stage training approach. In the first stage, we use all seen classes to train a Deforamble DETR framework that only contains the FB module. In the second stage, we replace the classifier in Deformable DETR with our current IICL module, ISCL module, and KT module. Then the model is finetuned based on first-stage parameters.

Inference Given a test image I, M object query embeddings $Q = \{\boldsymbol{q}_i\}_{i=1}^{M}$ are computed, and then bounding boxes are obtained with the regressor. Next, object query embeddings are mapped into the common embedding space and are used to predict the class by nearest neighbor search.

4 Experiments

4.1 Experimental Settings

Datasets We evaluate our method on MSCOCO 2014 [10] which contains 82,783 training images and 40,504 validation images. For MSCOCO 2014 with 80 categories, We follow the 65/15 split [16]. As for semantic embeddings in the classification subnet, we use 300-dimensional vectors from word2vec [16] for MSCOCO 2014.

Evaluation Protocol For MSCOCO 2014, we choose mAP and Recall@100 as our evaluation metrics. We conduct experiments under both standard and generalized settings and evaluate the Harmonic Mean (HM) to show the performance of GZSD.

Implementation Details We choose ResNet101 which is pretrained on ImageNet to extract multi-scale features. The transformer encoder-decoder structure is consistent with the standard Deformable DETR. The dimension of object queries is 512 and M is set to 100. The regressor consists of 3 multi-layer perceptrons (MLP), the FB module is one fully-connected layer and the mapping function E_s, E_v are accomplished by one fully-connected layer. Hyperparameters α, γ, λ in Eq.(10) are set as 0.2, 0.3, 0.1. And the temperature parameter τ_v, τ_s in Eq.(5), Eq.(6) and Hyperparameters $\lambda_{iou}, \lambda_{L1}$ in Eq.(9) are set to 0.1, 0.1, 2.0, 5.0, and The TZSDC framework is trained using SGD optimizer with the learning rate of 0.01 and momentum of 0.9 for 50 epochs in the first stage and the learning rate of 0.002 and momentum of 0.999 for 20 epochs in the second stage.

4.2 Comparison with Other Methods

As shown in Table 1, we compare the performance of the proposed model with TL-ZSD [15], PL [16], BLC [25], SU-ZSD [7] on MSCOCO for both ZSD and GZSD. As can be seen, our method achieves the best performance on both mAP and recall for ZSD. Compared with the second-best method SU-ZSD [7], the mAP of our method is improved from 19.00% to 19.58%, and the recall is improved from 54.00% to 56.45%, which indicates that our method improves the discriminatory ability for unseen classes. For GZSD, our method achieves the best performance in the unseen class. The unseen performance is improved without sacrificing the seen accuracy too much, and our HM value is competitive to the generative model SU-ZSD [7]. This shows that our model has learned a good visual-semantic alignment model, which realizes knowledge transfer from seen classes to unseen classes.

As can be seen from Table 2, which shows the class-wise AP performance for ZSD, our method improves the mAP of "mouse", "hotdog", "hairdrier" which are not similar to the seen classes at all, indicating that our model extracts the

Table 1. Comparison with other methods for ZSD/GZSD on MSCOCO dataset. We report both mAP(%) and recall@100. Bold represents the best result.

Method	S/U Split	ZSD		GZSD					
				Seen		Unseen		HM	
		mAP	Recall	mAP	Recall	mAP	Recall	mAP	Recall
TL-ZSD [15]	65/15	14.57	48.15	28.79	54.14	14.05	37.16	18.89	44.07
PL [16]	65/15	12.40	37.72	34.07	36.38	12.40	37.16	18.89	44.07
BLC [25]	65/15	13.10	51.65	36.00	56.39	13.10	51.65	19.20	53.92
SU-ZSD [7]	65/15	19.00	54.00	**36.90**	**57.70**	19.00	**53.90**	**25.08**	**55.74**
Ours	65/15	**19.58**	**56.45**	32.54	56.76	**19.20**	52.73	24.15	54.67

Table 2. Class-wise AP comparison with other methods on unseen classes of MSCOCO with 65/15 split for ZSD.

	airplane	train	Parking	cat	bear	suitcase	frisbee	snowboard	fork	sandwich	hotdog	toilet	mouse	toaster	hairdrier	mAP
TLZSD [15]	19.6	**63.4**	3.7	43.2	3.7	13.8	12.8	24.2	12.6	9.7	6.0	1.5	2.3	**2.0**	0.0	14.6
PL [16]	20	48.2	0.6	28.3	13.8	12.4	**21.8**	15.1	8.9	8.5	0.9	5.7	0.0	1.7	0.0	12.4
SU-ZSD [7]	10.1	48.7	1.2	**64.0**	**64.1**	12.2	0.7	28	**16.4**	**19.4**	0.1	18.7	1.2	0.5	0.2	19.0
Ours	**31.5**	45.0	**12.5**	55.1	42.3	**13.9**	5.5	**29.0**	6.40	15.9	**11.4**	**19.2**	**3.7**	1.2	**0.2**	**19.6**

contextual information of the images and is able to understand the scenario, for example, where there is a computer or keyboard, there is usually a mouse. What's more, our method achieves the best performance in 8 out of 15 categories, further demonstrating the superiority of our method.

4.3 Ablation Studies

To further verify the effectiveness of each component, we conduct ablation studies on the MSCOCO dataset with the 65/15 split. Table 3 shows the mAP of our model for ZSD and GZSD under different combinations of components. $\surd$ indicates the model with corresponding module loss.

The Effect of FB Module In order to verify the contribution of the FB module to the model, we remove the FB module during training. It can be observed that the performance of ZSD and the performance of unseen in GZSD drop from 19.58% to 19.25%, and 19.20% to 18.97% respectively, while the performance of seen is only improved by 0.06%. The result shows that after adding the FB module, the model can effectively reduce the confusion between unseen classes and backgrounds.

The Effect of KT Module During training, we remove the loss function $\mathcal{L}_{KT}$, that is, only visual features and semantic attributes of seen classes are used, while semantic features of unseen classes are not involved. The result in Table 3 shows that the performance of ZSD has dropped by 2.35% and the performance of unseen in GZSD has dropped sharply by 5.29%. If we don't explicitly transfer

Table 3. Effectiveness of each loss term for both ZSD and GZSD, measured by the mAP on MSCOCO 2014 with 65/15 split.

$\mathcal{L}_{reg} + \mathcal{L}_{IS}$	$\mathcal{L}_{FB}$	$\mathcal{L}_{KT}$	$\mathcal{L}_{II}$	ZSD	seen	unseen	HM
√	√	√	√	**19.58**	32.54	**19.20**	**24.15**
√		√	√	19.25	32.60	18.97	23.98
√	√		√	17.23	**35.72**	13.91	20.02
√	√	√		18.14	33.17	17.97	23.31

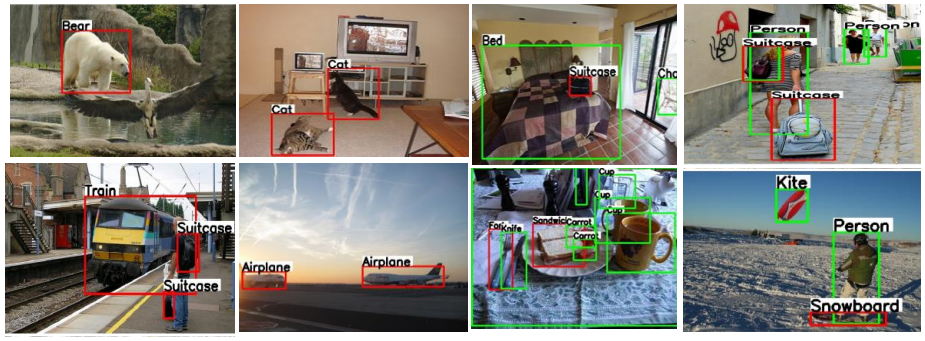

Fig. 2. Qualitative results on 65/15 split of MS COCO for ZSD and GZSD. The bounding boxes of seen classes and unseen classes are remarked as green and red respectively.

knowledge from seen classes to unseen classes through category similarity, both ZSD and GZSD performance will drop, and GZSD performance drops more sharply. It indicates that knowledge transfer has a greater impact on GZSD and can effectively alleviate the problem that the learned model will bias toward seen classes in GZSD.

The Effect of IICL Module After removing the IICL Module, ZSD performance and unseen performance in GZSD drop by 1.44% and 1.23%, respectively. The result shows that IICL can optimize the visual feature distribution in embedding space, enabling the model to learn more discriminative features.

4.4 Qualitative Result

In order to qualitatively evaluate our results, we show the detection results of our method on MSCOCO in Fig. 2. For ZSD, the image only contains unseen classes, for GZSD, the image may contain both seen classes and unseen classes. The results show that the proposed model is able to detect seen and unseen classes in different complex scenes, and it can detect multi-scale objects, such as large-scale "train", "bed" and small-scale "traffic light ", "suitcase", which verifies the effectiveness of the proposed model.

5 Conclusion

In this paper, we propose a novel framework for ZSD named Transformer-based Zero-Shot Detection via Contrastive Learning (TZSDC), which includes Deformable DETR, FB module, IICL module, and KT module. Deformable DETR extracts multi-scale contextual features, FB module separates the foreground objects from the background to alleviate the confusion of unseen classes and the background, IICL module optimizes the visual manifold structure in the embedding space to make the visual feature more discriminative, and KT module transfers knowledge from seen to unseen classes via category similarity. Experiments on MSCOCO well validate the effectiveness of the proposed method for ZSD and GZSD.

Acknowledgments This work is supported by the National Science Foundation of China (No. 62088102), China National Postdoctoral Program for Innovative Talents from China Postdoctoral Science Foundation (No. BX2021239)

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems **29** (2016)
4. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
5. Frome, A., Corrado, G., Shlens, J., et al.: A deep visual-semantic embedding model. Proceedings of the Advances in Neural Information Processing Systems pp. 2121–2129
6. Gupta, D., Anantharaman, A., Mamgain, N., Balasubramanian, V.N., Jawahar, C., et al.: A multi-space approach to zero-shot object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1209–1217 (2020)
7. Hayat, N., Hayat, M., Rahman, S., Khan, S., Zamir, S.W., Khan, F.S.: Synthesizing the unseen for zero-shot object detection. In: Proceedings of the Asian Conference on Computer Vision (2020)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
9. Li, Y., Shao, Y., Wang, D.: Context-guided super-class inference for zero-shot detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 944–945 (2020)
10. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

11. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
12. LIU Xin, LIU Xin, Z.W.W.J.W.F.: Parallel data: From big data to data intelligence. Pattern Recognition and Artificial Intelligence **30**(8), 9 (2017)
13. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra r-cnn: Towards balanced learning for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 821–830 (2019)
14. Rahman, S., Khan, S., Barnes, N.: Transductive learning for zero-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6082–6091 (2019)
15. Rahman, S., Khan, S., Barnes, N.: Transductive learning for zero-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6082–6091 (2019)
16. Rahman, S., Khan, S., Barnes, N.: Improved visual-semantic alignment for zero-shot object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11932–11939 (2020)
17. Rahman, S., Khan, S., Porikli, F.: Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In: Asian Conference on Computer Vision. pp. 547–563. Springer (2018)
18. Rahman, S., Khan, S.H., Porikli, F.: Zero-shot object detection: joint recognition and localization of novel concepts. International Journal of Computer Vision **128**(12), 2979–2999 (2020)
19. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
21. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. Advances in neural information processing systems **26** (2013)
22. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
23. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020)
24. Xie, G.S., Liu, L., Zhu, F., Zhao, F., Zhang, Z., Yao, Y., Qin, J., Shao, L.: Region graph embedding network for zero-shot learning. In: European conference on computer vision. pp. 562–580. Springer (2020)
25. Zheng, Y., Huang, R., Han, C., Huang, X., Cui, L.: Background learnable cascade for zero-shot object detection. In: Proceedings of the Asian Conference on Computer Vision (2020)
26. Zhu, P., Wang, H., Saligrama, V.: Don't even look once: Synthesizing features for zero-shot detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11693–11702 (2020)
27. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)