Built Year Prediction from Buddha Face with Heterogeneous Labels

Yiming Qian, Cheikh Brahim El Vaigh, Yuta Nakashima, Benjamin Renoust, Hajime Nagahara, Yutaka Fujioka

▶ **To cite this version:**

Yiming Qian, Cheikh Brahim El Vaigh, Yuta Nakashima, Benjamin Renoust, Hajime Nagahara, et al.. Built Year Prediction from Buddha Face with Heterogeneous Labels. SUMAC'21: 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents, Oct 2021, Chengdu, China. 10.1145/3475720.3484441 . hal-03520715

HAL Id: hal-03520715
https://inria.hal.science/hal-03520715

Submitted on 11 Jan 2022

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.

Built Year Prediction from Buddha Face with Heterogeneous Labels

Yiming Qian
yimingqian@ids.osaka-u.ac.jp
Osaka University
Osaka, Japan

Cheikh Brahim El Vaigh
cheikh-brahim.el-vaigh@irisa.fr
Univ. Rennes, CNRS, IRISA
Lannion, France

Yuta Nakashima
n-yuta@ids.osaka-u.ac.jp
Osaka University
Osaka, Japan

Benjamin Renoust
renoust@ids.osaka-u.ac.jp
Median Technologies, and Osaka University
Valbonne, France

Hajime Nagahara
nagahara@ids.osaka-u.ac.jp
Osaka University
Osaka, Japan

Yutaka Fujioka
fujioka@let.osaka-u.ac.jp
Osaka University
Osaka, Japan

ABSTRACT

Buddha statues are a part of human culture, especially of the Asia area, and they have been alongside human civilisation for more than 2,000 years. As history goes by, due to wars, natural disasters, and other reasons, the records that show the built years of Buddha statues went missing, which makes it an immense work for historians to estimate the built years. In this paper, we pursue the idea of building a neural network model that automatically estimates the built years of Buddha statues based only on their face images. Our model uses a loss function that consists of three terms: an MSE loss that provides the basis for built year estimation; a KL divergence-based loss that handles the samples with both an exact built year and a possible range of built years (e.g., dynasty or centuries) estimated by historians; finally a regularisation that utilises both labelled and unlabelled samples based on manifold assumption. By combining those three terms in the training process, we show that our method is able to estimate built years for given images with 37.5 years of a mean absolute error on the test set.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Neural networks**; • **Applied computing** → **Fine arts**.

KEYWORDS

Semi-supervised Learning, KL Divergence, Deep Learning, Regression

ACM Reference Format:
Yiming Qian, Cheikh Brahim El Vaigh, Yuta Nakashima, Benjamin Renoust, Hajime Nagahara, and Yutaka Fujioka. . Built Year Prediction from Buddha Face with Heterogeneous Labels. In *Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC '21), October 20, 2021, Virtual Event, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/

1 INTRODUCTION

Buddhism has started in India and spread all over the Asian subcontinent from India to, e.g. Japan through China. As Buddhism flourished along the centuries within several civilisations and dynasties, people started embracing the new religion, and now it is considered as the fourth largest religion in the world. They created their own Buddha statues, which are not only religion symbols but also art pieces that express their own culture.

The built information will help historians to connect the design of Buddha with the historical events such as dynasty change, revolution, etc. However, as history goes by, due to wars, natural disasters, and other reasons, the document that records the history of Buddha statues went missing.

It is an immense work for historians to estimate the built year of statues after the record is lost. Chemical analysis-based methods, such as radiocarbon dating [28], weathering-base dating [21], thermoluminescence dating [31], can be used to estimate Buddha statues' built years, but these methods have some drawbacks. For example, radiocarbon dating is only capable of identifying the date of organic components, which is not always the case for Buddha statues. The weathering-based method analyses the degrading of the surface due to weathering, being applicable to statues in limited environmental settings. Thermoluminescence dating is only applicable to statues made of a material that was heated during the building process, such as bronze, ceramic, or gold coating. Those methods are expensive, time consuming, and only applicable to limited situations.

In this work, we propose a method that estimates a Buddha statues built year leveraging only the image of Buddha faces. *To the best of our knowledge, we are the first to address this task in an automatic manner.* We use the dataset of Buddha statues presented by Renoust et al. [22], which, to the best of our knowledge, is the only dataset that comes with a rich set of annotations on built time, materials, etc. One major challenge in this dataset is its heterogeneous labels; that is, due to the inherent nature of cultural heritage as mentioned above, Buddha statue history can be lost, and thus some labels are completely missing or can only be roughly estimated by historians. As for their built time, many samples do not have relevant labels at all; some come with their exact built year, and the others only have more ambiguous expressions on the built time, such as the dynasty or century that they were established in.

In this paper, we are proposing a method that estimates the built year from Buddha statue's facial images using a deep learning-based model. We trained a regression model to compute the built year from the image embedding extracted with a convolutional neural network (CNN) backbone. To make full use of the heterogeneous

SUMAC '21, October 20, 2021, Virtual Event, China.
.

labels, we take advantage of weakly-labelled and unlabelled images in a semi-supervised manner and devise a loss function so that labels with different ambiguities can be incorporated into training. More specifically, our loss function consists of a MSE loss to reduce an image embedding to a scalar representing the built year, a Kullback-Leibler (KL) divergence-based loss that constraints the built year values to a certain given range, and a regularisation loss designed to unleash the information hidden in the unlabelled images.

Contribution. In addition to the above mentioned new loss functions, we propose to represent the built time as Gaussian functions. This unifies the labels in different ambiguity levels and provides a way to incorporate them in our KL divergence-based loss term. We experimentally show that our method can estimate the built year with 37.5 years of mean absolute error (MAE).

2 PRIOR WORK

In this section, we limit the discussion to the semi-supervised learning framework (Section 2.2) and how it can be used to address the fine-grained task of Buddha statue's built year prediction (Section 2.1).

2.1 Built Year Prediction

When literature about the Buddha statue is missing, their built year can be identified using some chemical analysis-based method, which analyses the chemical components of the material. Such methods are both time consuming and material-dependent since they can only be applied for some specific materials, such as wood or stone, in most cases demanding special equipment. The recent movement of digitisation of the humanities allows to build digital scans of Buddha statues, opening the door to automatic analysis that requires data and computers.

Machine learning, or more specifically CNNs, such as ResNet [11] and VGG [24], has been one of the main driving forces of this movement, allowing to automatically extract features from an image that give accurate representation for different kinds of entities, such as text, natural images, or art pieces. They are extensively used as off-the-shelf model in styles classification [2, 9, 10, 22], authorship identification [16], and artworks retrieval [17]. Mensink [18] proposed a method that applies max-margin regressor to SIFT features to identify painting built years. Strezoski [26] improved this built year estimation task with a multitask learning deep learning network. Classification of various aspects of Buddha statues is addressed in [9, 22], but their built years are predicted in the century basis, which only gives rough ideas about their establishment. The major drawback of the traditional deep learning approach is it requires an enormous amount of labelled data during the training process. Furthermore, the past approaches require built year labels in normalized form, which often leads to quantizate the label into centuries and reformulate the prediction problem as a classification problem.

In our Buddha project, we do not have the luxury of having a large labelled dataset. To handle this situation, we start to investigate the possibility of applying semi-supervised learning to relax this labelled data shortage.

2.2 Semi-supervised Learning

Semi-supervised learning is a way to combine labelled data and large amounts of unlabelled data into the same training process. In this context, several assumptions are adopted to make the best of unlabelled data.

The most widely used one is the cluster assumption, stating that: *if two samples are in the same cluster, they are likely to belong to the same class.* This is a strong assumption particularly for classification problems, which allows to give pseudo labels to unlabelled samples. In this way, the labelled data and pseudo-labelled data can be used in the training process. With inclusion of pseudo-labelled data, the training data pool significantly increases, which improves the data diversity and lowers the chance of over-fitting. There are two major ways of semi-supervised learning: transductive learning [6, 20] and inductive learning [5, 15, 25, 27, 33]. Transductive learning produces labels only for unlabelled data available in the training, while the inductive learning assigns new data with a label from prediction.

The manifold assumption is another popular assumption, which is applicable for both regression and classification problems [4, 12–14, 23, 32]. It states that: *the high-dimensional data lie roughly on a low-dimensional manifold*, which implies that in the manifold the densely sampled regions are smoother with smaller gradient. Under this assumption, a regularizer with radial basis function kernel can be used to enforce the smoothness among label and unlabelled data [4].

Our paper devises a semi-supervised regression model that predicts built years with high accuracy. Furthermore, we study how unlabelled data facilitate the regression task, taking into account the manifold assumptions in the context of semi-supervised learning.

3 DATASET

We obtained the dataset from Renoust et al. [22], which consists of 7,518 scanned Buddha images from 5 different books. The retina face detection algorithm [8] was deployed by the authors to extract the face of statues. These face images were then aligned and resize to 112×112, following the same process in [7]. The algorithm found 4,949 Buddha face images in the dataset. Among them, only 1,887 have built time labels, while the remaining do not provide any information about their built time. The built time labels associated with respective images fall into three types: *dynasty* in which the statue was built, ranging from 40 years to 700 years, *century*, and exact *year*. Hereinafter, we call these labels *built time* collectively, while using *built year* whenever it pinpoints a certain year. We picked out the label with the smallest range if multiple labels are available. The distribution of the labels is as follows:

- *dynasty*: 320 samples
- *century*: 316 samples
- *year*: 1,251 samples

We randomly split the dataset into 70% (i.e., 3,464 samples) for training and 30% (i.e., 1,485 samples) for testing, where 1,340 out of 3,464 samples have built time labels in the training set and 547 out of 1,485 samples in the testing set.

Figure 2 shows some sample Buddha face images after alignment. Many of them have missing facial parts (Fig. 2 (h)). All images of the dataset were collected by scanning printed books, which often leads

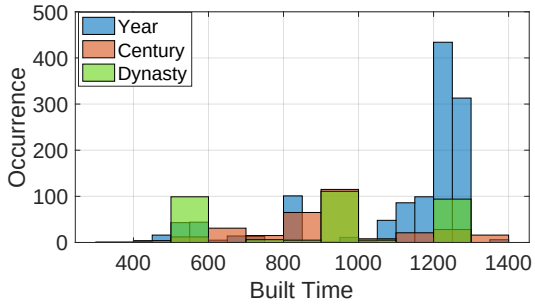


Figure 1: Distribution of built time on labelled data.

to large colour distortion from the original statues (Fig. 2 (c)). The original images in the books were captured by multiple cameras and their intrinsic parameters are unknown. The amount of object distortion caused by different cameras' focal lengths and capturing distances across images make the task particularly challenging. For example, there are different types of Buddha statues as shown in 2; there can be multiple Buddha statues of Amidanyorai, even from the same authors but established at different periods of time, and they only have a slight differences in their faces. Moreover, the original books come with artefacts due to the AM and FM screening halftone printing process (Fig. 2 (f)). The scanners used in the digitisation process can also introduce noises and artefacts known as Moiré patterns (Fig. 2 (e)).

4 METHOD

Our task is to estimate the built year t of a given image x. Although the transductive paradigm may also work for our task because the number of cultural heritages merely increases, we this time choose the inductive paradigm. More specifically, we denote our *labelled* training set $\mathcal{D}_{\mathrm{L}} = \{(x_i, t_i) \mid i = 0, \ldots, I_{\mathrm{L}}\}$ and *unlabelled* training set $\mathcal{D}_{\mathrm{U}} = \{x_i \mid i = 0, \ldots, I_{\mathrm{U}}\}$ for training, where x_i and t_i are the i-th image and the corresponding label; and I_{L} and I_{U} is the numbers of samples in the labelled and unlabelled training sets, respectively.

As mentioned previously, $\mathcal{D}_{\mathrm{L}}$ contains three types of labels, i.e., one that gives exact *year* of building and the others that give the range of built time (*dynasty* or *century*). We denote the corresponding sets by $\mathcal{D}_{\mathrm{L}}^{\mathrm{Y}}$, $\mathcal{D}_{\mathrm{L}}^{\mathrm{D}}$, and $\mathcal{D}_{\mathrm{L}}^{\mathrm{C}}$, respectively ($\mathcal{D}_{\mathrm{L}} = \mathcal{D}_{\mathrm{L}}^{\mathrm{Y}} \cup \mathcal{D}_{\mathrm{L}}^{\mathrm{D}} \cup \mathcal{D}_{\mathrm{L}}^{\mathrm{C}}$). In order to unify labels with different ambiguities, we represent the built time with Gaussian $\mathcal{N}(\mu, \sigma^2)$. For samples with exact built year, we use the built year as mean μ and 2.5 as standard deviation σ, covering 10 years centred at the year in the 95% confidence interval. A century spans 100 years, so the mean is set to the middle of the century (i.e., $\mu = 1,450$ for the 15th century) and 25 years as standard deviation. For dynasties, we use the middle year of the dynasty period as μ and the quarter of the dynasty period as σ, which covers the dynasty within the 95% confidence interval. Therefore, our label t is given by a tuple $t = (\mu_t, \sigma_t)$, which identifies a Gaussian. Figure 3 illustrates two examples Gaussian. The blue line represents the *Kamakura* dynasty, which starts at 1185 and lasts until 1333. The mean for this dynasty is 1259 and the standard deviation is 37. The red line is the 10th century (from 901 to 1000), where the mean and standard deviation are 950 and 25, respectively.

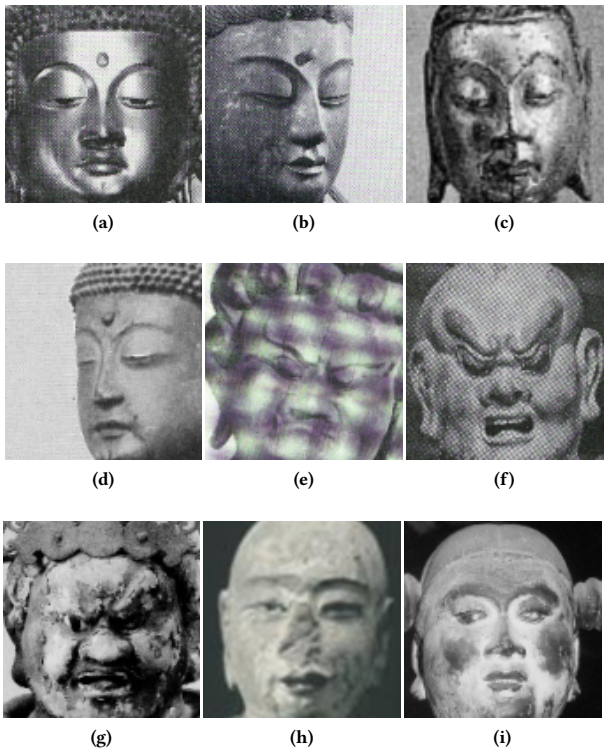


Figure 2: Some samples of Buddha face images in the dataset. (a) Amidanyorai (阿弥陀如来), (b) Seishibosatsu (勢至菩薩), (c) Mirokubutsu (弥勒仏), (d) Amidanyorai (阿弥陀如来), (e) Fudōmyōō (不動明王) (f) Kongōrikishi (金剛力士) (g) Zōchōten (増長天) (h) Jisha (侍者) (i) Zenzaidōji (善財童子).

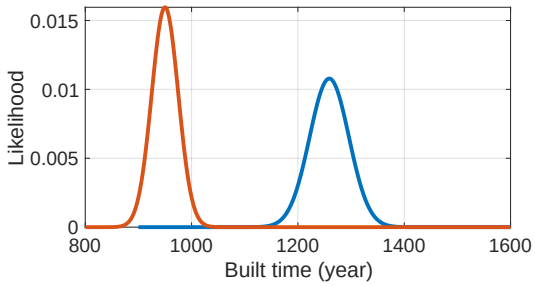


Figure 3: Two examples Gaussian to represent built times. The blue line is $\mathcal{N}(1259, 37^2)$, which represents the Kamakura dynasty (1185–1333). The red line is $\mathcal{N}(950, 25^2)$, covering the 10th century.

Our model consists of a ResNet50 variant of Arcface [7], pre-trained with 17 million human face datasets [1], as backbone, of which the last batch normalisation layer is connected to a fully connected layer to predict built year.

To make full use of the dataset with labelled and unlabelled images, we design a dedicated semi-supervised loss function, consisting of three terms, which are the mean squared error (MSE) loss, a KL divergence-based loss to better incorporate the different ambiguity levels, and a regularisation to make a smoother manifold with unlabelled samples.

MSE loss. The MSE loss provides basic supervision merely for samples with exact built years. Let $f(x)$ denote our model to regress the built year from image x. Our MSE loss E is defined as:

$$E = \frac{1}{|\mathcal{D}_{\mathrm{L}}^{\mathrm{Y}}|} \sum_{(x,t)\in\mathcal{D}_{\mathrm{L}}^{\mathrm{Y}}} \|f(x) - \mu_t\|_2, \tag{1}$$

where μ_t is the mean of t.

KL divergence-based loss. This loss utilises the different ranges of the built time labels in the dataset. Intuitively, an arbitrary pair of predicted built year must have a similar relationship (i.e., the closeness) to the ground-truth built time, which is not as straightforward as computing the distance between two samples, since the built time labels come with different ambiguities. Inspired by t-SNE [29], we encode such pairwise relationships into conditional probability $q(f(x') \mid f(x))$ of prediction $f(x')$ given $f(x)$ (or conditional probability $p(t' \mid t)$ of ground-truth t' given t), where (x, t) and (x', t') are in $\mathcal{D}_{\mathrm{L}}$. We enforce q being similar to p, so that the pairwise relationships in the ground-truth built year labels are maintained in the predictions.

More specifically, for prediction, we define the conditional probability $q(f(x') \mid f(x))$ as the likelihood of $f(x')$ assuming that it follows $N(f(x), \sigma_t^2)$, normalised over all samples in all training samples, where σ_t is borrowed from the ground truth-label t associated with x (i.e., $(x, t) \in \mathcal{D}_{\mathrm{L}}$). This can be formalised as:

$$q(f(x') \mid f(x)) = \frac{\frac{1}{\sigma_t\sqrt{2\pi}} \exp\left\{-\frac{(f(x')-f(x))^2}{2\sigma_t^2}\right\}}{\sum_{x'} \frac{1}{\sigma_t\sqrt{2\pi}} \exp\left\{-\frac{(f(x')-f(x))^2}{2\sigma_t^2}\right\}}. \tag{2}$$

Similarly, we can define the pairwise conditional probability $p(t' \mid t)$ from ground-truth labels t.

$$p(t' \mid t) = \frac{\frac{1}{\sigma_t\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(\mu_{t'}-\mu_t)^2}{\sigma_t^2}\right\}}{\sum_{t'} \frac{1}{\sigma_t\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(\mu_{t'}-\mu_t)^2}{\sigma_t^2}\right\}}. \tag{3}$$

The summations in the above two equations are computed over $(x', t') \in \mathcal{D}_{\mathrm{L}}$.

Our KL divergence-based loss C based on conditional probability q and p is calculated by

$$C = \sum_{(x,t)} \mathrm{KL}(p\|q) = \sum_{(x,t)} \sum_{(x',t')} p(t' \mid t) \log\left(\frac{p(t' \mid t)}{q(f(x') \mid f(x))}\right), \tag{4}$$

where (x, t) and (x', t') are in $\mathcal{D}_{\mathrm{L}}$.

Regularisation. Our regularisation loss is designed to unleash the information that is hidden in the unlabelled samples. It uses the manifold assumption, enforcing a smoother manifold by incorporating unlabelled data in the training process.

According to the manifold assumption, feature vectors from the backbone lie on a smooth low-dimensional manifold in the feature space. Berikov and Litvinenko [3] suggested that the decision function should change slowly in regions where feature vectors are densely distributed. From this assumption, the large amount of unlabelled data can help smooth the manifold and so regularise the training process. We adapted the radial basis function (RBF) [3, 23, 32] to approximate the smoothness of the manifold.

Let $g(x)$ denote feature vector v from our backbone g, where our model f is the composition of g and regressor (a fully connected layer) h (i.e., $f(x) = h(g(x))$. Following [3], the pairwise regularisation weight $\phi(g(x), g(x'))$ is applied between a pair of a labelled sample and an unlabelled sample, which is given by:

$$\phi(g(x), g(x')) = \exp\left\{-\frac{\|g(x) - g(x')\|^2}{2l^2}\right\}, \tag{5}$$

where $g(x)$ and $g(x')$ are the output from the last batch-normalisation layer, and l is a parameter to control the smoothness. We used $l = 0.75$ based on [3]. The regularisation term R, which is the mean of the regularisation weights, is given as follows:

$$R = \frac{1}{|\mathcal{D}_{\mathrm{L}}|\,|\mathcal{D}_{\mathrm{U}}|} \sum_{x} \sum_{x'} \phi(g(x), g(x')), \tag{6}$$

where the first and second summations are computed over the images x and x' in $\mathcal{D}_{\mathrm{L}}$ and $\mathcal{D}_{\mathrm{U}}$, respectively. This term enlarges data pools used in training by 200%, allowing our model to be robust to overfitting.

Overall loss function. Our overall loss function ℓ is a linear combination of three terms, given by

$$\ell = \alpha L + \beta C + \gamma R, \tag{7}$$

where we empirically set α to 1, β to 15, and γ to 0.1 which scales the three terms into similar weights.

5 EXPERIMENTS

We report on a set of experiments conducted to assess the benefit of our model. Section 5.1 describes the state-of-the-art (SOTA) methods, along with the different baselines we evaluated. Section 5.2 gives the implementation details. Moreover, we compare our model with the baseline methods in Section 5.3. Finally, Section 5.4.2 reports and analyses the errors of our model.

5.1 Baselines

Our model was compared against the following baseline and SOTA methods. For training these baselines, we used all samples in training set, where for the ones with dynasty or century labels, the middle years were employed instead of Gaussian function-based representation.

Nearest Neighbour Search. We store feature vector $g_0(x)$ of x and t for $(x, t) \in \mathcal{D}_{\mathrm{L}}$, where $g_0(x)$ is the feature embedding from a ResNet50 variant of ArcFace [7] (the subscript 0 is to emphasise that it is the original pretrained model). For a new Buddha face image x', we compute $g_0(x')$ and find the nearest neighbour from the stored feature vectors in terms of the cosine similarity. The prediction by this method is given as the label t associated with the nearest neighbour.

Gaussian Process Regression. Gaussian process regression [30] is a non-parametric kernel-based probabilistic model. We used g_0 as our feature extractor and MATLAB implementation with default parameters.

ResNet50 Regression w/o fine-tuning. We evaluate the same network structure as ours but without fine-tuning the backbone. The model is trained solely by our MSE loss.

Semi-supervised Deep Kernel Learning (SSDKL) [12]. SSDKL was proposed by Jean et al. [12] and utilises the power of deep learning and Gaussian process to learn a model from both labelled and unlabelled data. Their objective is to maximise the likelihood of labelled data and at the same time minimise the predictive variance in the unlabelled data. We used their original implementation.[1]

GCNBoost Regression. A GCN-based transductive semi-supervised learning classifier in [9] is adapted to the task of built year prediction. This method requires pseudo labels for unlabelled samples, and we used our full model to compute those pseudo labels. A knowledge graph is built by connecting samples that have the same built years. We adapted the GCN to make it an output layer a single scalar representing the built year. The MSE loss term is used to train the model in an end-to-end manner.

5.2 Implementation Details

Our model and its variants were implemented on PyTorch. For training, a A100 GPU with 40G of RAM was used. The batch size was set to 256, the learning rate was set to 0.003, and Adam was used for optimisation. For the MSE loss, we standardised the built year (i.e., μ's in the training set to have zero mean and unit variance). From our preliminary experiments, we found that data augmentation by only random horizontal flipping leads to the best performance; therefore, we used this in all experiments.

5.3 Results

We evaluated our method and its variants for ablation studies, as well as the baselines listed in Section 5.1. The test set of the dataset contains 1,485 samples, where 547 of them have a built time label, and 371 have an exactly built year label. Our evaluation uses only these 371 samples as it is not trivial to evaluate errors based on dynasty and century labels. We employed mean absolute error (MAE) as our error metric.

Table 1 shows the performances of all methods in comparison. Rows 1, 2, 3 and 6 are the inductive baselines. Nearest Neighbour Search (row 1) and Gaussian Process Regression (row 2) did not perform well, while ResNet50 Regression w/o fine-tuning (row 3) gave competitive performances. This implies that training with the MSE loss can benefit a lot, and our backbone provides rich cues about Buddha statues even though it is trained on human faces; yet end-to-end training helps. These results may also suggest that the relationship between the feature space and the built year space is not simple, which can support the use of our KL divergence-based loss and regularisation term.

Rows 4 and 5 are the performances of transductive baselines. GCNBoost Regression [9] shows a better performance when compared to SSDKL [12], but both methods did not perform as well as other methods. For SSDKL, we consider that the original feature vector obtained from our backbone did not have sufficient cues about the built years and the error accumulated in the iterations. GCNBoost Regressor suffers from the sparse connectivity between samples in the test set and those in the training set since we added edges between the samples with exactly same built years; therefore, the features cannot be well trained. Moreover, we do not have Buddha statues attributes other than the built years ones. Those attributes (if available) can improve GCNBoost accuracy as shown in [9] as they emphasis the relationship between the nodes in GCNBoost's knowledge graph.

Table 1: Comparison of different methods and some variants of ours. Only 371 samples in the test set that have exact built year labels were used.

	Methods	MAE (Year)
1	Nearest Neighbour Search	130.9 ± 9.8
2	Gaussian Process Regression	199.9 ± 5.4
3	ResNet50 Regression w/o fine-tuning	73.8 ± 4.0
4	GCNBoost Regression [9]	217± 15.5
5	SSDKL [12]	245.3± 4.0
6	Ours (MSE)	56.2 ± 3.7
7	Ours (KL+Reg.)	338.3 ± 33.1
8	Ours (MSE+KL)	*40.2 ± 3.62*
9	Ours (MSE+REG)	*39.3 ± 3.55*
10	Ours (MSE+KL+Reg.)	**37.5± 3.64**

Rows 6 to 10 show performances of our model on the ablation study over the loss terms, where "MSE", "KL", and "Reg." stand for the MSE loss, KL divergence-based loss, and regularisation terms, respectively. The results clearly show the importance of the MSE loss term, while the KL divergence-based loss and regularisation terms provide less impact on the performance. This is quite reasonable because the MSE loss term is the only one that gives direct supervision about the built time, while the others only tell the relationships among the samples. Yet, the best performance is obtained when we combined all three terms, which implies their complementarity.

For tasks that involve cultural heritage, the availability of samples (i.e., images in our case) is a critical problem. Cultural heritages are often stored in a secure place, and very limited access is allowed. Under this circumstance, the only way to acquire images of cultural heritages is to make digital scans of their images printed in catalogues or books, as Renoust et al. did in [22]. The image quality is thus affected by various factors, such as the quality of captured images, the quality of printing, and the quality of digital scans. Moreover, image quality degradation due to these factors may not necessarily be distinguishable from the texture of Buddha statues themselves. This is an inherent problem that particularly rises with cultural heritage. We therefore investigated the performance of our method with respect to the image quality.

We employed Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [19] to measure the image quality. BRISQUE uses an image quality score in $[0, 100]$, where a lower score indicates better quality. The distribution of the image quality scores in the entire dataset (including both training and test sets) is shown in Fig. 4. The correlation between the image quality scores and built year labels is -0.3021, which weakly indicates that a statue captured with a higher image quality tends to be established more recently. Example images with the highest, medium, and lowest qualities

[1] https://github.com/ermongroup/ssdkl

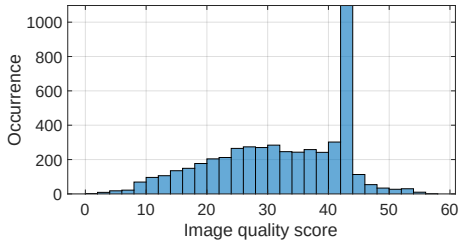


Figure 4: The distribution of image quality scores measured by BRISQUE.

Figure 5: Example images with highest (a-c), middle (d-f) and lowest (g-i) quality scores.

are shown in Fig. 5. The highest-quality images have a clear view of the Buddha face with less noises, while the medium ones show some noise or slight blur. The lowest-quality images suffer from a lower resolution (or severe blurring).

To see the impact of the image quality on our method, we plotted the relationship between the image quality score and prediction errors in Fig. 6. The figure shows that images with a higher quality (i.e., a lower score) got a lower error, and vice versa. From this, we can conclude that the image quality is an important factor to determine the prediction accuracy.

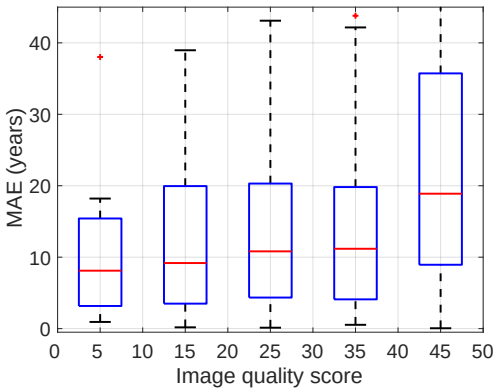


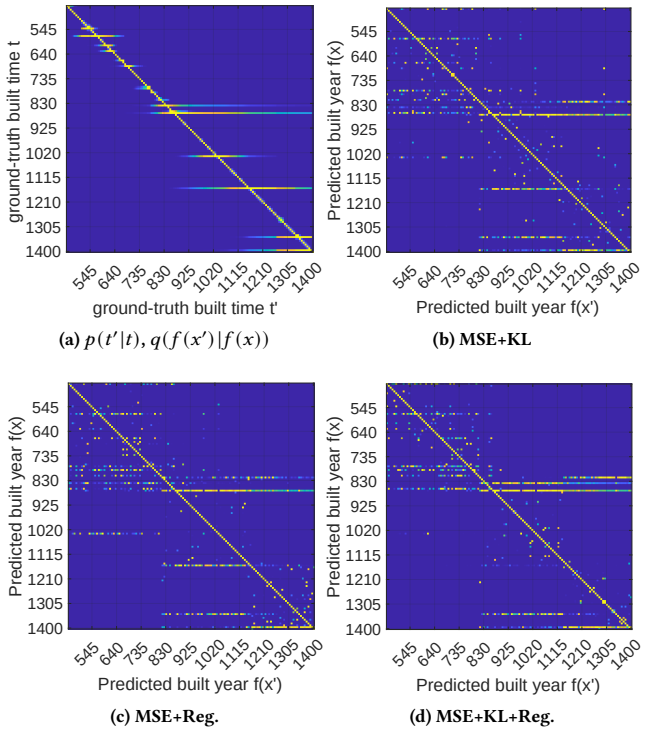
Figure 6: The relationship between image quality scores and prediction errors in MAE.

5.4 Discussion

5.4.1 With conditional probability $q(f(x')|f(x))$ and $p(t'|t)$. The conditional probabilities $q(f(x')|f(x))$ and $p(t'|t)$, which are used for our KL divergence-based loss term, encode the proximity between a pair of built times (either of the prediction and ground truth labels). This can be informative to comprehend the effect of our loss terms. We visualise $q(f(x')|f(x))$ and $p(t'|t)$ in Fig. 7. For $q(f(x')|f(x))$, we generated three visualisations for different combinations of loss terms, i.e., MSE+KL, MSE+Reg., and MSE+KL+Reg. (our full model). The yellow highlight of the built years indicates that the corresponding predictions ($f(x)$ and $f(x')$) or ground-truth labels (t and t') are close to each other. The conditional probabilities are computed over the test set, where the samples are in the chronological order based on the ground-truth labels (for dynasty and century labels, we take their middle years). Therefore, in the ideal case where an exact built year label is assigned to all samples, the highlight forms the diagonal line; however, as there are dynasty and century labels, we observe horizontal lines.

Comparisons between Fig. 7 (b) versus (c) and (d) demonstrate the importance of the regularisation term R. The regularisation term has two benefits: First it enriches our training process with large amount of unlabelled data. Second, it acts as a smoothing function that push the samples with similar properties closer in the manifold. Figure 7 (b) shows a more scattered distribution. On the other hand, Fig. 7 (c) shows a greater amount of estimation concentrated in a narrower band where its distribution shares more similarity with ground truth map.

The t-SNE visualisations of the feature vector extracted from both training and test sets before and after fine-tuning are shown in Figs. 8 and 9, respectively. For the labelled samples, the built years (the middle years for dynasty and century labels) are colour coded. Before fine-tuning, we can see that the year is almost randomly distributed. Whereas after fine-tuning, the distribution of the samples looks to be more structured, which indicates that our loss function effectively gives supervision to the backbone based on the similarity in the feature vectors and ground-truth labels.



(a) $p(t'|t)$, $q(f(x')|f(x))$ (b) MSE+KL

(c) MSE+Reg. (d) MSE+KL+Reg.

Figure 7: Visualisation of conditional probabilities (a) $p(t'|t)$, $q(f(x')|f(x))$ for (b) MSE+KL, (c) MSE+Reg., and (d) MSE+KL+Reg.

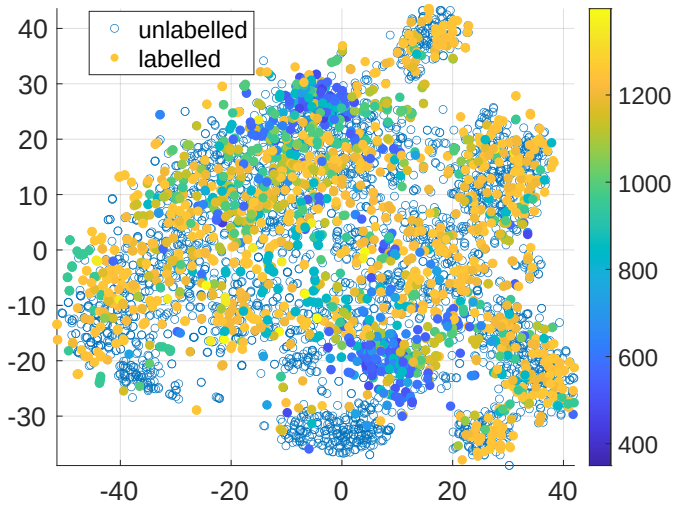


Figure 8: The t-SNE visualisation on the feature vectors before fine-tuning.

5.4.2 Qualitative analysis. The samples with low, medium and high prediction errors in the test set are shown in Fig. 10. The statues with low and medium errors contain clear facial features, while

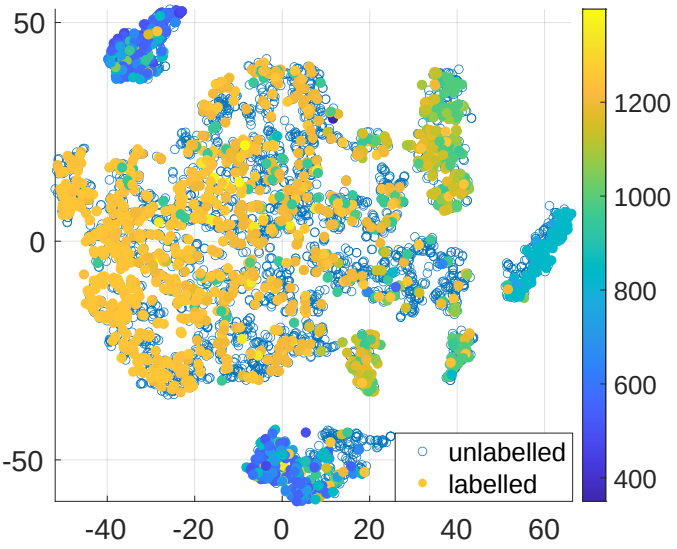
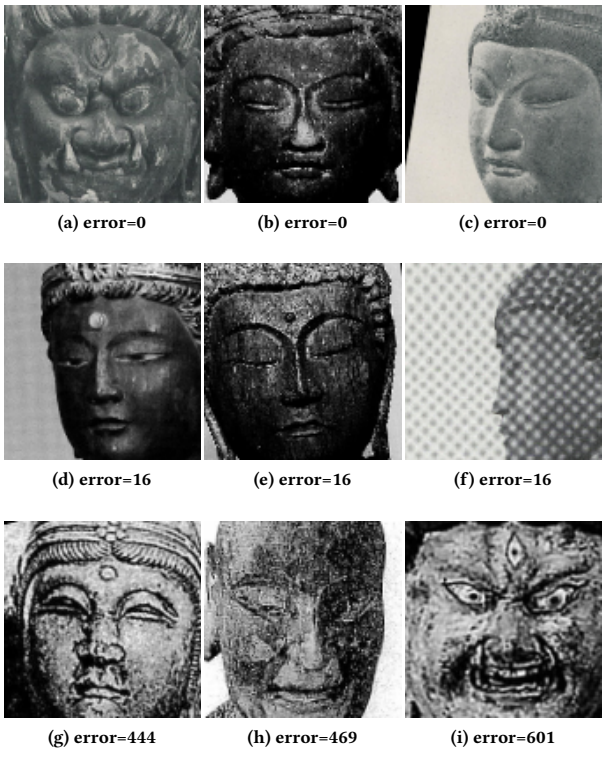


Figure 9: The t-SNE visualisation on the feature vectors after fine-tuning.

ones with high errors tend to have more visible damage and lower image quality. The statues with high errors were built in 1091, 1047, and 1241, respectively; however, our method predicted them as 647, 578, and 640. The source of the errors can come from the low image quality; for example, image (g) looks to have block noises due to compression, while (h) and (i) are dark and blurry. This is consistent with our result in Section 5.3.

6 CONCLUSION

In this work, we proposed a method to estimate the built year of Buddha statues from only their face images. We faced four main challenges to solve this task: First, the images in the dataset were collected from digital scans of 5 different books, which introduced many artefacts and distortions. Second, the dataset is quite small, containing only 4,949 facial images. Third, only 30% of samples in this small dataset are labelled. Finally, this dataset contains two types of labels: exact built years and the range of built years (i.e, dynasty and centuries) estimated by historians. To overcome those challenges, we modelled the labels in the form of Gaussian functions, which provided a unified built time representation for training. We also designed a new loss function to handle Gaussian function-based built time representation as well as unlabelled samples. Our experimental results showed that our method outperformed state-of-art methods and baselines by a significant margin. As our future work, we plan to use the different Buddha statues related information available in the dataset [22], such as built material, original location, which can be correlated with built time. For this, we need to handle highly heterogeneous data as the dataset has a lot of missing entries. The second direction is to collect more samples for better training.



(a) error=0 (b) error=0 (c) error=0

(d) error=16 (e) error=16 (f) error=16

(g) error=444 (h) error=469 (i) error=601

Figure 10: Examples of images with low (first row), medium (second row) and high error (third row) estimation error in our test set.

REFERENCES

[1] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Fu Ying. 2020. Partial FC: Training 10 Million Identities on a Single Machine. In *Arxiv 2010.05222*.

[2] Yaniv Bar, Noga Levy, and Lior Wolf. 2014. Classification of artistic styles using binarized features derived from a deep neural network. In *European conference on computer vision*. Springer, 71–84.

[3] Vladimir Berikov and Alexander Litvinenko. 2019. Semi-supervised regression using cluster ensemble and low-rank co-association matrix decomposition under uncertainties. *arXiv preprint arXiv:1901.03919* (2019).

[4] Vladimir Berikov and Alexander Litvinenko. 2021. Solving weakly supervised regression problem using low-rank manifold regularization. *arXiv preprint arXiv:2104.06548* (2021).

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. *Advances in Neural Information Processing Systems* 32 (2019).

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4690–4699.

[8] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641* (2019).

[9] Cheikh Brahim El Vaigh, Noa Garcia, Benjamin Renoust, Chenhui Chu, Yuta Nakashima, and Hajime Nagahara. 2021. GCNBoost: Artwork Classification by Label Propagation through a Knowledge Graph. In *ICMR 2021 - ACM International Conference on Multimedia Retrieval*. Taipei, Taiwan.

[10] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. 2019. Context-aware embeddings for automatic art analysis. In *Proceedings of the International Conference on Multimedia Retrieval*. 25–33.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Neal Jean, Sang Michael Xie, and Stefano Ermon. 2018. Semi-supervised Deep Kernel Learning: Regression with Unlabeled Data by Minimizing Predictive Variance. *Neural Information Processing Systems (NIPS)* (2018).

[13] John Lafferty and Larry Wasserman. 2007. Statistical Analysis of Semi-Supervised Regression. In *Proceedings of the 20th International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'07)*. Curran Associates Inc., Red Hook, NY, USA, 801–808.

[14] Shu Li, Wei Wang, Wen-Tao Li, and Pan Chen. 2021. Multi-View Representation Learning with Manifold Smoothness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8447–8454.

[15] Wei-Hong Li, Chuan-Sheng Foo, and Hakan Bilen. 2019. Learning to impute: A general framework for semi-supervised learning. *arXiv preprint arXiv:1912.10364* (2019).

[16] Daiqian Ma, Feng Gao, Yan Bai, Yihang Lou, Shiqi Wang, Tiejun Huang, and Ling-Yu Duan. 2017. From part to whole: who is behind the painting?. In *Proceedings of the 25th ACM international conference on Multimedia*. 1174–1182.

[17] Hui Mao, Ming Cheung, and James She. 2017. Deepart: Learning joint representations of visual arts. In *Proceedings of the 25th ACM international conference on Multimedia*. 1183–1191.

[18] Thomas Mensink and Jan Van Gemert. 2014. The rijksmuseum challenge: Museum-centered visual recognition. In *Proceedings of International Conference on Multimedia Retrieval*. 451–454.

[19] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* 21, 12 (2012), 4695–4708.

[20] Amit Moscovich, Ariel Jaffe, and Nadler Boaz. 2017. Minimax-optimal semi-supervised regression on unknown manifolds. In *Artificial Intelligence and Statistics*. PMLR, 933–942.

[21] Barbara A Purdy and David E Clark. 1987. Weathering of inorganic materials: dating and other applications. In *Advances in archaeological method and theory*. Elsevier, 211–253.

[22] Benjamin Renoust, Matheus Oliveira Franca, Jacob Chan, Noa Garcia, Van Le, Ayaka Uesaka, Yuta Nakashima, Hajime Nagahara, Jueren Wang, and Yutaka Fujioka. 2019. Historical and modern features for Buddha statue classification. In *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritAge Contents*. 23–30.

[23] Mugizi Robert Rwebangira and John Lafferty. 2009. Local linear semi-supervised regression. *School of Computer Science Carnegie Mellon University, Pittsburgh, PA* 15213 (2009).

[24] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.).

[25] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems* 33 (2020).

[26] Gjorgji Strezoski and Marcel Worring. 2018. Omniart: a large-scale artistic benchmark. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 4 (2018), 1–21.

[27] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 1195–1204.

[28] Royal Ervin Taylor and Ofer Bar-Yosef. 2016. *Radiocarbon dating: an archaeological perspective*. Routledge.

[29] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[30] Christopher K Williams and Carl Edward Rasmussen. 2006. *Gaussian processes for machine learning*. Vol. 2. MIT press Cambridge, MA.

[31] AG Wintle and DJ Huntley. 1982. Thermoluminescence dating of sediments. *Quaternary Science Reviews* 1, 1 (1982), 31–53.

[32] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*. 912–919.

[33] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. 2019. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5982–5991.