



Title	Built year prediction of buddha face with heterogeneous label modeled as probabilistic distribution
Author(s)	Qian, Yiming; Vaigh, Cheikh Brahim El; Nakashima, Yuta et al.
Citation	Multimedia Tools and Applications. 2025
Version Type	VoR
URL	https://hdl.handle.net/11094/101409
rights	This article is licensed under a Creative Commons Attribution 4.0 International License.
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka



Built year prediction of buddha face with heterogeneous label modeled as probabilistic distribution

Yiming Qian¹ · Cheikh Brahim El Vaigh² · Yuta Nakashima⁴ · Benjamin Renoust³ · Hajime Nagahara⁴ · Yutaka Fujioka⁴

Received: 13 April 2024 / Revised: 17 January 2025 / Accepted: 4 March 2025
© The Author(s) 2025

Abstract

Analysis of cultural heritages, particularly their construction years, provides new insights into human history. However, due to natural disasters, wars, material deterioration, and human errors, records documenting the construction years of many artifacts have often been lost. Historians and experts can estimate construction years within specific ranges using chemical-based analysis technologies or extensive historical research. Given the vast number of collected artifacts, applying these conventional methods to every artifact is impractical. To address this challenge, we developed a deep neural network model designed for Buddha statues to estimate an artifact's construction year from its image. One major challenge in this task is the heterogeneity of the labels: the training samples include both precise construction years and possible ranges (e.g., a dynasty or a century) estimated by historians. To unify these heterogeneous labels during training, we represent them as probabilistic distributions. In our previous work Qian et al. (2021), we assumed that the ambiguity in heterogeneous construction year labels followed a Gaussian distribution, assigning the highest likelihood to the midpoint of the designated time range. However, this assumption does not always hold. In this paper, we propose representing heterogeneous construction year labels as a uniform distribution, assigning equal probability to all points within the designated time range. Based on this label representation, we designed a semi-supervised learning loss function to leverage both labeled and unlabeled samples during training. Our experimental results demonstrate that our method achieves a mean absolute error of 34.3 years on a test set consisting of Buddha statues constructed between 400 and 1403. These results are further analyzed in two ways. First, we compared our model's performance to the image quality BRISQUE score, revealing a correlation between higher image quality and lower prediction error rates. Second, we validated our predictions with experts, assessing the level of agreement with our model, the challenges in determining construction years, and identifying features of interest in the artifacts.

Keywords Semi-supervised learning · KL divergence · Deep learning · Regression

Extended author information available on the last page of the article

1 Introduction

Buddhism originated in India and spread across the Asian subcontinent to East Asia. It is the fourth-largest religion in the world [2]. Throughout history, Buddhist practitioners have created Buddha statues to express their beliefs. Crafted in various regions and periods, these statues exhibit slight modifications to reflect local cultures. Over time, a wide spectrum of such statues has emerged. Unfortunately, many records documenting these statues have been lost due to natural disasters, wars, material deterioration, and human error, leaving the precise construction dates of many statues obscured. The cultural and historical significance of these artifacts has motivated historians to estimate their construction years using various technologies, including radiocarbon dating [3], weathering-based dating [4], and thermoluminescence dating [5]. However, each of these methods has its limitations.

Radiocarbon dating can determine the age of organic materials, making it applicable only to statues made from wood or other organic substances. Weathering-based dating estimates construction years by analyzing the rate of material surface deterioration, which is suitable for outdoor stone statues. However, environmental factors like global warming and air pollution have accelerated deterioration, complicating accurate estimations. Thermoluminescence dating applies to statues made of materials like bronze, ceramic, or gold coatings that were heated during their creation. Despite their utility, these methods are often inaccessible, costly, and have limited applicability.

Analyzing the visual appearance of statues [6, 7] offers another approach to dating. However, this method is time-consuming and requires experts with years of training. Alternatively, deep neural networks can be trained to assist with image analysis. Training such models necessitates a large-scale labeled dataset, but creating such datasets requires experts to annotate images or consult historical records, resulting in datasets that are often insufficient in size. For example, the dataset used in [8] contains only 4,949 Buddha face images, of which only 30% are labeled.

Beyond the challenge of dataset size, the inherent nature of historical heritage complicates precise dating. For statues lacking historical records, the visual analysis can only provide rough estimates of construction years, often represented as ranges (e.g., dynasties or centuries). Consequently, the available labels are *heterogeneous*, with some providing exact years and others indicating broader periods.

To address these challenges, we proposed a method in our previous work [1] for predicting construction years. In that work, we maximized the utility of both labeled and unlabeled samples in a small dataset through a semi-supervised learning framework, where unlabeled samples were used for regularization to smooth the feature space manifold. To better handle heterogeneous labels, we represented each label as a Gaussian distribution modeling the construction year and designed a loss function based on this representation. Specifically, we introduced a relationship supervision loss that constrains predictions to fall within the range specified by the label representation.

In this paper, we extend our previous workshop paper [1] into a journal version with new contributions. Our updated method represents heterogeneous labels as uniform distributions, assigning equal probability to all times within the specified range. This approach better aligns with actual labels where no preference exists for the range center (e.g., a label of *16th century* implies a uniform probability distribution across the century, as there is no information about when within the century the statue was built). We developed a Wasserstein distance [9]-based loss function to handle these uniform distributions effectively, accounting for both global and local relationships between samples. Additionally, we have conducted a deeper analysis of the relationship between image quality and prediction accuracy. Further-

more, we conducted an extensive expert survey involving four Buddha statue specialists to evaluate how well our method aligns with expert assessments.

Extensive historical studies and expert feedback highlight the rich information in various parts of Buddha statues for built-year prediction. While deep learning typically benefits from multiple features, we focus solely on the faces which is motivated by the fact that elements of the faces of Buddhas, Bodhisattvas, Hindu deities, Jain deities, and Taoist deities are common and that they can be compared across regions and eras [10]. In traditional Japanese craftship, the face sculptor carving was in most cases done by the head sculptor (*daibusshi*), which is often the name of the sculptor inscribed in the statues (as in the *Sanjūsangendō* statues, more details in Borengasser’s thesis [11]). Incidentally, this is also the source author retained in our database [8]. Given the maturity of deep learning applications to face analysis, this makes it an interesting ground for experiments.

Contribution For built-year prediction from Buddha faces, our semi-supervised framework achieved 34.3 years of mean absolute error (MAE). In contrast, the state-of-the-art in the Renoust et al. [8] casts the same problem in a classification task and only predicts up to centuries (100 years). In addition to the quantitative experiment against ground truth, we surveyed to study the opinions of history experts on our predictions. Our contribution can be summarized as follows:

1. We propose a dedicated semi-supervised method that represents each label as a uniform distribution, which better aligns with the actual knowledge conveyed by the labels compared to our previous Gaussian-based representation [1].
2. Through experimental and analytical evaluation, we demonstrate the benefits of our Wasserstein distance-based loss term for training with labels represented as uniform distributions.
3. We conducted a survey to gather the opinions of history experts on our prediction results, highlighting the positive influence of image quality on achieving expert consensus.

2 Related works

Predicting the built year of a Buddha statue is challenging due to the ambiguity and missing of built-year labels. Leveraging unlabeled statues or those with ambiguous labels is central to this task. In this section, we first review existing works that address built-year (or authored-year) prediction tasks (Section 2.1) and then introduce the semi-supervised learning framework (Section 2.2) designed to better model Buddha statues.

2.1 Built-year prediction

Classic and reliable methods for identifying the construction year of a Buddha statue primarily rely on analyzing the chemical components used in its creation. These methods are inherently material-dependent, working effectively only on specific materials such as wood, iron, and stone. Furthermore, these techniques often require specialized equipment, making them both expensive and time-consuming due to the chemical processes involved.

The growing availability of digitized images of Buddha statues [8, 12] makes a data-driven approach possible. The facial features of Buddha statues contain valuable cues that help researchers study cultural shifts throughout history. Traditional techniques, such as SIFT [13], cascades of features [14], rule-based systems [15], and texture feature-based methods [16],

have been popular for classifying Buddha statues. With the rise of deep learning, convolutional neural networks (CNNs) such as ResNet [17] and VGG [18] have become foundational tools for data-driven analysis, significantly improving the efficiency of Buddha statue analysis. Advanced approaches, including vanilla CNN models [19], ensembles of CNN and SVM models [20], and deep feature embedding [21], have been utilized for classifying Buddha statues.

CNNs have also been applied to a broad range of tasks, such as authorship identification [22] and artwork retrieval [23]. For the task of predicting the year of authorship, deep multitask learning approaches [24, 25] are widely used to leverage all available attributes during the learning process.

More closely related to our work is Buddha statue classification, as addressed in [8, 26], which focuses on roughly estimating the century of a statue's creation. However, a major limitation of these methods is their reliance on a substantial amount of labeled data. Predicting the construction or authorship year typically requires homogeneous labels, forcing researchers to quantize heterogeneous labels into coarser categories (e.g., grouping by century) and exclude samples with broader date ranges (e.g., the Heian period spans 391 years).

In this work, we address the challenge of predicting the construction year of Buddha statues without the advantage of a large labeled dataset. To overcome this limitation, we adopt a semi-supervised framework that leverages unlabeled samples.

2.2 Semi-supervised learning

Semi-supervised learning can partially mitigate the problem of a lack of labeled data. This learning paradigm leverages unlabeled data during training by relying on specific assumptions. For classification problems, semi-supervised learning often employs pseudo-labeling techniques for unlabeled samples. Self-training and co-training are two of the most widely used approaches.

Self-training [27, 28] begins by training a model on labeled samples and then assigns pseudo-labels to unlabeled samples based on high-confidence predictions, effectively expanding the training set. Data augmentation is commonly applied to enhance the self-training process [29]. In this approach, a model is first trained using labeled data, and then it predicts labels for multiple augmented versions of the same unlabeled sample. The average prediction is used as the pseudo-label for the sample. However, a significant drawback of this method is that random augmentations can often lead to incorrect predictions.

Berthelot et al. [30] proposed a method to address this issue by introducing a model that estimates the likelihood of augmented samples having correct classification labels. Augmented samples that fall outside a defined tolerance range are rejected. This augmentation strategy was later refined by Sohn et al. [31], who structured a two-step self-training process. In the first step, pseudo-labels for unlabeled data are generated using weakly augmented images (e.g., with flip and shift augmentations). In the second step, strongly augmented images, created using techniques such as Cutout [32], CTAugment [30], and RandAugment [33], are used to calculate losses for the unlabeled data during training.

Co-training [34, 35] divides the dataset into two groups, each assumed to contain sufficient data to train a classifier that provides pseudo-labels for the other group's unlabeled samples. A key limitation of co-training, as noted by Krogel and Scheffer [36], is that it only performs well when the two groups are independent and provide complementary information. If the classifiers consistently agree on all unlabeled data, they are deemed dependent, and the pseudo-labels fail to provide new information. When the dependency between classifiers

exceeds 60%, co-training performs worse than self-training. Du et al. [37] proposed systematic methods for splitting datasets to enhance the reliability of co-training.

For regression problems, manifold regularization techniques can be used to incorporate unlabeled data into the training process. Belkin et al. [38] introduced a manifold regularization loss, defined as the sum of ℓ_2 distances between the feature vectors of labeled and unlabeled data, weighted by a Mercer kernel function. Subsequently, Berikov and Litvinenko [39] proposed using a simpler radial basis function as a weighting mechanism. Recently, Li et al. [40] suggested utilizing learned attention coefficients to calculate weights dynamically, rather than relying on predetermined functions.

3 Dataset

The Buddha dataset (from [8]) contains 7,518 images of 1,788 Buddha statues, photo-scanned from five books. Focusing on the facial regions, we used a face detector [41] to extract 4,949 face images. The dataset includes 1,887 annotated images (with the exact year, century, or dynasty labels) and 3,062 unlabeled images. The labeled data comprises 320 *dynasty*, 316 *century*, and 1,251 *exact year* samples. The dataset was split into 70% training (3,464 samples) and 30% testing (1,485 samples), with 1,340 labeled samples in training and 547 in testing. Figure 1 illustrates sample images.

4 Method

We aim to predict the built year from a Buddha facial image x . To achieve this, we deploy a semi-supervised training scheme that utilizes both labeled and unlabeled samples. The *labeled* and *unlabeled* samples in the training set are denoted as $\mathcal{D}_L = \{(x_i, t_i) | i = 0, \dots, I_L\}$ and $\mathcal{D}_U = \{x_i | i = 0, \dots, I_U\}$, respectively, where x_i and t_i are the image and label of the i -th sample. The built time label t_i is heterogeneous. It may represent an exact year (as a scalar) or a range of years, such as a dynasty or a century.

The core of our approach is to effectively utilize these diverse label types in training the model. To address this heterogeneity, we propose representing labels as two types of probabilistic distributions: Gaussian and uniform distributions. We incorporate both labeled and unlabeled samples into training by designing three distinct loss terms:

1. **Direct supervision** loss for samples with exact built-year labels.
2. **Relationship supervision** loss, which handles all three types of built-time labels to force the prediction to follow similar neighborhood identity [42] pattern as ground truth.
3. **Regularization** loss to incorporate unlabeled samples in training. Our overall loss function is a linear combination of these three losses.

The overall loss function is formulated as a linear combination of these three losses.

To determine the optimal cost term for each distribution representation, we define multiple cost functions, summarized in Table 1. Details of each cost function are provided in Sections 4.2.1, 4.3, and 4.4. A comprehensive evaluation of mix-and-match experiments with different cost functions is presented in Section 5.4.

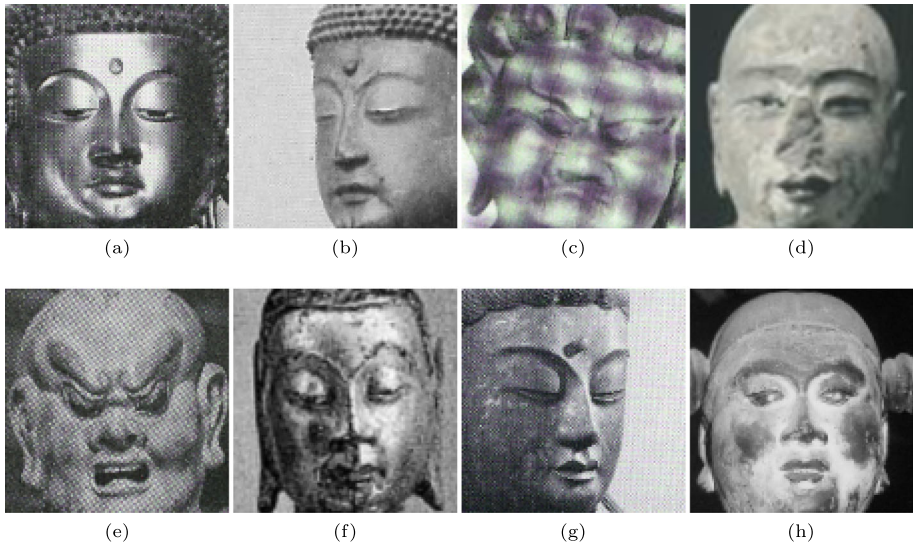


Fig. 1 Some samples of Buddha face images are in the dataset. (a-b) Amidanyorai, (c) Fudōmyōō, (d) Jisha, (e) Kongōrikishi, (f) Mirokubutsu, (g) Seishibosatsu, (h) Zenzaidōji

4.1 Representation of heterogeneous labels

\mathcal{D}_L contains three types of built-time labels, i.e., (i) exact built year and rough built time categorized by (ii) dynasty or (iii) century. Let \mathcal{D}_L^Y , \mathcal{D}_L^D , and \mathcal{D}_L^C denote the sets of samples for respective label types, where $\mathcal{D}_L = \mathcal{D}_L^Y \cup \mathcal{D}_L^D \cup \mathcal{D}_L^C$. To uniformly handle the ambiguities in there label types, we use probability distributions (either Gaussian or uniform) to represent them.

Label representation by Gaussian distributions A Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ is used to represent a range of time. For exact built-year labels, we use the labeled year as mean μ and 2.5 as standard deviation σ . Our choice of a 2.5-year standard deviation gives a 10-year time span for the 95% confidence interval. A century covers 100 years, and thus we set the middle of the century as the mean (e.g., $\mu = 1550$ for the 16th century) and 25 years as the standard deviation to cover the entire 100-year time span with a 95% confidence interval. Similarly, for dynasty labels, the middle year of the dynasty and the 25% of the dynasty time span are set as mean μ and standard deviation σ of the Gaussian. With this, all labels are represented by Gaussian $\mathcal{N}(\mu, \sigma^2)$.

Table 1 Information of the cost functions configuration for two proposed methods

	Method 1: Gaussian	Method 2: Uniform
Direct supervision	Conditional Probability (3) MSE (2)	MSE (2)
Relationship supervision	Conditional Probability (5) Wasserstein (7)	Conditional Probability (6) Wasserstein (9)
Regularization	RBF (Eqn. 11)	RBF (11)

Label representation by uniform distribution Gaussians give the highest probability to the middle of a dynasty or century, giving a preference for the middle of the dynasty or century. Such behavior is not desirable as we have no prior knowledge of the built-year distribution within the period (i.e., century or dynasty). To mitigate this problem, we represent a built-time label with a uniform distribution, i.e., $\mathcal{U}(a, b)$. The start and end years of a dynasty or century are the boundaries a and b of the uniform distribution. For exact built-year labels, we adopt a 10-year tolerance, which leads the boundary at ± 5 years from the annotated year.

4.2 Model

We use Arcface [43] as the backbone pretrained with 17 million human faces. Let g denote the last batch normalization layer of the pretrained Arcface. Our built-year predictor y for Buddha face image x is given by a fully connected layer f' on top of g , i.e.,

$$y = f(x) = f'(g(x)), \quad (1)$$

where f denotes the entire model. We fine-tune f with our direct supervision, relationship supervision, and regularization.

4.2.1 Direct supervision

MSE loss E provides basic supervision for samples with exactly built years, which is formulated as:

$$E = \frac{1}{|\mathcal{D}_L^Y|} \sum_{(x,t) \in \mathcal{D}_L^Y} \|f(x) - \mu_t\|_2, \quad (2)$$

where $\mu(t)$ is the exact built year, corresponding to the mean of the Gaussian- or Uniform distribution-based representation.

For Gaussian-based representation, we can alternatively use average negative log-likelihood between prediction and ground truth as loss:

$$E_{CP} = -\frac{1}{|\mathcal{D}_L^Y|} \sum_{(x,t) \in \mathcal{D}_L^Y} \log \mathcal{N}(f(x) | \mu_t, \sigma_t^2) \quad (3)$$

where σ_t^2 is the standard distribution associated with the label t . The conditional probability of uniform distributions does not provide an informative gradient for training; therefore, we do not use this loss with the uniform distribution-based representation.

4.3 Relationship supervision

Intuitively, an arbitrary pair $(f(x), f(x'))$ of predicted built years must have a similar neighborhood to the corresponding pair (t, t') of ground-truth labels, taking into account different ambiguity levels of ground-truth labels.

We instantiate this by computing the pairwise distances among predictions and ground truths and aligning them via the KL divergence loss. Specifically, let $\chi(s, s')$ and $\tau(s, s')$ denote the pairwise distances of predictions and ground truths for samples s and s' , respectively, where s and s' are (x, t) and (x', t') . We define the KL divergence loss by:

$$C = \sum_s \sum_{s'} \tau(s|s') \log \frac{\tau(s|s')}{\chi(s|s')}, \quad (4)$$

where the summations are computed over \mathcal{D}_L .

However, quantifying these pairwise distances (i.e., χ and τ) is not as straightforward as computing the built time difference between two samples when they are represented by probabilistic distributions. We propose two measurements to quantify the pairwise relationships, i.e., the conditional probability and the Wasserstein distance [9].

Conditional probability The conditional probability can be a natural candidate for quantifying distances between z and z' , where z is either $f(x)$ or μ_t (for uniform distribution $\mathcal{U}(a_t, b_t)$, $\mu_t = (a_t + b_t)/2$).

For the Gaussian-based label representation, we compute conditional probability ψ_G^{CP} of s given s' by:

$$\psi_G^{\text{CP}}(s|s') = \mathcal{N}(z|z', \sigma_{t'}^2), \quad (5)$$

which can be interpreted as a similarity between $f(x)$ and $f(x')$ normalized by $\sigma_{t'}^2$.

Similarly, when t' is represented by uniform distribution, the conditional probability ψ_U^{CP} can be calculated as:

$$\psi_U^{\text{CP}}(s|s') = \begin{cases} \frac{1}{b_{t'} - a_{t'}} & \text{for } a_{t'} \leq z \leq b_{t'} \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

where $a_{t'}$ and $b_{t'}$ are the boundaries of the uniform distribution associated with t' . Again, ψ_U^{CP} is not informative for training.

Wasserstein distance The second measure is the Wasserstein distance, which can quantify the closeness of two non-overlapping distributions. This measure allows learning the relationship from the uniform distribution-based label representation. The Wasserstein distance between two Gaussians $\mathcal{N}(z, \sigma_t^2)$ and $\mathcal{N}(z', \sigma_{t'}^2)$ can be formulated as [44]:

$$\psi_G^{\text{W}}(s|s') = \sqrt{(z - z')^2 + (\sigma_t - \sigma_{t'})^2}. \quad (7)$$

We can also calculate the Wasserstein distance for uniform distributions. For sample s , we use uniform distribution $\mathcal{U}(a_s(z), b_s(z))$, where

$$a_s(z) = z - \frac{b_t - a_t}{2}, \quad b_s(z) = z + \frac{b_t - a_t}{2}. \quad (8)$$

When $z = t$, this distribution is the same as the label representation, whereas when $z = f(x)$, it is the uniform distribution whose mean is at $f(x)$ but the length of the interval is still the same as ground-truth one.

The Wasserstein distance between $\mathcal{U}(a_s(z), b_s(z))$ and $\mathcal{U}(a_{s'}(z'), b_{s'}(z'))$ can be formulated as:

$$\psi_U^{\text{W}}(s|s') = \sqrt{\frac{m^2}{3} + 3mn + 3n^2}, \quad (9)$$

where $m = (b_t - a_t) - (b_{t'} - a_{t'})$ and $n = z - z' - m/2$. The derivative of this equation is found in the appendix.

We use either $\psi_U^{\text{CP}}(s|s')$ or $\psi_U^{\text{W}}(s|s')$ as our pairwise distance. That is, $\chi(s|s')$ and $\tau(s|s')$ correspond either $\psi_U^{\text{CP}}(s|s')$ or $\psi_U^{\text{W}}(s|s')$, but evaluated for z (and z') being $f(x)$ and t (and $f(x')$ and z'), respectively, where they are normalized in the same way as [42].

4.4 Regularization

Our regularization loss is designed based on the smoothness assumption, which incorporates unlabeled samples into the training process to enforce a smoother manifold [45].

Following the approach in [45–47], a radial basis function (RBF), denoted as ϕ , is employed to calculate the regularization loss between labeled and unlabeled samples. Let g represent the feature vector from the last batch normalization layer in the model f . The loss term is then defined as:

$$\phi(g(x), g(x')) = \exp \left\{ -\frac{\|g(x) - g(x')\|^2}{2l^2} \right\}, \quad (10)$$

where x and x' are samples in \mathcal{D}_L and \mathcal{D}_U , respectively; l is a constant to control the smoothness. We use $l = 0.75$ as in [45]. The regularization loss R , which is the mean of the regularization weight over all pairs of labeled and unlabeled samples, is formalized as:

$$R = \frac{1}{|\mathcal{D}_L| |\mathcal{D}_U|} \sum_{x \in \mathcal{D}_L} \sum_{x' \in \mathcal{D}_U} \phi(g(x), g(x')). \quad (11)$$

4.5 Overall loss function

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

$$\ell = \alpha L + \beta C + \gamma R, \quad (12)$$

where we empirically set α to 1, β to 15, and 0.2 for the Gaussian distribution- and uniform distribution-based label representations, respectively, and γ to 0.1. This hyperparameter is defined by trial and error.

5 Results

5.1 Competing models

We compared our model with baselines and SOTA models. Since these models cannot handle heterogeneous labels at the training time, we instead use the middle year of the dynasty or century as their built-year labels.

Nearest Neighbour Search computes and stores the feature vector $g_0(x)$ of x in \mathcal{D}_L , where $g_0(x)$ is our pretrained backbone (i.e., the ResNet50 variant of ArcFace [43]). At the inference time, for a new image x' , we extract $g_0(x')$ and retrieve its nearest neighbor according to the cosine similarity and pick its label t as the built year prediction.

Gaussian Process Regression [48] is a non-parametric kernel-based probabilistic model, which uses the pretrained Arcface model to extract a feature vector $g_0(x)$. We use the MATLAB built-in implementation with default parameters.

Semi-supervised Deep Kernel Learning (SSDKL) was proposed by Jean et al. [49] that combines deep neural networks and Gaussian processes. It incorporates the unlabeled data into the training process by minimizing predictive variance in the posterior regularization framework. We use the original implementation¹, where the input feature vector $g_0(x)$ is extracted from the pretrained Arcface model.

¹ <https://github.com/ermongroup/ssdkl>

GCNBoost Regression is a graph convolutional network (GCN)-based transductive semi-supervised learning classifier for automatic art analysis [26]. This model requires pseudo-labels for unlabeled samples to create a knowledge graph connecting the samples with cosine similarity greater than 0.8 computed from feature vectors $g_0(x)$. The pseudo-labels are generated from our model's predictions (trained with configuration in Table 3 row 16). The classification layer of GCNBoost is replaced with a linear layer with a single scalar output to predict the built year, where the MSE loss is used for training.

5.2 Implementation details

We used PyTorch to implement our model and its variants. We deploy pre-trained the human face recognition model arcface (ResNet50 based) [43] as our backbone then attach one fully connected layer to conduct regression task. For training, an A100 GPU with 40G of RAM was used. We used a batch size of 256, a learning rate of 0.003, and Adam as the optimizer. We tried different data augmentation strategies and found that random horizontal flipping leads to the best performance; therefore, we used it in all our experiments.

5.3 Evaluation

The methods described in Section 5.1 are compared with our model. Among our test set with 1,485 samples, 547 samples have built time range labels (i.e., either dynasty or century) and 371 samples have exact built year labels. We evaluated all models only on these 371 samples with exact built year labels as it is not trivial to define errors for built time range labels. The mean absolute error (MAE) is used to evaluate the performance of the different approaches.

The performances of all models are shown in Table 2. In the first section of the table, we used feature vectors extracted by g_0 from the pretrained Arcface [43]. The SOTA models (GCNBoost regression and SSDKL) yielded high MAEs; surprisingly, the simple nearest neighbor search achieved the best performance among the four with the MAE of 130.9 ± 9.8 years.

In the second and third sections, we supplied the feature vectors extracted from our fine-tuned backbone g to the existing four models. We tried feature vectors from both Gaussian- and uniform-variant of our model (their configurations correspond to rows 7 and 16 in Table 3, respectively). With the feature vectors from our fine-tuned model, MAEs are significantly reduced by 67–77%. This result indicates that the feature vector pretrained on human faces does not capture the essential information for Buddha statues' built year prediction. Our approach fine-tunes the model to focus on the feature that is useful for built year prediction.

The last section shows the performance of our model with Gaussian and uniform distribution-based label representations (again correspond to row 7 and 16 of Table 3). They delivered the highest performance with MAE of 37.5 ± 3.64 and 34.3 ± 3.47 for the Gaussian- and uniform-variant of our model.

5.4 Ablation study

We conducted an ablation study to show the impact of the loss terms and the model configurations on the performance. Table 3 summarizes the results. G and U denote the label representation with Gaussian and uniform distributions. The MSE and CP in the Direct Sup. column indicate the loss calculated from (2) and (3), respectively. In the Relational Sup. column, CP, and WD denote the distance measured by conditional probabilities and Wasserstein distances, respectively.

Table 2 Comparison of different models

Models	Feature vector	MAE (year)
Nearest neighbour search	pretrained	130.9 ± 9.8
Gaussian process regression	pretrained	199.9 ± 5.4
GCNBoost regression [26]	pretrained	217 ± 15.5
SSDKL [49]	pretrained	245.3 ± 4.0
Nearest neighbour search	Ours (Gaussian)	58.1 ± 5.1
Gaussian process regression	Ours (Gaussian)	54.3 ± 3.6
GCNBoost regression [26]	Ours (Gaussian)	322.5 ± 17.2
SSDKL [49]	Ours (Gaussian)	59.4 ± 3.4
Nearest neighbour search	Ours (uniform)	43.4 ± 4.4
Gaussian process regression	Ours (uniform)	54.8 ± 3.4
GCNBoost regression [26]	Ours (uniform)	345.1 ± 19.9
SSDKL [49]	Ours (uniform)	57.1 ± 3.2
Ours (Gaussian)	N/A	37.5 ± 3.64
Ours (uniform)	N/A	34.3 ± 3.47

Table 3 shows the direct supervision loss to provide the dominant impact in the training process. The relationship supervision learns the neighborhood identity between samples from the ground truth labels and enforces the model prediction to follow the same pattern. This neighborhood identity information assisted by direct supervision achieved higher performance. The optimal performance is achieved when we combine all three loss terms.

Rows 13 and 16 show the Wasserstein distance as a better choice when using uniform distribution as a label. The conditional probability for uniform distributions only gives a non-zero probability when two samples with an overlap which can not measure the amount of dissimilarity between two non-overlapping samples. On the other hand, row 7 and 10

Table 3 Comparison of different model configurations

	Label	Direct Sup.	Relational Sup.	Reg.	MAE (year)
1	G	CP			54.7 ± 3.97
2	G	CP	CP		58.2 ± 4.14
3	G	CP	CP	✓	60.3 ± 3.38
4	G	MSE			56.2 ± 3.70
5	G		CP	✓	205.9 ± 5.25
6	G	MSE	CP		40.2 ± 3.62
7	G	MSE	CP	✓	37.5 ± 3.64
8	G	MSE	WD		46.8 ± 3.88
9	G		WD	✓	199.2 ± 5.53
10	G	MSE	WD	✓	41.1 ± 3.51
11	U	MSE	CP		46.0 ± 4.03
12	U		CP	✓	204.7 ± 5.32
13	U	MSE	CP	✓	43.7 ± 3.51
14	U	MSE	WD		45.4 ± 4.87
15	U		WD	✓	201.2 ± 5.35
16	U	MSE	WD	✓	34.3 ± 3.47

G and U represent the samples labeled in Gaussian and uniform distribution respectively. CP denotes conditional probability. WD denotes Wasserstein distance

show the labels represented by Gaussian distributions work better with relational supervision calculated from conditional probability.

In conclusion, the configuration of using uniform distributions as sample labels and Wasserstein distances as distance measure obtained the lowest MAE at 34.3 ± 3.47 years.

6 History expert survey

Protocol description In addition to the qualitative experiment, we surveyed experts, graduate students, and professors specializing in Buddhist cultural studies at Osaka University. We selected 200 Buddha face images (100 images each for labeled and unlabeled samples) with the highest image quality. Image quality was assessed using the **Blind/Referenceless Image Spatial Quality Evaluator** (BRISQUE) index [50], which assigns a score to each image. A lower score indicates higher image quality. For the survey, we selected the top 200 images based on quality (i.e., those with the lowest scores, ranging from 5.51 to 53.6).

Half of the statues provided had a ground truth year associated with them (hereafter representing the *labeled* group) and another half had no ground truth attached (hereafter the *unlabeled* group). Data collection was conducted in two sessions, each consisting of 50 labeled and 50 unlabeled images (a total of 100 images per session and 200 images surveyed across both sessions). There was no overlap of images between the two sessions or between the labeled and unlabeled sets.

We collected responses from four participants for these 200 images, resulting in a total of 800 survey responses. Each participant submitted their selected statue images individually, along with an estimated construction year. To investigate expert agreement and the features influencing construction year estimation, three questions were asked:

Q1 : Do you agree with the estimation? (5 choices Likert)

Q2 : If you select neutral please indicate the reason (4 choices)

Q3 : What features did you focus on to determine the built year? (5 choices)

6.1 Q1: expert agreement

Answers to the expert agreement on Q1 follow a Likert scale, from *1: Strongly disagree* to *5: Strong agree*. Looking at the overall results from Fig. 2 (a), we notice that we have roughly a balance between agreement (in total 358) and disagreement (in total 333). The agreement between labeled and unlabeled data follows a similar distribution. It indicates the labeled and unlabeled data have similar data quality. Our dataset considered is small, and an investigation on a larger dataset should settle this question.

Relationship between agreement and prediction error On labeled data, we compared errors between the predicted date and expert agreement (Fig. 2 (b)). We found that lower error tends to display higher agreement, i.e. confirming the goal of our model. Interestingly, expert users also frequently disagree with low error, emphasizing the subjective nature of dating historical art and the need for a reliable predictive system.

We further categorized the prediction error into three groups, low prediction error (the error is below 10 years), medium (error is between 10 and 99 years), and high (error is from and beyond 100 years) (Fig. 3). The tendency of lower predicted error tending to

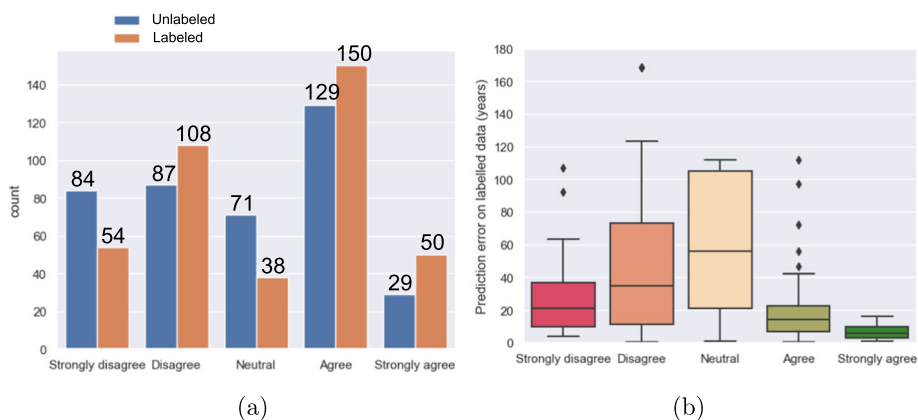


Fig. 2 (a) Results of agreement (Q1) on labeled and unlabeled data. (b) Distribution of agreement (Q1) across prediction error

higher expert agreement is confirmed. Furthermore, high error images tend to display more disagreement. We finally investigated the Pearson correlation coefficient ρ between error groups and agreement (considering the Likert scale score from 1 to 5) among the 100 labeled images, as reported in Table 4. Year prediction errors (supposedly good predictions) are anti-

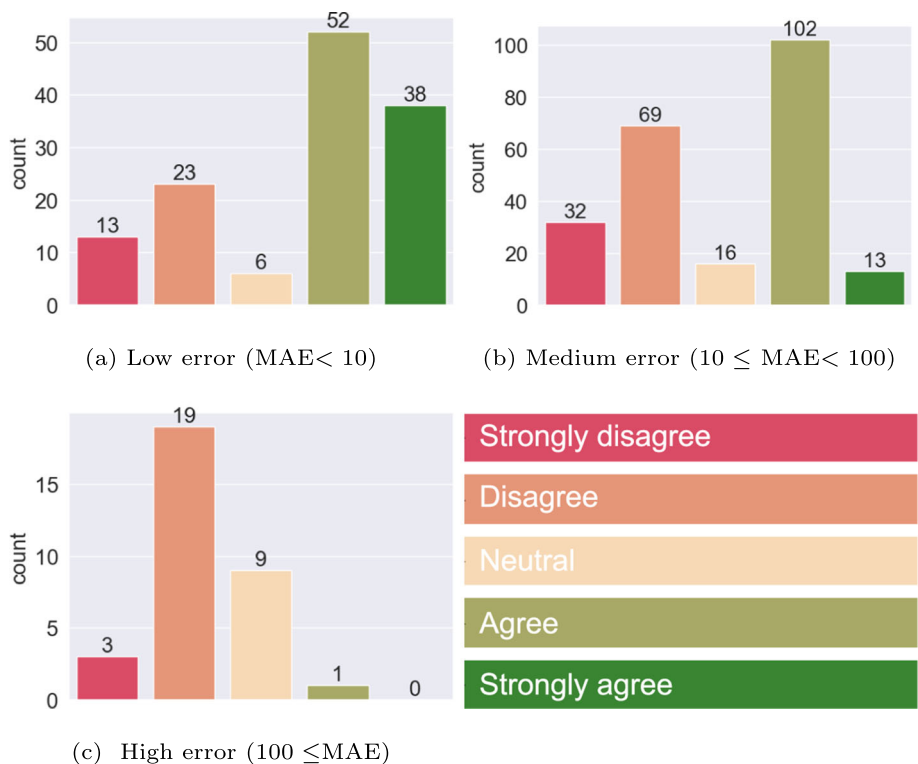


Fig. 3 Prediction error on the human expert agreement (Q1)

Table 4 Correlation between survey agreement and prediction error

	Low error MAE < 10	Medium error $10 \leq \text{MAE} < 100$	High error $100 \leq \text{MAE}$
# images	33	58	8
# responses	132	232	32
Pearson ρ	-0.196	-0.302	-0.147

correlated to expert agreement across all subgroups, once more confirming the intuition of our model.

Relationship between agreement and image quality We employed BRISQUE to measure the image quality of the test set, where a lower score indicates better quality. The relationship between the image quality score and prediction errors is shown in Fig. 4 (a). This experiment confirmed our observation that images with higher quality are more likely to receive a lower error. It shows that the image quality is an important factor that influences the prediction accuracy.

If we compare the results with the BRISQUE image quality on survey options (Fig. 4 (b)). We found that people tend to have very slightly more difficulty emitting a recommendation

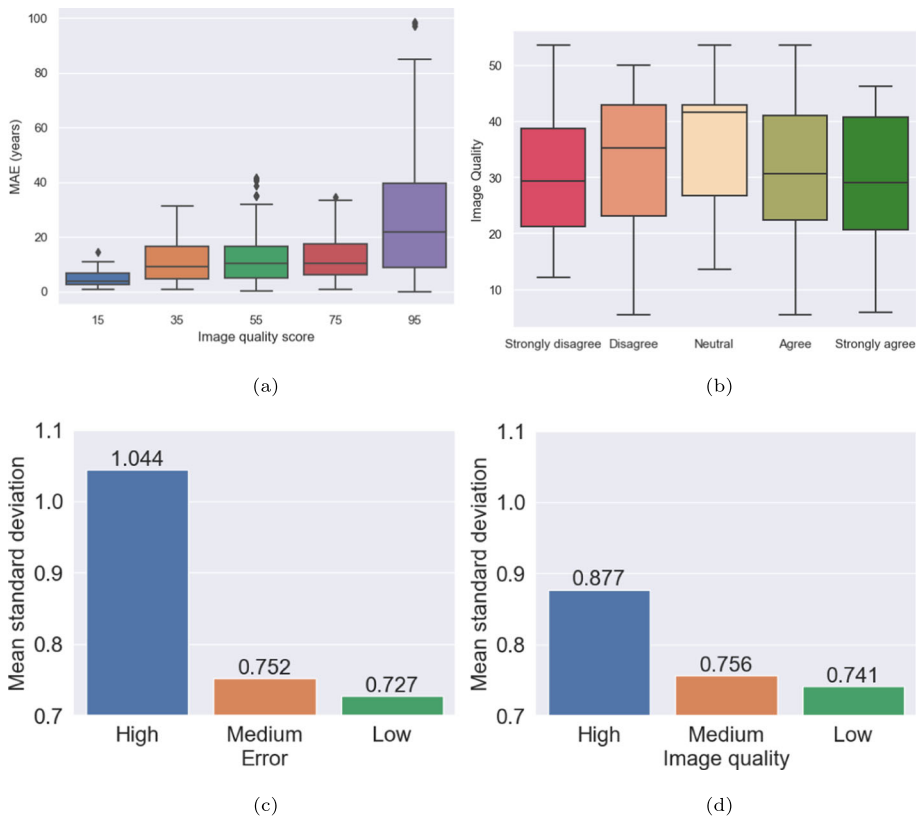


Fig. 4 The relationship between image quality scores (smaller score indicates better quality) on (a) prediction errors in MAE and (b) the expert agreement (Q1). Measure the mean standard deviation between expert opinions as a function of (c) prediction error and (d) image quality on Q1

with lower quality images: the median value of Neutral agreement is slightly over 40 in quality, while other advice ranges between 29 and 35. Note that this effect is limited since we are already studying images pre-filtered because of their top-quality images.

We further investigate this relationship by dividing image quality into three groups: lower quality images with a score beyond 45, medium quality images with a score between 25 and 35, and higher quality images with a score below 25. Among the 200 images, the Pearson correlations between agreement in the Likert scale and image quality are reported in Table 5. Since the Likert scale and the BRISQUE index progress in opposite directions, higher agreement in the lowest quality group is positively associated with better image quality (but negatively with the BRISQUE score).

We measured the standard deviation across the four participants for each question (based on their Likert score 1 to 5). For the labeled images, we calculated the mean standard deviation for high, medium, and low error conditions as reported in Fig. 4 (c). The disagreement appears higher on the images that have higher errors. We also studied standard deviation across agreement among the three groups of image quality overall labeled and unlabeled images (Fig. 4 (d)): interestingly the images that had higher quality received higher disagreement between survey participants, probably because lower and medium quality images gathered mostly neutral agreements.

6.2 Q2: difficulties in determination

Q2 consists of four options investigating the cause of the impossibility of determining the built year of a statue. Multiple choices are possible:

Q2.A : I don't know this statue.

Q2.B : I know this Buddha statue, but there is no established or well-known hypothesis/theory regarding the built year.

Q2.C : Unable to identify which Buddha statue due to poor image quality.

Q2.D : Unable to guess the built year due to poor image quality.

From Fig. 5 (a), we can first conclude that there is almost no statue that was unknown to our experts, validating the relevance of our panel of experts. The main reason for the inability to determine a year appears to be a lack of consensus among experts on the given statutes, confirming a need for a fully validated automatized model to help build a consensus. In addition, we can note that the image quality was an issue more in determining which statue the face proposed belonged to, rather than the actual year itself. This shows that experts dig into their contextual knowledge of a statue to determine its built year, rather than limiting their judgment to the style of its face.

Table 5 Correlation between survey agreement (Q1) and image quality

	High quality score < 25	Medium quality $25 \leq \text{score} < 45$	Low quality $45 \leq \text{score}$
# images	62	131	7
# responses	248	524	28
Pearson ρ	0.051	-0.016	-0.383



Fig. 5 Statistics on Q2 and Q3

6.3 Q3: features of interest

Q3 investigates the features in the statue's face that have helped the expert to determine its built year. Five choices are available and multiple answers are allowed:

Q3.A : Shape of the face.

Q3.B : Shape of the eyes.

Q3.C : Texture.

Q3.D : Head shape.

Q3.E : Hair style.

Figure 5 (b) shows that without appeal, the top indicators for our experts to identify the built year are: the face shape and its eye shape. This helps to determine the construction method, which is strongly dependent on the construction year and localization. This agrees with the principles of a prior study from Renoust et. al. [51] that tried to regress the construction rules of a statute. Surprisingly, our results show that texture, hair, and head style are not so important to date a statue in our sample. One might have expected that texture would correlate to the construction material, linked to the construction method and location. However, since most of the statues in this dataset are made of wood with/without lacquer (as reported in the prior work [51]) this did not appear as a determining feature in our sample.

We have further investigated the breakdown of Q3 answers across the different low/mid/high classes of prediction error and image quality, and the same observation is consistent across all classes.

7 Discussion

7.1 General performance of our approach

Our proposed method of formulating the heterogeneous label from the Buddha dataset into a probabilistic distribution. In this paper, we evaluated two types of distributions, namely, Gaussian distribution and uniform distribution. Additionally, we proposed a set of loss functions that utilize both labeled and unlabeled data in the semi-supervised learning process which leads to significant performance gain. Our evaluation shows the combination of applying uniform distribution and semi-supervised learning delivers the highest performance with a mean absolute error of 34.3 years on the test set.

7.2 On the relationship supervision with $\chi(s|s')$ and $\tau(s|s')$

Our distance measures $\chi(s|s')$ and $\tau(s|s')$ encode the proximity between prediction pairs and label pairs. Our relational supervision loss term is designed to force χ to follow a similar neighborhood identity as τ . A visualization of τ is shown in Fig. 6 (a), and χ in different model configurations in Fig. 6 (b)–(f). All figures are generated over the test set, where the samples are chronologically ordered based on the ground-truth labels (for dynasty and century labels, we take their middle years). In the ideal situation where each sample has an exact built-year label, a higher similarity should be assigned only to the diagonal elements. However, our test set has labels that only give time ranges, leading to horizontal lines.

Figure 6 (b) is generated from f trained only with the MSE loss term, which has a larger number of off-diagonal samples with higher similarities compared with (c) and (d). It aligns with our quantitative results in Table 2, where the relational supervision loss and regularization loss lead to better prediction accuracy. The importance of the regularization loss can be seen by comparing Fig. 6 (c) and (e), where the number of scattered distributions was reduced in (e). The regularization loss term enlarges our training data pool with unlabeled samples. Figure 6 (e) and (f) show that the Gaussian distribution-based label representation shows a more smooth relationship between samples.

7.3 Qualitative analysis

The quality of the image is one of the factors to impact the prediction results. Figure 7 shows samples with low (a)–(c), medium (d)–(f), and high (g)–(i) prediction errors from the test

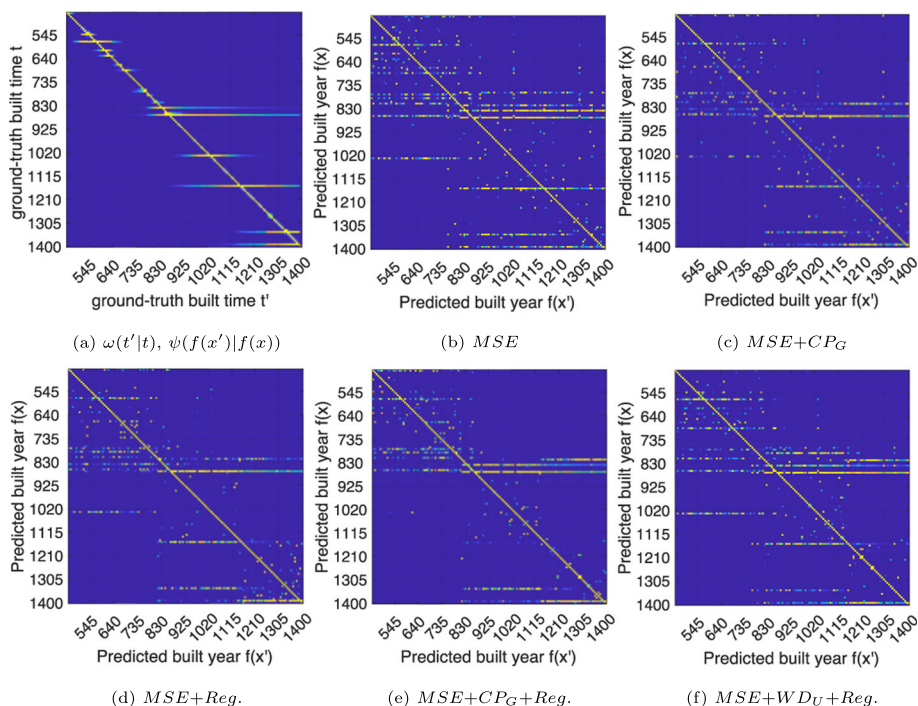


Fig. 6 Visualization of conditional probabilities. G and U denote the labels modeled with Gaussian and uniform distribution respectively



(a) error = 0,
predicted = 839,
ground truth = 839



(b) error = 0,
predicted = 1256,
ground truth = 1256



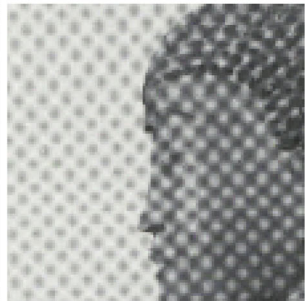
(c) error = 0,
predicted = 839,
ground truth = 839



(d) error = 16,
predicted = 1266,
ground truth = 1250



(e) error = 16,
predicted = 1239,
ground truth = 1255



(f) error = 16,
predicted = 1131,
ground truth = 1147



(g) error = 444,
predicted = 647,
ground truth = 1091



(h) error = 469,
predicted = 578,
ground truth = 1047



(i) error = 601,
predicted = 640,
ground truth = 1241

Fig. 7 Examples of images with low (first row), medium (second row), and high error (third row) estimation error (rounded to the nearest integer) in our test set

set. The samples with low prediction errors have a higher visual quality, which can provide cleaner features, while the samples with medium prediction errors have more noise, but still, the facial features are visible to some extent. The samples with high errors have a lower visual quality, and it is hard to distinguish actual features, including damage marks, from noise. The ground-truth labels for (g), (h), and (i) are 1091, 1047, and 1241, respectively, whereas our

method estimates them as 647, 578, and 640. We suspect that dark and noisy images often appear in older statues.

8 Conclusion

In this paper, we address the task of Buddha statues' built year prediction merely based on Buddha face images. This task is challenging since our target dataset is small with very few labeled samples and the images in the dataset have high noise and distortions. Moreover, some samples have a range approximating the built years rather than the exact built years as their label, making the labels heterogeneous. To overcome those challenges, we proposed to represent the labels by Gaussian or uniform distributions, providing a unified representation for these heterogeneous labels. Finally, we also designed three loss terms to handle our new label representations and to incorporate unlabeled samples in the training process, making use of all available samples. Our experimental evaluation showed the benefit of the three aforementioned loss functions and our model is better than the state-of-art models on this task. We additionally showed the negative impact of low image quality on the precision of predictions. Our extensive expert survey has shown encouraging results on the relevance of our prediction, while it underlines the difficulty of obtaining consensus from an image only even between experts: our prediction can then become another tool to help establish a shared estimate of a year of construction. In future work, we are interested in identifying from the picture which features of interest motivate a specific year prediction. We also plan to incorporate multiple attributes a Buddha statue may have, such as style, material, statue height [8], and descriptive text to build a multi-task/multi-modal system. This system can then be used to further help build consensus across experts and fill in missing information in a dataset given an image and the partially available attributes.

A Journal Statement

This paper is the journal version of our conference paper. The published paper is attached at the end of the file. In this journal paper, we have added a new method to enhance our approach. Additionally, we surveyed with history experts to evaluate the performance of our algorithm.

B Wasserstain distance for uniform distribution

Let P and Q be two uniform distributions, where $P \sim U(a, b)$, $Q \sim U(c, d)$. Let F and G be the cumulative distribution functions (CDFs) of P and Q , which can be given by:

$$F = \begin{cases} 0 & \text{for } z < a \\ \frac{z-a}{b-a} & \text{for } a \leq z \leq b \\ 1 & \text{for } z > b \end{cases}, \quad G = \begin{cases} 0 & \text{for } z < c \\ \frac{z-c}{d-c} & \text{for } c \leq z \leq d \\ 1 & \text{for } z > d \end{cases}. \quad (\text{B1})$$

The inverses of F and G for $0 < z < 1$, denoted as F^{-1} and G^{-1} , can be computed by

$$\begin{aligned} F^{-1}(z) &= z(b-a) + a \\ G^{-1}(z) &= z(d-c) + c. \end{aligned} \quad (\text{B2})$$

From the definition of the Wasserstein distance [9], we have:

$$W_p(P, Q) = \left(\int_0^1 |F^{-1}(z) - G^{-1}(z)|^p dz \right)^{1/p}, \quad (\text{B3})$$

where p is a constant. Combing this with (B2) we have:

$$W_p(P, Q) = \left(\int_0^1 |z(b - a - d + c) + a - c|^p dz \right)^{1/p}. \quad (\text{B4})$$

Letting $m = b - a - d + c$, $n = a - c$, and $p = 2$, we can simplify (B4) into:

$$W_2(P, Q) = \sqrt{\frac{m^2}{3} + 3mn + 3n^2}, \quad (\text{B5})$$

We use this in (9), i.e., $\psi_W(f(x') | f(x))_{\mathcal{U}} = W_2(P, Q)$, where

$$P = \mathcal{U} \left(f(x) - \frac{b_x - a_x}{2}, f(x) + \frac{b_x - a_x}{2} \right) \quad (\text{B6})$$

$$Q = \mathcal{U} \left(f(x') - \frac{b_{x'} - a_{x'}}{2}, f(x) + \frac{b_{x'} - a_{x'}}{2} \right). \quad (\text{B7})$$

C Sample survey

The estimated construction year is **1209**. Do you agree with this estimated result?



Question 1:

1. Strongly disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly agree

Question 2:

If you selected **Neutral**, please explain why.

1. I don't know this statue.

2. I know this Buddha statue, but there is no established or well-known hypothesis/theory regarding the built year.
3. Unable to identify which Buddha statue due to poor image quality.
4. Unable to guess the built year due to poor image quality.

Question 3:

What did you focus on to determine the year of construction?

1. Shape of the face.
2. Shape of the eyes.
3. Texture.
4. Head shape.
5. Hairstyle.

Author Contributions Yiming Qian: the main author of the paper, designed and implemented the experiments and prepared the manuscript. Cheikh Brahim El Vaigh: Contributed to the experiment as well as proofread of the paper. Yuta Nakashima: suggested new ideas and experiment designs in the paper, and revised the manuscript. Benjamin Renoust: conducted analysis on the experiment data, and revised manuscript. Hajime Nagahara: contributed ideas and suggestion on the algorithm design, revised the manuscript. Yutaka Fujioka: contributed ideas and suggestion on the algorithm design, revised the manuscript.

Funding Open Access funding provided by The University of Osaka. This work was supported by JSPS KAKENHI Grant No. JP23H05427.

Data Availability Statement Data will be made available at a reasonable request.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Our experience does not collect a person's private information. All participants are treated equally and ethically.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.


References

1. Qian Y, El Vaigh CB, Nakashima Y, Renoust B, Nagahara H, Fujioka Y (2021) Built year prediction from buddha face with heterogeneous labels. In: Proceedings of the 3rd workshop on structuring and understanding of multimedia heritage contents, pp 5–12
2. Hackett C, Stonawski M, McClendon D (2017) The changing global religious landscape. Pew Res Center pp 1–45
3. Taylor R, Bar O (2018) Radiocarbon Dating: An Archaeological Perspective vol. 122
4. Purdy B.A, Clark DE (1987) Weathering of inorganic materials: dating and other applications. In: Advances in archaeological method and theory, pp 211–253
5. Wintle A, Huntley D (1982) Thermoluminescence dating of sediments. Quat Sci Rev 1(1):31–53
6. Wisetchat S (2013) Visualizing the evolution of the sukhothai buddha. Southeast Asian Stud 2(3):559–582

7. Karlsson K (2000) Face to face with the absent buddha: The formation of buddhist aniconic art. PhD thesis, Acta Universitatis Upsaliensis
8. Renoust B, Oliveira Franca M, Chan J, Garcia N, Le V, Uesaka A, Nakashima Y, Nagahara H, Wang J, Fujioka Y (2019) Historical and modern features for Buddha statue classification. In: Proceedings of the 1st workshop on structuring and understanding of multimedia heritage contents, pp 23–30
9. Vallender S (1974) Calculation of the wasserstein distance between probability distributions on the line. *Theor Probability Appl* 18(4):784–786
10. Shimizu M (2013) Butsuzo no Kao: Katachi to Hyojo wo Yomu (The faces of Buddhist statues: reading shapes and expressions). Iwanami Shinsho
11. Borengasser DP (2024) Hall of the Lotus King: sculpture and multiplicity in early medieval Japan. PhD thesis, Harvard University
12. Rienjang W, Stewart P (2019) The Geography of Gandhāran Art: Proceedings of the Second International Workshop of the Gandhāra Connections Project, University of Oxford, 22nd-23rd March, 2018. Archaeopress, ???
13. Marlinda L, Rustad S, Basuki RS, Budiman F, Fatchan M (2020) Matching images on the face of a buddha statue using the scale invariant feature transform (sift) method. In: 2020 7th International conference on information technology, computer, and electrical engineering (ICITACEE), IEEE, pp 169–172
14. Pornpanomchai C, Arpamong V, Iamvisetchai P, Pramanus N (2011) Thai buddhist sculpture recognition system (tburs). *Int J Eng Technol* 3(4):342
15. Pornpanomchai C, Srisupornwattana N (2013) Buddhist amulet coin recognition by genetic algorithm. In: 2013 International computer science and engineering conference (ICSEC), pp 324–327. <https://doi.org/10.1109/ICSEC.2013.6694802>
16. Kompreyarat W, Bunnam T (2015) Robust texture classification using local correlation features for thai buddha amulet recognition. In: Advanced Engineering Research. Applied Mechanics and Materials, 781:531–534. Trans Tech Publications Ltd, ??? <https://doi.org/10.4028/www.scientific.net/AMM.781.531>
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
18. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds.) Proceedings of International Conference on Learning Representations (ICLR)
19. Dalara A, Sindhu C, Vasanth R (2022) Entity recognition in indian sculpture using clahe and machine learning. In: 2022 First International conference on electrical, electronics, information and communication technologies (ICEEICT), IEEE, pp 1–12
20. Rao A, Sindhu C, Suhail A, Mehta A, Dube S et al (2023) Panchadeva: Sculpture image classification using cnn-svm. *J Population Therapeutics Clinical Pharmacol* 30(9):332–344
21. Shivanee, Rajput NK, Jaiswal A (2021) A machine learning approach for the classification of the buddha statues of borobudur (indonesia). In: Data analytics and management: proceedings of ICDAM, Springer, pp 891–900
22. Ma D, Gao F, Bai Y, Lou Y, Wang S, Huang T, Duan L-Y (2017) From part to whole: who is behind the painting? In: Proceedings of the 25th ACM international conference on multimedia (ACMM), pp 1174–1182
23. Mao H, Cheung M, She J (2017) Deepart: Learning joint representations of visual arts. In: Proceedings of the 25th ACM international conference on multimedia (ACMM), pp 1183–1191
24. Strezoski G, Worring M (2018) Omniart: a large-scale artistic benchmark. *ACM transactions on multimedia computing, communications, and applications (TOMM)* 14(4):1–21
25. Khan SJ, Noord N (2021) Stylistic multi-task analysis of ukiyo-e woodblock prints. In: Proceedings of The British Machine Vision Conference (BMVC)
26. El Vaigh CB, Garcia N, Renoust B, Chu C, Nakashima Y, Nagahara H (2021) GCNBoost: Artwork Classification by Label Propagation through a Knowledge Graph. In: Proceedings of the international conference on multimedia retrieval (ICMR), Taipei, Taiwan
27. Yarowsky D (1995) Unsupervised word sense disambiguation rivaling supervised methods. In: 33rd Annual meeting of the association for computational linguistics, pp 189–196
28. Zou Y, Yu Z, Liu X, Kumar B, Wang J (2019) Confidence regularized self-training. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 5982–5991
29. Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel CA (2019) Mixmatch: A holistic approach to semi-supervised learning. *Adv Neural Inf Process Syst* 32
30. Berthelot D, Carlini N, Cubuk ED, Kurakin A, Sohn K, Zhang H, Raffel C (2020) Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In: International conference on learning representations (ICLR). <https://openreview.net/forum?id=HkIkeR4KPB>

31. Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, Cubuk ED, Kurakin A, Li C.-L (2020) Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv Neural Inf Process Syst* 33
32. DeVries T, Taylor GW (2017) Improved regularization of convolutional neural networks with cutout. [arXiv:1708.04552](https://arxiv.org/abs/1708.04552)
33. Cubuk ED, Zoph B, Shlens J, Le QV (2020) Randaugment: Practical automated data augmentation with a reduced search space. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp 702–703
34. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: *Proceedings of annual conference on computational learning theory*, pp 92–100
35. Qiao S, Shen W, Zhang Z, Wang B, Yuille A (2018) Deep co-training for semi-supervised image recognition. In: *Proceedings of the European conference on computer vision (ECCV)*, pp 135–152
36. Krogel M-A, Scheffer T (2004) Multi-relational learning, text mining, and semi-supervised learning for functional genomics. *Mach Learn* 57(1):61–81
37. Du J, Ling CX, Zhou Z-H (2010) When does cotraining work in real data? *IEEE Trans Knowl Data Eng* 23(5):788–799
38. Belkin M, Niyogi P, Sindhvani V (2006) Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J Mach Learn Res* 7(11)
39. Berikov V, Litvinenko A (2021) Solving weakly supervised regression problem using low-rank manifold regularization. [arXiv:2104.06548](https://arxiv.org/abs/2104.06548)
40. Li S, Wang W, Li W-T, Chen P (2021) Multi-view representation learning with manifold smoothness. *Proceedings of the AAAI conference on artificial intelligence* 35:8447–8454
41. Deng J, Guo J, Zhou Y, Yu J, Kotsia I, Zafeiriou S (2019) Retinaface: Single-stage dense face localisation in the wild. [arXiv:1905.00641](https://arxiv.org/abs/1905.00641)
42. Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(11)
43. Deng J, Guo J, Xue N, Zafeiriou S (2019) Arcface: additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 4690–4699
44. Salmona A, Delon J, Desolneux A (2021) Gromov-wasserstein distances between gaussian distributions. [arXiv:2104.07970](https://arxiv.org/abs/2104.07970)
45. Berikov V, Litvinenko A (2019) Semi-supervised regression using cluster ensemble and low-rank co-association matrix decomposition under uncertainties. [arXiv:1901.03919](https://arxiv.org/abs/1901.03919)
46. Rwebangira MR, Lafferty J (2009) Local linear semi-supervised regression. *School of Computer Science Carnegie Mellon University, Pittsburgh, PA*, p 15213
47. Zhu X, Ghahramani Z, Lafferty JD (2003) Semi-supervised learning using gaussian fields and harmonic functions. In: *Proceedings of the 20th international conference on machine learning (ICML)*, pp 912–919
48. Williams CK, Rasmussen CE (2006) *Gaussian Processes for Machine Learning* vol. 2. MIT press Cambridge, MA, ???
49. Jean N, Xie SM, Ermon S (2018) Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. *Neural Inf Process Syst (NIPS)*
50. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. *IEEE Trans Image Process* 21(12):4695–4708
51. Renoust B, Oliveira Franca M, Chan J, Garcia N, Le V, Uesaka A, Nakashima Y, Nagahara H, Wang J, Fujioka Y (2019) Historical and modern features for buddha statue classification. In: *Proceedings of the 1st workshop on structuring and understanding of multimedia heritage contents. SUMAC '19, Association for Computing Machinery, New York, NY, USA*, pp 23–30

Authors and Affiliations

Yiming Qian¹  · Cheikh Brahim El Vaigh² · Yuta Nakashima⁴ · Benjamin Renoust³ · Hajime Nagahara⁴ · Yutaka Fujioka⁴

✉ Hajime Nagahara
nagahara@ids.osaka-u.ac.jp

Yiming Qian
qiany@ihpc.a-star.edu.sg

Cheikh Brahim El Vaigh
Cheikh-Brahim.El-Vaigh@u-bourgogne.fr

Yuta Nakashima
n-yuta@ids.osaka-u.ac.jp

Benjamin Renoust
renoust@ids.osaka-u.ac.jp

Yutaka Fujioka
fujioka@let.osaka-u.ac.jp

¹ Institute of High Performance Computing, and Agency for Science, Technology and Research, 1 Fusionopolis Way, 16-16 Connexis, Singapore 138632, Singapore

² University of Burgundy / CIAD, Maison de l'Université, Esp. Erasme, Dijon, France

³ Median Technologies, and Osaka University, Valbonne, France

⁴ Osaka University, Osaka, Japan