Image-Generation AI Model Retrieval by Contrastive Learning-based Style Distance Calculation

Vu Thi Ngoc Anh[1], Yoshiyuki Shoji[1][0000−0002−7405−9270], Yuma Oe[1], Huu-Long Pham[2][0009−0002−4857−7004], and Hiroaki Ohshima[2][0000−0002−9492−2246]

[1] Shizuoka University, Hamamatsu, Shizuoka 432–8011, Japan
oe.yuma.21@shizuoka.ac.jp, vu.thi.ngoc.anh.20@shizuoka.ac.jp, shojiy@inf.shizuoka.ac.jp
[2] University of Hyogo, Kobe, Hyogo 651–2197, Japan
af23a009@guh.u-hyogo.ac.jp, ohshima@ai.u-hyogo.ac.jp

Abstract. This paper proposes a method for retrieving trained image-generation LoRA (Low-Rank Adaptation) models. This search algorithm takes a single arbitrary image input and then ranks the models in the order in which they will likely transform the image to the same style as the input image. We adopted a contrastive learning approach using a Triplet Network (Siamese network with triplet loss). We created a sample image set and performed style transfers on the pre-collected LoRA models to be retrieved. Using these transferred images, the network was fine-tuned to calculate the distance by their style rather than by their subject; the distance becomes large for pairs of images of the same subject transformed by different LoRA models and small for pairs of images of different subjects transformed by the same LoRA model. The search algorithm was evaluated through accuracy assessment tasks that estimated whether two images were transformed by the same model and user experiments that ranked the models. The experimental results demonstrated that fine-tuning is crucial and that the diversity of the sample image set is also important.

Keywords: Metric Learning · Triplet Network · LoRA Search

1 Introduction

Despite only a few years since Stable Diffusion [19] was released in 2022, the community of image-generation AI users has rapidly developed. Image-generation AI models, such as Stable Diffusion, are used for various purposes, including generating images from text and transforming images. These image-generation AIs can be fine-tuned with many images to learn specific styles or specialize in particular applications. Many people have started using these image-generation AIs not only as end-users but also to publish and share fine-tuned models. Nowadays, numerous fine-tuned LoRA (Low-Rank Adaptation) models are available for download on platforms like Hugging Face[3] and Civit AI[4], for good or for evil.

[3] Hugging Face: https://huggingface.co/
[4] civitai: https://civitai.com/

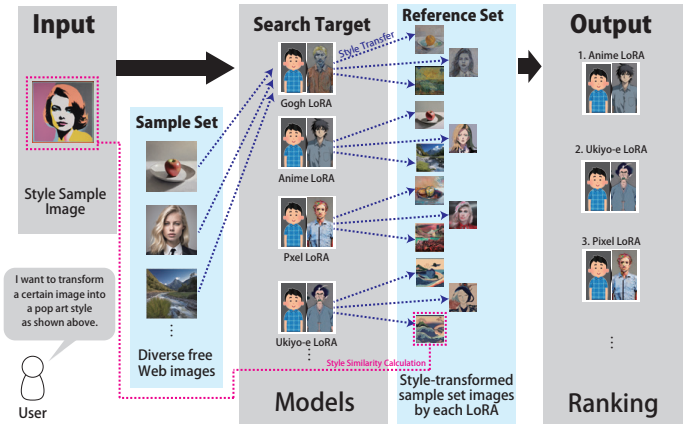

Fig. 1. System input/output and algorithm overview. Method inputs a desired style sample image, and returns ranked fine-tuned LoRA models. It calculates the style distance between the input image and images in the reference set (transformed images in the prepared sample set).

Even though many models have become publicly available, there is currently no established method for searching for models that fit a user's specific purpose. Consider an example: a user who wants to perform an image style transfer. This user wants to transform their own images into the style of a particular artist. The user has only a few images of that artist's work. Fine-tuning a model by the users themselves with just these few images is difficult. Therefore, the user needs to find an existing model that can transform images into a style similar to the images they have. In current search systems, the user must search these sharing sites using keyword queries to find models based on metadata. To determine whether the search results meet their needs, users must visit each model's description page and judge by a few thumbnail images.

To address this search difficulty, we propose a new search algorithm that uses an image as input to find the fine-tuned LoRA model. The system accepts a single sample image of the style the user wants to transform their images into. It then ranks and outputs the registered LoRA models most likely to transform any given image into the desired style. Figure 1 shows the outline of our proposed algorithm[5].

To implement this algorithm, we first collected numerous models from LoRA model-sharing websites. Next, we gathered images from free image sites on the web, ensuring they covered a wide range of categories, to create the sample set. Then, we performed style transfer on the sample set images using each of the collected LoRA models. This process created a reference set, consisting of the images in the sample set transformed by each model.

[5] All the images used in the figures are generated by Stable Diffusion for explanatory purposes and were not used in the actual training.

Using the images in the reference set, we trained a Triplet Network (*i.e.*, Siamese Network with triplet loss) [7]. This network was fine-tuned using sets of two reference images transformed by the same LoRA model, and one image transformed by another LoRA model. The actual learning task was to estimate the distance so that two images generated by the same LoRA model are close, and images generated by different LoRA models are far apart. In this process, the distance is calculated based on the style (*i.e.*, the LoRA used for the transformation) rather than the image's subject (*i.e.*, the original sample image). Given any two images, this approach provides a style distance score that indicates the likelihood that the same LoRA model generated them.

With this fine-tuned triplet network, the algorithm can rank LoRA models based on any given input image. The style distance between the input image and the reference images of each model is calculated in a pairwise manner. Our algorithm then outputs the LoRA model ranking in order of increasing closeness of average style distance.

This search algorithm assists general users who want to transform their images into a desired style but cannot draw or fine-tune models by themselves. Additionally, this technology can help artists. Using this algorithm, artists can find a model that can generate images in their style, even if they do not wish for it. It is commonly reported that artists are concerned about their own work being used for fine-tuning without their permission [20]. The ability to search for fine-tuned models is crucial for detecting infringing models that have been trained without permission using their own images or artistic styles.

We implemented a retrieval system for 100 LoRA models and conducted evaluation experiments. An automatic evaluation task, which determines whether images are transformed by the same model for an arbitrary image pair, showed that the proposed triplet network could sufficiently measure the distance of styles. Furthermore, in the task of ranking models, both automatic and user evaluations demonstrated that the models most likely to transform an arbitrary image into the desired style were ranked higher.

The contributions of this research are

1. Demonstrating that a Triplet network can discover models capable of transforming images into the style of a given image, and
2. Clarifying that fine-tuning and diversifying sample images contribute to the accuracy of the final ranking.

On the other hand, challenges such as improving accuracy and addressing the considerable computational cost were also identified.

This paper is structured into six sections. Section 2 positions this study within the context of existing research. Section 3 details the proposed method, while Section 4 evaluates its implementation. Section 5 discusses the results based on the evaluation, and Section 6 concludes the paper.

2 Related Work

This research focuses on a search algorithm that ranks models based on their ability to generate an input query image. Therefore, it is related to information

retrieval research concerning images and searches of machine learning models. Our method implements a network learned through metric learning: when given two images, it returns the likelihood that the same model generated them. Thus, this study is also related to image feature extraction and metric learning.

2.1 Information Retrieval on Images

Similar to this research, search technologies using images as input have been widely studied. Flickner *et al.* [6] proposed using images and videos as database queries. Recent research has focused on advanced information retrieval from images considering their semantics [14], with many utilizing deep learning for content-based retrieval [5]. Notable examples using contrastive learning with Siamese networks include work by Qi *et al.* [18], and Chung *et al.* [4]. However, unlike our study, which focuses on style similarity, these aim to find images with the same subject or similar images.

Another related field is estimating the creator of an image. Mohsen *et al.* proposed predicting the author of handwritten text using DNNs [16]. In the art domain, research on predicting the creator of paintings has been increasing [1, 22]. While the author is part of what we call "style," it is not the entirety. Our research defines the style of an image by the image generation model, which is highly novel.

2.2 Machine Learning Model Retrieval

In response to the rapid proliferation of AI in recent years, there has been an increase in research focused on searching for pre-trained AI models. For example, Pham *et al.* [17] proposed a search algorithm for language AI models. Many studies focus on selecting machine learning models trained for specific tasks [15, 25]. For instance, Karimi *et al.* [8] proposed a method for real-time selecting models.

There also exists research on searching for training datasets instead of the models themselves [2, 10, 12]. While this study is similar to these works, no research has taken an image as input and searched for models specifically based on style.

2.3 Contrastive Learning for Images

Our method fine-tuned a Triplet Network [7] using contrastive learning to compare images based on style rather than subject matter. Contrastive learning methods like the Triplet Network are self-supervised deep metric learning forms, which learn representations by bringing similar data points closer together and pushing dissimilar data points further apart [3, 9].

Recently, there has been a surge in research using contrastive learning with applications in information recommendation [23] and search [13]. Our study is also an application, focusing on searching for models based on style.

3 Contrastive Learning-based Method for Image Generative AI Model Retrieval

This section explains a method for searching trained image-generation LoRA models by calculating style distance using contrastive learning. It covers the

overview of the search algorithm, the definition of the style transfer being targeted, the process of creating the training dataset, the structure and training method of the network used for determining style distance, and the ranking method.

3.1 Overview

Our search algorithm takes a single image as input and outputs a ranking of pre-trained image-generation LoRA models that are likely capable of transforming images into the style of the input image. As shown in Figure 1, the algorithm's input is a single image query. The user wants to use this image as a style reference for transforming other images, which we call a "style sample image." We first collected many pre-trained image-generation models to implement this search system, which serves as the actual search targets. Next, we collected diverse images from free image sites on the internet, referred to as the "sample set". Each image in the sample set was then style-transformed using all the models, resulting in $|S| \times |M|$ transformed images, where $|S|$ is the size of the sample set, and $|M|$ is the number of models. These transformed images constitute the reference set. Our system computes the style distance between the style sample image and each image in the reference set in a pairwise manner. style distance refers to the likelihood of being generated by the same model. Models that generate images with a high style distance to the query image are ranked as capable of transforming images into the desired style. We employed a triplet network, a contrastive learning approach using triplet loss to calculate style distance. This network is trained to output a distance score between 0 and 1 when given two images. During training, the network is fine-tuned to compute distance based on style rather than subject or pixel similarity. By using style distance, the system can rank models that are most likely to generate images in the style closest to any given input image.

3.2 Style Transfer Task Targeted by Our Search

Various methods exist for the style transfer of images using image-generation AI, and the definition of the styles imparted can vary widely. Here, we define the style-transfer task targeted by our search as Figure 2. This study's style transfer begins with converting the input image into a black-and-white line drawing using Holistically-nested Edge Detection (HED). This line drawing displays the rough shapes in black on a white background, ignoring colors and details. Next, Stable Diffusion completes the image by generating a colored version from the line drawing. Specifically, we utilized ControlNet's "Reference Only" function, an extension of Stable Diffusion [24]. The ControlNet "Reference Only" function allows for adding a reference image when performing typical image generation tasks (*i.e.*, txt2img). Leaving the text prompt empty can generate images solely based on the reference image. This approach allows the model's style to be applied to the image while retaining the minimal features of the input image. During this transformation, using LoRA models specialized in particular styles or artistic expressions can impart these features to the image.

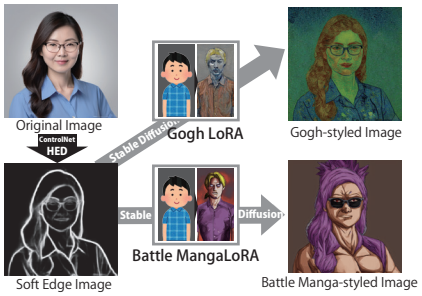

Fig. 2. Our target style-transfer task: Extract a line drawing and complete it by Stable Diffusion. Result image may have changes in color, texture, or even shape.

Next, we define "style" in the context of this study. "Style" encompasses all characteristics of the model that manifest in the image changes during the above transformation process. The currently available pre-trained LoRA models are diverse. Some models specialize in replicating the artistic style of specific creators, primarily altering colors and surface textures to resemble those artists. Other models may alter the shapes of the image's subjects, such as transforming the input into abstract art, adding animal ears to human images, or converting images into stylized cartoons. All these unique modifications made by each model to the images are considered part of the "style."

3.3 Creating Training Dataset

A classifier that can determine whether two given images are transformed by the same model is necessary to calculate the style distance. A training dataset, comprising images of various subjects transformed by various LoRA models, is constructed to create such a classifier.

Initially, a sample set was created by collecting many images from several royalty-free image resources on the web. This sample set includes diverse images, including photos of faces and objects, landscapes, abstract art, geometric patterns, and paintings.

Next, all images in the collected sample set were transformed by all models to create the reference set. The pre-trained LoRA models (*i.e.*, the search targets) were collected from generative AI model-sharing sites. Each model can perform distinct image transformations, such as anime-style or Van Gogh-style. Each image in the sample set was transformed using all collected models, as described in Subsection 3. During this process, each image in the reference set was assigned an ID indicating the sample image and the LoRA model used for the transformation.

Finally, a cleansing process was applied to the style-transformed images. Sometimes, the result of style transfer using the line drawing restoration approach was almost indistinguishable from the original image. For example, simple subject photos or images with symbols could be almost completely restored by some models. To remove such images that failed to impart a style, the cleansing

process involved comparing pre- and post-transformation images on a pixel level and excluding images with a pixel change rate below a certain threshold. The mean squared error $\mathrm{mse}(a, o)$ between the original image o and the transformed image a is defined as:

$$\mathrm{mse}(a,o) = \frac{1}{h \times w} \sum_{x=1}^{w} \sum_{y=1}^{h} (r(a_{xy}) - r(o_{xy}))^2 + (g(a_{xy}) - g(o_{xy}))^2 + (b(a_{xy}) - b(o_{xy}))^2, \quad (1)$$

where the x-th horizontal and y-th vertical pixel of image i with width w and height h is denoted as i_{xy}, and the RGB intensity (0 to 255) of each pixel is denoted as $r(i_{xy})$, $g(i_{xy})$, and $b(i_{xy})$ (the image size does not change before and after the transformation). Images for which this $\mathrm{mse}(a, o)$ was less than 0.01 were removed from the data set.

3.4 Finetuning of Triplet Network

The goal is to train a classifier to determine whether two images share the same style. In this context, we call the "same style" as a possibility of two images are transformed by the same model. To achieve this, we employed a fine-tuned triplet network.

The triplet network is an extension of a Siamese network [11] that uses triplet loss [7]. It is a prominent contrastive learning method that calculates image distance based on any desired aspect. Figure 3 illustrates the schematic diagram of the network used in this study. The network consists of three Convolutional neural networks (CNNs) that share their parameters, making them triplets. Each CNN functions as an encoder, transforming images into compressed vectors. In the underlying Siamese network, pairs of images are encoded by the CNNs, and distances are calculated in Euclidean space. The encoding is designed such that vectors for similar image pairs are close together, while vectors for dissimilar pairs are farther apart. The triplet network extends this by accepting three input images: an anchor, a positive, and a negative image. The triplet loss is calculated to ensure that the distance between the anchor and positive images is always smaller than between the anchor and negative images.

Our network aims to correctly vectorize images so that images transformed by the same LoRA model are close together in vector space, regardless of the subject. Conversely, images transformed by different LoRA models should be far apart in vector space, even if the subject is the same.

For this purpose, sets of three images are created from two different LoRA models. The anchor image and the positive image are transformed using the same LoRA model, while the negative image is transformed using a different LoRA model. The original three images are randomly selected from the sample set (Note that using only two original images and treating the style transfer results by different LoRA models as positive and negative images could be considered; however, due to a shortage of data, we adopted the current approach).

Given a triplet of images a, p, and n, the triplet loss $\mathcal{L}(a, p, n)$ is expressed as the non-negative difference in distances as follows:

$$\mathcal{L}(a,p,n) = max(||\overrightarrow{\boldsymbol{a}} - \overrightarrow{\boldsymbol{p}}||_2 - ||\overrightarrow{\boldsymbol{a}} - \overrightarrow{\boldsymbol{n}}||_2 + m, 0), \quad (2)$$

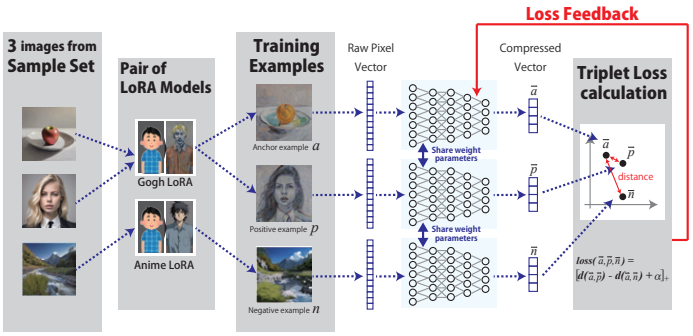

Fig. 3. Fine-tuning of the network using triplet loss. Sample images are style-transformed by two models, generating three types of images: anchor examples, positive examples, and negative examples. The CNN converts these images into compressed vectors. Images generated by the same model must always be vectorized to have smaller distances between them compared to images transformed by different models.

where a, p, and n represent the anchor, positive, and negative images, respectively, while $\overrightarrow{\mathbf{a}}$, $\overrightarrow{\mathbf{p}}$, and $\overrightarrow{\mathbf{n}}$ denote the encoded vectors of these images. Parameter m is the margin, set to ensure that the distance between the anchor and the positive sample is at least m greater than the distance between the anchor and the negative image.

The CNN for vectorization is fine-tuned using this loss function. A pre-trained image network can be utilized directly for the CNN. In this study, we used VGG19 [21], a CNN trained on over a million ImageNet images for object recognition, as the base model. Our fine-tuning involves adjusting the weights of this base network to fit the current task.

In practice, training was performed using Stochastic Gradient Descent with a learning rate of 0.001, a batch size of 20, and 12 epochs. The dimension of the generated vector is 128. The modified network is expected to generate vectors that reflect the style of the images more than their superficial appearance or subject matter. By vectorizing two images using the trained network and calculating the distance between them, it can be determined whether the images were generated by the same model based on the proximity of the vectors.

3.5 Ranking LoRA Models

When an arbitrary style sample image is input, the LoRA models are ranked based on their likelihood of generating images in that style. The given style sample is vectorized using the network trained as described above. Then, the distance is calculated in a pairwise manner for all pre-vectorized images in the reference set. Each image in the reference set has been transformed by one of the LoRA models being searched. The distances to the style sample are averaged for each LoRA model. Namely, given an image $s \in S$ from the sample set S, the transformation of image i by LoRA model m as $t(i, m)$, and the vectorization of an image i by the network as $\overrightarrow{\boldsymbol{i}}$, the average distance $d_{avg}(q, m)$ between the

given style sample image q and a model m can be expressed as

$$d_{avg}(q, m) = \frac{1}{|S|} \sum_{s \in S} ||\overrightarrow{\boldsymbol{q}} - \overrightarrow{\boldsymbol{t(s, m)}}||_2. \quad (3)$$

The LoRA models are then ranked based on the shortest $d_{avg}(q, m)$, which constitutes the search results. The models ranked highest in the results are considered more likely to transform any image into the same style as the style sample image.

4 Experiment

To ensure that the network trained to calculate style distance and to verify if the desired models could be effectively retrieved using the network, we conducted three evaluation experiments: an automatic evaluation using a binary classification task, an automatic ranking evaluation of estimating the model used for style transfer, and a user evaluation of the rankings.

4.1 Dataset and Variant Methods

We created two datasets for implementation and experiments. The first dataset is for training the network and consists of 30 LoRA models and a sample set of 238 images. All LoRA models were collected from Civitai. We gathered 238 diverse images and 238 human face photos from royalty-free image sites.

The second dataset is for evaluation experiments and consists of 101 base images and 100 LoRA models. The models and images in this dataset are independent of those used in the training data.

Our method includes several key variations: diversifying the images in the sample set, performing both transfer learning and fine-tuning during the triplet network's training, using a large number of LoRA models during training, and performing a cleansing process. To compare the effects of these variations, we created five variant methods as follows:

- **Proposed**: Diverse images were transformed using 20 models and fine-tuned,
- **Proposed+**: The same process with 30 models,
- **NoCleansing**: Images failed to style-transfer were not removed from the reference set,
- **NoFine-Tuning**: The proposed method without fine-tuning, using only transfer learning to measure style distance, and
- **FaceOnly**: Training was conducted using a sample set consisting solely of face images, equivalent in number to the original sample set.

4.2 Automatic Evaluation with Binary Classification

First, we monitored the loss progression during training to confirm that the style distance was correctly learned and analyzed accuracy using a classification task. As shown in Figure 4, the loss for each model converged correctly. The **FaceOnly** model converged the fastest, likely because the lower diversity of training images

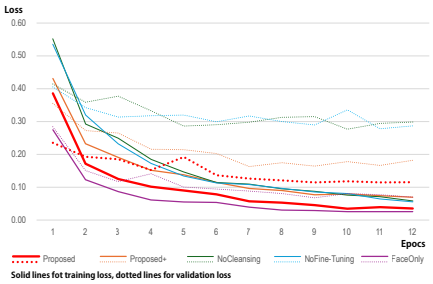

Fig. 4. Losses during the training for each method.

Table 1. F_1 score of each method in the binary classification task (**: $p < 0.01$ to Proposed).

Method	F1	Precision	Recall
Proposed	70.30	0.78	0.69
Proposed+	****72.3**	**0.79**	**0.69**
NoCleansing	70.14	0.75	0.68
NoFine-Tuning	**62.00	0.72	0.60
FaceOnly	**50.88	0.54	0.51

made learning easier. Conversely, the **NoCleansing** model, which used a diverse dataset without cleansing, converged more slowly due to the presence of many similar images from different models.

Next, we extracted 10,000 correct and 10,100 incorrect pairs from the evaluation data set. The correct pairs consisted of images from different two images transformed by the same LoRA model, while the incorrect pairs consisted of images from the same image transformed by different LoRA models . A binary classification was performed based on whether the style distance between paired images exceeded a threshold, and the F_1 score was calculated. The threshold was automatically set at the point on the ROC curve where the F_1 score was maximized.

As shown in Table 1, the **Proposed**+ method, trained with 30 models' images, achieved the highest accuracy. The difference in accuracy compared to training with 20 models was not so large but statistically significant ($p < 0.01$ on Dunnett's test). Among the four methods trained with images from the same 20 models, **FaceOnly** and **NoFine-Tuning** had particularly low accuracy ($p < 0.01$). The results indicate that diverse sample images and fine-tuning are crucial during the training of the style distance calculation network. Meanwhile, the effectiveness of cleansing showed differences in learning efficiency but did not significantly affect its accuracy ($p = 0.93$).

4.3 Automatic Evaluation with LoRA Ranking Task

Having confirmed the correct calculation of the style distance, we employed this distance to rank the models and evaluate the system. In this evaluation task, the system inputs a single image as a style sample. This style sample is an image that transferred its style using one of the target LoRA models. The search system ranks 100 LoRA models in order of increasing style distance. The position of the model used to generate the style sample in the ranking is evaluated using Mean Reciprocal Rank (MRR).

The target models for this task consist of 100 different LoRA models, independent from those used in training. The queries consist of 40 images distinct from the training images and have been transformed using 40 different LoRA

Table 2. Ranking metric (MRR) from automated evaluation and Precision@1, @5, @10 from human evaluation.

Method	MRR (Auto)	p@1	p@5	p@10
Proposed	0.44	0.55	**0.42**	**0.35**
Proposed+	**0.47**	0.55	0.33	0.27
NoCleansing	0.39	**0.60**	0.38	0.30
NoFine-Tuning	0.35	0.25	0.21	0.25
FaceOnly	0.26	0.20	0.21	0.17

models. For the reference set used in the actual calculation, 30 sample images were used, which are unrelated to the training and query images. Therefore, $d_{avg}(q, m)$ represents the average distance between the input image and these 30 reference images.

As shown in Table 2, similar to the evaluation results of the style distance, the proposed method demonstrated higher ranking performance compared to comparative methods that omitted some of the refinements. Furthermore, the system with the **Proposed**+ method, trained with images from 30 models, achieved the highest accuracy.

4.4 Subject Evaluation with LoRA Search Task

The automated ranking evaluation focused solely on the position at which the model that generated the image appeared in the ranking. However, in the natural search situation, other models might also generate images similar to the given one. Therefore, we showed the participants the top 10 images, transferred by the top 10 ranked models. Three participants evaluated whether the purpose of creating the query image could be achieved using the model that performed this transformation using a Likert scale from one to four. It determines if images perceived as having a similar style by humans were also calculated to be similar in style distance. Due to the large number of evaluations required, we randomly selected 20 queries from those used in the automated ranking; the participants evaluated 1,000 images, which covered the top 10 models across five methods for these 20 queries.

The experimental results are shown in Table 2. The Krippendorff's alpha is 0.52, indicating a reasonable agreement among participants and making the results acceptable. Regarding p@k values, **NoCleansing**, **Proposed**, and **Proposed**+ all demonstrated high accuracy. There were no significant differences in accuracy among these three methods (Dunnett's $p > 0.05$). In contrast, **NoFine-Tuning** ($p < 0.01$) and **FaceOnly** ($p \approx 0.05$) clearly showed lower accuracy compared to these methods. It indicates that transfer learning alone is insufficient, and fine-tuning and diversity in the training dataset are necessary.

5 Discussion

Overall, the proposed method performed well. The triplet network, fine-tuned using a diverse dataset, correctly inferred the model used to transform the im-

ages. Furthermore, ranking the models based on these inferences demonstrated that the method could generally search for the correct model with reasonable accuracy.

The comparison between methods revealed several important processes and others that are less so. It became clear that transfer learning alone is insufficient for calculating distances based on style rather than the subject of the images. Similarly, the diversity of the training images seems essential. While the number of different LoRA models used in training had some effect, it was less significant than expected (no statistically significant difference was found between the **Proposed** trained with 20 models and the **Proposed**+ with 30 models). On the other hand, the dataset cleansing process showed limited effectiveness. This process was intended to remove images that showed little change after style transformation, but since such images were relatively few, the impact on accuracy was minimal.

Looking at the subject experiment's overall results, all methods demonstrated a high precision. Specifically, although the correct LoRA model was only one out of 100, participants claimed there were around three relevant LoRA models among the top 10 search results. It suggests that the concept of "style" perceived by humans is somewhat ambiguous, and with careful use, a LoRA model with a slightly different style could still achieve the desired style transfer.

Finally, we discuss the feasibility of this method. The proposed method is a highly computationally inefficient search method. For each input image, one vector transformation is required. Then, for a search with $|M|$ target models and $|S|$ sample images, $|M| \times |S|$ vector similarity comparisons are necessary. Additionally, every time a new model is registered, $|S|$ vector transformations of images are required. Moreover, when models are sequentially added, periodic retraining of the Triplet Network is necessary. A more efficient search algorithm is required to make it feasible to search across many models continuously.

6 Conclusion

This paper proposed a method for searching for image transformation LoRA models from any given image using a fine-tuned triplet network. This approach allows for the calculation of a style distance where images transformed by the same model are closer together. The experimental results demonstrated that the proposed method could effectively retrieve the desired models.

However, the current accuracy is still insufficient, presenting a challenge for future work. Enhancements in the algorithm are needed to improve accuracy. Additionally, the current method involves exhaustive pairwise calculations, making it impractical for real-world deployment as a service. Therefore, developing faster alternative methods is necessary.

Acknowledgements

This work was supported by JSPS KAKENHI Grants Number 24K03228, 22H03905, JP21H03554, and 21H03775.

References

1. Akter, M., Akther, M.R., Khaliluzzaman, M.: Recognizing art style automatically in painting using convolutional neural network. In: Computational Intelligence in Machine Learning: Select Proceedings of ICCIML 2021, pp. 221–236. Springer (2022)
2. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., Groth, P.: Dataset search: a survey. The VLDB Journal **29**(1), 251–272 (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. Chung, Y.A., Weng, W.H.: Learning deep representations of medical images using siamese cnns with application to content-based image retrieval. arXiv preprint arXiv:1711.08490 (2017)
5. Dubey, S.R.: A decade survey of content based image retrieval using deep learning. IEEE Transactions on Circuits and Systems for Video Technology **32**(5), 2687–2704 (2021)
6. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., et al.: Query by image and video content: The qbic system. computer **28**(9), 23–32 (1995)
7. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3. pp. 84–92. Springer (2015)
8. Karimi, M.R., Gürel, N.M., Karlaš, B., Rausch, J., Zhang, C., Krause, A.: Online active model selection for pre-trained classifiers. In: International Conference on Artificial Intelligence and Statistics. pp. 307–315. PMLR (2021)
9. Kaya, M., Bilge, H.Ş.: Deep metric learning: A survey. Symmetry **11**(9), 1066 (2019)
10. Kern, D., Mathiak, B.: Are there any differences in data set retrieval compared to well-known literature retrieval? In: Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings 19. pp. 197–208. Springer (2015)
11. Koch, G., Zemel, R., Salakhutdinov, R., et al.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop. vol. 2, pp. 1–30. Lille (2015)
12. Kunze, S.R., Auer, S.: Dataset retrieval. In: 2013 ieee seventh international conference on semantic computing. pp. 1–8. IEEE (2013)
13. Lei, Y., Ding, L., Cao, Y., Zan, C., Yates, A., Tao, D.: Unsupervised dense retrieval with relevance-aware contrastive pre-training. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 10932–10940. ACL (Jul 2023)
14. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern recognition **40**(1), 262–282 (2007)
15. Madani, O., Lizotte, D.J., Greiner, R.: Active model selection. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. pp. 357–365 (2004)
16. Mohsen, A.M., El-Makky, N.M., Ghanem, N.: Author identification using deep learning. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 898–903 (2016). https://doi.org/10.1109/ICMLA.2016.0161

17. Pham, H.L., Mibayashi, R., Yamamoto, T., Kato, M.P., Yamamoto, Y., Shoji, Y., Ohshima, H.: Inference-based no-learning approach on pre-trained bert model retrieval. In: 2024 IEEE International Conference on Big Data and Smart Computing (BigComp). pp. 234–241 (2024). https://doi.org/10.1109/BigComp60711.2024.00044
18. Qi, Y., Song, Y.Z., Zhang, H., Liu, J.: Sketch-based image retrieval via siamese convolutional neural network. In: 2016 IEEE international conference on image processing (ICIP). pp. 2460–2464. IEEE (2016)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
20. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 2187–2204 (2023)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Valencia, J., Pineda, G.G., Pineda, V.G., Valencia-Arias, A., Arcila-Diaz, J., de la Puente, R.T.: Using machine learning to predict artistic styles: an analysis of trends and the research agenda. Artificial Intelligence Review **57**(5), 118 (2024)
23. Yu, J., Yin, H., Xia, X., Chen, T., Li, J., Huang, Z.: Self-supervised learning for recommender systems: A survey. IEEE Transactions on Knowledge & Data Engineering **36**(01), 335–355 (2024)
24. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
25. Zhang, Y.K., Huang, T.J., Ding, Y.X., Zhan, D.C., Ye, H.J.: Model spider: Learning to rank pre-trained models efficiently. Advances in Neural Information Processing Systems **36** (2024)