Contrastive Learning using Synthetic Images Generated from Real Images

Tenta Sasaya[1], Shintaro Yamamoto[1], Takashi Ida[1], and Takahiro Takimoto[1]

Toshiba Corporation, Kawasaki, Japan
tenta1.sasaya@toshiba.co.jp

Abstract. The effectiveness of pre-training using large-scale natural image datasets has been demonstrated for situations in which there are limited available real images. However, some research has shown that models pre-trained using natural images cannot achieve sufficient performance on non-natural images taken under special circumstances or with special measurement devices. Although more general pre-training methods that use synthetic images such as random pattern images or noise images are a promising approach for such cases, their effectiveness depends on downstream tasks. To deal with this problem, we propose a contrastive learning framework using synthetic images generated from real images of downstream tasks to directly learn feature representations suitable for downstream tasks of non-natural images. Image classification experiments are performed on five non-natural image datasets mimicking real-world application with little available data, and these demonstrate that the proposed method achieves higher average classification accuracy compared with pre-training using ImageNet-1k or existing synthetic images with an improvement of over 6.5 points.

Keywords: Synthetic images · Contrastive learning · Non-natural images

1 Introduction

Although deep learning is widely used for various classification tasks, most research has used only well-organized public datasets that contain pairs of data and labels. However, such an ideal dataset is not available in most real-world applications owing to annotation costs. For this reason, unsupervised learning[1] (*e.g.* contrastive learning, self-supervised learning) [5,9–14,21–23,28,42] in which models are trained using a large number of unlabeled images is becoming a common approach for pre-training.

However, since there is a lack of data in the early stages of system operation with deep learning algorithms, it is not easy to prepare large quantities of even unlabeled images. In such cases, it might not be possible to achieve the full potential of unsupervised learning, which requires a large number of unlabeled images.

[1] Self-supervised learning and contrastive learning are form of unsupervised learning. Their distinction is informal in the existing literature. In this paper, we use the more classical term of "unsupervised learning", in the sense of "not supervised by human-annotated labels".

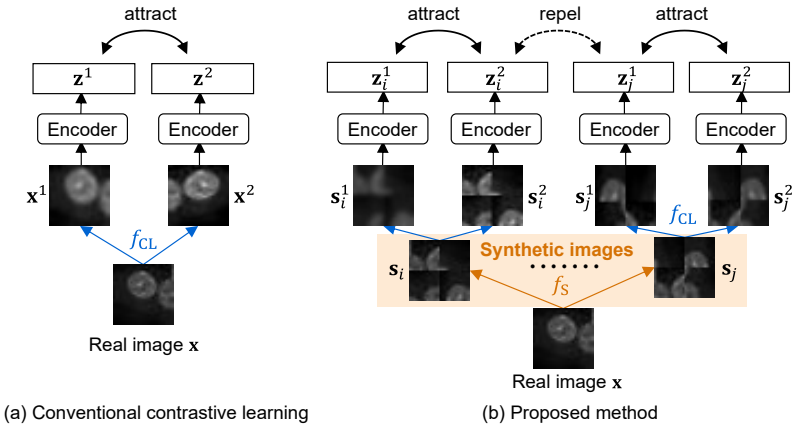

Fig. 1: Conceptual comparison of conventional contrastive learning and proposed method. f_{CL} is data augmentation for generating paired images for contrastive learning, f_{S} is an image generator for generating synthetic images.

Although pre-training using large-scale public datasets of natural images (*e.g.* ImageNet) is a common approach for scenarios in which data are limited, a previous study [7] showed that models pre-trained using natural images cannot achieve sufficient performance on non-natural images captured in special environments or by special measurement devices in real-world applications.

Several recent studies have proposed another approach that does not need real images by employing a novel pre-training scheme that uses synthetic images such as random pattern images or noise images to learn more general feature representations. Their effectiveness has been demonstrated on various downstream tasks such as image classification [6, 7, 26, 27, 36, 37], semantic segmentation [35], and image denoising [1, 2].

While pre-training using synthetic images achieve high performance on natural images regardless of the target dataset, Baradad *et al.* [7] showed that pre-training using synthetic images sometimes underperform in cases without pre-training, depending on the combination of synthetic images and target dataset.

Based on this fact we assume that, unlike natural images, non-natural images have characteristics unique to the individual image depending on the measurement environment and measurement device. This hypothesis implies that it is unrealistic to design universal synthetic images suitable for pre-training of every non-natural image. In this paper, we propose more specialized synthetic images generated from real images of downstream tasks in order to directly learn feature representations suitable for non-natural images of individual downstream tasks (Fig. 1 (b)) and present an application example to contrastive learning framework by following recent studies [6, 7] that investigated pre-training using synthetic images on image classification tasks. Although the generation of images based on real images may appear similar to well-known data augmentation, we emphasize that the concept of the proposed image generator (f_{S} in Fig. 1 (b)) differs from data augmentation in that the data augmentation commonly used

for image generation is not suitable for data augmentation in contrastive learning (f_{CL} in Fig. 1 (a)) and causes the contrastive learning scheme to collapse. To avoid this, we propose using patch-based transformations as an image generator (f_{S}) by utilizing the property that convolutional neural networks (CNNs) are known to learn object features based on local structures. Our main contributions are as follows:

- We introduce the new concept of using a more specialized pre-trained model trained with synthetic images tailored for individual non-natural images, unlike previous studies that have aimed at building a universal pre-trained model.
- We empirically show the existence of key factor for the success of pre-training using synthetic images other than diversity of synthetic images (**Hypotheses 1 and 2** in Section 3.1).
- We propose a synthetic image generation process from real images (simple, yet different from well-known data augmentation) that satisfies the aforementioned key factors while considering compatibility with CNNs and contrastive learning (Section 3.2, and see Fig. 3).
- We perform experiments of image classification task on five non-natural image datasets mimicking real-world applications with little available data and demonstrate that the proposed method achieves higher average classification accuracy compared with pre-training using ImageNet-1k or existing synthetic images, with an improvement of over 6.5 points (see Table 2).

2 Related Work

2.1 Unsupervised Learning

Several studies have demonstrated the effectiveness of unsupervised learning that pre-trains feature extractors based on large amounts of unlabeled data. In particular, contrastive learning that trains a feature extractor based on the similarity of a pair of input images achieves higher performance regardless of the downstream task compared with self-supervised learning using a pseudo task (pretext task) [19, 31] different from the downstream task.

Recently, MoCo v1 [23] and v2 [12], SimCLR [11], BYOL [21], and SwAV [9] have used the new paradigm of contrastive learning as shown in Fig. 1 (a) of preparing image pairs by augmenting a single image to improve the performance of downstream tasks and training stability. Subsequently, Chen *et al*. [13] found the learning mechanism in contrastive learning, and proposed a simplified framework called SimSiam based on MoCo [23] and SwAV [9]. Furthermore, MoCo v3 [14], DINO [10], and EsViT [28] were proposed as vision transformers. Although these promising results demonstrated the potential of contrastive learning as a common pre-training method, such methods still have the limitation of requiring a large number of unlabeled images. More recently, several studies [5, 22, 42] proposed masked image modeling (MIM), a new unsupervised approach different from contrastive learning inspired by masked language modeling

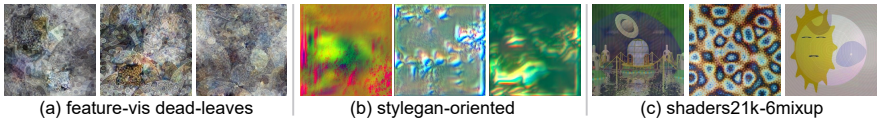

Fig. 2: Examples of existing synthetic images for pre-training that achieved high performance in downstream tasks. (a) Using feature visualization of a classifier pre-trained with dead-leaves images (a kind of synthetic images) [7]. (b) Using untrained StyleGANv2 initialized to output oriented structures [7]. (c) Using OpenGL based on short code snippets curated from Twitter and Shadertoy [6].

(MLM) [15, 32]. However, a comprehensive study about data scaling by Xie *et al*. [43] showed MIM requires at least 20% of ImageNet-1k images (26,000 images) to avoid overfitting phenomenon on ImageNet-1k image classification task, even smaller models.

Although using data augmentation to increase the number of input images from a small number of available unlabeled images may appear to be a simple solution for dealing with this issue, applying the data augmentation that is commonly used in supervised learning to the input images of contrastive learning leads to a collapse of the contrastive learning scheme because the data augmentation of the input images ($\mathbf{x}$ in Fig. 1 (a)) causes conflicts with one of the paired images ($\mathbf{x}^1$ and $\mathbf{x}^2$ in Fig. 1 (a)). As another pre-training approach, Asano *et al*. [3] proposed a data preparation method for unsupervised learning that utilizes a large number of patches extracted from a small number of reference images as training data. However, its effectiveness has been confirmed only with RotNet [19], BiGAN [16], and DeepCluster [8], which are based on different learning principles from the aforementioned contrastive learning [9–14, 21, 23, 28]. In addition, since it is necessary to manually select reference images containing diverse textures and subjects from external datasets other than the target dataset, this method cannot be applied in cases in which a suitable external dataset does not exist for the target dataset.

2.2 Pre-training using Synthetic Images

As a method to obtain a large number of labeled images without human annotation, Kataoka *et al*. [27] proposed a synthetic image dataset (Fractal DB) consisting of fractal images generated based on mathematical formulas, motivated by the insight that public datasets of natural images include images with a fractal structure. Their experiment on the natural image recognition task with CNNs showed that pre-training using Fractal DB achieved performance comparable to that when pre-training using a large-scale public natural image dataset such as ImageNet, despite not using any real images. In addition, subsequent studies [26, 36] empirically found that object contours in the fractal images play an important role in training vision transformers and proposed specifically tailored synthetic images consisting of contour components.

Inspired by the concept of Fractal DB [27], Baradad *et al*. [7] found that noise captures certain structural properties of real images, and through an in-

vestigation of a suite of image generation models, they proposed synthetic images (*e.g.* Fig. 2 (a), (b)) produced by simple random processes. Their comprehensive experiments on VTAB benchmark consisting of 19 datasets including some non-natural image datasets demonstrated that pre-training using their synthetic images outperforms pre-training using large-scale natural image datasets for some non-natural image datasets. However, they pointed out the limitation that the performance saturates despite increasing the number of synthetic images owing to lack of diversity of synthetic images generated from a random process. To overcome this limitation, Baradad *et al.* [6] proposed generating a diverse set of synthetic images (Fig. 2 (c)) using OpenGL based on short code snippets curated from Twitter and Shadertoy, and showed that the performance does not saturate with increasing number of synthetic images. However, Baradad *et al.* [7] showed that pre-training using synthetic images sometimes underperforms in cases without pre-training, depending on the combination of synthetic images and target dataset.

2.3 Simulating Desired Synthetic Images

Several researchers have taken a different perspective and attempted to explicitly generate synthetic images for pre-training suitable for the downstream task. Tu *et al.* [37] proposed optimizing the parameters for generating fractal images [27] by minimizing mean squared error between generated images and target images of the downstream tasks, and demonstrated that pre-training using the tailored fractal images outperforms pre-training using original fractal images. However, the effectiveness of this approach was confirmed only for low-resolution images such as three MNIST-related datasets owing to the lack of fine structure in fractal images during the optimization process and the high computational cost.

Based on similar motivations, Mishra *et al.* [29] proposed generating synthetic natural images using an image generation simulator called TDW [17] designed for the specific downstream task. First, they trained an estimator that outputs the optimal parameters to feed into TDW to generate synthetic images mimicking real images in several datasets, and then they generated realistic synthetic images using TDW by inputting a few of the available images in the target task into the trained estimator. However, TDW is capable of generating only natural images, and therefore it cannot be applied to non-natural images.

Recent text-to-image models (*e.g.* stable diffusion [33]) also have the potential to generate synthetic images suitable for downstream tasks by inputting a prompt describing the target images of the downstream tasks. However, since such text-to-image models are pre-trained using pairs of natural images and captions (text), it is difficult to generate non-natural images that are not included in the training data. Even if text-to-image models fine-tuned using non-natural images are available, it is unrealistic to prepare many prompts because most non-natural images cannot be described by natural languages, unlike natural images.

Major limitations of related works applied to non-natural images in real-world applications including the following:

- Contrastive learning requires a large number of unlabeled images. Data augmentation of input images does not help owing to conflicts between data augmentation of the input image and one of the paired images.
- Pre-training using synthetic images shows poor performance depending on the combination of synthetic images and target dataset. In the worst case, suitable synthetic images do not exist.
- Generating realistic synthetic images was originally proposed for natural images. Building a simulator specialized for non-natural images is technically possible, but requires a large number of unlabeled images (and many prompts for text-to-image models).

3 Proposed Method

3.1 Concept

Baradad *et al*. [7] claimed that diversity of synthetic images is a key property for learning good representations and their comprehensive experiments showed the existence of a sweet spot for diversity within the synthetic images. In the same context, we found that one synthetic image (Fig. 2 (a) [7]) showed effectiveness despite having less diversity compared with other synthetic images (Fig. 2 (b) [7], (c) [6]). This observation suggests that diversity of synthetic images is not a dominant factor for success of pre-training, and implies the existence of other factors.

In this paper, we assume that the learning process for discriminating similar images as shown in Fig. 2 (a) dominantly contributes to acquisition of more informative feature representation for various downstream tasks, rather than learning of diverse local structures (**Hypothesis 1**). Moreover, as Tu *et al*. [37] pointed out, since existing synthetic images do not necessarily contain local structures in real images of the target dataset, we assume that the generation of synthetic images which explicitly contain local structures or partial image of real images offers more efficient pre-training for downstream tasks (**Hypothesis 2**).

As a pre-training method suitable for non-natural images based on these two hypotheses, we propose a contrastive learning framework using synthetic images obtained by applying an image generator f_S to the real images of the target dataset, as shown in Fig. 1 (b).

General data augmentations perform transformation such that the images before and after transformation are in the same image or class. In contrast, f_{S} no longer needs to satisfy this kind of constraint and it is acceptable to generate synthetic images that do not belong to any class in the target dataset, even if at first glance it is a meaningless image with no semantic information. For this reason, in order to explicitly distinguish between an image transformed via general data augmentation and an image generated from f_{S}, we refer to

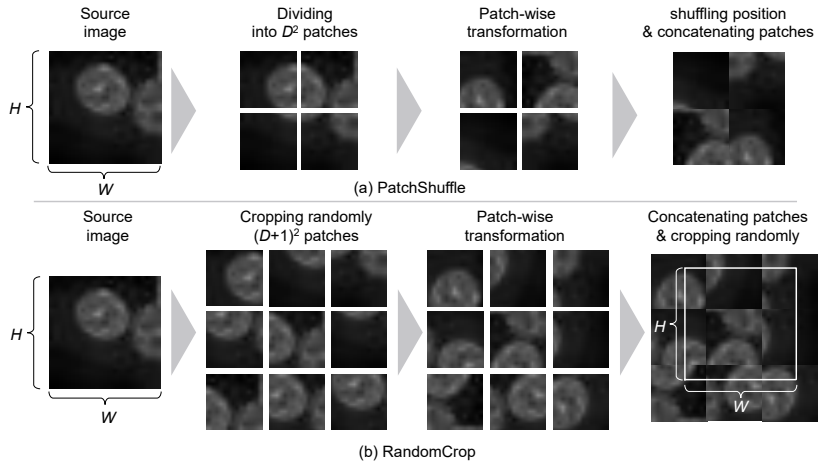

Fig. 3: Proposed two patch-based image generator (f_S).

the latter as a synthetic image[2]. In conventional contrastive learning (Fig. 1 (a)), the feature extractor is trained so that the feature vectors $\mathbf{z}^1$ and $\mathbf{z}^2$ of the paired images $\mathbf{x}^1$ and $\mathbf{x}^2$ obtained by data augmentation f_{CL} consisting of image cropping and color conversion to the input image $\mathbf{x}$ are close to each other.

In contrast, the proposed method (Fig. 1 (b)) first generates M synthetic images $\mathbf{s}_{m\in M}$ from the input image $\mathbf{x}$ using the image generator f_{S} to generate similar images as discussed in **Hypothesis 1**, and then trains a feature extractor in the same way as conventional contrastive learning. However, the proposed method differs from conventional contrastive learning in that the feature extractor is trained to repel pairs of features obtained from different synthetic images (for example, pair of $\mathbf{z}_i^2$ and $\mathbf{z}_j^1$ shown in Fig. 1 (b)), even if the synthetic images $\mathbf{s}_i$ and $\mathbf{s}_j$ are generated from the same input image $\mathbf{x}$.

3.2 Image Generator

The requirements that f_{S} must satisfy for this learning scheme to work properly are shown below.

Requirement 1: $f_{\mathbf{S}}$ is not included in $f_{\mathbf{CL}}$. If f_{S} is included in f_{CL}, paired images generated from different synthetic images may unintentionally become identical, making it impossible to discriminate them.

Requirement 2: The local structure of the original image is preserved. CNNs are known to learn object features based on local structures (*e.g.* high frequency components, textures) in the input image [18, 25, 30, 38]. Therefore, rescaling and filtering processes that change the frequency characteristics of images, or processes that change the aspect ratios of images are not suitable,

[2] Our definition of "synthetic image" may differ from one of existing papers. Generally, "synthetic image" means an image generated from generative models (*e.g.* GAN, diffusion model). In contrast, our definition means "non-real images" which is broader concept than general one. So, f_{S} does not necessarily have to be generative models.

because they destroy the local structure contributing to classifying images in the target dataset.

Although existing image generators from single images (*e.g.* SinGAN [34], GPNN [20]) seem to be a promising candidate as f_S, these methods does not satisfy **Requirement 2** perfectly in some cases. Because these methods generate diverse images from a single input image by swapping similar patches after adding noise and overlaying swapped patches at multiple resolutions, overlaying patches partially destroy local structure of the overlayed region if a less similar patch is found in the swapping process. As a simple yet universal f_{S} that satisfies both requirements, we propose two patch-based transformations (Fig. 3) inspired from self-supervised learning by solving jigsaw puzzles [31].

Image generator 1: PatchShuffle (Fig. 3 (a)). First, this transformation divides a source image of size $W \times H$ into D^2 patches of size $W/D \times H/D$, where D is a hyperparameter defining the number of divisions per side. Next, each patch is applied by a rigid transformation that maintains its characteristics. In this paper, we used random rotations of [0°, 90°, 180°, 270°] and random horizontal flips as rigid transformations that can be applied regardless of the characteristics of the subject. Finally, all patches are shuffled randomly and concatenated into a single image of size $W \times H$. This transformation is beneficial for completely preserving statistics of pixel values (mean and standard deviation) because the included pixels in the synthetic image are exactly the same as those in the source image. One drawback is that there is an upper limit to the number of synthetic images generated from one source image, because the total combination of patches is D^2 ! (permutation of patch position) $\times$ 8 (patch orientation).

Image generator 2: RandomCrop (Fig. 3 (b)). RandomCrop is an extended version of PatchShuffle that generates more diverse synthetic images. First, $(D+1)^2$ patches are randomly cropped from the source image. Next, these patches are transformed following the same protocol as PatchShuffle. Finally, all patches are concatenated into a single image and the concatenated image is randomly cropped to $W \times H$.

Confirmation of satisfaction of the requirements. Since most recent contrastive learning for CNNs [9,11–13,21,23] use color distortion, Gaussian blur, rescaling, and cropping as data augmentation (f_{CL}), the proposed image generator (f_{S}) using patch-based transformation is not included in f_{CL} (**satisfying Requirement 1**). The proposed image generator also uses only a patch-based transformation via rigid transformation (no use of rescaling and filtering processes), so it can clearly preserve local structures in source images (**satisfying Requirement 2**).

4 Experiments

4.1 Experimental Setup

Task setting. Figure 4 shows an overview of our experiments. Using the public dataset of non-natural images shown in Table 1, we conducted an image classification task under conditions that mimic real-world applications with little

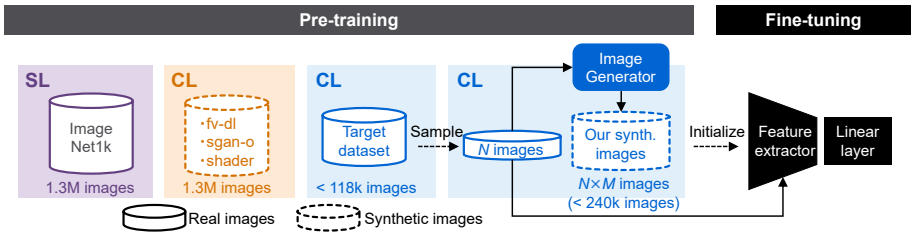

Fig. 4: An overview of our experiments (SL, supervised learning; CL, contrastive learning).

Table 1: Detail of datasets; Thermal Image dataset (`FLIR`), TissueMNIST (`Tissue`), WM811k (`WM`), PatchCamelyon (`PCAM`) and Diabetic Retinopathy (`DR`). Note that the image sizes for `FLIR`, `WM`, `DR` are the sizes of the images after resizing.

	FLIR [4]	Tissue [44]	WM [41]	PCAM [39]	DR [40]
Image type	Infrared	Microscopic	Wafer map	Pathological	Retinal
Image size	250×190	28×28	45×45	96×96	224×224
# class	3	8	9	2	5
# train image	645	165,466	54,345	262,144	28,100
# test image	225	47,280	118,595	32,782	7,026

available training data and evaluated the classification accuracy on each test set. We randomly extracted N labeled images for training from a training set of original datasets so that each class has the same number of images. We took half of the N labeled images as a training set and the remain as a validation set.

Pre-training. For training the feature extractor of the classifier, we used contrastive learning using the three existing synthetic images shown in Fig. 2 that showed high performance in previous works [6, 7] and supervised learning using ImageNet-1k as comparison methods, and contrastive learning using the entire training set in the original datasets as a reference method. Note that we used pre-trained models for existing synthetic images and ImageNet-1k.

Training detail. After pre-training, we fine-tuned the parameters of the entire classifier using N labeled images of the target dataset. For contrastive learning, we used MoCo v2 [12] following previous works [6, 7] and trained the feature extractor for 400 epochs with Adam optimizer and learning rate annealing by cosine decay. For supervised learning, we used the data augmentation described in the original papers [39–41, 44] that proposed each dataset, and then trained the entire classifier for 100 epochs with SGD optimizer and learning rate annealing by step decay multiplying by 0.1 every 30 epochs. Through all experiments, we used ResNet50 [24] as a backbone architecture and tuned mini-batch size $\in$ [64, 128, 256], initial learning rate $\in$ [0.01, 0.03, 0.05].

Synthetic images. We generated M synthetic images using the proposed image generators (PatchShuffle and RandomCrop) described in the previous section as f_{S} with $D \in [2, 4, 8]$. In addition, we also used an existing image generator

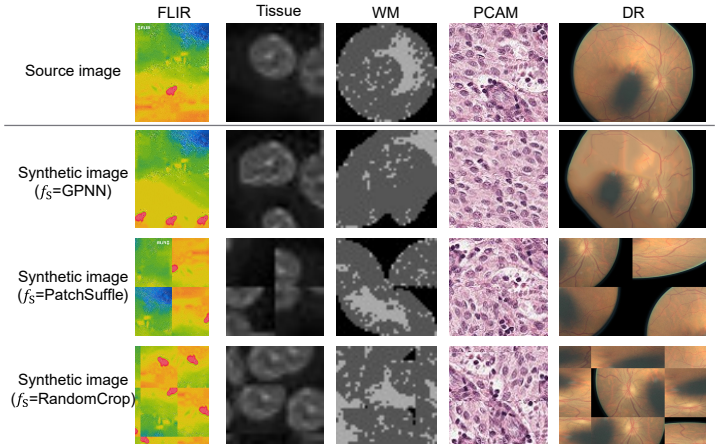

Fig. 5: Examples of source images and synthetic images.

called GPNN [20], which is a lightweight version of SinGAN [34] that replaces the neural network with a non-learnable module, to confirm the effectiveness of the proposed image generator. Note that although image generation with f_S can be applied on-the-fly during the training process, we instead randomly selected a synthetic image from pre-generated synthetic images to compare the number of pre-learning images with ImageNet-1k and existing synthetic images. Figure 5 shows examples of the synthetic images.

4.2 Results

Comparison result (Table 2). Pre-training using the proposed method achieved the highest average classification accuracy regardless of f_S, especially using RandomCrop as f_S achieved the highest one among the proposed method. Even though pre-training using existing synthetic images and ImageNet-1k used 1,300,000 images, the proposed method achieved a comparable performance to the other methods despite using a smaller number of images (less than 240,000 images). These results suggest that the proposed synthetic images based on the real images of the target dataset contributed to efficient pre-training for downstream task (**confirming Hypothesis 2**).

Although it is known that contrastive learning using a larger number of data tends to bring higher performance on downstream tasks, the proposed method using only a small number of real images outperformed contrastive learning using the entire training set (top row in Table 2) on all datasets except `PCAM`. Since `PCAM` is a class-balanced dataset that has the same amount of data in each class, unlike the other class-imbalanced datasets that have a different amount of data in each class, we assume that this is caused by the class balance of the dataset used for pre-training. More specifically, although it is known that general contrastive learning requires special aids to obtain the full potential in a class-imbalanced dataset, we did not apply any aid. In contrast, since the proposed method used

Table 2: Classification accuracy on each dataset [%]. (SL, supervised learning; CL, contrastive learning; IN1k, ImageNet-1k; GP, GPNN; PS, PatchShuffle; RC, RandomCrop). **Bold** indicates the best result in each column except contrastive learning using the entire training set in the original dataset (top row; reference method).

Pre-training	#imgs Real	#imgs Synth.	FLIR $N60$	FLIR $N300$	Tissue $N160$	Tissue $N100$	WM $N180$	WM $N720$	PCAM $N40$	PCAM $N400$	DR $N100$	DR $N1000$	Ave.
CL (w/ all)	<118k	–	51.6	77.3	34.7	45.3	71.7	67.0	76.3	79.6	31.1	36.9	57.1
Scratch	–	–	41.8	67.6	32.7	41.1	31.2	62.5	65.8	74.7	18.6	30.5	46.6
SL (IN1k)	1.3M	–	75.1	90.2	34.5	43.0	78.9	79.5	72.5	78.6	30.9	33.0	61.6
CL	N	–	66.7	84.4	35.5	40.2	46.2	77.9	62.4	74.5	15.3	35.7	53.9
CL (fv-dl)	–	1.3M	71.6	82.7	35.3	44.1	66.8	73.1	70.0	77.5	29.0	52.6	60.3
CL (sgan-o)	–	1.3M	63.6	78.7	33.6	46.5	57.6	71.8	**75.5**	76.9	28.1	39.6	57.2
CL (shader)	–	1.3M	68.0	79.1	36.3	45.5	61.3	70.2	73.9	77.8	27.1	41.7	58.1
Ours (GP)	N	<240k	**85.3**	96.4	37.0	41.4	65.6	90.0	72.7	76.4	32.4	43.6	64.1
Ours (PS)	N	<240k	78.7	95.6	37.7	45.9	**82.8**	90.0	73.8	78.5	**34.1**	53.8	67.1
Ours (RC)	N	<240k	81.6	**98.7**	**39.4**	**47.0**	77.2	**91.3**	75.2	**79.5**	33.7	**57.4**	**68.1**

Table 3: Comparison of classification accuracy [%] on downstream tasks between conventional contrastive learning using f_{CL} and f_{S} as data augmentation and the proposed method. **Bold** indicates the best result on each dataset and f_{S}.

f_{S}	Method	FLIR $N60$	Tissue $N160$	WM $N180$	Ave.
GPNN	CL (f_{CL}+f_{S})	79.6	33.8	56.7	56.7
	Ours	**85.3**	**37.0**	**65.6**	**62.6**
Patch Shuffle	CL (f_{CL}+f_{S})	68.9	35.5	76.1	60.2
	Ours	**78.7**	**37.7**	**82.8**	**66.4**
Random Crop	CL (f_{CL}+f_{S})	70.7	35.5	61.4	55.9
	Ours	**81.6**	**39.4**	**77.2**	**66.1**

class-balanced synthetic images generated equally from each class, we assume that the image generation process itself implicitly acted as a countermeasure against class imbalancement. Note that we aware of the existence of situations in which we cannot control the measurement data (for example, when some of the classes correspond to anomaly data). Therefore, investigation of the class-imbalanced setting is left as future work.

Ablation study. While conventional contrastive learning uses only data augmentation f_{CL}, the proposed method uses not only f_{CL} but also an image generator f_{S} during training. To confirm that the effect of the proposed method is not simply due to diversification of data augmentation, we evaluated another configuration that uses both f_{CL} and f_{S} as data augmentation. As shown in Table 3, the average classification accuracy of the proposed method exceeded that of conventional contrastive learning. Therefore, we empirically demonstrate that the learning algorithm itself that discriminates similar synthetic images generated by f_{S} plays a key role in the proposed method (**confirming Hypothesis 1**).

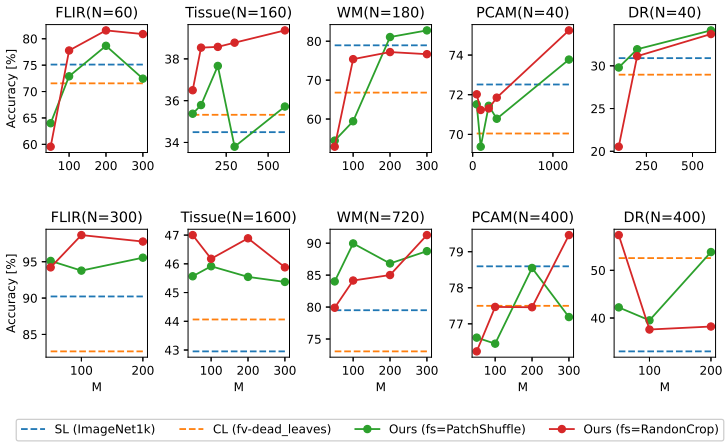

Fig. 6: Effect of number of generated synthetic images from each source image (M). Each column corresponds to a dataset. The top and bottom rows differ in terms of the number of real images (N). Dashed lines indicate strong comparison methods that achieved higher performance in Table 2 (pre-training using ImageNet-1k and feature-vis dead-leaves).

Hyperparameters of synthetic images. Figure 6 shows the effect of the number of generated synthetic images from each source image (M). We found that more synthetic images tend to bring higher classification accuracy, particularly in situations in which the number of real images is small (top row of Fig. 6). However, since the number of synthetic images that can be generated by PatchShuffle without duplicates has an upper bound determined by D, some cases in the results of the proposed method using PatchShuffle do not contribute to performance gain even with increasing M (see `FLIR` at N=60, `Tissue` at N=160, and `PCAM` at N=400 in Fig. 6). Therefore, RandomCrop gives a suitable f_S for maximizing the potential of the proposed method by increasing the number of synthetic images.

Figures 7 and 8 show the effects of the number of divisions per side in source image (D) and optimal D against image size. These results show that the optimal D depends not only on the dataset but also on the number of real images (N). Although a larger D results in more diverse synthetic images, the occurrence of flat patches that do not contain texture increases because a larger D decreases the patch size. Moreover, Fig. 8 indicates that the optimal D and images are not correlated. Based on this fact, we assume that the optimal D depends on scene composition, object size, or frequency of local structures of each image rather than on image size. Therefore, determining D for each image may bring performance improvement.

Combination of different f_S. As mentioned in the discussion of Fig. 6, the number of synthetic images related to diversity of synthetic images is a key factor in the proposed method. This observation suggests that combination of different f_S has the potential to boost performance particularly in situations in which the

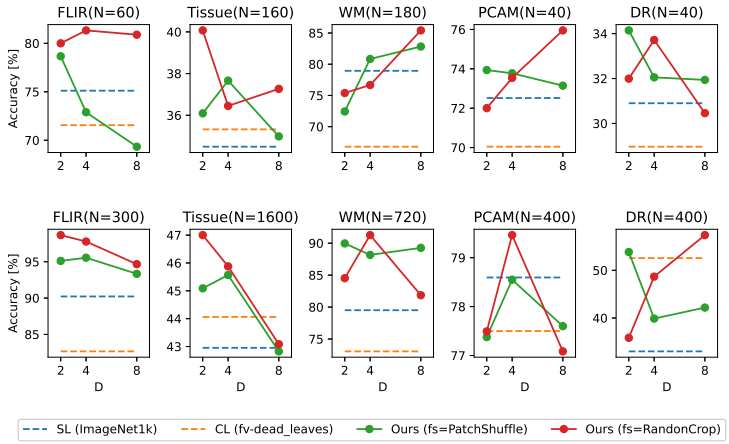

Fig. 7: Effect of number of divisions per side in source image (D). Each column corresponds to a dataset. The top and bottom rows differ in terms of the number of real images (N). Dashed lines indicate strong comparison methods that achieved higher performance in Table 2 (pre-training using ImageNet-1k and feature-vis dead-leaves).

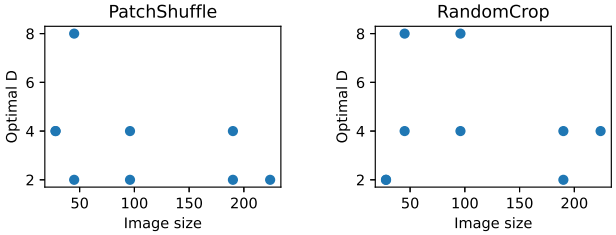

Fig. 8: Optimal number of divisions per side in source image (D) versus image size.

number of real images is small. Table 4 demonstrates that the combination of different f_S brings further improvement.

5 Discussion and Conclusion

In this paper, we proposed a contrastive learning framework using synthetic images generated from real images of downstream tasks to directly learn feature representations suitable for downstream tasks of non-natural images. Experiments using image classification tasks mimicking real-world application with little available data on five non-natural image datasets demonstrated that the proposed method achieves higher average classification accuracy compared with pre-training methods using ImageNet or existing synthetic images. We now summarize our observations through exploration as follows.

Effect of number of real and synthetic images. Although the proposed method achieved the highest classification accuracy in most situations, it slightly

Table 4: Comparison of classification accuracy [%] on downstream tasks between different configuration of f_{S}. **Bold** indicates higher performance than using each f_{S} alone.

f_{S}	FLIR $N60$	Tissue $N160$	WM $N180$	Ave.
GPNN	85.3	37.0	65.6	62.6
PatchShuffle	78.7	37.7	82.8	66.4
PatchShuffle+GPNN	73.8	37.2	**84.6**	65.2
RandomCrop	81.6	39.4	77.2	66.1
RandomCrop+GPNN	78.2	**40.3**	**85.5**	**68.0**

underperformed pre-training using existing synthetic images (stylegan-o) in the PCAM dataset where N=40 in contrast to the same dataset where N=400 (Table 2). This result means that the number of real images used as the source of synthetic images a crucial factor for determining the performance of the proposed method. However, we empirically found that more synthetic images brought higher classification accuracy even situations with few real images (see Fig. 6).

Limitations. In terms of pre-training cost, the proposed method requires additional cost due to the necessity of pre-training for each dataset, unlike models pre-trained using ImageNet-1k or existing synthetic images. However, as mentioned in the comparison results (Table 2), the proposed method can train a relatively smaller number of images (less than 240,000 images) compared with ImageNet-1k (1,300,000 images) or existing synthetic images (1,300,000 images), so pre-training of the proposed method takes only 1-3 days with a single GPU (using NVIDIA A100 or NVIDIA A6000 Ada). From another perspective, since the image generator (f_{S}) differed depending on the target dataset, the design of a more general image generator remains as a future challenge. However, we found that one promising solution is combination with a different image generator (Table 4). In this paper, we focused mainly on a relatively standard setting in real-world applications with a well-known architecture (CNNs) and task (image classification) as a first step. In future work, we intend to reveal the effectiveness on other tasks besides image classification and extend the concept of the proposed method to other architectures (*e.g.* vision transformers).

Towards a better synthetic image. Based on the fact that contrastive learning using the entire training set (top row in Table 2) achieved the highest classification accuracy on the PCAM dataset, we infer that the individual images in PCAM were more diverse than in other datasets. For this kind of dataset, we believe that image generation from multiple real images to generate more diverse synthetic images capturing the essential structure of target dataset is a possible solution. However, we stress that the proposed method still outperformed training from scratch and contrastive learning using only real images and showed comparable performance to existing pre-training. Therefore, considering the average classification accuracy, we believe that the proposed method is worth applying to most non-natural images in real-world applications.

References

1. Achddou, R., Gousseau, Y., Ladjal, S.: Synthetic images as a regularity prior for image restoration neural networks. In: Scale Space and Variational Methods in Computer Vision. pp. 333–345 (2021) 2
2. Achddou, R., Gousseau, Y., Ladjal, S.: Fully synthetic training for image restoration tasks", computer vision and image understanding. Computer Vision and Image Understanding **233**, 103723 (2023) 2
3. Asano, Y.M., Rupprecht, C., Vedaldi, A.: A critical analysis of self-supervision, or what we can learn from a single image. In: ICLR. pp. 1–16 (2020) 4
4. Ashfaq, Q., Akram, U., Zafar, R.: Thermal image dataset for object classification. https://data.mendeley.com/datasets/btmrycjpbj (2021) 9
5. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: ICLR. pp. 1–18 (2022) 1, 3
6. Baradad, M., Chen, C.F., Wulff, J., Wang, T., Feris, R., Torralba, A., Isola, P.: Procedural image programs for representation learning. In: NeurIPS. pp. 1–13 (2022) 2, 4, 5, 6, 9
7. Baradad, M., Wulff, J., Wang, T., Isola, P., Torralba, A.: Learning to see by looking at noise. In: NeurIPS. pp. 2556–2569 (2021) 2, 4, 5, 6, 9
8. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV. pp. 1–18 (2018) 4
9. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS. pp. 9912–9924 (2020) 1, 3, 4, 8
10. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021) 1, 3, 4
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020) 1, 3, 4, 8
12. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. In: arXiv:2003.04297. (2020) 1, 3, 4, 8, 9
13. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021) 1, 3, 4, 8
14. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: ICCV. pp. 9640–9649 (2021) 1, 3, 4
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171–4186 (2019) 3
16. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. In: ICLR. pp. 1–18 (2017) 4
17. Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., Freitas, J.D., Kubilius, J., Bhandwaldar, A., Haber, N., Sano, M., Kim, K., Wang, E., Mrowca, D., Lingelbach, M., Curtis, A., Feigelis, K., Bear, D.M., Gutfreund, D., Cox, D., DiCarlo, J.J., Tenenbaum, J.B., McDermott, J.H., Yamins, D.L.K.: Threedworld: A platform for interactive multi-modal physical simulation. In: NeurIPS. pp. 1–13 (2021) 5
18. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: ICLR. pp. 1–22 (2019) 7

19. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR. pp. 1–16 (2018) 3, 4
20. Granot, N., Feinstein, B., Shocher, A., Bagon, S., Irani, M.: Drop the gan: In defense of patches nearest neighbors as single image generative models. In: CVPR. pp. 13460–13469 (2022) 8, 10
21. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learningl representations. In: NeurIPS. pp. 21271–21284 (2020) 1, 3, 4, 8
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022) 1, 3
23. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9279–9738 (2020) 1, 3, 4, 8
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 9
25. Hermann, K.L., Chen, T., Kornblith, S.: The origins and prevalence of texture bias in convolutional neural networks. In: NeurIPS. pp. 1–16 (2020) 7
26. Kataoka, H., Hayamizu, R., Yamada, R., Nakashima, K., Takashima, S., Xinyu Zhang, E.J.M.N., Inoue, N., Yokota, R.: Replacing labeled real-image datasets with auto-generated contours. In: CVPR. pp. 21232–21241 (2022) 2, 4
27. Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., Satoh, Y.: Pre-training without natural images. In: ACCV. pp. 1–17 (2020) 2, 4, 5
28. Li, C., Yang, J., Zhang, P., Gao, M., Xiao, B., Dai, X., Yuan, L., Gao, J.: Efficient self-supervised vision transformers for representation learning. In: ICLR. pp. 1–27 (2022) 1, 3, 4
29. Mishra, S., Panda, R., Phoo, C.P., Chen, C.F.R., Karlinsky, L., Saenko, K., Saligrama, V., Feris, R.S.: Task2sim : Towards effective pre-training and transfer from synthetic data. In: CVPR. pp. 9194–9204 (2022) 5
30. Naseer, M., Ranasinghe, K., Salman Khan, M.H., Khan, F.S., Yang, M.H.: Intriguing properties of vision transformers. In: NeurIPS. pp. 1–13 (2021) 7
31. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV. pp. 69–84 (2016) 3, 8
32. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018) 3
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. pp. 10684–10695 (2022) 5
34. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. In: ICCV. pp. 4570–4580 (2019) 8, 10
35. Shinoda, R., Hayamizu, R., Nakashima, K., Inoue, N., Yokota, R., Kataoka, H.: Segrcdb: Semantic segmentation via formula-driven supervised learning. In: ICCV. pp. 20054–20063 (2023) 2
36. Takashima, S., Hayamizu, R., Inoue, N., Kataoka, H., Yokota, R.: Visual atoms: Pre-training vision transformers with sinusoidal waves. In: CVPR. pp. 18579–18588 (2023) 2, 4
37. Tu, C.H., Chen, H.Y., Carlyn, D., Chao, W.L.: Learning fractals by gradient descent. In: AAAI. pp. 2456–2464 (2023) 2, 5, 6
38. Tuli, S., Dasgupta, I., Grant, E., Griffiths, T.L.: Are convolutional neural networks or transformers more like human vision? In: Annual Meeting of the Cognitive Science Society. pp. 1844–1850 (2021) 7

39. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 210–218. Springer International Publishing (2018) 9
40. Wang, Z., Yang, J.: Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation. In: AAAI Workshop. pp. 514–521 (2018) 9
41. Wu, M.J., Jang, J.S.R., Chen, J.L.: Wafer map failure pattern recognition and similarity ranking for large-scale data sets. IEEE Transactions on Semiconductor Manufacturing **28**(1), 1–12 (2015) 9
42. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: CVPR. pp. 9653–9663 (2022) 1, 3
43. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Wei, Y., Dai, Q., Hu, H.: On data scaling in masked image modeling. In: CVPR. pp. 10365–10374 (2023) 4
44. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023) 9