

Inter-Instance Similarity Modeling for Contrastive Learning

Chengchao Shen¹, Dawei Liu¹, Hao Tang¹, Zhe Qu¹, Jianxin Wang¹

¹Central South University

{scc.cs, tanghao, zhe_qu}@csu.edu.cn, liudw0702@163.com, jxwang@mail.csu.edu.cn

Abstract

The existing contrastive learning methods widely adopt one-hot instance discrimination as pretext task for self-supervised learning, which inevitably neglects rich inter-instance similarities among natural images, then leading to potential representation degeneration.

In this paper, we propose a novel image mix method, PatchMix, for contrastive learning in Vision Transformer (ViT), to model inter-instance similarities among images. Following the nature of ViT, we randomly mix multiple images from mini-batch in patch level to construct mixed image patch sequences for ViT. Compared to the existing sample mix methods, our PatchMix can flexibly and efficiently mix more than two images and simulate more complicated similarity relations among natural images. In this manner, our contrastive framework can significantly reduce the gap between contrastive objective and ground truth in reality. Experimental results demonstrate that our proposed method significantly outperforms the previous state-of-the-art on both ImageNet-1K and CIFAR datasets, e.g., 3.0% linear accuracy improvement on ImageNet-1K and 8.7% kNN accuracy improvement on CIFAR100. Moreover, our method achieves the leading transfer performance on downstream tasks, object detection and instance segmentation on COCO dataset. The code is available at <https://github.com/visresearch/patchmix>.

1. Introduction

Massive contrastive learning studies [33, 4, 16, 6] have achieved impressive performance on unsupervised visual representation learning. Although various designs are proposed to improve the representation performance, these methods can be summarized into an instance discrimination task. In this pretext task, the unlabeled images are augmented into several views with different appearances. Under the contrastive objective, the model is trained to capture appearance-invariant representations and understand the semantics consistency between positive pairs in the different views. Specifically, this objective aims to maximize the

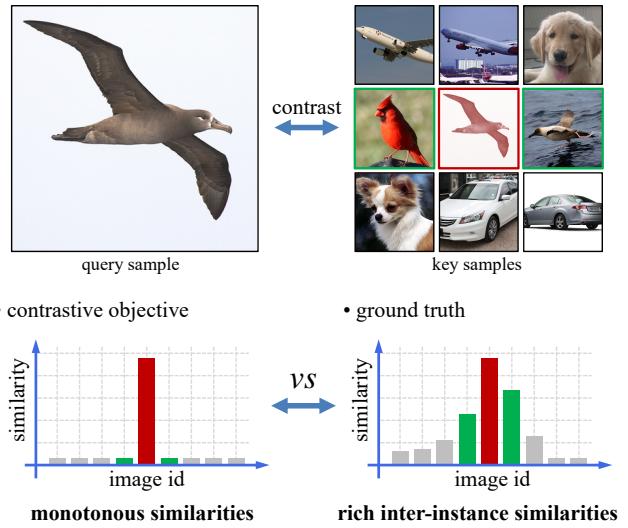


Figure 1: Monotonous similarity in contrastive objective vs rich inter-instance similarities among natural images. Simply contrasting different views from the same image may miss potential positive samples and thus degrade representation performance.

similarity between views augmented from the same images, meanwhile minimize the one between samples from different images. However, as shown in Figure 1, there are rich similarities among different image samples in reality, which is ignored by the existing contrastive frameworks. In other words, this issue overlooks massive potential positive samples, thus introducing inaccurate targets into contrastive learning. The discrepancy between one-hot targets and real label distributions hurts the performance of unsupervised representation.

Supervised classification tasks suffer a similar issue, where one-hot classification targets omit potential inter-class similarity. Image mix strategies, such as Mixup [37] and CutMix [36], are proposed to soften one-hot labels by simply mixing image pairs. These strategies are also applied to contrastive learning [21, 32, 18] to alleviate inaccurate target issue. However, for contrastive learning task,

inter-instance similarity is significantly richer than inter-class similarity in supervised classification task, not only including inter-class similarity as classification task but also intra-class similarity (image instances from the same class are more likely to be similar to each other). In other words, the real targets for contrastive learning are softer than the ones for supervised classification. The plain Mixup and CutMix strategy only supports similarity modeling between two images, but fails to construct similarities among more images.

Vision Transformer (ViT) [13] takes image as the sequence of image patches, which provides a novel perspective for image modeling. Furthermore, the success of masked image modeling [1, 34, 15] demonstrates that small portion of patches from image, e.g., 25%, can well represent the semantics of the original image, which allows mixing more images into a hybrid one.

Following the nature of ViT, we propose a novel image mix method, *PatchMix*, which randomly mixes several patch sequences of images, to simulate training samples with rich inter-instance similarities in an unsupervised manner. For example, we mix four images, including dog, bird, airplane and car, into a hybrid patch sequence of image, whose patches contain dog head, bird wing, tail plane and wheel. This hybrid sample can effectively encourage the trained model to explore rich component similarities among multiple images and improve the generalization of representations. Moreover, in the hybrid sequence, the inconsecutive patches from the same image regularize the trained model to construct long distance dependency among patches.

To model inter-instance similarities, we conduct contrastive learning from three aspects. First, *mix-to-origin contrast* is conducted to construct similarities between the mixed images and the original ones, where each mixed image corresponds to multiple original images. Second, *mix-to-mix contrast* provides more positive pairs than mix-to-origin contrast to further model more complicated relations between the mixed images. Third, *origin-to-origin contrast* is used to eliminate potential representation gap produced by the domain gap between mixed images and original ones.

In summary, the main contribution of our paper is a novel image mix method dedicated to ViT, PatchMix, which can effectively mixes multiple images to simulate rich inter-instance similarities among natural images. Experimental results demonstrate that our method effectively captures rich inter-instance similarities and significantly improves the quality of unsupervised representations, achieving state-of-the-art performance on image classification, object detection and instance segmentation.

2. Related Work

In this section, we discuss related work from two aspects, contrastive learning and image mix strategies and analyze

the main differences between our method and related work.

2.1. Contrastive Learning

To learn meaningful representations from unlabeled image data, contrastive learning methods [33, 4, 16] conduct instance discrimination by maximizing representation similarity between positive pairs and minimizing the one between negative pairs. Several key components are proposed to improve the performance of contrastive representations.

First, more and stronger data augmentation techniques, such as random crop and color jitter, are applied to positive pairs [4, 14]. Hence, the trained model is encouraged to learn appearance-invariant representations from samples under different distortions. Second, additional modules, such as projection [4, 16] and prediction head [14], are used to improve the transferability of contrastive representations by decoupling the representations from contrastive pretext task. To construct a more stable dynamic look-up dictionary, momentum encoder is introduced by MoCo series [16, 6, 8]. Third, clustering-based methods [2] and prediction-based methods [14, 7] are proposed to explore contrastive learning without negative pairs.

In spite of encouraging performance achieved, these works neglect the potential similarities among image instances, where massive potential positive samples are ignored during training and the trained model is hurt by misleading targets. In this work, we efficiently construct learning targets with multi-instance similarities by mixing multiple images. In this manner, the trained network is encouraged to model rich inter-instance similarities and then improves the quality of unsupervised representations.

2.2. Image Mix

Image mix strategies are widely adopted in image classification to reduce the overfitting risk by softening the classification labels. Mixup [37] forms a mixed sample by applying a weighted sum between two randomly sampled images, where the weights for sum are used as the classification labels. Following the idea of Mixup, CutMix [36] pastes the patch from image to another one, to regularize the training of model. To alleviate uninformative patch mix introduced by CutMix, saliency-guided mix strategies [29, 10, 19] are proposed to mix informative parts from two images according to saliency maps. To reduce computation cost of saliency-guided mix strategies, TokenMixup [9] mixes informative parts from two images by off-the-shelf attention scores of Transformer architecture. Apart from mix on raw images, TokenMix [24] and TokenMixup [9] also mix the tokens from two images under Transformer architecture. Manifold Mixup [31] and Alignmixup [30] also explore image mix on intermediate representations.

Image mix strategies are also be researched in contrastive learning [21, 32, 18]. SDMP [26] conducts a se-

symbol	description
x_i	the i -th image sample in image batch x
x_i^{shuffle}	the image x_i whose patch sequence is randomly shuffled.
x_i^{smix}	the image applied patch mix in a shuffled patch order
p_{ij}	the j -th image patch from sample x_i
G_{im}	the m -th patch group of x_i
$k(j)$	the index item after shuffling $\{j\}_{j=0}^{T-1}$
$u(i, m)$	the index transform for patch mix
$r(j)$	the index after recovering the order of patch sequence
$v(i, j)$	the index transform from group index to patch index
$w(i, j)$	the index transform from shuffled index to unshuffled index
l	1-d index for the original patch sequence of mini-batch
q	1-d index for patch mix in mini-batch

Table 1: List of symbols for patch mix strategy.

ries of image mix augmentation strategies: Mixup, CutMix and ResizeMix [25] to obtain a novel mixed image from two given images, which form a positive triplet to perform contrastive learning. RegionCL [35] swaps the region of two images to obtain additional positive pairs for contrastive learning.

Although these methods effectively alleviate the risk of overfitting, they only consider similarity between two image instances and fail to simulate richer inter-instance similarities in reality. The proposed PatchMix can be easily extended to a more general case, where any number of images can be mixed to model more sophisticated similarity relations among images in reality.

3. The Proposed Method

In this section, we present our patch mix strategy to construct hybrid image instances for the simulation of rich inter-instance similarities in reality. Then, we conduct mix-to-origin, mix-to-mix and origin-to-origin contrast using hybrid samples constructed by the proposed patch mix to model inter-instance similarities.

3.1. Patch Mix

To simulate rich inter-instance similarities in reality, we construct image instances and soft training targets by mixing multiple images in patch level. In this way, the mixed patch sequence can possess features from several instances and provide a more practical target for contrastive learning.

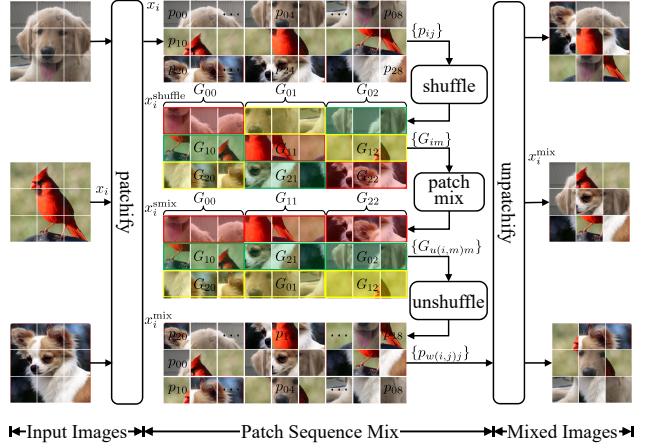


Figure 2: The overview of patch mix strategy. First, the patch sequence $\{p_{ij}\}_{j=0}^{T-1}$ is shuffled into x_i^{shuffle} to conduct unbiased patch sampling. Second, the shuffled patch sequences are mixed in a group-wise fashion to reform novel mixed image x_i^{smix} (e.g., patch group $\{G_{00}, G_{11}, G_{22}\}$ is combined into mixed image x_0^{smix} .). Third, the patches in the shuffled mixed image x_i^{smix} are recovered into original positions to obtain the final mixed image x_i^{mix} .

The overview of PatchMix is shown in Figure 2. For better illustration, the symbols used in this section are listed in Table 1. In spite of somewhat complicated description, our method provides a general-purpose but efficient solution to implement patch mix with any number of images.

Let x_i denote the i -th image instance in the image batch $x = \{x_i\}_{i=0}^{N-1}$,¹ where N denotes batch size of x . Without loss of generality, we patchify x_i into an image patch sequence $\{p_{ij}\}_{j=0}^{T-1}$, which can be written as

$$\{p_{ij}\}_{j=0}^{T-1} = \text{patchify}(x_i), \quad (1)$$

where p_{ij} denotes the j -th image patch from sequence and T is the length of sequence.

To alleviate the bias of image patch mix, we conduct uniform sampling strategy on the patches of each image. So each patch of the original image has equal opportunity to be selected in the mixed one. For easier implementation, we replace uniform sampling with random shuffle operation. In this way, we obtain the indices of sampled patches by $\{k(j)\}_{j=0}^{T-1} = \text{shuffle}(\{j\}_{j=0}^{T-1})$. Then the indices are used to obtain the uniformly sampled patches by

$$x_i^{\text{shuffle}} = \text{shuffle}(\{p_{ij}\}_{j=0}^{T-1}) = \{p_{ik(j)}\}_{j=0}^{T-1}. \quad (2)$$

For convenience of patch mix, we divide the shuffled sequence $\{p_{ik(j)}\}_{j=0}^{T-1}$ into M groups, where M denotes the

¹In this paper, “{·}” denotes ordered sequence, instead of unordered set.

number of images for patch mix. Each group is denoted as $G_{im} = \{p_{ik(j)}\}_{j=0}^{(m+1) \cdot S - 1}$, where $S = \lfloor T/M \rfloor$ denotes the number of patches in group G_{im} . For brevity, the shuffled image x_i^{shuffle} can be also presented as

$$x_i^{\text{shuffle}} = \{p_{ik(j)}\}_{j=0}^{T-1} = \{G_{im}\}_{m=0}^{M-1}. \quad (3)$$

To mix patches from different images, we randomly replace the patches of image with the ones of other images in the same position. Specifically, we rearrange the patches from different images to form mixed image by

$$x_i^{\text{smix}} = \{G_{u(i,m)m}\}_{m=0}^{M-1}, \quad (4)$$

where $u(i, m) = (i + m) \bmod N$ conducts patch mix operation in image batch.² For example, as shown in Figure 2, $\{G_{00}, G_{11}, G_{22}\}$, $\{G_{10}, G_{21}, G_{02}\}$ and $\{G_{20}, G_{01}, G_{12}\}$ respectively reform the corresponding mixed image.

To recover the order of patch sequence, we introduce additional unshuffle operation. To this end, we compute the indices for the rearrangement of patch sequence by

$$\{r(j)\}_{j=0}^{T-1} = \text{argsort}(\{k(j)\}_{j=0}^{T-1}), \quad (5)$$

which obtains the element indices by sorting $\{k(j)\}_{j=0}^{T-1}$ in ascending order. Hence, the unshuffle operation can be implemented by $\text{unshuffle}(\cdot) = \text{index}(\cdot, \{r(j)\}_{j=0}^{T-1})$, where $\text{index}(\cdot, \cdot)$ rearranges elements of the given sequence according to indices $\{r(j)\}_{j=0}^{T-1}$.

For convenience, we expand the patch group $G_{u(i,m)m}$ in the shuffled mixed image x_i^{smix} and simplify the Eq. 4 as

$$\begin{aligned} x_i^{\text{smix}} &= \left\{ \{p_{u(i,m)k(j)}\}_{j=m \cdot S}^{(m+1) \cdot S - 1} \right\}_{m=0}^{M-1} \\ &= \{p_{v(i,j)k(j)}\}_{j=0}^{T-1}, \end{aligned} \quad (6)$$

where $v(i, j)$ denotes the index mapping function for simplification. Then, we recover the order of patch indices by $\{j\}_{j=0}^{T-1} = \text{unshuffle}(\{k(j)\}_{j=0}^{T-1})$ and $w(i, j) = \text{unshuffle}(v(i, j))$. Finally, we obtain the mixed image by

$$x_i^{\text{mix}} = \text{unshuffle}(x_i^{\text{smix}}) = \{p_{w(i,j)j}\}_{j=0}^{T-1}. \quad (7)$$

To implement the efficient of tensor operation, we flatten the indices of patch groups in image batch x^{shuffle} as

$$l = \left\{ \{i \cdot M + m\}_{m=0}^{M-1} \right\}_{i=0}^{N-1}. \quad (8)$$

The indices for patch mix operation are rewritten as

$$q = (l + (l \bmod M) \cdot M) \bmod L, \quad (9)$$

²More explanations can be found in the supplementary material.

Algorithm 1 Patch Mix Strategy

Input: Input image batch x ; batch size N ; the image number for mix M .
Output: The mixed image batch x^{mix} ; the mixed target batch y^{mix} .

```

1: function PATCHMIX(Batch  $x$ )  $\triangleright$  PatchMix operation
2:   Obtain shuffled image batch by  $x^{\text{shuffle}}$  by Eq. 2;
3:   Divide into  $M$  groups,  $x^{\text{shuffle}} = \{\{G_{im}\}_{m=0}^{M-1}\}_{i=0}^{N-1}$ ;
4:   Flatten patch groups  $\{\{G_{im}\}_{m=0}^{M-1}\}_{i=0}^{N-1}$  as  $\{G_l\}_{l=0}^{L-1}$ ;
5:   Obtain the indices for patch mix  $q$  by Eq. 9;
6:   Conduct patch mix by  $x^{\text{smix}} = \text{index}(x^{\text{shuffle}}, q)$ ;
7:   Recover the patch order  $x^{\text{mix}} = \text{unshuffle}(x^{\text{smix}})$ ;
8:   Compute mixed targets  $y^{\text{mix}}$  by Eq. 10;
9:   return  $x^{\text{mix}}, y^{\text{mix}}$ 
10: end function
```

where $L = N \cdot M$ denotes the number of patches in image batch. The patch mix operation for the shuffled image can be presented as $x^{\text{smix}} = \text{index}(x^{\text{shuffle}}, q)$ and the unshuffle operation as $x^{\text{mix}} = \text{unshuffle}(x^{\text{smix}})$. The labels for mixed image batch x^{mix} with respect to original image batch x can be obtained by

$$y^{\text{mto}} = \left\{ \{u(i, m)\}_{m=0}^{M-1} \right\}_{i=0}^{N-1}. \quad (10)$$

The labels for contrast between mixed image batch and mixed image batch can be written as

$$y^{\text{mtm}} = \left\{ \{j\}_{j=(i-M+1+N) \bmod N}^{(i+M-1) \bmod N} \right\}_{i=0}^{N-1}. \quad (11)$$

In other words, each image in mixed image batch has $(2M - 1)$ positive pairs in another augmented view. Moreover, due to similarity difference between mixed images, the weights (similarity scores) for each item in label y^{mtm} can be written as

$$\omega^{\text{mtm}} = \left\{ \{1 - |M - j - 1| / M\}_{j=0}^{2M-2} \right\}_{i=0}^{N-1}. \quad (12)$$

The algorithm of patch mix can be summarized as Algorithm 1

3.2. Inter-Instance Similarity Modeling

In this section, we apply the obtained mixed images and soft targets to guide the trained model to capture potential rich similarities among massive image instances.

As shown in Figure 3, the input image batch x are processed by two random data augmentations $\mathcal{T}_1(\cdot)$ and $\mathcal{T}_2(\cdot)$ to obtain transformed ones $x^{(1)} = \mathcal{T}_1(x)$ and $x^{(2)} = \mathcal{T}_2(x)$. To construct similarity-rich positive pairs, patch mix strategy is respectively conducted on image batch $x^{(1)}$ and $x^{(2)}$ to obtain mixed image batch x^{mix1} and x^{mix2} .

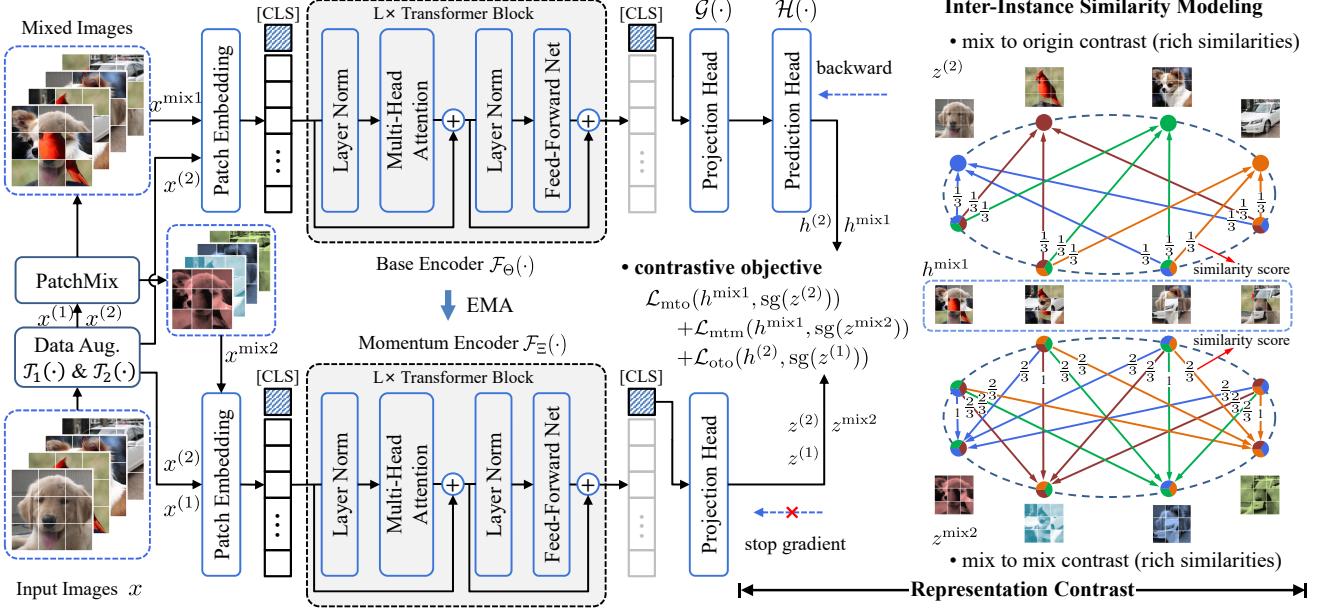


Figure 3: The framework of inter-instance similarity modeling. The patch mix strategy is applied to image batch to construct positive pairs with rich instance similarities among images in the batch. Then ViT model is trained by mix to mix contrastive objective and mix to original one, which guides the model to capture rich inter-instance similarities. To reduce the potential representation gap between mixed images and original images, additional contrastive objective between original images is also applied.

More inter-instance similarities introduced by patch mix is two fold. (1) Different from the plain contrastive framework, mixed image $x_i^{\text{mix}1}$ contains M positive samples in image batch $x^{(2)}$, namely $\{x_{u(i,m)}^{(2)}\}_{m=0}^{M-1}$, each of which is partially similar to mixed image $x_i^{\text{mix}1}$. The same case is also applicable to mixed image batch $x^{\text{mix}2}$. We call this contrastive framework *mix-to-origin contrast*. (2) Mixed image $x_i^{\text{mix}1}$ has $(2M - 1)$ positive samples in mixed image batch $x^{\text{mix}2}$, namely $\{x_j^{\text{mix}2}\}_{j=(i-M+1+N) \bmod N}^{(i+M-1) \bmod N}$, which provides more inter-instance similarities than mix-to-origin contrast. We call this contrastive framework *mix-to-mix contrast*.

To stabilize the training of vision Transformer model $\mathcal{F}_\Theta(\cdot)$, we introduce additional momentum encoder $\mathcal{F}_\Xi(\cdot)$, whose parameters Ξ are updated by the parameters of base encoder $\mathcal{F}_\Theta(\cdot)$ in an EMA (exponential moving average) manner. To improve the transferability of contrastive representations, projection head $\mathcal{G}(\cdot)$ and prediction head $\mathcal{H}(\cdot)$ [8, 14, 7] are also applied.

Combining module $\mathcal{G}(\cdot)$ and $\mathcal{H}(\cdot)$, the image batch $x^{\text{mix}1}$ is fed into model $\mathcal{F}_\Theta(\cdot)$ to obtain predicted representation $h^{\text{mix}1} = \mathcal{H}(\mathcal{G}(\mathcal{F}_\Theta(x^{\text{mix}1})))$. Then, both $x^{\text{mix}2}$ and $x^{(2)}$ are fed into momentum encoder $\mathcal{F}_\Xi(\cdot)$ to obtain projected representation $z^{\text{mix}2} = \mathcal{G}(\mathcal{F}_\Xi(x^{\text{mix}2}))$ and $z^{(2)} = \mathcal{G}(\mathcal{F}_\Xi(x^{(2)}))$. To optimize the model, we introduce mix-to-

origin contrastive objective as

$$\mathcal{L}_{\text{mto}}(h^{\text{mix}1}, z^{(2)}) = -\frac{1}{N \cdot M} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \log \frac{\exp \left(\text{sim} \left(h_i^{\text{mix}1}, z_{y_{ij}^{\text{mto}}}^{(2)} \right) / \tau \right)}{\sum_{t=0}^{N-1} \exp \left(\text{sim} \left(h_i^{\text{mix}1}, z_t^{(2)} \right) / \tau \right)}, \quad (13)$$

where $\text{sim}(h, z) = \frac{h^T z}{\|h\| \|z\|}$ denotes cosine similarity between h and z ; τ denotes temperature coefficient.

Furthermore, mix-to-mix contrastive objective is used to model richer inter-instance similarity by

$$\mathcal{L}_{\text{mtm}}(h^{\text{mix}1}, z^{\text{mix}2}) = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{2M-2} \omega_{ij}^{\text{mtm}} \cdot \log \frac{\exp \left(\text{sim} \left(h_i^{\text{mix}1}, z_{y_{ij}^{\text{mtm}}}^{\text{mix}2} \right) / \tau \right)}{\sum_{t=0}^{N-1} \exp \left(\text{sim} \left(h_i^{\text{mix}1}, z_t^{\text{mix}2} \right) / \tau \right)}. \quad (14)$$

Since there is potential domain gap between mixed images and original images, we further introduce plain contrastive objective between original images, which can be presented as

$$\mathcal{L}_{\text{oto}}(h^{(2)}, z^{(1)}) = -\frac{1}{N} \sum_{i=0}^{N-1} \log \frac{\exp \left(\text{sim} \left(h_i^{(2)}, z_i^{(1)} \right) / \tau \right)}{\sum_{t=0}^{N-1} \exp \left(\text{sim} \left(h_i^{(2)}, z_t^{(1)} \right) / \tau \right)}, \quad (15)$$

Algorithm 2 Inter-Instance Similarity Modeling

Input: Training dataset \mathcal{X} ; ViT model \mathcal{F}_* ; projection head \mathcal{G} ; prediction head \mathcal{H} ; momentum coefficient μ .
Output: The parameters Θ of ViT model \mathcal{F}_* .

- 1: Initialize the parameters of \mathcal{F}_* , \mathcal{G} and \mathcal{H} ;
- 2: Initialize the momentum parameters $\Xi \leftarrow \Theta$;
- 3: **for** batch x in dataset \mathcal{X} **do**
- 4: Augment image batch $x^{(1)}, x^{(2)} = \mathcal{T}_1(x), \mathcal{T}_2(x)$;
- 5: Obtain $x^{\text{mix}1}, y^{\text{mto}1} = \text{PATCHMIX}(x^{(1)})$;
- 6: Obtain $x^{\text{mix}2}, y^{\text{mto}2} = \text{PATCHMIX}(x^{(2)})$;
- 7: Obtain label y^{mtm} and weight ω^{mtm} by Eq. 11&12;
- 8: $h^{\text{mix}1}, h^{(2)} = \mathcal{H}(\mathcal{G}(\mathcal{F}_\Theta(x^{\text{mix}1}))), \mathcal{H}(\mathcal{G}(\mathcal{F}_\Theta(x^{(2)})))$;
- 9: $z^{(1)}, z^{(2)}, z^{\text{mix}2} = \mathcal{G}(\mathcal{F}_\Xi(x^{(1)})), \mathcal{G}(\mathcal{F}_\Xi(x^{(2)})), \mathcal{G}(\mathcal{F}_\Xi(x^{\text{mix}2}))$;
- 10: Compute loss $\mathcal{L}_{\text{total}}$ and update the parameters Θ ;
- 11: Update momentum params $\Xi \leftarrow \mu \cdot \Xi + (1 - \mu) \cdot \Theta$;
- 12: **end for**

where $h^{(2)} = \mathcal{H}(\mathcal{G}(\mathcal{F}_\Theta(x^{(2)})))$ and $z^{(1)} = \mathcal{G}(\mathcal{F}_\Xi(x^{(1)}))$. To further alleviate representation degeneration, stop gradient operation $\text{sg}(\cdot)$ is also applied to z and the total contrastive objective can be presented as

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{mto}}(h^{\text{mix}1}, \text{sg}(z^{(2)})) + \mathcal{L}_{\text{mtm}}(h^{\text{mix}1}, \text{sg}(z^{\text{mix}2})) \\ &\quad + \mathcal{L}_{\text{oto}}(h^{(2)}, \text{sg}(z^{(1)})). \end{aligned} \quad (16)$$

Overall, our approach can be summarized as Algorithm 2.

4. Experiments

In this section, we first introduce the experimental settings, including datasets, network architecture, optimization and evaluation. Then, we report the main experimental results of our proposed method on ImageNet-1K, CIFAR10 and CIFAR100 to evaluate its effectiveness. Finally, ablation study is conducted in detail to validate the effectiveness of each component in our approach.

4.1. Experimental Settings

4.1.1 Datasets

To validate the effectiveness of our proposed method, we evaluate the performance on small-scale datasets: CIFAR10 (60,000 colour images with 10 classes including 50,000 images as training set and 10,000 images as validation set) and CIFAR100 [20] (60,000 colour images with 100 classes including 50,000 images as training set and 10,000 images as validation set) and large-scale dataset, ImageNet-1K [27] (including 1.2 million images as training set and 50,000 images as validation set, from 1,000 categories). For CIFAR and ImageNet-1K, the size of image is set to 32×32 and 224×224 , respectively. Additionally, detailed data augmentation techniques for contrastive learning are depicted in the supplementary material.

To validate the transferability of unsupervised representation, we also finetune the ImageNet-1K pretrained model on COCO [23], which contains 118,000 images as training set and 5,000 images as validation set. The images are padded as 1024×1024 size during training and testing.

4.1.2 Networks and Optimization

We adopt basic ViT [13] as backbone. The patch sizes for patchify are respectively 2×2 and 16×16 for CIFAR and ImageNet-1K. Specifically, we evaluate our method on ViT-T/2³, ViT-S/2 and ViT-B/2 for CIFAR, then ViT-S/16 and ViT-B/16 for ImageNet-1K. Following MoCo v3 [8], additional projection and prediction module are applied.

For ImageNet-1K, we pretrain the ViT-S/16 model with AdamW optimizer with learning rate 2×10^{-3} for 300 epochs and 800 epochs. And we pretrain ViT-B/16 model with learning rate 3×10^{-3} for 300 epochs. For the pre-training of all models, the batch size is 1024. We conduct linear warmup on learning rate for 10 epochs and then follow a cosine learning rate decay schedule for the rest 260 epochs. The weight decay follows a cosine schedule from 0.04 to 0.4. And momentum coefficient μ follows a cosine schedule from 0.996 to 1. By default, the number of images for PatchMix is set to 3. The temperature τ is set to 0.2.

For CIFAR datasets, the model is pretrained for 800 epochs, where 100 epochs for warmup. The learning rates for ViT-S/2 and ViT-B/2 model are 1×10^{-3} and 1.5×10^{-3} . The batch size is 512. Other hyperparameters are consistent with ImageNet-1K. For ViT-T/2, the initial learning rate is set to 4×10^{-3} and the weight decay follows a cosine schedule from 0.02 to 0.2. Other hyperparameters are consistent with ImageNet-1K.

4.1.3 Evaluation

To evaluate the representation performance, we adopt three common evaluation protocols: finetune evaluation, linear evaluation and k-NN (k-nearest neighbor) evaluation. For finetune evaluation, we initialize the model with the pre-trained weights and then adapt them to downstream tasks by finetuning. This protocol can be flexibly applicable to various downstream tasks by transfer learning, such as objection detection, semantic segmentation and so on. For linear evaluation, we fix the pretrained model and feed the predicted representation to a linear classifier for image recognition. This protocol evaluates the representation quality without changing the original representations, which avoids disturbance due to additional training. However, this protocol is not suitable for the evaluation of non-linear representations [1, 11, 15]. For k-NN evaluation, it doesn't require

³The details are clarified in the supplementary material

Method	Backbone	#Views	Batch size	#Epochs	Finetune (%)	Linear (%)	kNN (%)
SimCLR [4]	ViT-S/16	4(224)	4096	300	NA	69.0	NA
SwAV [2]	ViT-S/16	4(224)	4096	300	NA	67.1	NA
BYOL [14]	ViT-S/16	4(224)	4096	300	NA	71.0	NA
MoCo v3 [8]	ViT-S/16	4(224)	4096	300	81.4	72.5	67.8
DINO [3]	ViT-S/16	4(224)	1024	300	NA	72.5	67.9
DINO* [3]	ViT-S/16	4(224)+8(96)	1024	300	NA	76.1	72.8
iBOT [38]	ViT-S/16	4(224)	1024	800	NA	76.2	72.4
SDMP [26]	ViT-S/16	4(224)	1024	300	79.1	73.8	NA
PatchMix (ours)	ViT-S/16	4(224)	1024	300	82.8	77.4	73.3
PatchMix (ours)	ViT-S/16	4(224)	1024	800	83.4	77.9	74.3
SimCLR [4]	ViT-B/16	4(224)	4096	300	NA	73.9	NA
SwAV [2]	ViT-B/16	4(224)	4096	300	NA	71.6	NA
BYOL [14]	ViT-B/16	4(224)	4096	300	NA	73.9	NA
MoCo v3 [8]	ViT-B/16	4(224)	4096	300	83.2	76.5	70.7
DINO [3]	ViT-B/16	4(224)	1024	400	NA	72.8	68.9
DINO* [3]	ViT-B/16	4(224)+8(96)	1024	400	82.3	78.2	76.1
iBOT [38]	ViT-B/16	4(224)	1024	400	NA	76.0	71.2
SDMP [26]	ViT-B/16	4(224)	1024	300	NA	77.2	NA
PatchMix (ours)	ViT-B/16	4(224)	1024	300	84.1	80.2	76.2

Table 2: Performance comparison on ImageNet-1K dataset under finetune, linear and kNN evaluation protocols. “NA” denotes that the result is not available in original paper. “*” denotes the model pretrained with multi-crop strategy [2]. “4(224)+8(96)” denotes 4 images with size 224×224 and 8 images with size 96×96 . “# Epochs” denotes the number of pretraining epochs. Both finetune accuracy and linear accuracy are evaluated by finetuning the pretrained model for 100 epochs.

any learnable parameters and provides more stable evaluation results. However, it is also limited to simple representation evaluation. Hence, we report all three protocols in our experiments for a more comprehensive evaluation. For the fairness of comparative experiments, all pretrained model are finetuned for 100 epochs under both finetune and linear evaluation protocols.

4.2. Experimental Results

4.2.1 Image Classification on ImageNet-1K

In this experiment, we pretrain the model on the training set of ImageNet-1K and then evaluate the pretrained model on the validation set of ImageNet-1K under the finetune, linear and kNN evaluation protocols.

As shown in Table 2, our proposed PatchMix achieves 82.8%, 77.4% and 73.3% accuracy with ViT-S/16 pretrained for 300 epochs under finetune, linear and kNN evaluation protocol, respectively. It significantly outperforms DINO counterpart by a large margin, even DINO with multi-crop strategy. When pretrained for 800 epochs, our method with ViT-S/16 respectively achieves 83.4%, 77.9% and 74.3% accuracy under finetune, linear and kNN evaluation protocol, substantially superior to iBOT counterpart. For ViT-B/16 pretrained for 300 epochs, our method respec-

tively reaches 84.1%, 80.2% and 76.2% under the finetune, linear and kNN evaluation protocol, which surpasses other self-supervised methods, including DINO pretrained for 400 epochs with multi-crop strategy. Especially under linear evaluation protocol, the model pretrained by our method for 300 epochs outperforms the previous state-of-the-art: SDMP, by 3.0% accuracy using ViT-B/16 backbone. The above experimental results strongly support the effectiveness of our proposed PatchMix on unsupervised representation learning. We believe that inter-instance similarity modeling introduced by our proposed PatchMix can effectively improve the representation generalization among different images.

4.2.2 Image Classification on CIFAR10 and CIFAR100

In this experiment, we respectively pretrain the model on the training set of CIFAR10 and CIFAR100, and then respectively evaluate the corresponding pretrained model on the validation set of CIFAR10 and CIFAR100, under the finetune, linear and kNN evaluation protocols.

As shown in Table 3, under all three evaluation protocols, our approach consistently outperforms the previous contrastive methods by a significant margin on both CIFAR10 and CIFAR100. Specially under kNN evaluation proto-

Method	Backbone	#Epochs	Batch	#FLOPs	CIFAR10			CIFAR100		
					Tune	Linear	kNN	Tune	Linear	kNN
MoCo v3 [8]	ViT-T/2	800	512	44.4G	95.5	89.8	88.1	78.6	67.1	58.9
DINO [3]	ViT-T/2	800	512	80.2G	93.5	88.4	87.3	75.4	61.8	57.4
iBOT [38]	ViT-T/2	800	512	142.8G	96.3	93.0	92.3	82.0	63.6	58.2
SDMP [26]	ViT-T/2	800	512	50.0G	96.4	93.2	92.2	82.3	72.4	65.7
PatchMix (ours)	ViT-T/2	800	512	50.0G	97.5	94.4	92.9	84.9	74.7	68.8
PatchMix (ours)	ViT-T/2	1600	512	50.0G	97.8	94.8	93.5	84.9	76.0	70.1
MoCo v3 [8]	ViT-S/2	800	512	175.6G	95.8	90.2	89.1	78.8	66.4	62.3
DINO [3]	ViT-S/2	800	512	311.4G	96.3	92.6	91.8	75.9	63.8	60.6
iBOT [38]	ViT-S/2	800	512	571.4G	97.0	94.8	93.1	82.8	67.8	64.2
SDMP [26]	ViT-S/2	800	512	197.6G	96.9	94.2	92.1	85.0	77.1	66.7
PatchMix (ours)	ViT-S/2	800	512	197.6G	98.1	96.0	94.6	86.0	78.7	75.4
PatchMix (ours)	ViT-S/2	1600	512	197.6G	98.2	96.4	95.1	86.1	78.9	74.6
MoCo v3 [8]	ViT-B/2	800	512	700.0G	96.1	90.9	89.5	79.5	67.3	63.4
DINO [3]	ViT-B/2	800	512	1237.2G	96.8	92.8	92.1	76.4	65.6	62.7
iBOT [38]	ViT-B/2	800	512	2014.1G	97.3	94.9	93.2	83.3	68.7	65.5
SDMP [26]	ViT-B/2	800	512	787.3G	97.2	94.3	92.4	85.1	77.3	68.3
PatchMix (ours)	ViT-B/2	800	512	787.3G	98.3	96.6	95.8	86.0	79.7	75.7

Table 3: Performance (%) comparison on CIFAR datasets using finetune, linear and kNN evaluation protocols, respectively. “Tune” denotes classification accuracy using finetune protocol. “# FLOPs” denotes the number of floating-point operations per iteration during training, where batch size is 2. “# Epochs” denotes the number of pretraining epochs. Both finetune accuracy and linear accuracy are evaluated by finetuning the pretrained model for 100 epochs.

col, our PatchMix achieves significantly large improvement over the existing contrastive methods, e.g. 8.7% and 5.1% kNN accuracy improvement on CIFAR100 using ViT-S/2 and ViT-B/2, respectively. The reason is analyzed as follows. Since kNN protocol doesn’t introduce any additional learnable parameters, the evaluation results reflect, to some degree, the intrinsic quality of representations. The superiority of our PatchMix, better inter-instance similarity modeling, is fully presented. Moreover, in spite of data-hungry ViT architecture, our powerful capacity for inter-instance similarity modeling significantly reduces the risk of overfitting and achieves excellent performance on small datasets, CIFAR10 and CIFAR100.

4.3. Transfer Learning on Downstream Tasks

To further evaluate the transferability of our proposed PatchMix, we conduct transfer learning experiments on downstream tasks: object detection and instance segmentation on COCO [23] by Mask RCNN [17] with FPN [22] as MAE [15]. The results are reported in Table 4. With ViT-B/16 backbone pretrained on ImageNet-1K for 300 epochs, our PatchMix achieves 51.9 on AP^{box} and 46.1 on AP^{mask}, surpassing the previous state-of-the-art iBOT by 0.7 on AP^{box} and 1.9 on AP^{mask}. Meanwhile, compared to the previous leading self-supervised learning methods, MAE and SIM, our method requires significantly fewer pretraining epochs but achieves significantly transfer learning perfor-

mance on object detection and instance segmentation task. The experimental results demonstrate that our PatchMix can effectively model image semantics and structure. We believe that our method can effectively balance the global and local features under the guidance of patch-mixed images.

Method	#Pre-Epochs	AP ^{box}	AP ^{mask}
MoCo v3 [8]	300	47.9	42.7
MAE [15]	1600	50.3	44.9
CAE [5]	300	48	42.3
PeCo [12]	300	43.9	39.8
SIM [28]	1600	49.1	43.8
PatchMix (ours)	300	51.9	46.1

Table 4: Transfer learning performance on object detection and instance segmentation task using Mask RCNN with ViT-B/16 backbone (pretrained on ImageNet-1K for 300 epochs) on COCO dataset. All pretrained models are fine-tuned on COCO dataset for 100 epochs.

4.4. Ablation Study

To evaluate the effectiveness of each module in our method, we implement ablation study on CIFAR10 and CIFAR100 datasets using ViT-S/2 backbone. In this section, we mainly analyze the effect of our proposed patch mix, the image number for patch mix, the loss function items and the

number of pretraining epochs.

4.4.1 The Effect of Patch Mix

To validate the effectiveness of our proposed patch mix strategy, we compare it with popular image mix methods applied on contrastive learning. As shown in Table 5, image mix strategies, including CutMix+Mixup [37], RegionSwap [35] and ResizeMix [26], indeed improve the representation quality of contrastive learning. Notably, our proposed PatchMix achieves 96.0% linear accuracy and 94.6% kNN accuracy on CIFAR10, 78.7% linear accuracy and 75.4 kNN accuracy on CIFAR100, consistently outperforms other image mix strategies by a significantly large margin, and effectively boosts the performance of contrastive representations. We owe the improvement to more complicated inter-instance similarity modeling capacity introduced by our PatchMix strategy.

Method	CIFAR10		CIFAR100	
	Linear	kNN	Linear	kNN
Baseline	89.8	88.2	66.0	61.2
CutMix+Mixup [37]	90.6	88.7	66.8	62.3
RegionSwap [35]	92.1	91.3	70.4	64.2
ResizeMix [26]	94.2	92.1	77.1	66.7
PatchMix (ours)	96.0	94.6	78.7	75.4

Table 5: The effect of different mix strategies on CIFAR10 and CIFAR100 datasets. The performance (%) is measured by linear and kNN evaluation protocols.

4.4.2 The Effect of Image Number for Patch Mix

To determine the best image number for patch mix, we evaluate the representation performance on CIFAR datasets under different image numbers for patch mix. In Table 6, our method achieves the best performance under both linear and kNN evaluation protocols when the number of images for patch mix is 3. More or fewer image number for patch mix doesn't introduce additional performance gain. We explain this result as follows. Due to more images for patch mix, the number of patches from the same one image is significantly reduced. Hence, the informative patches in mixed images also decrease, making the corresponding targets not so accurate to boost performance. Meanwhile, fewer images for patch mix can't well establish inter-instance similarities among natural images.

4.4.3 The Effect of Loss Function

To investigate the effectiveness of loss items in our contrastive objective, we evaluate the performance as single loss item or their combination is applied. The results are

#Mix (M)	CIFAR10		CIFAR100	
	Linear	kNN	Linear	kNN
1	89.8	88.2	66.0	61.2
2	95.6	94.3	77.6	71.1
3	96.0	94.6	78.7	75.4
4	95.7	94.3	75.2	70.9

Table 6: The effect of image number for PatchMix on CIFAR10 and CIFAR100 datasets. The performance (%) is measured by linear and kNN evaluation protocols.

presented in Table 7. First, we can observe that simply utilizing single loss item can not achieve excellent performance. Second, when original images to original ones contrast item \mathcal{L}_{oto} and mixed images to original ones contrast item \mathcal{L}_{mto} are jointly applied, the representation performance is significantly improved, e.g. from 59.7% to 69.2% on CIFAR100 under the kNN evaluation protocol. Third, combining \mathcal{L}_{oto} , \mathcal{L}_{mto} and \mathcal{L}_{mtm} achieves the best performance, surpassing the baseline with only \mathcal{L}_{oto} by 4.9% linear accuracy and 7.6% kNN accuracy improvement on CIFAR10, 9.0% linear accuracy and 15.7% kNN accuracy improvement on CIFAR100. We analyze the reason as follows. The item \mathcal{L}_{oto} , \mathcal{L}_{mto} and \mathcal{L}_{mtm} respectively establishes image relations between natural images and natural ones, mixed images and natural ones, mixed images and mixed one. As all above items work, the trained model can well balance the capacity of natural image representations and multi-image relation modeling to achieve high-quality unsupervised representation performance.

\mathcal{L}_{oto}	\mathcal{L}_{mto}	\mathcal{L}_{mtm}	CIFAR10		CIFAR100	
			Linear	kNN	Linear	kNN
✓	-	-	91.1	87.0	69.7	59.7
-	✓	-	NA	NA	61.3	57.2
-	-	✓	NA	NA	52.6	47.4
✓	✓	-	95.4	93.8	77.9	69.2
✓	✓	✓	96.0	94.6	78.7	75.4

Table 7: The effect of loss items on CIFAR10 and CIFAR100 datasets. The performance (%) is measured by linear and kNN evaluation protocols. “NA” indicates that the model fails to converge.

4.4.4 The Effect of Pretraining Epochs

To validate the effect of training epochs during self-supervised pretraining, we conduct experiments with different numbers of pretraining epochs on CIFAR10 and CIFAR100, respectively. As shown in Figure 5, we report performance on both ViT-T/2 and ViT-S/2 under all three evaluation protocols, finetune evaluation protocol, linear

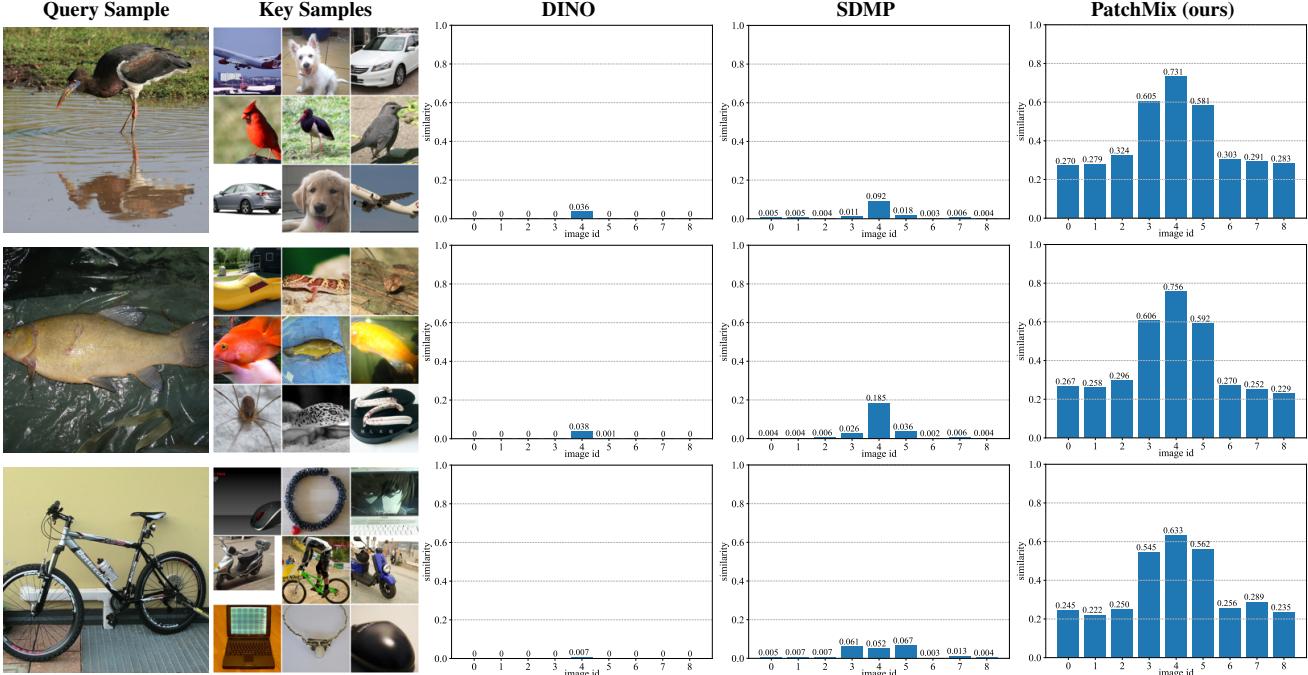


Figure 4: The visualization of inter-instance similarities using ViT-S/16 pretrained on ImageNet-1K dataset for 300 epochs. The query sample and the image with id 4 in key samples are from the same category. The images with id 3 and 5 come from category similar to query sample. We compare inter-instance similarities constructed by our method with two representative methods, DINO and SDMP.

evaluation protocol and kNN evaluation protocol. We can observe that the proposed method achieves better performance with longer training schedule, especially on CIFAR100. Meanwhile, the performance improvement progressively reaches to saturation with the increment of pre-training epochs. Training the model using our proposed PatchMix for 800 epochs can achieve the balance between performance and time consumption.

4.5. Inter-Instance Similarity Visualization

To evaluate inter-instance similarities constructed by our method, we visualize the similarity distribution among multiple images. Specifically, we follow the similarity metric adopted in kNN evaluation protocol and normalize this metric as $\frac{\exp(\text{sim}(a,b)/\tau')}{\exp(1/\tau')} \in (0, 1]$, where the temperature coefficient $\tau' = 0.07$.

In Figure 4, we measure the similarities between query sample and key samples using the outputs of backbone encoder, ViT-S/16 pretrained on ImageNet-1K for 300 epochs. For DINO and SDMP, the distribution of similarities to key samples is significantly sharp and the similarity peak concentrates on the sample with the same category as the query image. SDMP achieves higher similarity on the category-related samples than DINO, but also only a slight one. In contrast, our proposed PatchMix establishes significantly

rich similarity relations among natural images, not only the corresponding category but also the related categories. This experimental result demonstrates that our method can effectively improve the generalization of unsupervised representations. More results can be found in the supplementary material.

5. Discussion

In this section, we analyze the properties of the proposed PatchMix as follows.

- The proposed PatchMix constructs hybrid image, which includes parts from several image instances, to simulate rich inter-instance similarities among natural images. The model pretrained by PatchMix can effectively capture the inter-instance similarities, especially for the images from the related categories as Figure 4, thus improving the generalization ability of representations among different instances and significantly outperforming the previous unsupervised learning methods as Table 2.
- Our PatchMix also presents excellent performance on small-scale datasets, such as CIFAR10 and CIFAR100, as Table 3. Despite data-hungry architecture of ViT, the model pretrained by PatchMix significantly al-

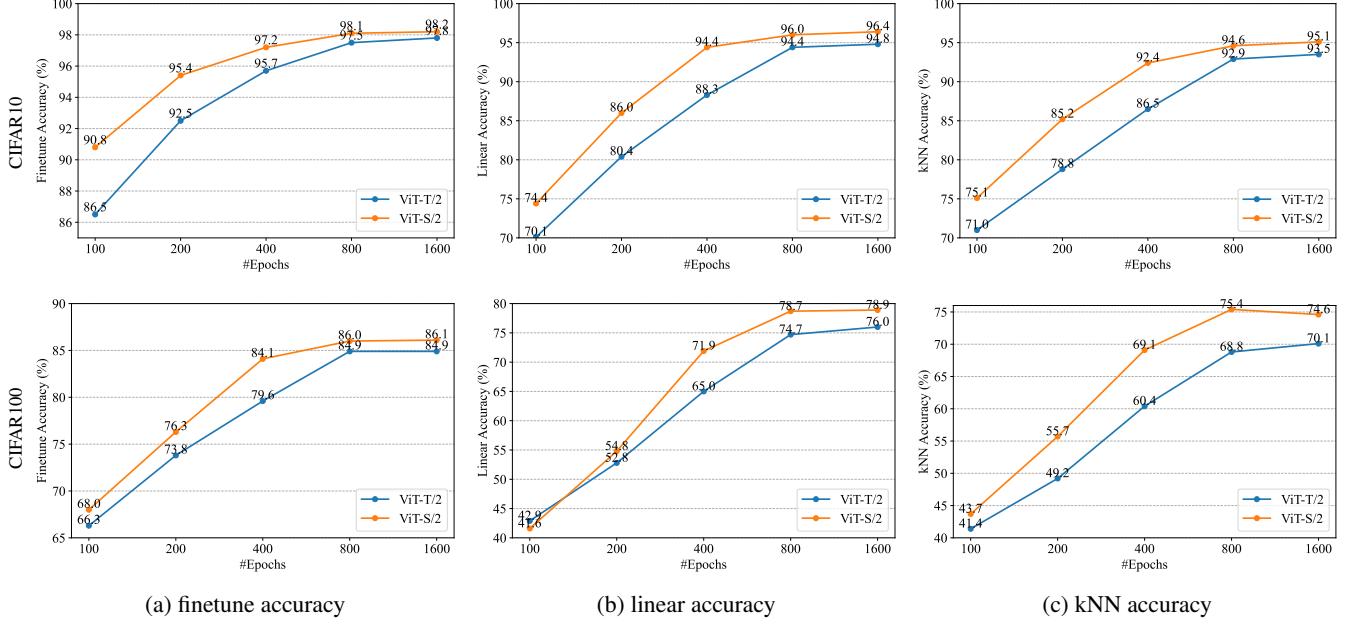


Figure 5: Pretraining with different numbers of epochs on CIFAR10 and CIFAR100 dataset. The accuracies are evaluated under the finetune protocol, linear protocol and kNN protocol, respectively.

leviates potential overfitting and representation degeneration without additional training data, such as ImageNet-1K.

- Due to the similarities introduced by local patches, our PatchMix encourages the model to capture local structures of images. Hence, the pretrained model presents the excellent transferability on downstream tasks, which are sensitive to local structures, such as object detection and instance segmentation in Table 4.

6. Conclusion

In this paper, we address monotonous similarity issue suffered by contrastive learning methods. To this end, we propose a novel image mix strategy, PatchMix, which mixes multiple images in patch level. The mixed image contains massive local components from multiple images and efficiently simulates rich similarities among natural images in an unsupervised manner. To model rich inter-instance similarities among images, the contrasts between mixed images and original ones, mixed images to mixed ones, and original images to original ones are conducted to optimize the ViT model. Experimental results illustrate that our method significantly improves the quality of unsupervised representations, achieving state-of-the-art performance on image classification task of ImageNet-1K, CIFAR10 and CIFAR100 datasets, object detection and instance segmentation tasks of COCO dataset. Extensive experiments support that our proposed PatchMix can effectively model the rich similiari-

ties among natural images and improve the generalization of unsupervised representations on various downstream tasks.

In future work, we plan to explore more general and accurate inter-instance modeling pretext task to further improve the representation quality of contrastive learning.

References

- [1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022. [2](#), [6](#)
- [2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9912–9924, 2020. [2](#), [7](#)
- [3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. [7](#), [8](#)
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020. [1](#), [2](#), [7](#)
- [5] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. [8](#)

- [6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2
- [7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. 2, 5
- [8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9640–9649, 2021. 2, 5, 6, 7, 8
- [9] Hyeong Kyu Choi, Joonmyung Choi, and Hyunwoo J Kim. Tokenmixup: Efficient attention-guided token-level data augmentation for transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2
- [10] Ali Dabouei, Sobhan Soleymani, Fariborz Taherkhani, and Nasser M Nasrabadi. Supermix: Supervising the mixing data augmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13794–13803, 2021. 2
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (ACL)*, pages 1877–1901, 2019. 6
- [12] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 8
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 6
- [14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21271–21284, 2020. 2, 5, 7
- [15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 2, 6, 8
- [16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 1, 2
- [17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2961–2969, 2017. 8
- [18] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21798–21809, 2020. 1, 2
- [19] Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. In *International Conference on Learning Representations (ICLR)*, 2021. 2
- [20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6
- [21] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-mix: A domain-agnostic strategy for contrastive representation learning. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2
- [22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017. 8
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 6, 8
- [24] Jihao Liu, Boxiao Liu, Hang Zhou, Hongsheng Li, and Yu Liu. Tokenmix: Rethinking image mixing for data augmentation in vision transformers. In *European Conference on Computer Vision (ECCV)*, pages 455–471. Springer, 2022. 2
- [25] Jie Qin, Jiemin Fang, Qian Zhang, Wenyu Liu, Xingang Wang, and Xinggang Wang. Resizemix: Mixing data with preserved object information and true labels. *arXiv preprint arXiv:2012.11101*, 2020. 3
- [26] Sucheng Ren, Huiyu Wang, Zhengqi Gao, Shengfeng He, Alan Yuille, Yuyin Zhou, and Cihang Xie. A simple data mixing prior for improving self-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14595–14604, 2022. 2, 7, 8, 9
- [27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6
- [28] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022. 8
- [29] AFM Uddin, Mst Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae, et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. 2021. 2
- [30] Shashanka Venkataramanan, Ewa Kijak, Laurent Amsaleg, and Yannis Avrithis. Alignmixup: Improving representations by interpolating aligned features. In *Proceedings of the*

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.* 2
- [31] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najaifi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning (ICML)*, pages 6438–6447. PMLR, 2019. 2
 - [32] Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc Le. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning (ICML)*, pages 10530–10541. PMLR, 2021. 1, 2
 - [33] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018. 1, 2
 - [34] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 2
 - [35] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Regioncl: Can simple region swapping contribute to contrastive learning? In *European Conference on Computer Vision (ECCV)*, 2022. 3, 9
 - [36] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6023–6032, 2019. 1, 2
 - [37] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2017. 1, 2, 9
 - [38] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022. 7, 8

A. Patch Mix

In this section, we supplement more explanations about patch mix strategy. As shown in Figure 2 (a), we mix the given image $x_i^{\text{shuffle}} = \{G_{im}\}_{m=0}^{M-1}$ by rearranging the indices of the first dimension $u(i, m) = (i + m) \bmod N$ as $x_i^{\text{smix}} = \{G_{u(i,m)m}\}_{m=0}^{M-1}$. When the index $(i + m)$ overflows the boundary along batch dimension, we conduct modular operation on index $(i + m)$ as $((i + m) \bmod N)$ to cyclically mix the image patch group G_{im} in the image batch x^{shuffle} .

To fit our patch mix strategy into efficient implement of tensor operation, we flatten the indices of patch groups in image batch $x^{\text{shuffle}} = \left\{ \{G_{im}\}_{m=0}^{M-1} \right\}_{i=0}^{N-1}$ as $l = \left\{ \{i \cdot M + m\}_{m=0}^{M-1} \right\}_{i=0}^{N-1} = \{i\}_{i=0}^{M \cdot N - 1}$ as Figure 2 (b). To

mix the image patch group G_{im} , a similar index transformation on 1-d indices as $u(i, m)$ is implemented by

$$q = (l + (l \bmod M) \cdot M) \bmod L, \quad (17)$$

where $L = N \cdot M$ denotes the number of image patch groups in image batch. This index transformation can be efficiently implemented by tensor operation. Our proposed patch mix strategy can be efficiently implemented by $x^{\text{smix}} = \text{index}(x^{\text{shuffle}}, q)$.

We present the computataion of mix-to-mix targets and mix-to-mix similarity scores in Figure 7. For example, we analyze the mix-to-mix targets for the second sample in view 1, including the patches from image (1, 2, 3). The mixed samples from view 2, which contains the patches from image (1, 2, 3), includes 5 samples (8, 0, 1), (0, 1, 2), (1, 2, 3), (2, 3, 4) and (3, 4, 5). Hence, the mix-to-mix target between view 1 and view 2 is $((0 - 2, 0 - 1, 0, 0 + 1, 0 + 2) + 9) \bmod 9 = (7, 8, 0, 1, 2)$. The general formula for mix-to-mix targets can be written as

$$y^{\text{mtm}} = \left\{ \{j\}_{j=(i-M+1+N) \bmod N}^{(i+M-1) \bmod N} \right\}_{i=0}^{N-1}, \quad (18)$$

which contains $(2M - 1)$ positive samples for view 1. As shown in Figure 7, the number of overlapping patches between positive samples and the sample in view 1 are different, $(\frac{1}{3}, \frac{2}{3}, 1, \frac{2}{3}, \frac{1}{3})$ for image (7, 8, 0, 1, 2), respectively. More general formula for mix-to-mix scores can be obtained by

$$\omega^{\text{mtm}} = \left\{ \left\{ 1 - \frac{|M - j - 1|}{M} \right\}_{j=0}^{2M-2} \right\}_{i=0}^{N-1}, \quad (19)$$

which provides more inter-instance similarities for contrastive learning.

B. Network Structures

As shown in Table 8, we give the details of backbones used in our experiments. For ImageNet-1K with input size 224×224 , we adopt standard vision transformer architectures, ViT-Small and ViT-Base, where the patch size for tokenization is 16×16 . For CIFAR10 and CIFAR100 with input size 32×32 , we modify the patch size of standard vision transformer architectures from 16×16 to 2×2 , to adapt the small input images. We also introduce a more lightweight vision transformer architecture, ViT-Tiny, which only has half head number and half token dimension of ViT-Small.

As shown in Table 9, we further present the structure of projection and prediction head adopted during self-supervised pretraining. For ImageNet-1K, there are 3 linear layers in projection head and 2 linear layers in prediction heads. The first two linear layers are followed by batch normalization and rectify linear unit in turn, and the output

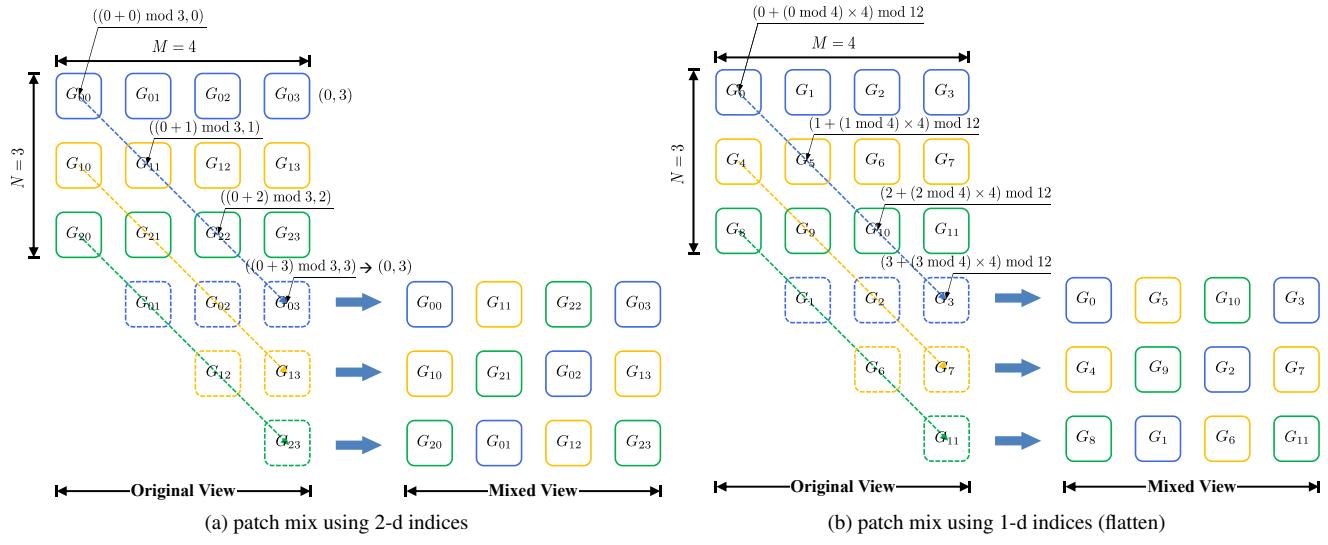


Figure 6: The illustration for patch mix strategy, where the mix number $M = 4$ and the batch size $N = 3$.

Mix View 1	Mix View 2	Mix-to-Mix Contrast Targets	Mix-to-Mix Similarity Scores
(0, 1, 2)	(8, 0, 1)	((0-2, 0-1, 0, 0+1, 0+2) + 9) mod 9	(1/3, 2/3, 1, 2/3, 1/3)
(1, 2, 3)	(0, 1, 2)	((1-2, 1-1, 1, 1+1, 1+2) + 9) mod 9	(1/3, 2/3, 1, 2/3, 1/3)
(2, 3, 4)	(1, 2, 3)	((2-2, 2-1, 2, 2+1, 2+2) + 9) mod 9	(1/3, 2/3, 1, 2/3, 1/3)
(3, 4, 5)	(2, 3, 4)	((3-2, 3-1, 3, 3+1, 3+2) + 9) mod 9	(1/3, 2/3, 1, 2/3, 1/3)
(4, 5, 6)	(3, 4, 5)	((4-2, 4-1, 4, 4+1, 4+2) + 9) mod 9	(1/3, 2/3, 1, 2/3, 1/3)
(5, 6, 7)	(4, 5, 6)	((5-2, 5-1, 5, 5+1, 5+2) + 9) mod 9	(1/3, 2/3, 1, 2/3, 1/3)
(6, 7, 8)	(5, 6, 7)	((6-2, 6-1, 6, 6+1, 6+2) + 9) mod 9	(1/3, 2/3, 1, 2/3, 1/3)
(7, 8, 0)	(6, 7, 8)	((7-2, 7-1, 7, 7+1, 7+2) + 9) mod 9	(1/3, 2/3, 1, 2/3, 1/3)
(8, 0, 1)	(7, 8, 0)	((8-2, 8-1, 8, 8+1, 8+2) + 9) mod 9	(1/3, 2/3, 1, 2/3, 1/3)
	(8, 0, 1)		

Figure 7: The illustration of mix-to-mix targets and mix-to-mix similarity scores, where the mix number $M = 3$ and the batch size $N = 9$.

sizes of them are both 4096. The last linear of both heads are followed by only batch normalization, and output sizes of them are 256. For CIFAR10 and CIFAR100, the configurations of projection and prediction head are consistent with the ones of ImageNet-1K.

C. Data Augmentations

As shown in Table 10, we describe the parameters of data augmentations used during self-supervised pretraining. For ImageNet-1K, 6 data augmentation techniques are applied to the input images, including random crop and resize, horizontal flip, color jittering, gray scale, Gaussian blurring, as well as solarization. The same data augmentation techniques are applied to CIFAR10 and CIFAR100, except area of the crop.

D. Inter-Instance Similarity Visualization

To further validate the effectiveness of our proposed PatchMix on inter-instance similarity modeling, we provide more visualization results the similarity distribution among multiple images. The results are presented in Figure 8, Figure 9 and Figure 10. For DINO, only the samples with the same category present instance similarity with the corresponding query samples. In some cases of Figure 10, even the samples with the same category only have negligible similarities with the corresponding query samples. For SDMP, it achieves richer inter-instance similarities among images, especially the samples from the similar categories. However, there are also some cases, such as the second and third row in Figure 4, where the samples with a similar category but significantly low similarity score. In contrast,

Dataset	Network	Patch Size	#Blocks	#Heads	Token Dim	#Params (M)
ImageNet-1K	ViT-Small	16	12	6	384	21.6
	ViT-Base	16	12	12	768	85.7
CIFAR	ViT-Tiny	2	12	3	192	5.4
	ViT-Small	2	12	6	384	21.3
	ViT-Base	2	12	12	768	85.1

Table 8: The structure of visual transformer backbones. “#Blocks” denotes the number of standard transformer blocks in backbone. “Token Dim” denotes the dimension of visual token vector.

Dataset	Layer	Projection Head	Prediction Head
ImageNet-1K	1	Linear (4096) + BN + ReLU	Linear (4096) + BN + ReLU
	2	Linear (4096) + BN + ReLU	Linear (256) + BN*
	3	Linear (256) + BN*	
CIFAR	1	Linear (4096) + BN + ReLU	Linear (4096) + BN + ReLU
	2	Linear (4096) + BN + ReLU	Linear (256) + BN *
	3	Linear (256) + BN*	

Table 9: The structure of projection and prediction heads. “Linear (m)” denotes linear layer with output size m . “BN” and “ReLU” denote batch normalization and rectify linear unit operation, respectively. “BN*” denotes batch normalization without learnable parameters.

Augmentation	Parameter	ImageNet-1K		CIFAR	
		Aug. $\mathcal{T}_1(\cdot)$	Aug. $\mathcal{T}_2(\cdot)$	Aug. $\mathcal{T}_1(\cdot)$	Aug. $\mathcal{T}_2(\cdot)$
random crop and resize	area of the crop	[0.05, 1.0]	[0.05, 1.0]	[0.1, 1.0]	[0.1, 1.0]
	aspect ratio of the crop	[$\frac{3}{4}, \frac{4}{3}$]			
random horizontal flip	horizontal flip probability	0.5	0.5	0.5	0.5
random color jittering	color jittering probability	0.8	0.8	0.8	0.8
	max brightness adjustment intensity	0.4	0.4	0.4	0.4
	max contrast adjustment intensity	0.4	0.4	0.4	0.4
	max saturation adjustment intensity	0.2	0.2	0.2	0.2
	max hue adjustment intensity	0.1	0.1	0.1	0.1
random gray scale	color dropping probability	0.2	0.2	0.2	0.2
random Gaussian blurring	Gaussian blurring probability	1.0	0.1	1.0	0.1
	sigma of Gaussian blurring	[0.1, 2.0]	[0.1, 2.0]	[0.1, 2.0]	[0.1, 2.0]
random solarization	solarization probability	0.0	0.2	0.0	0.2

Table 10: The parameters of data augmentations applied during self-supervised training. “[\cdot, \cdot]” denotes the range for uniform sampling.

our proposed PatchMix consistently achieves richer inter-instance similarities as expected in Figure 8, Figure 9 and Figure 10.

E. Self-Attention Map Visualization

To analyze the representations learned by our proposed method, we visualize the self-attention scores of [CLS] token from multiple heads of the last transformer block, where the ViT-S/16 model is pretrained on ImageNet-1K for 800 epochs. Specifically, for each head of the last transformer block, the attention scores between [CLS] token and

all patch token of the input image is reshaped into the size of 2D map for better visualization, illustrating the importance of patches in the final decision layer. For example, the input image patch sequence with 196 tokens (not including [CLS] token) is reshaped into the 2D map with size of 14×14 . The visualization results are shown in Figure 11, where each sample contains attention maps from 12 heads of [CLS] token.

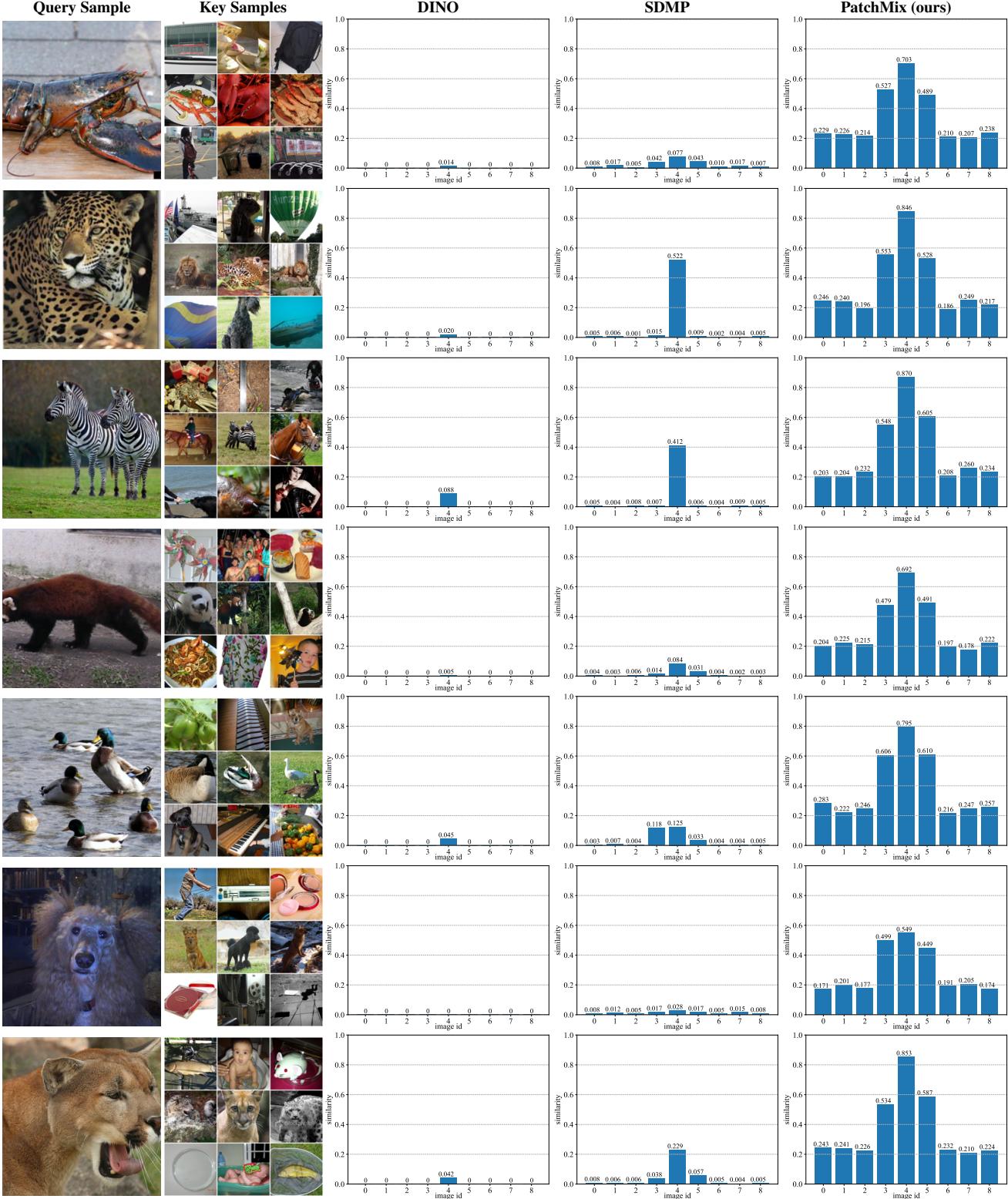


Figure 8: The visualization of inter-instance similarities on ImageNet-1K. The query sample and the image with id 4 in key samples are from the same category. The images with id 3 and 5 come from category similar to query sample.

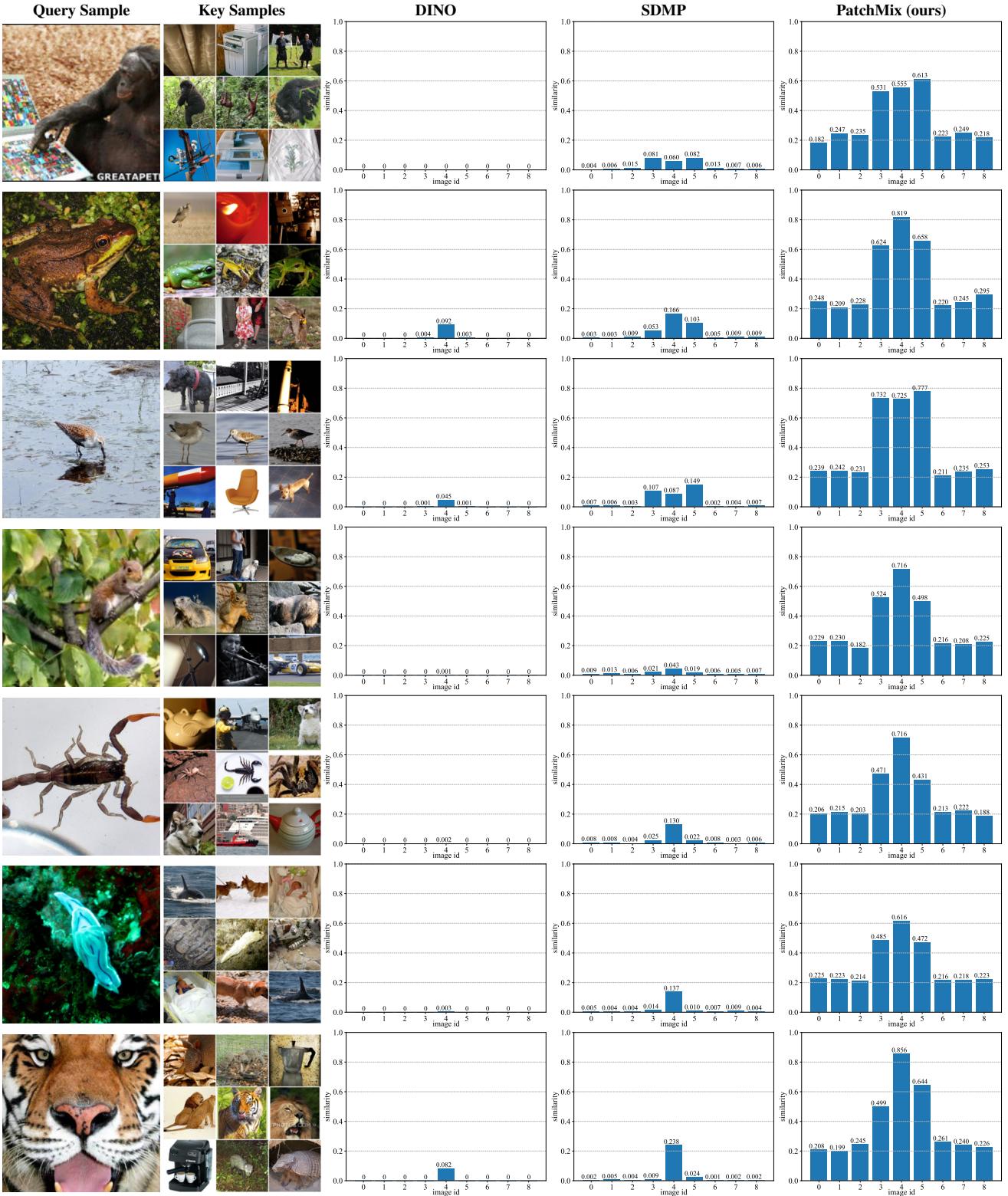


Figure 9: The visualization of inter-instance similarities on ImageNet-1K. The query sample and the image with id 4 in key samples are from the same category. The images with id 3 and 5 come from category similar to query sample.

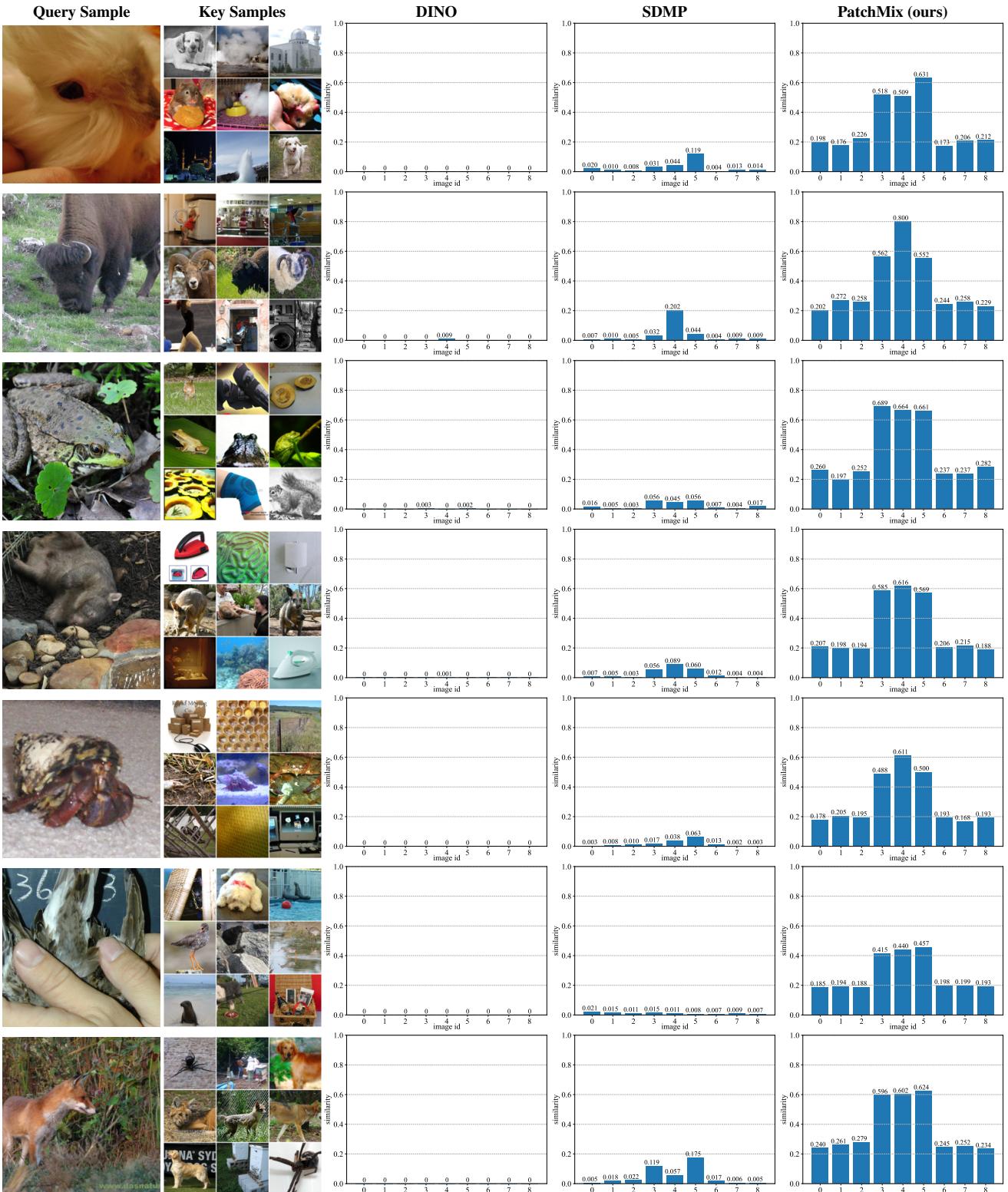


Figure 10: The visualization of inter-instance similarities on ImageNet-1K. The query sample and the image with id 4 in key samples are from the same category. The images with id 3 and 5 come from category similar to query sample.

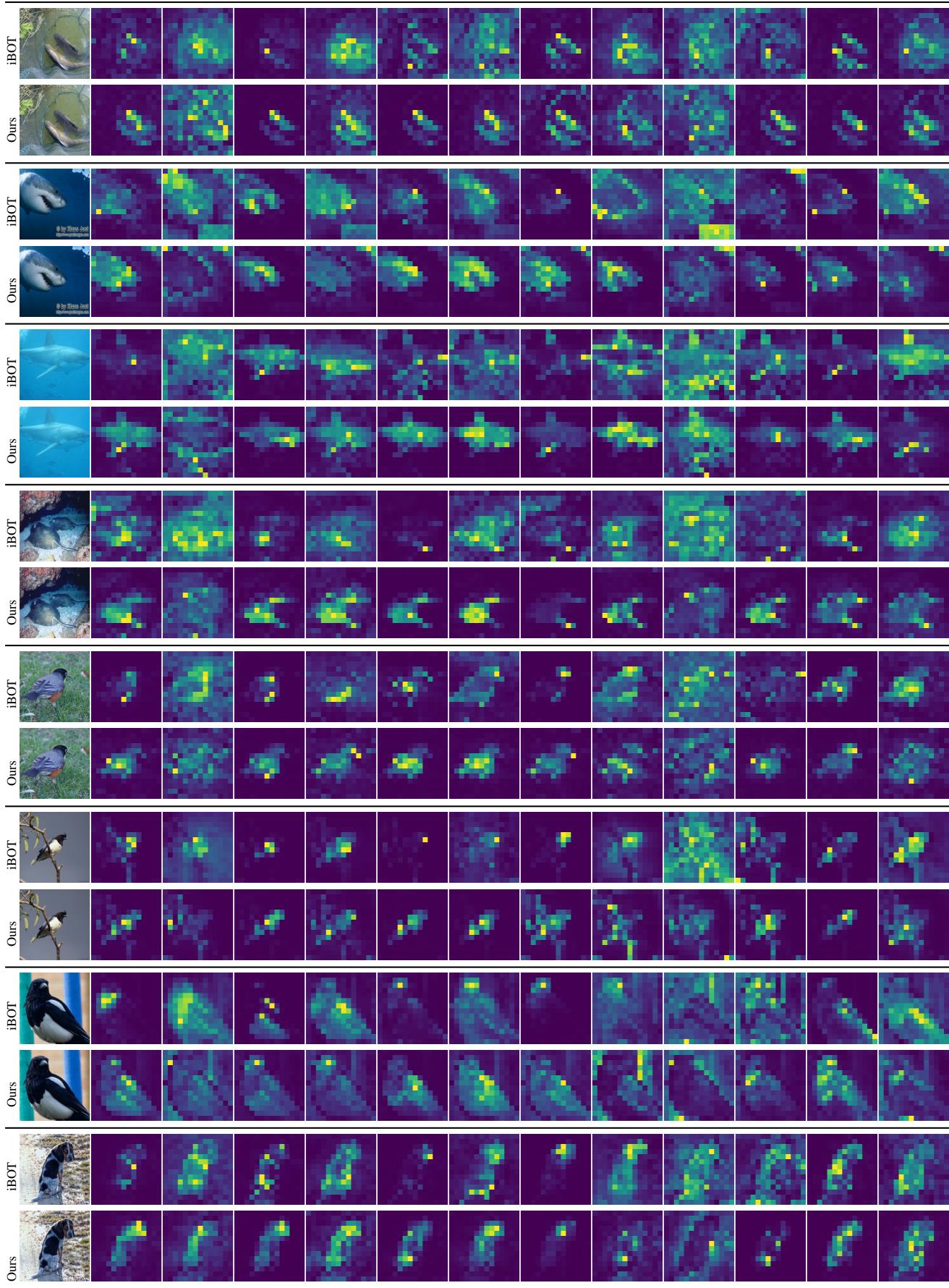


Figure 11: The visualization of self-attention maps from 12 heads of the last transformer block of ViT-S/16 models, which are pretrained on ImageNet-1K for 800 epochs using iBOT and our proposed PatchMix.