SERE: Exploring Feature Self-relation for Self-supervised Transformer

Zhong-Yu Li, Shanghua Gao, Ming-Ming Cheng

Abstract—Learning representations with self-supervision for convolutional networks (CNN) has been validated to be effective for vision tasks. As an alternative to CNN, vision transformers (ViT) have strong representation ability with spatial self-attention and channel-level feedforward networks. Recent works reveal that self-supervised learning helps unleash the great potential of ViT. Still, most works follow self-supervised strategies designed for CNN, *e.g.,* instance-level discrimination of samples, but they ignore the properties of ViT. We observe that relational modeling on spatial and channel dimensions distinguishes ViT from other networks. To enforce this property, we explore the feature **SE**lf-**RE**lation (SERE) for training self-supervised ViT. Specifically, instead of conducting self-supervised learning solely on feature embeddings from multiple views, we utilize the feature self-relations, *i.e.,* spatial/channel self-relations, for self-supervised learning. Self-relation based learning further enhances the relation modeling ability of ViT, resulting in stronger representations that stably improve performance on multiple downstream tasks. Our source code is publicly available at: https://github.com/MCG-NKU/SERE.

Index Terms—feature self-relation, self-supervised learning, vision transformer

1 Introduction

Supervised training of neural networks thrives on many vision tasks at the cost of collecting expensive human-annotations [1], [2], [3]. Learning visual representations from un-labeled images [4], [5], [6], [7], [8] has proven to be an effective alternative to supervised training, *e.g.,* convolutional networks (CNN) trained with self-supervision have shown comparable or even better performance than its supervised counterparts [9], [10]. Recently, vision transformers (ViT) [11], [12] have emerged with stronger representation ability than CNN on many vision tasks. Pioneering works have shifted the methods designed for self-supervised CNN to ViT and revealed the great potential of self-supervised ViT [13], [14], [15]. Typical self-supervised learning methods designed for ViT, *e.g.,* DINO [13]and MoCoV3 [15], send multiple views of an image into a ViT network to generate feature representations. Self-supervisions, *e.g.,* contrastive learning [15], [16], [17] and clustering [13], [18], are then implemented on these representations based on the hypothesis that different views of an image share similar representations. However, the widely used feature representations are still limited to feature embedding used by CNN based methods, *e.g.,* image-level embeddings [6], [7], [19] and patch-level embeddings [20], [21]. But the properties of ViT, *e.g.,* the self-relation modeling ability, are less considered by existing self-supervised methods. We wonder if other forms of representations related to ViT can benefit the training of self-supervised ViT.

We seek to improve the training of self-supervised ViT by exploring the properties of ViT. ViT models the feature relations on spatial and channel dimensions with the multi-head self-attention (MHSA) and feedforward network (FFN) [11], [22], [23], respectively. The MHSA aggregates the spatial information with the extracted relations among patches, resulting in stronger spatial relations among patches with similar semantic contexts (see Fig. 1(c)). The FFN combines features from different channels, implicitly modeling the feature self-relation in the channel dimension. For instance, Fig. 1(c) reveals that channels learn diverse patterns, and there are varying degrees of relations between different channels. Feature self-relation modeling enables ViT with strong representation ability, motivating us to use self-relation as a new representation form for self-supervision.

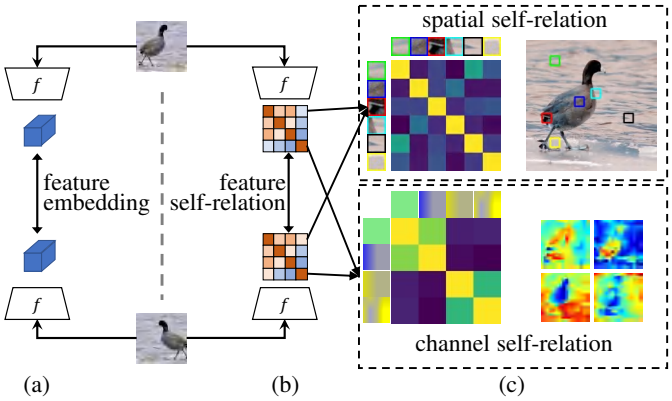


Fig. 1. The illustration of self-supervised learning using feature embeddings and our proposed feature self-relation. (a) Typical self-supervised learning methods process the feature embeddings of the image views. (b) We propose to model the feature self-relation that measures the relation inside an image view from different dimensions. (c) Two specific forms of self-relation, *i.e.,* the spatial and channel self-relations. For spatial self-relation, we select 6 patches indicated by differently colored boxes (top right) and visualize their self-relation (top left). For channel self-relation, we show visualized feature maps of 4 channels (bottom right) and the corresponding self-relation (bottom left).

In this work, we propose to utilize the feature **SE**lf-**RE**lation (SERE) for self-supervised training, enhancing the self-relation modeling properties in ViT. Following the spatial relation in MHSA and channel relation in FFN, we form the spatial and channel self-relations as representations. The spatial self-relation extracts the relations among patches within an image. The channel self-relation models the connection of different channels, where each channel in feature embeddings highlights unique semantic information. Feature self-relation is the representation

• *The authors are with TMCC, CS, Nankai University, Tianjin 300350, China. S. Gao is the corresponding author (shgao@mail.nankai.edu.cn).*

in a new dimension and is compatible with existing representation forms, *e.g.,* image-level and patch-level feature embeddings. As shown in Fig. 1, we can easily replace the feature embeddings with the proposed feature self-relation on existing self-supervised learning methods. We demonstrate that utilizing feature self-relation could stably improve multiple training methods for self-supervised ViT, *e.g.,* DINO [13], iBOT [18], and MoCoV3 [15], on various downstream tasks, *e.g.,* object detection [2], [24], semantic segmentation [3], [25], semi-supervised semantic segmentation [26] and image classification [1]. To our best knowledge, we are the first to study self-relations in self-supervised learning. Our major contributions are summarized as follows:

- We propose to utilize the self-relations (SERE) of ViT, *i.e.,* spatial and channel self-relations that fit well with the relation modeling property of ViT, as the representations for self-supervised learning.
- The proposed SERE method is compatible with existing self-supervised methods and stably boosts ViT on various downstream tasks.

2 Related Work

2.1 Self-Supervised Learning

Self-supervised learning aims at learning rich representations without any human annotations. Early works utilized hand-crafted pretext tasks, *e.g.,* coloration [27], [28], jigsaw puzzles [29], rotation prediction [30], autoencoder [31], [32], image inpainting [33] and counting [34] to learn representations based on heuristic cues [19], but only achieved limited generalization ability. Recently, self-supervised learning has shown great breakthroughs due to new forms of self-supervisions, *e.g.,* contrastive learning [7], [35], [36], [37], [38], [39], [40], self-clustering [41], [42], [43], and representation alignment [5], [6], [44], [45], [46], [47]. These methods directly utilize the feature embeddings as representations to generate self-supervisions. For example, many of these methods utilize image-level feature embeddings [19], [41], [48] as representations. And some methods explore using embeddings in more fine-grained dimensions, *e.g.,* pixel [20], [49], patch [50], [51], object [21], and region [21], [52] dimensions. However, these representations are still embeddings corresponding to different regions of input images. Compared to these embedding based methods that only constrain individual embedding, we further transform the feature embedding to self-relation as a new representation dimension, which adds the constraint to the relation among embeddings. The self-relation provides rich information for self-supervised training and fits well with the relation modeling properties of ViT, thus further boosting the representation quality of ViT. Meanwhile, the self-relation is orthogonal to embedding based methods and consistently improves the performance of multiple methods.

2.2 Self-Supervised Vision Transformer

Transformers have been generalized to computer vision [11], [53] and achieved state-of-the-art performance on many tasks, *e.g.,* image classification [12], semantic segmentation [53], [54], and object detection [55]. Due to a lack of inductive bias, training ViT requires much more data and tricks [11], [56]. Recent works have been working on training ViT with self-supervised learning methods [16], [57], [58], [59] to meet the data requirement of ViT with low annotation costs. Many instance discrimination based methods use feature embeddings as the representation for self-supervised learning. For instance, Chen *et al.* [15] and Caron *et al.* [13] implement contrastive learning and self-clustering with image-level embeddings, respectively. Zhou *et al.* [18] develop self-distillation with patch-level embeddings. However, these methods still follow the pretext task of instance discrimination initially designed for CNNs, where representations with invariance to transformation are learned by maximizing the similarity among positive samples. New properties in ViT may help the self-supervised training but are ignored by these methods. We explore spatial self-relation and channel self-relation, which are proven more suitable for the training of ViT.

2.3 Masked Image Modeling

Concurrent with our work, self-supervised learning by masked image modeling (MIM) [14], [33], [60], [61] has become a popular alternative to instance discrimination (ID) for self-supervised ViT. MIM reconstructs masked patches from unmasked parts, with different forms of reconstruction targets, *e.g.,* discrete tokenizer [60], [62], raw pixels [14], [59], [63], [64], [65], HOG features [66], patch representations [18], *etc*. Compared to ID, patch-level reconstruction in MIM enhances token-level representations [18], [61]. Differently, the proposed SERE enhances the ability to model inter-token relations. Experiments also demonstrate that SERE can outperform and complement various MIM-based methods. Additionally, we strengthen the ability to model inter-channel relations, which MIM is missing.

2.4 Property of Vision Transformer

Recent works have shown that the remarkable success of ViT on many vision tasks [12], [54], [67] relies on their strong ability to model spatial relations. Dosovitskiy *et al.* [11] and Kim *et al.* [23] find that attention attends to semantically relevant regions of images. Raghu *et al.* [22] reveal the representations of ViT preserve strong spatial information even in the deep layer. They also observe that patches in ViT have strong connections to regions with similar semantics. Caron *et al.* [13] find that self-supervised ViT captures more explicit semantic regions than supervised ViT. These observations indicate that ViT has a strong ability to model relations, which is quite different from the pattern-matching mechanisms of CNNs. In this work, we propose to enhance such ability by explicitly using spatial and channel feature self-relations for self-supervised learning.

2.5 Relation Modeling

Relation modeling, which has different forms such as pairwise relation and attention, has facilitated various vision tasks, *e.g.,* knowledge distillation [68], [69], [70], [71], [72], [73], metric learning [74], semantic segmentation [75], [76], [77], unsupervised semantic segmentation [78], object localization [79], [80], [81], contrastive learning [82], masked image modeling [83], feature aggregation [84] and texture descriptor [85], [86]. In self-supervised learning, early work [87] proposes to utilize relation modeling by calculating channel relations in the whole batch, *i.e.,* batch-relation. In comparison, we explore self-relation, which is the spatial or channel relations for features within an image and fits well with the relation modeling property of ViT.

3 Method

3.1 Overview

In this work, we focus on the instance discriminative self-supervised learning pipeline [4], [13]. First, we briefly revisit

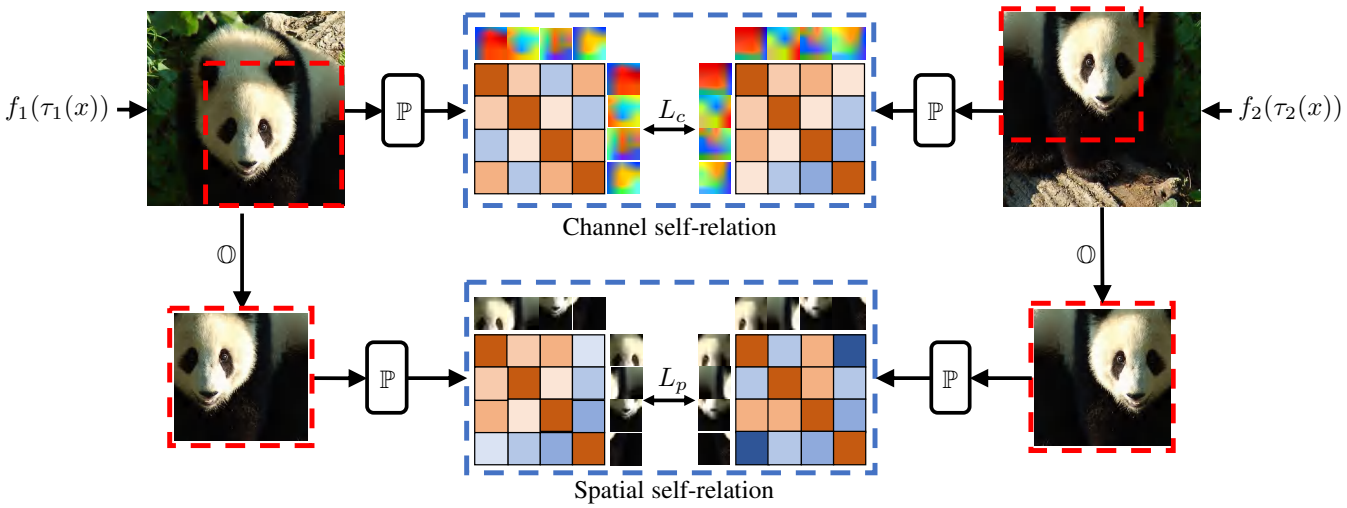


Fig. 2. Our method models self-relation from spatial and channel dimensions. Given an image x, two views are generated by two random data augmentations. Here the image patches represent the feature embeddings extracted by the encoder. The feature embeddings are transformed by representation transformation $\mathbb{P}$ to generate spatial or channel self-relations. L_p and L_c, *i.e.,* the loss functions defined in Equ. (3) and Equ. (5), enforce consistency between self-relations of different views. For spatial self-relation, only the features in the overlapping region are considered. $\mathbb{O}$ means the operation of extracting features from the overlapping region between two views in Equ. (2), where the red dotted box indicates the overlapping region.

the framework of common instance discriminative self-supervised learning methods. Given an un-labeled image x, multiple views are generated by different random data augmentations, *e.g.,* generating two views $\tau_1(x)$ and $\tau_2(x)$ with augmentations τ_1 and τ_2. Under the assumption that different views of an image contain similar information, the major idea of most instance discriminative methods is to maximize the shared information encoded from different views. Firstly, two views are sent to the encoder network to extract the feature embeddings $r_1 \in \mathbb{R}^{C\times HW}$ and $r_2 \in \mathbb{R}^{C\times HW}$ with $H \cdot W$ local patches and C channels. According to the training objective of self-supervised learning methods, the feature embeddings are then transformed with transformation $\mathbb{P}$ to obtain different representations, *e.g.,* image-level and patch-level embeddings. Different self-supervised optimization objectives utilize the obtained representations to get the loss as follows:

$$L_I = \mathrm{R}(\mathbb{P}(r_1), \mathbb{P}(r_2)), \tag{1}$$

where R means the function that maximizes the consistency across views and can be defined with multiple forms, *e.g.,* contrastive [7], non-contrastive [6], and clustering [4] losses.

Our main focus in this work is exploring new forms of representation transformation $\mathbb{P}$. Motivated by the relation modeling properties in ViT, instead of directly using feature embeddings, we utilize feature self-relation in multiple dimensions as the representations for self-supervised learning on ViT. In the following sections, we introduce two specific self-relation representations for self-supervised ViT, *i.e.,* spatial and channel self-relations.

3.2 Spatial Self-relation

Prior works [11], [13], [22], [23] have observed that ViT has the property of modeling relations among local patches by the MHSA module. Meanwhile, modeling more accurate spatial relations is crucial for many dense prediction tasks [20], [21], *e.g.,* object detection and semantic segmentation. So we propose to enhance the relation modeling ability of ViT by cooperating spatial self-relation for self-supervised training. In the following part, we first give details of the transformation $\mathbb{P}$ that transforms the feature embeddings encoded by ViT to spatial self-relation. Then, we give the self-supervision loss utilizing spatial self-relation as the representation.

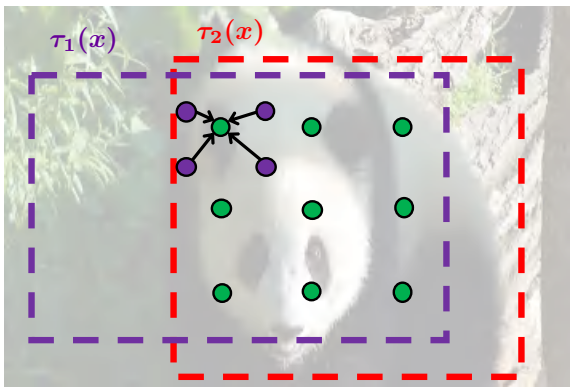


Fig. 3. The region-aligned sampling operation for spatial self-relation. $\tau_1(x)$ and $\tau_2(x)$ are the different views of an image, and the dotted boxes indicate their regions in the original image. The points in green mean the uniformly sampled points in the overlapped regions. And the points in purple mean the patch features in ViT.

Generating spatial self-relation representation. Given the feature embeddings $r_1 = f_1(\tau_1(x)) \in \mathbb{R}^{C\times HW}$ and $r_2 = f_2(\tau_2(x)) \in \mathbb{R}^{C\times HW}$ from the ViT backbone, a projection head h_p, which consists of a batch normalization [88] layer and a ReLU [89] activation layer, processes these embeddings to obtain $p_1 = h_p(r_1)$ and $p_2 = h_p(r_2)$. Then, we separately calculate their spatial self-relation.

In contrast to the image-level embedding, the supervision between spatial self-relation of different views should be calculated between patches at the same spatial positions. However, p_1 and p_2 are not aligned in the spatial dimension due to the random crop and flip in data augmentations. To solve the misalignment issue, we apply a region-aligned sampling operation $\mathbb{O}$ [26] to uniformly sample $H_s \times W_s$ points from the overlapping region of p_1 and

p_2.[1] As shown in Fig. 3, we localize the overlapping region in the raw image and split the region into $H_s \times W_s$ grids, which are not essentially aligned with the patches in ViT. For the center of each grid, we calculate its spatial coordinates in feature maps of each view and then sample its features by bi-linear interpolation. The details of this operation $\mathbb{O}$ are shown in the supplementary. For one view, *e.g.,* $p_1 \in \mathbb{R}^{C\times HW}$, we calculate the spatial self-relation $\mathbb{A}_p(p_1) \in \mathbb{R}^{H_sW_s\times H_sW_s}$ as follows:

$$\mathbb{A}_p(p_1) = \mathrm{Softmax}\left(\frac{\mathbb{O}(p_1)^T \cdot \mathbb{O}(p_1)}{\sqrt{C}} \Big/ t_p\right), \tag{2}$$

where $\mathbb{O}(p_1) \in R^{C\times H_sW_s}$ is the feature sampled in the overlapping region, T is the matrix transpose operation, and t_p is the temperature parameter that controls the sharpness of the $\mathrm{Softmax}$ function. In the spatial self-relation, each row represents the relation of one local patch to other patches and is normalized by the $\mathrm{Softmax}$ function to generate probability distributions.

Self-supervision with spatial self-relation. Spatial self-relation can be used as the representation of many forms of self-supervisions. For simplicity, we give an example of using self-relation for asymmetric non-contrastive self-supervision loss [5], [6] as follows:

$$L_p = \mathrm{R}_e(\mathcal{G}(\mathbb{A}_p(p_1)), \mathbb{A}_p(g_p(p_2))), \tag{3}$$

where R_e is the cross-entropy loss, $\mathcal{G}$ is the stop-gradient operation to avoid training collapse following [5], and g_p is the prediction head for asymmetric non-contrastive loss [5], [6] consisting of a fully connected layer, a batch normalization layer, and a ReLU layer.

Multi-head spatial self-relation. In ViT, the MHSA performs multiple parallel self-attention operations by dividing the feature into multiple groups. It is observed that different heads might focus on different semantic patterns [13]. Inspired by this, we divide the feature embeddings into M groups along the channel dimension and calculate the spatial self-relation within each group, obtaining M spatial self-relations for each view. By default, we choose $M = 6$, as shown in Tab. 12.

3.3 Channel Self-relation

In neural networks, each channel represents some kind of pattern within images. Different channels encode diverse patterns [90], [91], providing neural networks with a strong representation capability. The FFN [11] in ViT combines patterns across channels and implicitly models the relation among channels [90], *i.e.,* the pattern encoded in one channel has different degrees of correlation with the patterns encoded by other channels, as shown in Fig. 2. This mechanism motivates us to form channel self-relation as the representation for self-supervised learning to enhance self-relation modeling ability in the channel dimension. Specifically, we transform the feature embedding of ViT to channel self-relation and then use the channel self-relation as the representation for self-supervision.

Generating channel self-relation representation. Here, we give the details of the transformation $\mathbb{P}$ that transforms the feature embeddings to channel self-relation. As in Equ. (2), given the feature embeddings of two views, *i.e.,* r_1 and r_2, a projection head h_c with the same structure as h_p processes these embeddings and obtains $c_1 = h_c(r_1)^T$ and $c_2 = h_c(r_2)^T$. Then we separately calculate the channel self-relation for each view. For one view, *e.g.,* $c_1 \in \mathbb{R}^{HW\times C}$, we calculate its channel self-relation $\mathbb{A}_c(c_1) \in \mathbb{R}^{C\times C}$ as follows:

$$\mathbb{A}_c(c_1) = \mathrm{Softmax}\left(\frac{c_1^T \cdot c_1}{H \cdot W} \Big/ t_c\right), \tag{4}$$

where the Softmax function normalizes each row of the self-relation to get probability distributions, and t_c is the temperature parameter controlling the sharpness of probability distributions.

Self-supervision with channel self-relation. The channel self-relation can also be utilized as a new form of representation for many self-supervised losses. Similar to the spatial self-relation based loss in Equ. (3), we give the non-contrastive loss using channel self-relation as follows:

$$L_c = \mathrm{R}_e(\mathcal{G}(\mathbb{A}_c(c_1)), \mathbb{A}_c(g_c(c_2))), \tag{5}$$

where the R_e is the cross-entropy loss, and g_c is a prediction head with the same structure as g_p in Equ. (3). This loss function enforces the consistency of channel self-relations among views and thus enhances the channel self-relation modeling ability of the model. Unlike spatial self-relation, we do not need to consider the spatial misalignment between different views. Because we enforce the consistency between channel self-relations, not the channel features, and the channel self-relation defined in Equ. (4) has no spatial dimension.

3.4 Implementation Details

Loss function. By default, we apply our proposed spatial/channel self-relations and image embeddings as representations for self-supervision losses, as these representations reveal different properties of features. The summarized loss function is as follows:

$$L = L_I + \alpha L_p + \beta L_c, \tag{6}$$

where the spatial and channel losses are weighted by α and β, and L_I is the loss using image-level embeddings, *e.g.,* the clustering-based loss in DINO [13]. We show in Tab. 8 that solely using our proposed self-relation could achieve competitive or better performance than using image-level embeddings. Combining these three representations results in better representation quality, showing self-relation is a complementary representation form to image-level embeddings. To increase the training efficiency and make fair comparisons, we utilize the multi-crop [4], [13] augmentation to generate global and local views. For local views, we follow [4], [13] to calculate the loss between each global and local view but ignore the loss among local views.

Architecture. We use the Vision Transformer [11] as the encoder network. Following [7], [13], the representations r_1 and r_2 of two views $\tau_1(x)$ and $\tau_2(x)$ are extracted by a momentum-updated encoder network f_1 and the encoder network f_2. During training, the parameters θ_2 of f_2 are updated by gradient descent. And the parameters θ_1 of f_1 are updated as $\theta_1 = \lambda\theta_1 + (1-\lambda)\theta_2$, where $\lambda \in [0, 1]$ is the momentum coefficient. Following DINO [13], the λ is set to 0.996 and is increased to 1.0 during training with a cosine schedule. Accordingly, we denote the projections following f_1 and f_2 as h_p^1/h_c^1 and h_p^2/h_c^2, respectively. The parameters

1. In this work, we combine the proposed spatial self-relation with existing methods due to the orthogonality of self-relation. Since existing methods do not restrict that different views must overlap, we only add spatial self-relation to the views with overlapping regions.

TABLE 1
Fully fine-tuning classification on ImageNet-1K and semi-supervised semantic segmentation on ImageNet-S. For ImageNet-S, we report the mIoU on the val and test set. The PT means loading self-supervised pre-trained weights for initialization and FT means loading fully fine-tuned weights on classification labels of ImageNet-1K for initialization, respectively.

	Backbone	Epochs	Classification ImageNet-1K		Segmentation ImageNet-S_{PT}		ImageNet-S_{FT}	
			Top-1	Top-5	val	test	val	test
DINO [13]	ViT-S/16	100	79.7	95.1	35.1	34.4	54.6	54.4
+SERE	ViT-S/16	100	**80.9**	**95.5**	**36.9**	**36.0**	**57.3**	**56.2**
iBOT [18]	ViT-S/16	100	80.9	95.4	38.1	37.8	57.9	57.4
+SERE	ViT-S/16	100	**81.5**	**95.8**	**41.0**	**40.2**	**58.9**	**57.8**
iBOT [18]	ViT-B/16	100	83.3	96.6	48.3	47.8	62.6	63.0
+SERE	ViT-B/16	100	**83.7**	**96.7**	**48.6**	**48.2**	**63.0**	**63.3**

TABLE 2
Transferring learning on semantic segmentation, object detection, and instance segmentation. The AP^b means the bounding box AP for object detection (DET), and AP^m means the segmentation mask AP for instance segmentation (SEG).

	VOC SEG		ADE20K SEG	
	mIoU	mAcc	mIoU	mAcc
DINO [13]	77.1	87.5	42.6	53.4
+SERE	**79.7**	**88.8**	**43.8**	**54.6**

	COCO DET			COCO SEG		
	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
DINO [13]	46.0	64.9	49.7	40.0	62.0	42.8
+SERE	**46.6**	**65.9**	**50.2**	**40.5**	**62.9**	**43.5**

of h_p^1/h_c^1 are also momentum-updated by h_p^2/h_c^2, following the updating scheme of f_1. Only the encoder network is used for transfer learning on downstream tasks after pre-training.

4 Experiments

This section verifies the effect of using proposed spatial and channel self-relations as representations for self-supervised learning. We give the pre-training settings in Section 4.1. In Section 4.2, we compare our method with existing methods on multiple evaluation protocols, showing stable improvement over multiple methods. In Section 4.3, we conduct ablations to clarify design choices.

4.1 Pre-training Settings

Unless otherwise stated, we adopt the ViT-S/16 as the backbone network. DINO [13]is selected as our major baseline method. The model is trained by an AdamW [92] optimizer with a learning rate of 0.001 and a batch size of 512. We pre-train models for 100 epochs on the ImageNet-1K [1] dataset for performance comparison. For ablation, the ImageNet-S_{300} dataset [26] is used to save training costs. Following [13], we apply the multi-crop training scheme where 2 global views with the resolution of 224×224 and 4 local views with the resolution of 96×96 are adopted. The global views are cropped with a ratio between 0.35 and 1.0. And the local views are cropped with a ratio between 0.05 and 0.35. For spatial self-relation, the H_s/W_s of the operation $\mathbb{O}$ in Equ. (2) are set to 13/13 for global views and 6/6 for local views. The number of heads M in spatial self-relation is set to 6 by default. The t_p in Equ. (2) and t_c in Equ. (4) are set to 0.5 and 0.1 for the encoder network. For the momentum encoder, we set the t_p and t_c to 1.0 and 1.0. The α and β in Equ. (6) are set to 1.0 and 1.0, respectively.

For iBOT [18], 10 local views are used for a fair comparison. And we crop images with a ratio between 0.4 and 1.0 for global views and between 0.05 and 0.4 for local views. A gradient clip of 0.3 is used for optimization. The α and β in Equ. (6) are set to 0.2 and 0.5. Additionally, we provide experiments with ViT-B/16 as the backbone and show the pre-training and fine-tuning details in the supplementary.

TABLE 3
Comparison with longer pre-training epochs.

(a) Semantic segmentation on the ADE20K dataset.

	Backbone	Epochs	mIoU	mAcc
iBOT [18]	ViT-S/16	800	45.4	56.2
+SERE	ViT-S/16	100	**45.8**	**56.8**
iBOT [18]	ViT-B/16	400	50.0	60.3
+SERE	ViT-B/16	200	**50.0**	**60.9**

(b) Classification on the ImageNet-1K dataset.

	Backbone	Epochs	Top-1	Top-5
iBOT [18]	ViT-S/16	300	81.1	-
+SERE	ViT-S/16	100	**81.5**	**95.8**

TABLE 4
Semi-supervised classification on ImageNet-1K. We fine-tune the models with 1%/10% training labels and evaluate them with 100% val labels.

	1%		10%	
	Top-1	Top-5	Top-1	Top-5
DINO [13]	52.1	77.8	70.0	89.8
+SERE	**55.9**	**81.0**	**71.5**	**90.6**

4.2 Performance and Analysis

We verify the effectiveness of self-relation for self-supervised learning by transferring the pre-trained models to image-level classification tasks and dense prediction downstream tasks. Models are pre-trained with 100 epochs on ImageNet-1k unless otherwise stated. For easy understanding, models pre-trained with self-relation representations are marked as SERE.

Fully fine-tuning classification on ImageNet-1K. We compare the fully fine-tuning classification performance on the ImageNet-1K dataset. When utilizing ViT-S/16, the pre-trained model is fine-tuned for 100 epochs with the AdamW [92] optimizer and a batch size of 512. The initial learning rate is set to 1e-3 with a layer-wise decay of 0.65. After a warmup of 5 epochs, the learning rate gradually decays to 1e-6 with the cosine decay schedule. We report the Top-1 and Top-5 accuracy for evaluation on the ImageNet-1k val set. As shown in Tab. 1, SERE advances DINO and iBOT by 1.2% and 0.6% on Top-1 accuracy. Even compared to iBOT of 300 epochs, SERE can improve 0.4% Top-1 accuracy with a third of the pre-training time (100 epochs), as shown in Tab. 3 (b). Moreover, using ViT-B/16, SERE surpasses iBOT by 0.4% in Top-1 accuracy, as shown in Tab. 1. These results demonstrate that SERE enhances the category-related representation ability of ViT.

Semi-supervised classification on ImageNet-1K. We also evaluate the classification performance in a semi-supervised fashion. Following the setting of [18], we fully fine-tune the pre-trained

TABLE 5
Transfer learning on the classification task. We fine-tune the pre-trained models on multiple datasets and report the Top-1 accuracy.

	Cifar_{10}	Cifar_{100}	INat_{19}	Flwrs	Cars
DINO [13]	98.8	89.6	76.9	97.8	93.5
+SERE	**98.9**	**90.0**	**77.5**	**98.0**	**93.5**

TABLE 6
Compared with masked image modeling on the ImageNet-1K dataset. † means effective pre-training epochs [18] that account for actually used images during pre-training. ‡ means the models are fine-tuned for 200 epochs on ImageNet-1K, while others are fine-tuned for 100 epochs.

	Architecture	Pre-training Epochs†	Top-1
DINO [13]	ViT-S/16	300	79.7
MAE‡ [14]		800	80.9
iBOT [18]		400	80.9
DINO [13]+SERE		300	80.9
iBOT [18]+SERE		400	**81.5**
BEiT [60]	ViT-B/16	800	83.2
MAE [14]		800	83.3
iBOT [18]		400	83.3
iBOT [18]+SERE		400	**83.7**

models with 1% and 10% training labels on the ImageNet-1K dataset for 1000 epochs. We use the AdamW optimizer to train the model with a batch size of 1024 and a learning rate of 1e-5. Tab. 4 reports the Top-1 and Top-5 accuracy on the ImageNet-1K val set. SERE consistently achieves better accuracy with 1% and 10% labels. With only 1% labels, there is a significant improvement of 3.8% in Top-1 accuracy, showing the advantage of our method in the semi-supervised fashion.

Semi-supervised semantic segmentation for ImageNet-S. The ImageNet-S dataset [26] extends ImageNet-1K with pixel-level semantic segmentation annotations on almost all val images and parts of training images. Evaluating semantic segmentation on the ImageNet-S dataset avoids the potential influence of domain shift between pre-training and fine-tuning datasets. We fine-tune the models with the semantic segmentation annotations in the ImageNet-S training set and evaluate the performance on the val and test sets of ImageNet-S. The ViT-S/16 model is initialized with self-supervised pre-trained weights ($\text{ImageNet-S}_{\text{PT}}$) or fully fine-tuned weights on classification labels ($\text{ImageNet-S}_{\text{FT}}$) of the ImageNet-1K dataset. A randomly initialized 1×1 conv is attached to the model as the segmentation head. We fine-tune models for 100 epochs with an AdamW optimizer, using a batch size of 256 and a weight decay of 0.05. The learning rate is initially set to 5e-4 with a layer-wise decay of 0.5. After a warmup of 5 epochs, the learning rate decays to 1e-6 by the cosine decay schedule. The images are resized and cropped to 224×224 for training and are resized to 256 along the smaller side for evaluation.

As shown in Tab. 1, compared to DINO and iBOT, SERE improves the val mIoU by 1.8% and 2.9% when initializing the model with self-supervised pre-trained weights. When loading weights of the fully fine-tuned classification model for initialization, SERE brings a 2.7%/1.0% gain on mIoU over DINO/iBOT. We conclude that SERE enhances the relation modeling ability, enabling ViT with much stronger shape-related representations.

Transferring learning on the classification task. To evaluate the transferring ability on classification tasks, we fine-tune pre-trained models on multiple datasets, including CIFAR [93], Flowers [94], Cars [95], and iNaturalist19 [96]. The training details are summarized in the supplementary. Tab. 5 shows that SERE performs better on Top-1 accuracy over DINO, demonstrating that SERE benefits the transferring learning on classification tasks.

TABLE 7
Cooperating SERE with multiple self-supervised learning methods. Models are pre-trained on the ImageNet-S_{300} dataset with 100 epochs.

	VOC SEG		$\text{ImageNet-S}^{\text{PT}}_{300}$	
	mIoU	mAcc	val	test
MoCov3 [15]	65.7	78.7	24.0	24.8
+SERE	**67.5**	**80.6**	**29.1**	**29.9**
DINO [13]	68.1	81.1	28.8	29.6
+SERE	**73.5**	**84.7**	**41.2**	**42.0**
iBOT [18]	74.5	85.5	41.5	42.0
+SERE	**75.9**	**86.3**	**45.3**	**45.6**

Transfer learning on semantic segmentation. We also evaluate the transfer learning performance on the semantic segmentation task using PASCAL VOC2012 [25] and ADE20K [3] datasets. The UperNet [97] with the ViT-S/16 backbone is used as the segmentation model. Following the training setting in [18], we fine-tune models for 20k and 160k iterations on PASCAL VOC2012 and ADE20K datasets, with a batch size of 16. Tab. 2 reports the mIoU and mAcc on the validation set. The self-relation improves the DINO by 2.6% on mIoU and 1.3% on mAcc for the PASCAL VOC2012 dataset. On the ADE20K dataset, there is also an improvement of 1.2% on mIoU and 1.2% on mAcc compared to DINO. Tab. 3 (a) shows that SERE even outperforms iBOT with much fewer pre-training epochs. Therefore, semantic segmentation tasks benefit from the stronger self-relation representation ability of SERE.

Transfer learning on object detection and instance segmentation. We use the Cascade Mask R-CNN [24] with ViT-S/16 to evaluate the transfer learning performance on object detection and instance segmentation tasks. Following [18], the models are trained on the COCO train2017 set [2] with the $1\times$ schedule and a batch size of 16. Tab. 2 reports the bounding box AP (AP^{b}) and the segmentation mask AP (AP^{m}) on the COCO val2017 set. Compared to DINO, SERE improves by 0.6% on AP^{b} and 0.5% on AP^{m}, showing that SERE facilitates the model to locate and segment objects accurately.

Comparison with masked image modeling (MIM). We also demonstrate that our proposed method, SERE, outperforms and complements various masked image modeling (MIM) based methods. As shown in Tab. 6, SERE can significantly enhance contrastive learning based approach (e.g., DINO). DINO+SERE achieves comparable performance compared to MIM based methods (iBOT and MAE), requiring less pre-training/fine-tuning epochs. Meanwhile, SERE and MIM can be complementary. For instance, cooperating with SERE further improves iBOT by 0.4% Top-1 accuracy. Moreover, qualitative results in Fig. 4 show that SERE produces more precise and less noisy attention maps than iBOT. These results strongly confirm the effectiveness of SERE compared to MIM-based methods.

TABLE 8
Ablation of using different representations for self-supervised training. The L_I, L_p, and L_c denote the loss functions using image-level embedding [13], spatial self-relation, and channel self-relation, respectively. The model without these three losses is randomly initialized when fine-tuned on downstream tasks.

L_I	L_p	L_c	VOC SEG		ImageNet-S^{PT}_{300}	
			mIoU	mAcc	val	test
✗	✗	✗	25.6	35.7	0.2	0.2
✓			68.1	81.1	28.8	29.6
	✓		71.5	83.0	23.7	23.7
		✓	61.4	75.6	22.5	22.3
✓	✓		70.7	82.6	33.3	34.5
✓		✓	69.8	82.9	36.5	38.3
	✓	✓	71.5	83.3	30.6	30.3
✓	✓	✓	**73.5**	**84.7**	**41.2**	**42.0**

TABLE 9
Segmentation F-measure [98] on the PASCAL VOC dataset. The F-measure ignores semantic categories.

	L_p	$L_p + L_I$	$L_p + L_I + L_c$
IoU	87.1	86.7	87.7

Cooperating with more self-supervised learning methods. The self-relation representation is orthogonal to the existing feature representations. Therefore, it can be integrated into various self-supervised learning methods. To demonstrate this, we combine the SERE with MoCo v3 [15], DINO, and iBOT, *i.e.,* utilizing the self-supervision of these methods as the L_I in Equ. (6). We pre-train models on the ImageNet-S_{300} dataset with 100 epochs to save computation costs, and other training settings are constant with baseline methods. As shown in Tab. 7, using SERE consistently improves baseline methods, verifying its generalization to different methods. For example, SERE improves the MoCo v3 by 1.8% on mIoU and 2.0% on mAcc for semantic segmentation on the Pascal VOC dataset. For the semi-supervised semantic segmentation on the ImageNet-S_{300} dataset, SERE gains 5.1% on mIoU over MoCo v3.

4.3 Ablation Studies

To save computational costs for the ablation study, we pre-train all models on the ImageNet-S_{300} [26] dataset with two global views for 100 epochs. We evaluate models with semantic segmentation on the PASCAL VOC dataset and semi-supervised semantic segmentation on the ImageNet-S_{300} dataset.

Effect of spatial and channel self-relation. We compare the effectiveness of different representation forms for self-supervised learning, *i.e.,* our proposed spatial/channel self-relations and image-level feature embeddings used by DINO. As shown in Tab. 8, the spatial self-relation improves the mIoU by 3.4% and mAcc by 1.9% on the PASCAL VOC dataset compared to the feature embedding. These results show that training self-supervised ViT with spatial self-relation further enhances the spatial relation modeling ability of ViT, benefiting dense prediction tasks. Although inferior to the other two representation forms, channel self-relation still improves the representation quality of ViT. The model pre-trained with channel self-relation performs much better than the randomly initialized model on segmentation and classification tasks.

TABLE 10
Cooperating self-relations with patch-level embeddings. DINO+ indicates adding the clustering loss using patch-level embeddings to DINO [13].

DINO	DINO+	SERE	VOC SEG		ImageNet-S^{PT}_{300}	
			mIoU	mAcc	val	test
✓			68.1	81.1	28.8	29.6
	✓		72.6	84.3	40.0	40.4
		✓	73.5	84.7	41.2	42.0
	✓	✓	**75.0**	**86.1**	**44.8**	**46.0**

TABLE 11
Comparison with Barlow [87] that utilizes the batch-relation based loss.

	VOC SEG		ImageNet-S^{PT}_{300}	
	mIoU	mAcc	val	test
Barlow [87]	69.5	82.2	33.2	32.9
SERE	**69.8**	**82.9**	**36.5**	**38.3**

Cooperating with image-level embeddings. We verify the orthogonality between self-relations and image-level embeddings, as shown in Tab. 8. When combined with the image-level feature embedding, the spatial and channel self-relations improve the mIoU by 2.6% and 1.7% on the PASCAL VOC dataset. On the ImageNet-S_{300} dataset, there is also an improvement of 4.5% and 7.7% on mIoU over feature embedding. And cooperating three representations further boosts the performance on all tasks, indicating that self-relations are orthogonal and complementary to image-level feature embeddings for self-supervised learning.

Cooperation between L_I and L_c. Tab. 8 shows that L_p alone performs better than $L_p + L_I$ or $L_p + L_c$ on the PASCAL VOC dataset. However, using $L_p + L_I + L_c$ performs better than L_p. This phenomenon is because utilizing image-level embedding (L_I) and channel self-relation (L_c) have their limits, while their cooperation can mitigate them. The details are as follows: 1) Regarding L_c, modeling channel self-relations requires meaningful and diverse channel features as the foundation. However, solely relying on L_c cannot adequately optimize the channel features and may lead to model collapse, where an example is that each channel encodes the same features. In comparison, L_I facilitates learning diverse and meaningful channel features, thus addressing the limitation mentioned above of L_c. 2) The L_I harms spatial features. We validate this by examining the F-measure [98] that ignores the semantic categories. Tab. 9 shows a decrease in IoU when comparing L_p+L_I with L_I, indicating that L_I impairs spatial features. We assume L_I makes representations less discriminable in the spatial dimension than L_p. However, by using L_c simultaneously, we promote learning more accurate spatial features, mitigating the drawback caused by using L_I.

Cooperating with patch-level embeddings. We also verify the orthogonality of self-relation representation to patch-level embeddings in Tab. 10. As a baseline, we add a clustering loss using patch-level embeddings to DINO, denoted by DINO+. DINO+ consistently advances DINO, showing the effectiveness of patch-level embedding. Compared to DINO+, the self-relation improves the mIoU by 0.9% and 1.2% on PASCAL VOC and ImageNet-S datasets. Cooperating two representations further brings constant improvements over DINO+, *e.g.,* achieving 2.4% and 4.8% gains on mIoU for PASCAL VOC and ImageNet-S datasets. These

TABLE 12
The effect of different numbers of heads M for spatial self-relation.

M	VOC SEG		ImageNet-S^{PT}_{300}	
	mIoU	mAcc	val	test
1	72.4	84.0	38.7	39.3
3	72.7	84.8	38.9	39.4
6	**73.5**	84.7	**41.2**	**42.0**
12	73.4	**85.1**	40.8	41.7
16	72.5	84.3	39.3	39.8

TABLE 13
The effect of different t_p and t_c in Equ. (2) and Equ. (4).

t_p	t_c	VOC SEG		ImageNet-S^{PT}_{300}	
		mIoU	mAcc	val	test
0.50	0.50	72.0	84.2	36.7	36.7
0.50	**0.10**	**73.5**	**84.7**	**41.2**	**42.0**
0.50	0.01	70.4	82.7	33.6	34.6
1.00	0.10	70.2	83.1	36.7	38.2
0.50	**0.10**	73.5	84.7	**41.2**	**42.0**
0.10	0.10	**73.7**	**85.0**	39.9	40.8

TABLE 14
The effect of different α and β in Equ. (6) when cooperating the SERE with iBOT [18]. All models are pre-trained for 100 epochs on ImageNet-1K.

α	β	Classification ImageNet-1K		Segmentation VOC		ImageNet-S_{PT}	
		Top-1	Top-5	mIoU	mAcc	val	test
0.20	0.20	81.3	95.7	80.7	89.9	39.9	39.3
0.20	0.50	**81.5**	**95.8**	**81.2**	**90.0**	41.0	40.3
0.20	1.00	81.3	95.8	80.9	89.8	**41.7**	**41.8**
0.10	0.50	81.3	95.8	80.9	89.5	40.7	40.5
0.20	0.50	**81.5**	**95.8**	**81.2**	**90.0**	**41.0**	**40.3**
0.80	0.50	81.3	95.8	80.8	89.7	40.3	40.1

TABLE 15
The effect of the asymmetric losses in Equ. (3) and Equ. (5).

	VOC SEG		ImageNet-S^{PT}_{300}	
	mIoU	mAcc	val	test
DNIO baseline	68.1	81.1	28.8	29.6
+SERE symmetry	72.1	84.4	37.1	37.9
+SERE asymmetric	**73.5**	**84.7**	**41.2**	**42.0**

results indicate that the self-relation is complementary to patch-level embedding for self-supervised ViT.

Comparison between self-relation and batch-relation. A related work, Barlow [87], models channel relation in the whole batch, *i.e.,* batch-relation. In comparison, the proposed SERE computes self-relation within a single image. To verify the advantage of self-relation over batch-relation, we pre-train the ViT-S/16 with the two forms of relation, respectively. As shown in Tab. 11, compared to the batch-relation, the self-relation improves mIoU by 0.3% and 3.3% on the PASCAL VOC and ImageNet-S_{300} datasets. These results show that self-relation is more suitable for the training of ViT over batch-relation.

Effect of multi-head. We utilize the multi-head spatial self-relation following the MHSA module in ViT. Tab. 12 shows the effect of different numbers of heads M in spatial self-relation. Compared to the single-head version, increasing M to 6 brings the largest performance gain of 1.1% on mIoU for the PASCAL VOC dataset. $M = 12$ achieves limited extra gains, while $M = 16$ suffers a rapid performance drop. More heads enable diverse spatial self-relation, but the number of channels used for calculating each self-relation is reduced. Too many heads result in inaccurate estimation of self-relation, hurting the representation quality. So we default set the number of heads to 6 to balance the diversity and quality of spatial self-relation.

Effect of sharpness. The temperature terms in Equ. (2) and Equ. (4) control the sharpness of the self-relation distributions. A small temperature sharpens the distributions, while a large temperature softens the distributions. In Tab. 13, we verify the effectiveness of temperatures for both spatial and channel self-relations. For the channel self-relation, decreasing temperature from 0.1 to 0.01 results in a rapid performance drop from 73.5% to 70.4% on mIoU for the PASCAL VOC dataset. And increasing it from 0.1 to 0.5 also degrades the mIoU from 73.5% to 72.0%. Therefore, we choose 0.1 as the default temperature for the channel self-relation. For the spatial self-relation, the temperature 0.5 performs better than 1.0, and changing the temperature from 0.5 to 0.1 has a limited difference. We set the default temperature of spatial self-relation to 0.5 because a temperature of 0.5 achieves slightly better performance on the large-scale ImageNet-S dataset.

Effect of loss weights. The α and β in Equ. (6) determine the relative importance of spatial and channel self-relations, respectively. Tab. 14 shows that the SERE is robust to different α and β. Among different weights, the combination of $\alpha = 0.2$ and $\beta = 0.5$ achieves the best performance on the classification task and competitive performances on the segmentation task. Therefore, we use this combination as the default setting.

Effect of asymmetric loss. The asymmetric structure has been proven effective for non-contrastive loss [5], [6] when using image-level embedding as the representation. To verify if self-relation representations also benefit from the asymmetric structure, we compare the asymmetric and symmetry structures for the self-relation based loss in Tab. 15. Self-relation improves the DINO baseline with both asymmetric and symmetry structures. The symmetrical structure outperforms the DINO on PASCAL VOC and ImageNet-S_{300} datasets with 4.0% and 8.3% on mIoU. The asymmetric structure further advances symmetric structure by 1.4% and 4.1% on mIoU for the PASCAL VOC and ImageNet-S_{300} datasets. Therefore, though the asymmetric structure is not indispensable for self-relation, it still benefits the pre-training with self-relation.

Adaptability to convolutional neural networks. Using self-relation for self-supervised learning is inspired by the properties of ViT. Still, we wonder if the self-relation representation could benefit self-supervised learning on convolutional neural networks (CNN). To verify this, we pre-train the ResNet-50 [9] with DINO and SERE, respectively. The training details are shown in the supplementary. As shown in Tab. 16, SERE improves DINO by 0.7% and 0.8% on mIoU for the semantic segmentation task on the PASCAL VOC and ImageNet-S_{300} datasets compared to DINO. Though designed for ViT, the self-relation still improves the representation quality of the CNN. Meanwhile, the improvement on CNN is relatively small compared to that on ViT, showing that the self-relation is more suitable for ViT.

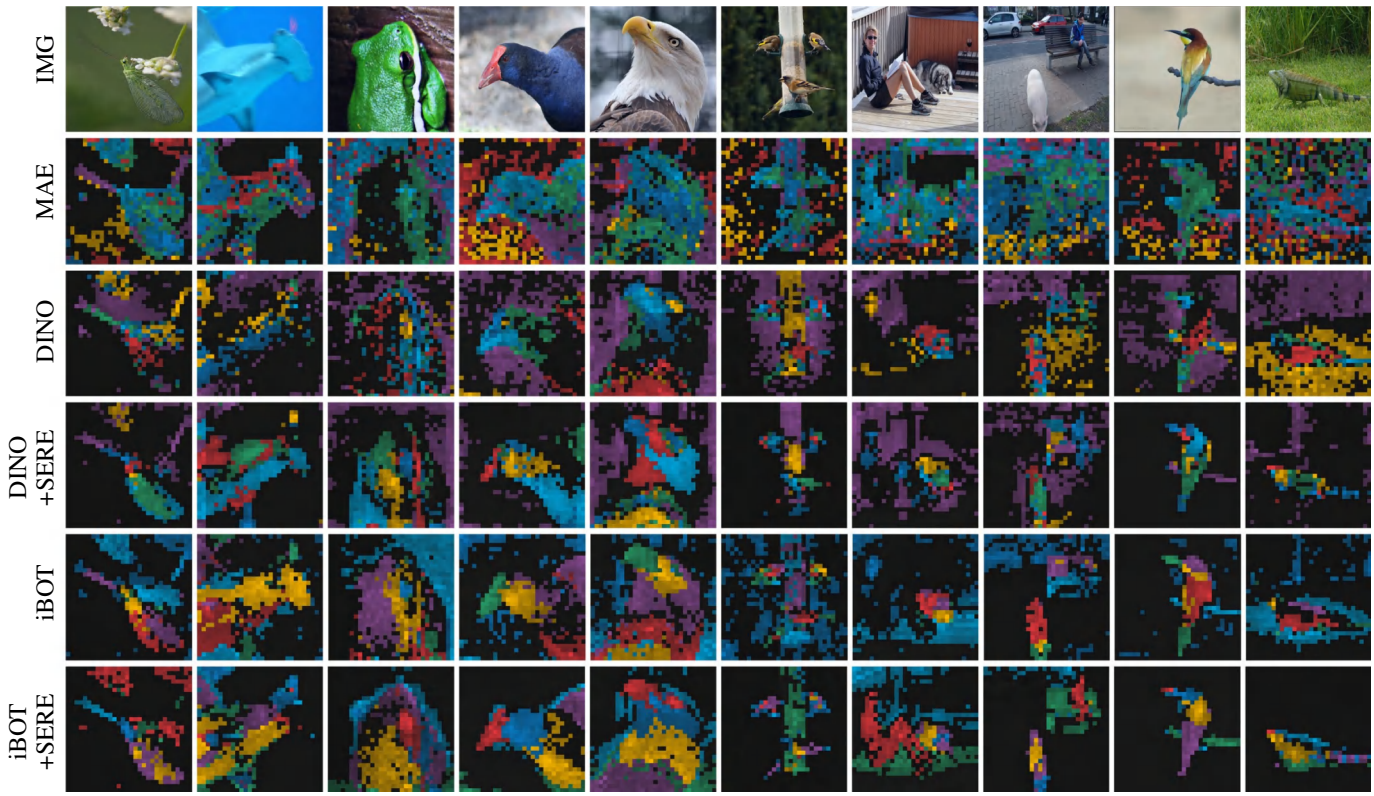


Fig. 4. Visualization for attention maps from the last block of the pre-trained ViT-S/16. We extract the attention maps of the CLS token on other patch-level tokens. Different colors indicate the regions focused by different heads.

TABLE 16
The effect of self-relation representation on CNN. DINO and SERE are trained with the ResNet-50 network.

	VOC SEG		ImageNet-S^{PT}_{300}	
	mIoU	mAcc	val	test
DINO (ResNet-50)	61.6	74.6	20.2	19.9
+SERE (ResNet-50)	**62.5**	**75.0**	**20.9**	**20.7**

4.4 Analysis and Visualization

Invariance on self-relations. The importance of learning representations invariant to image augmentations, *e.g.,* scaling, shifting, and color jitter, has been validated in self-supervised learning [99], [100], [101], [102], [103], [104]. However, existing methods focus on the invariance of feature embeddings but do not consider the invariance of spatial/channel relations, which are also important properties of ViT. In contrast, our proposed SERE can enhance the invariance of spatial/channel relations. To verify this, we measure the averaged differences between self-relations of different views. As shown in Fig. 6, we obverse that SERE significantly narrows the self-relation differences in both the spatial and channel dimensions. The visualizations in Fig. 5 also show that the SERE pre-trained model produces smaller spatial self-relation differences on the overlapping regions of two views. A smaller difference means a higher invariance. Thus, these results indicate that SERE makes the ViT capture self-relations with stronger invariance to image augmentations.

Visualization of attention maps. In Fig. 4, we visualize the attention maps from the last block of ViT. These visualizations demonstrate that SERE produces more precise and less noisy attention maps than various methods, including MIM-based methods, *i.e.,* MAE [14]and iBOT [18]. MAE produces noisy attention maps that highlight almost all tokens in an image. In comparison, the attention maps of SERE mainly focus on semantic objects. For instance, the third column of Fig. 4 shows that SERE can locate the frog, but MAE primarily focuses on the background. Moreover, compared to iBOT and DINO, SERE generates atten-

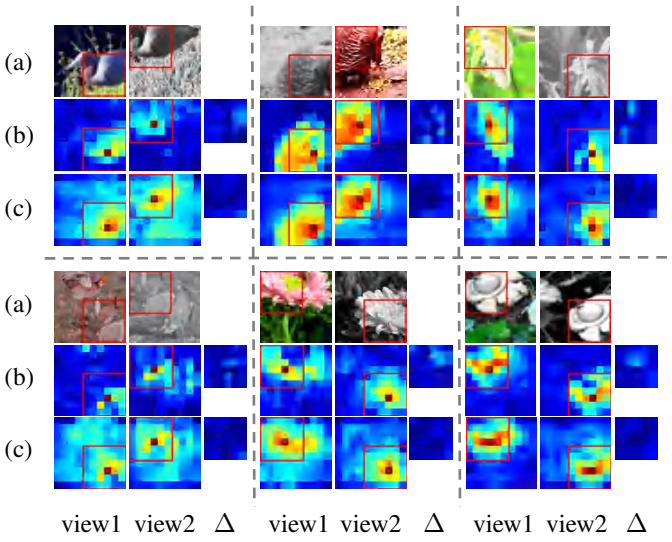


Fig. 5. The differences between spatial self-relations of two views. (a) Two views from each image. (b) The spatial self-relation generated by DINO. (c) The spatial self-relation generated by SERE. View1 and view2 mean the self-relations of two views generated from an image. The Δ is the difference between self-relations in the overlapping region, which is indicated by red boxes. We give the details of the visualization method in the supplementary.

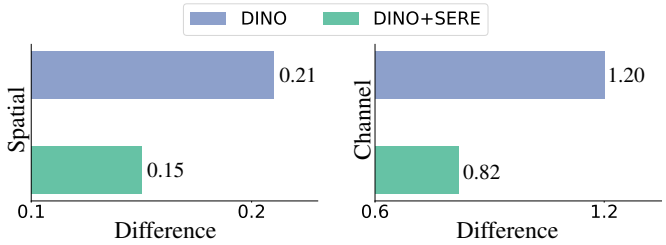


Fig. 6. The average differences of spatial (left) and channel (right) self-relations between two views on the val set of ImageNet-S. We show the calculation details in the supplementary.

tion maps that locate objects more accurately. For instance, in the seventh and eighth columns of Fig. 4, SERE discovers the persons missed by iBOT.

Comparison between spatial self-relation and MIM. Both spatial self-relation and MIM act on the spatial dimension, but their effects significantly differ. MIM enhances the token-level representations, while spatial self-relation focuses on improving the ability to model inter-token relations. We support this argument with the following points: 1) As depicted in Fig. 4, SERE generates more precise and less noisy attention maps than MAE [14]and iBOT [18]. The attention maps of ViT can reflect the ability to model inter-token relations because attentions are calculated as token-level relations between query and key. Thus this observation indicates that SERE provides models with a stronger ability to capture inter-token relations. In Fig. 6, we show that SERE enhances the invariance of spatial self-relation to different image augmentations. 3) As shown in Tab. 6, SERE achieves consistent improvements compared to different MIM-based methods, strongly confirming the effectiveness of SERE compared to MIM. For example, cooperating with SERE improves iBOT by 0.4% Top-1 accuracy, as shown in Tab. 1.

5 Conclusions

In this paper, we propose a feature self-relation based self-supervised learning scheme to enhance the relation modeling ability of self-supervised ViT. Specifically, instead of directly using feature embedding as the representation, we propose to use spatial and channel self-relations of features as representations for self-supervised learning. Self-relation is orthogonal to feature embedding and further boosts existing self-supervised methods. We show that feature self-relation improves the self-supervised ViT at a fine-grained level, benefiting multiple downstream tasks, including image classification, semantic segmentation, object detection, and instance segmentation.

Acknowledgements. This work is funded by NSFC (NO. 62225604, 62176130), and the Fundamental Research Funds for the Central Universities (Nankai Universitiy, 070-63233089). Computation is supported by the Supercomputing Center of Nankai University.

References

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[3] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *IEEE Conf. Comput. Vis. Pattern Recog.*, July 2017.

[4] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Adv. Neural Inform. Process. Syst.*, 2020.

[5] X. Chen and K. He, "Exploring simple siamese representation learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2021.

[6] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Ávila Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in *Adv. Neural Inform. Process. Syst.*, 2020.

[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020.

[8] H. SUN and M. LI, "Enhancing unsupervised domain adaptation by exploiting the conceptual consistency of multiple self-supervised tasks," *SCIENCE CHINA Information Sciences*, vol. 66, no. 4, pp. 142 101–, 2023.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[10] S. Gao, Z.-Y. Li, Q. Han, M.-M. Cheng, and L. Wang, "Rf-next: Efficient receptive field search for convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *Int. Conf. Learn. Represent.*, 2021.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Int. Conf. Comput. Vis.*, 2021.

[13] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Int. Conf. Comput. Vis.*, 2021.

[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2022, pp. 16 000–16 009.

[15] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Int. Conf. Comput. Vis.*, October 2021.

[16] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, "Self-supervised learning with swin transformers," *arXiv preprint arXiv:2105.04553*, 2021.

[17] H. Lu, Y. Huo, M. Ding, N. Fei, and Z. Lu, "Cross-modal contrastive learning for generalizable and efficient image-text retrieval," *Machine Intelligence Research*, pp. 1–14, 2023.

[18] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "ibot: Image bert pre-training with online tokenizer," *Int. Conf. Learn. Represent.*, 2022.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020.

[20] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[21] O. J. Hénaff, S. Koppula, J.-B. Alayrac, A. van den Oord, O. Vinyals, and J. a. Carreira, "Efficient visual pretraining with contrastive detection," in *Int. Conf. Comput. Vis.*, October 2021, pp. 10 086–10 096.

[22] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Adv. Neural Inform. Process. Syst.*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[23] K. Kim, B. Wu, X. Dai, P. Zhang, Z. Yan, P. Vajda, and S. J. Kim, "Rethinking the self-attention in vision transformers," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, June 2021, pp. 3071–3075.

[24] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018.

[25] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2009.

[26] S. Gao, Z.-Y. Li, M.-H. Yang, M.-M. Cheng, J. Han, and P. Torr, "Large-scale unsupervised semantic segmentation," *arXiv preprint arXiv:2106.03149*, 2021.

[27] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 649–666.

[28] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *IEEE Conf. Comput. Vis. Pattern Recog.*, July 2017.

[29] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Eur. Conf. Comput. Vis.*, 2016.

[30] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Int. Conf. Learn. Represent.*, 2018.

[31] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Int. Conf. Comput. Vis.*, December 2015.

[32] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *International Conference on Machine Learning (ICML)*, 2008.

[33] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2016.

[34] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *Int. Conf. Comput. Vis.*, Oct 2017.

[35] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018.

[36] Y. Zhao, G. Wang, C. Luo, W. Zeng, and Z.-J. Zha, "Self-supervised visual representations learning by contrastive mask prediction," in *Int. Conf. Comput. Vis.*, October 2021, pp. 10 160–10 169.

[37] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," in *Int. Conf. Comput. Vis.*, October 2021, pp. 9588–9597.

[38] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, and Y. LeCun, "Decoupled contrastive learning," *arXiv preprint arXiv:2110.06848*, 2021.

[39] W.-C. Wang, E. Ahn, D. Feng, and J. Kim, "A review of predictive and contrastive self-supervised learning for medical images," *Machine Intelligence Research*, pp. 483–513, 2023.

[40] L. Wang, H. Xu, and W. Kang, "Mvcontrast: Unsupervised pretraining for multi-view 3d object recognition," *Machine Intelligence Research*, pp. 1–12, 2023.

[41] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Eur. Conf. Comput. Vis.*, 2018.

[42] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020.

[43] A. YM., R. C., and V. A., "Self-labelling via simultaneous clustering and representation learning," in *Int. Conf. Learn. Represent.*, 2020.

[44] S. A. Koohpayegani, A. Tejankar, and H. Pirsiavash, "Mean shift for self-supervised learning," in *Int. Conf. Comput. Vis.*, October 2021, pp. 10 326–10 335.

[45] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for self-supervised representation learning," in *International Conference on Machine Learning (ICML)*, 2021, pp. 3015–3024.

[46] Y. Tian, X. Chen, and S. Ganguli, "Understanding self-supervised learning dynamics without contrastive pairs," in *International Conference on Machine Learning (ICML)*, 2020.

[47] C. Ge, Y. Liang, Y. Song, J. Jiao, J. Wang, and P. Luo, "Revitalizing cnn attentions via transformers in self-supervised visual representation learning," in *Adv. Neural Inform. Process. Syst.*, 2021.

[48] Q. Hu, X. Wang, W. Hu, and G.-J. Qi, "Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2021, pp. 1074–1083.

[49] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2021, pp. 16 684–16 693.

[50] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, and P. Luo, "Detco: Unsupervised contrastive learning for object detection," in *Int. Conf. Comput. Vis.*, October 2021, pp. 8392–8401.

[51] Z. Dai, B. Cai, Y. Lin, and J. Chen, "Up-detr: Unsupervised pre-training for object detection with transformers," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2021, pp. 1601–1610.

[52] B. Roh, W. Shin, I. Kim, and S. Kim, "Spatilly consistent representation learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[53] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Int. Conf. Comput. Vis.*, 2021.

[54] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Adv. Neural Inform. Process. Syst.*, 2021.

[55] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2T: Pyramid pooling transformer for scene understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[56] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 10 347–10 357.

[57] C. Li, J. Yang, P. Zhang, M. Gao, B. Xiao, X. Dai, L. Yuan, and J. Gao, "Efficient self-supervised vision transformers for representation learning," in *Int. Conf. Learn. Represent.*, 2022.

[58] P. Zhou, Y. Zhou, C. Si, W. Yu, T. K. Ng, and S. Yan, "Mugs: A multi-granular self-supervised learning framework," in *arXiv preprint arXiv:2203.14415*, 2022.

[59] Z. Li, Z. Chen, F. Yang, W. Li, Y. Zhu, C. Zhao, R. Deng, L. Wu, R. Zhao, M. Tang, and J. Wang, "MST: Masked self-supervised transformer for visual representation," in *Adv. Neural Inform. Process. Syst.*, 2021.

[60] H. Bao, L. Dong, S. Piao, and F. Wei, "BEit: BERT pre-training of image transformers," in *Int. Conf. Learn. Represent.*, 2022.

[61] S. Gao, P. Zhou, M.-M. Cheng, and S. Yan, "Towards sustainable self-supervised learning," *arXiv preprint arXiv:2210.11016*, 2022.

[62] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *arXiv preprint arXiv:2202.03026*, 2022.

[63] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2022.

[64] L. Wang, F. Liang, Y. Li, H. Zhang, W. Ouyang, and J. Shao, "Repre: Improving self-supervised vision transformer with reconstructive pre-training," *arXiv preprint arXiv:2201.06857*, 2022.

[65] S. Atito, M. Awais, and J. Kittler, "Sit: Self-supervised vision transformer," *arXiv preprint arXiv:2104.03602*, 2021.

[66] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," *arXiv preprint arXiv:2112.09133*, 2021.

[67] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 213–229.

[68] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Int. Conf. Comput. Vis.*, October 2019.

[69] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2019.

[70] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Eur. Conf. Comput. Vis.*, 2018.

[71] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Int. Conf. Comput. Vis.*, October 2019.

[72] X. Li, J. Wu, H. Fang, Y. Liao, F. Wang, and C. Qian, "Local correlation consistency for knowledge distillation," in *Eur. Conf. Comput. Vis.*, 2020.

[73] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Int. Conf. Learn. Represent.*, 2017.

[74] Y. Chen, N. Wang, and Z. Zhang, "Darkrank: Accelerating deep metric learning via cross sample similarities transfer," *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 32, no. 1, Apr. 2018.

[75] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2019.

[76] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2019.

[77] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2022, pp. 12 319–12 328.

[78] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," in *Int. Conf. Learn. Represent.*, 2022.

[79] O. Siméoni, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Unsupervised object discovery for instance recognition," in *Winter Conference on Applications of Computer Vision*, 2018.

[80] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce, "Localizing objects with self-supervised transformers and no labels," in *Brit. Mach. Vis. Conf.*, November 2021.

[81] Y. Wang, X. Shen, S. X. Hu, Y. Yuan, J. L. Crowley, and D. Vaufreydaz, "Self-supervised transformers for unsupervised object discovery using normalized cut," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2022.

[82] M. Ki, Y. Uh, J. Choe, and H. Byun, "Contrastive attention maps for self-supervised co-localization," in *Int. Conf. Comput. Vis.*, October 2021, pp. 2803–2812.

[83] I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bursuc, K. Karantzalos, and N. Komodakis, "What to hide from your students: Attention-guided masked image modeling," *arXiv preprint arXiv:2203.12719*, 2022.

[84] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Eur. Conf. Comput. Vis. Worksh.*, 2016, pp. 685–701.

[85] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Adv. Neural Inform. Process. Syst.*, vol. 28, 2015.

[86] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Int. Conf. Comput. Vis.*, December 2015.

[87] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," *arXiv preprint arXiv:2103.03230*, 2021.

[88] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.

[89] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[90] L. Liu, Q. Huang, S. Lin, H. Xie, B. Wang, X. Chang, and X. Liang, "Exploring inter-channel correlation for diversity-preserved knowledge distillation," in *Int. Conf. Comput. Vis.*, October 2021, pp. 8271–8280.

[91] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inform. Process. Syst.*, 2012.

[92] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Int. Conf. Learn. Represent.*, 2019.

[93] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep. 0, 2009.

[94] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729, 2008.

[95] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Int. Conf. Comput. Vis. Worksh.*, 2013.

[96] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018.

[97] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Eur. Conf. Comput. Vis.*, September 2018.

[98] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2015.

[99] S. Purushwalkam Shiva Prakash and A. Gupta, "Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases," *Adv. Neural Inform. Process. Syst.*, vol. 33, 2020.

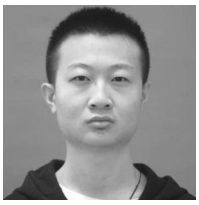
[100] M. Patrick, Y. M. Asano, P. Kuznetsova, R. Fong, J. a. F. Henriques, G. Zweig, and A. Vedaldi, "On compositions of transformations in contrastive self-supervised learning," in *Int. Conf. Comput. Vis.*, 2021.

[101] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020.

[102] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *Int. Conf. Learn. Represent.*, 2022.

[103] L. Ericsson, H. Gouk, and T. M. Hospedales, "Why do self-supervised models transfer? investigating the impact of invariance on downstream tasks," 2022.

[104] X. Wang, K. He, and A. Gupta, "Transitive invariance for self-supervised visual representation learning," in *Int. Conf. Comput. Vis.*, 2017.



Zhong-Yu Li is a Ph.D. student from the college of computer science, Nankai university. He is supervised via Prof. Ming-Ming cheng. His research interests include deep learning, machine learning and computer vision.

Shanghua Gao is a Ph.D. candidate in Media Computing Lab at Nankai University. He is supervised via Prof. Ming-Ming Cheng. His research interests include computer vision and representation learning.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012, and then worked with Prof. Philip Torr in Oxford for 2 years. Since 2016, he is a full professor at Nankai University, leading the Media Computing Lab. His research interests include computer vision and computer graphics. He received awards, including ACM China Rising Star Award, IBM Global SUR Award, *etc.* He is a senior member of the IEEE and on the editorial boards of IEEE TPAMI and IEEE TIP.