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 self-relation (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.

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

I. 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.

In this work, we propose to utilize the feature **S**elf-RElation (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 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], 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.

II. RELATED WORK

A. Self-Supervised Learning

Self-supervised learning aims at learning rich representations without any human annotations. Early works utilized handcrafted 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.

B. 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.

C. 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.

D. 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.

E. 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.

III. Method

A. Overview

In this work, we focus on the instance discriminative self-supervised learning pipeline [4], [13]. First, we briefly revisit 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 \( r_1(x) \) and \( r_2(x) \) with augmentations \( r_1 \) and \( r_2 \). 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. First, two views are sent to the encoder network to extract the feature embeddings \( r_1 \) and \( r_2 \) with \( H \times W \) local patches and \( C \) channels. According to the training objective of self-supervised learning methods, the feature embeddings are then transformed with \( P \) 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:

\[
L_I = R(P(r_1), P(r_2)),
\]

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 \( P \). 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-relations for self-supervised ViT, i.e., spatial and channel self-relations.

B. 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 \( P \) 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.

Generating Spatial Self-Relation Representation: Given the feature embeddings \( r_1 = f_1(r_1(x)) \in \mathbb{R}^{C \times HW} \) and \( r_2 = f_2(r_2(x)) \in \mathbb{R}^{C \times HW} \) 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 \( \mathcal{O} \) [26] to uniformly sample \( H_s \times W_s \) points from the overlapping region of \( p_1 \) and \( p_2 \). 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 \( \mathcal{O} \) are shown in the supplementary. For one view, e.g., \( p_1 \in \mathbb{R}^{C \times HW} \), we calculate the spatial self-relation \( A_p(p_1) \) as follows:

\[
A_p(p_1) = \text{Softmax} \left( \frac{\mathcal{O}(p_1)^T \cdot \mathcal{O}(p_1)}{\sqrt{C}} / t_p \right),
\]

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

\[ L_p = R_c(\mathcal{G}(A_p(p_1)), A_p(g_p(p_2))), \]

where \( R_c \) is the cross-entropy loss, \( \mathcal{G} \) is the stop-gradient operation to avoid training collapse following [5], and \( g_p \) is the projection 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 Table XII.

### C. 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 (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 \( A_c(c_1) \in \mathbb{R}^{C \times C} \) as follows:

\[ A_c(c_1) = \text{Softmax} \left( \frac{c_1^T \cdot c_1}{H \cdot W} / t_c \right), \]

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

---

**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 (3) and (5), enforce consistency between self-relations of different views. For spatial self-relation, only the features in the overlapping region are considered. \( \mathcal{O} \) means the operation of extracting features from the overlapping region between two views in (2), where the red dotted box indicates the overlapping region.

---

**Fig. 3.** The region-aligned sampling operation for spatial self-relation. \( r_1(x) \) and \( r_2(x) \) 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.
self-relation based loss in (3), we give the non-contrastive loss using channel self-relation as follows:

\[ L_c = R_c(G(A_c(e_1)), A_c(g_c(e_2))), \]  

(5)

where the \( R_c \) is the cross-entropy loss, and \( g_c \) is a prediction head with the same structure as \( g_p \) in (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 (4) has no spatial dimension.

D. 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:

\[ L = L_I + \alpha L_p + \beta L_c, \]  

(6)

where the spatial and channel losses are weighted by \( \alpha \) and \( \beta \), and \( L_I \) is the loss using image-level embeddings, e.g., the clustering-based loss in DINO [13]. We show in Table VIII 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 representations for self-supervision losses, as these representations reveal different properties of features. The summarized loss function is as follows:

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 \( \alpha \) and \( \beta \) 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.

B. 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 Table I, 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 Table III(b). Moreover, using ViT-B/16, SERE surpasses iBOT by 0.4% in Top-1 accuracy, as shown in Table I. These results demonstrate that SERE enhances the category-related representation ability of ViT.

IV. EXPERIMENTS

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

A. 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-S300 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 \( H/IV \), the number of heads \( M \) in spatial self-relation is set to 6 by default. The \( t_p \) and \( t_c \) in (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 \( \alpha \) and \( \beta \) in (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 \( \alpha \) and \( \beta \) 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.
pre-trained 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. Table IV 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.

**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. Table V shows that SERE performs better on Top-1 accuracy over DINO/iBOT. We conclude that SERE enhances the relation modeling ability, enabling ViT with much stronger shape-related representations.

**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. Table II reports the mIoU and mAcc on the validation set.
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. Table III(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× schedule and a batch size of 16. Table II reports the bounding box AP (APb) and the segmentation mask AP (APm) on the COCO val2017 set. Compared to DINO, SERE improves by 0.6% on APb and 0.5% on APm, 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 Table VI, 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.

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 $L_2$ in (6). We pre-train models on the ImageNet-S300 dataset with 100 epochs to save computation costs, and other training settings are constant with baseline methods. As shown in Table VII, 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-S300 dataset, SERE gains 5.1% on mIoU over MoCo v3.

C. Ablation Studies
To save computational costs for the ablation study, we pretrain all models on the ImageNet-S300 [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-S300 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 Table VIII, 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.

Cooperating With Image-Level Embeddings: We verify the orthogonality between self-relations and image-level embeddings, as shown in Table VIII. 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-S300 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 $L_c$: Table VIII 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. Table IX 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 Table X. As a baseline, we add a clustering loss using patch-level embeddings to DINO [13]. The boldface values represent the optimal values.

Table IX

|                     | L_p       | L_p + L_I | L_p + L_I + L_c |
|---------------------|-----------|-----------|-----------------|
| IoU                 | 87.1      | 86.7      | 87.7            |

The F-measure ignores semantic categories.

Table X

|                     | VOC SEG   | ImageNet-S^PT_100 |
|---------------------|-----------|-------------------|
|                     | mIoU      | mAcc              |
|                     | val       | test              |
| DINO                | 68.1      | 81.1              |
| DINO+               | 72.6      | 84.3              |
| SERE                | 73.5      | 84.7              |
| DINO + SERE         | 75.0      | 86.1              |

DINO+ indicates adding the clustering loss using patch-level embeddings to DINO [13]. The boldface values represent the optimal values.

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.
The temperature terms in $t(4)$ to datasets. These results show that self-relation is shows (4) in spatial self-

| $M$   | VOC SEG | ImageNet-S$^m_{100}$ |
|-------|---------|----------------------|
|       | mIoU    | mAcc                 |
| 1     | 72.4    | 84.0                 |
| 3     | 72.7    | 84.8                 |
| 6     | 73.5    | 84.7                 |
| 12    | 73.4    | 85.1                 |
| 16    | 72.5    | 84.3                 |

The boldface values represent the optimal values.

**TABLE XIV**

The effect of different $\alpha$ and $\beta$ in (6) when cooperating the SERE with iBOT [18]

| $\alpha$ | $\beta$ | Classification | Segmentation |
|----------|---------|----------------|--------------|
| 0.20     | 0.50    | 81.5           | 95.8         |
| 0.20     | 1.00    | 81.3           | 95.8         |
| 0.10     | 1.00    | 81.3           | 95.8         |
| 0.20     | 5.00    | 81.3           | 95.8         |

All models are pre-trained for 100 epochs on ImageNet-1K. The boldface values represent the optimal values.

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 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 Table XI, compared to the batch-relation, the self-relation improves mIoU by 0.3% and 3.3% on the PASCAL VOC and ImageNet-S 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. Table XII 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 (2) and (4) control the sharpness of the self-relation distributions. A small temperature sharpens the distributions, while a large temperature softens the distributions. In Table XIII, 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 $\alpha$ and $\beta$ in (6) determine the relative importance of spatial and channel self-relations, respectively. Table XIV shows that the SERE is robust to different $\alpha$ and $\beta$. 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 Table XV. Self-relation improves the DINO baseline with both asymmetric and symmetry structures. The symmetrical structure outperforms the DINO on PASCAL VOC and ImageNet-S 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 datasets. Therefore, though the asymmetric
TABLE XV
THE EFFECT OF THE ASYMMETRIC LOSSES IN (3) AND (5)

| Method                     | VOC SEG | ImageNet-S | PASCAL VOC-2012 |
|----------------------------|---------|------------|-----------------|
|                            | mIoU    | mAcc       | val             |
| DINO baseline              | 68.1    | 81.1       | 28.8            |
| +SERE symmetry             | 72.1    | 84.4       | 37.1            |
| +SERE asymmetric           | 73.5    | 84.7       | 41.2            |

The boldface values represent the optimal values.

TABLE XVI
THE EFFECT OF SELF-RELATION REPRESENTATION ON CNN

| Method                      | VOC SEG | ImageNet-S | PASCAL VOC-2012 |
|-----------------------------|---------|------------|-----------------|
|                            | mIoU    | mAcc       | val             |
| DINO (ResNet-50)            | 61.6    | 74.6       | 20.2            |
| +SERE (ResNet-50)           | 62.5    | 75.0       | 20.9            |

DINO and SERE are trained with the ResNet-50 network. The boldface values represent the optimal values.

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 $\Delta$ 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.

D. 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 observe 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 attention 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

... additional content...

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
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 Table VI, 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 Table I.

V. CONCLUSION

In this article, 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.

REFERENCES

[1] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.
[2] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. 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 Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5122–5130.
[4] M. Caron, I. Misra, J. Mairal, P. Goyal, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 9912–9924.
[5] X. Chen and K. He, “Exploring simple siamese representation learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 15745–15753.
[6] J.-B. Grill et al., “Bootstrap your own latent - a new approach to self-supervised learning,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 21271–21284.
[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 9726–9735.
[8] H. Sun and M. Li, “Enhancing unsupervised domain adaptation by exploiting the conceptual consistency of multiple self-supervised tasks,” Sci. China Inf. Sci., vol. 66, no. 4, 2023, Art. no. 142101.
[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 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., vol. 45, no. 3, pp. 2984–3002, Mar. 2023.
[11] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Representations, 2021.
[12] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 9992–10002.
[13] M. Caron et al., “Emerging properties in self-supervised vision transformers,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 9630–9640.
[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 16000–16009.
[15] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers,” in Proc. Int. Conf. Comput. Vis., 2021, pp. 9620–9629.
[16] Z. Xie et al., “Self-supervised learning with swin transformers,” 2021, arXiv:2105.04152.
[17] H. Lu, Y. Huo, M. Ding, N. Fei, and Z. Lu, “Cross-modal contrastive learning for generalizable and efficient image-text retrieval,” Mach. Intell. Res., vol. 20, pp. 569–582, 2023.
[18] J. Zhou et al., “iBOT: Image bert pre-training with online tokenizer,” in Proc. Int. Conf. Learn. Representations, 2022.
[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 1597–1607.
[20] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, “Dense contrastive learning for self-supervised visual pre-training,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 3023–3032.
[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 Proc. Int. Conf. Comput. Vis., 2021, pp. 10086–10096.
[22] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, “Do vision transformers see like convolutional neural networks?,” in Proc. Adv. Neural Inf. Process. Syst., 2021.
[23] K. Kim et al., “Rethinking the self-attention in vision transformers,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2021, pp. 3071–3075.
[24] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6154–6162.
[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,” 2021, arXiv:2106.03149.
[27] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 649–666.
[28] G. Larsson, M. Maire, and G. Shakhnarovich, “Colorization as a proxy task for visual understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 840–849.
[29] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 69–84.
[30] S. Giradis, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” in Proc. Int. Conf. Learn. Representations, 2018.
[31] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in Proc. Int. Conf. Comput. Vis., 2015, pp. 1422–1430.
[32] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in Proc. Int. Conf. Mach. Learn., 2008, pp. 1096–1103.
[33] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2536–2544.
[34] M. Noroozi, H. Pirsiavash, and P. Favaro, “Representation learning by learning to count,” in Proc. Int. Conf. Comput. Vis., 2017, pp. 5899–5907.
[35] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3733–3742.
[36] Y. Zhao, G. Wang, C. Luo, W. Zeng, and Z.-J. Zha, “Self-supervised visual representations learning by contrastive mask prediction,” in Proc. Int. Conf. Comput. Vis., 2021, pp. 10160–10169.
[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 Proc. Int. Conf. Comput. Vis., 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,” 2021, arXiv:2110.06848.
[39] W.-C. Wang, E. Ahn, D. Feng, and J. Kim, “A review of predictive and self-supervised learning for medical images,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 1597–1607.
[40] L. Wang, H. Xu, and W. Kang, “MVContrast: Unsupervised pretraining for multi-view 3D object recognition,” Mach. Intell. Res., vol. 20, pp. 1–12, 2023.
[94] M.-E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” in Proc. Indian Conf. Comput. Vis. Graph. Image Process., 2008, pp. 722–729.

[95] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3D object representations for fine-grained categorization,” in Proc. Int. Conf. Comput. Vis. Workshop, 2013, pp. 554–561.

[96] G. Van Horn et al., “The inaturalist species classification and detection dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8769–8778.

[97] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in Proc. Eur. Conf. Comput. Vis., 2018, sp:418–434.

[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, Mar. 2015.

[99] S. Purushwalkam Shiva Prakash and A. Gupta, “Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases,” in Proc. Adv. Neural Inform. Process. Syst., 2020, pp. 3407–3418.

[100] M. Patrick et al., “On compositions of transformations in contrastive self-supervised learning,” in Proc. Int. Conf. Comput. Vis., 2021, sp:1726–1736.

[101] I. Misra and L. van der Maaten, “Self-supervised learning of pretext-invariant representations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6707–6717.

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

[103] L. Ericsson, H. Gouk, and T. M. Hospedales, “Why do self-supervised models transfer? Investigating the impact of invariance on downstream tasks,” in Proc. Brit. Mach. Vis. Conf., 2022.

[104] X. Wang, K. He, and A. Gupta, “Transitive invariance for self-supervised visual representation learning,” in Proc. Int. Conf. Comput. Vis., 2017, pp. 1329–1338.

Zhong-Yu Li is currently working toward the PhD degree 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 currently working toward the PhD degree in media computing lab with Nankai University. He is supervised via prof. Ming-Ming Cheng. His research interests include computer vision and representation learning.

Ming-Ming Cheng (Senior Member, IEEE) received the 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 with 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 editorial boards of IEEE TPAMI and IEEE TIP.