Exploring Feature Self-relation for Self-supervised Transformer

Zhong-Yu Li*, Shanghua Gao*, and Ming-Ming Cheng†

TMCC, College of Computer Science, Nankai University
lizhongyu@mail.nankai.edu.cn, shgao@mail.nankai.edu.cn, cmm@nankai.edu.cn

Abstract. Learning representations with self-supervision for convolutional networks (CNN) has proven effective for vision tasks. As an alternative for CNN, vision transformers (ViTs) emerge strong representation ability with the pixel-level self-attention and channel-level feed-forward networks. Recent works reveal that self-supervised learning helps unleash the great potential of ViTs. Still, most works follow self-supervised strategy designed for CNNs, *e.g.,* instance-level discrimination of samples, but they ignore the unique properties of ViTs. We observe that modeling relations among pixels and channels distinguishes ViTs from other networks. To enforce this property, we explore the feature self-relations for training self-supervised ViTs. Specifically, instead of conducting self-supervised learning solely on feature embeddings from multiple views, we utilize the feature self-relations, *i.e.,* pixel/channel-level self-relations, for self-supervised learning. Self-relation based learning further enhance the relation modeling ability of ViTs, resulting in strong representations that stably improve performance on multiple downstream tasks. Our source code will be made publicly available.

Keywords: 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 [39,31,56]. Learning visual representations from un-labeled images has proven to be an effective alternative to the supervised training [54,28,34,17,36,13,35,41], *e.g.,* convolutional networks (CNNs) trained with self-supervision have shown comparable or even better performance than its supervised counterparts [21,**?**,6,53,10,18,19]. Recently, vision transformer (ViTs) [14] emerges stronger representation ability than CNNs on many vision tasks. Pioneering works shift the self-supervision methods designed for CNNs and reveal the great potential of self-supervised ViTs. However, the unique properties of ViTs, *e.g.,* the self-relation modeling ability, are rarely considered by existing methods. We seek to explore the properties of ViTs to improve the training of self-supervised ViTs.

* Equal contribution
† M.M. Cheng (cmm@nankai.edu.cn) is the corresponding author.

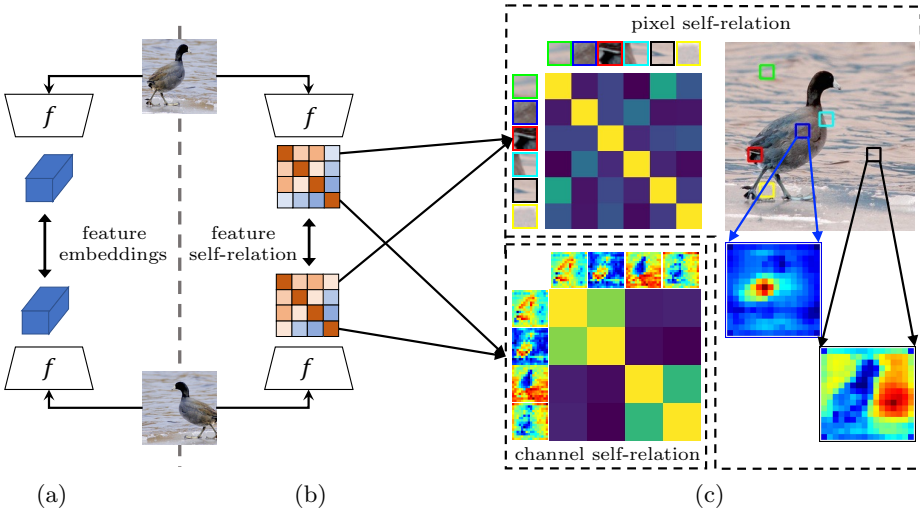

Fig. 1. The illustration of the self-supervised learning using feature embeddings and our proposed feature self-relation. (a) Typical self-supervised learning methods process the feature embeddings in multiple views. (b) We propose to model the feature self-relation that measures the relation inside a view across different dimensions. (c) Two specific forms of self-relation, *i.e.,* the pixel/channel-level self-relation. For pixel self-relation, we select 6 patches that are indicated by differently colored boxes and give their visualized relations and relation matrix of embeddings. For channel self-relation, we show visualized feature maps of 4 channels and the corresponding self-relation matrix.

Typical self-supervised learning methods for ViTs, *e.g.,* DINO [7] and MoCoV3 [11], send multiple views of an image into a ViT network to generate feature representations. Self-supervisions are then implemented on these representations based on the hypothesis that different views of an image share similar representations. Most works present various self-supervision schemes, *e.g.,* contrastive [21], non-contrastive [19], and clustering [6] types, but the forms of feature representations are less explored. Existing feature representations correspond to different granularities of the input image according to task requirements. For example, image-level embeddings are utilized as representations of instance discrimination [9,6,19], and dense tasks, *e.g.,* semantic segmentation, require pixel-level embeddings to learn dense representations [44,22]. We wonder if there exist other forms of representations for ViTs that encode semantics in other dimensions?

The self-relation modeling properties of ViTs inspire us. ViTs model the feature relations on pixel and channel dimensions with the multi-head self-attention (MHSA) and feed forward network (FFN) [37,14,26], respectively. The MHSA aggregates the pixel-level information with the extracted relations among pixels, resulting in stronger relations among pixels with similar semantic context

(see Figure 1(c)). The FFN combines features from different channels, implicitly modeling the feature self-relations in the channel dimension. For instance, Figure 1(c) reveals that feature patterns across channels have different degrees of relations. Feature relation modeling enables ViTs with strong representation ability, which motivates us to use self-relation as a new representation form for self-supervision.

In this work, we propose to utilize the feature self-relation for self-supervised training, enhancing the self-relation modeling properties in ViTs. Following the pixel relation in MHSA and channel relation in FFN of ViTs, we form the pixel-level/channel-level self-relations as representations. Pixel-level self-relation extracts the similarity among pixels of the feature map, representing the relations among semantic regions. Channel-level self-relation models the connection of different channels, where each channel in feature embeddings highlights unique semantic information. Feature self-relation is the representation of a new dimension, which is compatible with existing representation forms, *e.g.,* image-level and pixel-level feature embeddings. As shown in Figure 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 stably improve multiple self-supervised ViT training methods, *e.g.,* DINO [7] and MoCoV3 [11] on various downstream tasks, *e.g.,* object detection [3,31], semantic segmentation [15,56], semi-supervised semantic segmentation [16] and image classification [39]. Our major contributions are summarized as follows:

- We propose to utilize feature self-relations, *i.e.,* pixel/channel-level self-relations, as the representation for self-supervised learning, which fits well with the relation modeling property of ViTs.
- Cooperating with exiting self-supervised methods, our proposed self-relation based method stably boosts ViTs on both image-level classification tasks and pixel-level segmentation tasks.

2 Related Work

2.1 Self-Supervised Learning

Self-supervised learning aims at learning rich representation without any human annotations by training with different pretext tasks, *e.g.,* instance discrimination in contrastive methods [21,46,23,9,24,55,22], sample self-clustering [5,53] and representation alignment [18,10,27]. These methods are based on the feature representations extracted from the encoder network. And self-supervision losses update the encoder according to the representations. A large amount of methods uses image-level feature embeddings [21,9,5]. While some methods explore using representation in more fine-grained dimensions, *e.g.,* pixel [44,50], patch [48], object [22] and region [38] dimension. However, these representations are still the feature embeddings corresponding to different regions of input images. Instead, Zbontar *et al.* [52] explore cross-relations over all samples of a

batch. Different from these lines of works, we further explore the self-relations of samples for self-supervised learning. The self-relations model the relation among pixels and channels of an image, which is a new dimension of representation that is orthogonal to feature embeddings.

2.2 Self-Supervised Vision Transformer

Recently, transformers are generalized to computer vision [14,43] and achieve state-of-the-art performance on many tasks, *e.g.,* image classification, semantic segmentation and object detection. Due to lack of indicative bias, training ViTs requires much more data and tricks [14,40]. To meet the data requirement of ViTs with low annotation cost, recent works have been working on training ViTs with self-supervised learning methods. Chen *et al.* [11] empirically explore the contrastive learning with ViTs. Caron *et al.* [7] propose a framework of self-distillation for ViTs, and find that self-supervised ViTs have the potential to capture explicit information of semantic regions. Xie *et al.* [49] implement contrastive learning in Swin Transformer [32]. These methods follow the pretext of instance discrimination where representations with invariance to transformation are learned by maximizing the similarity among positive samples. Li *et al.* [29] propose to use multi-stage architectures and fine-grained region-level matching to develop efficient self-supervised ViTs. Reconstruction from noisy images to learn representations [8,2,20,57,51,42,30,45,1] has became another popular pipeline for self-supervised ViTs. However, most self-supervised ViTs still follow the framework designed for CNNs and model the relation between feature embeddings of images. Instead, we explore another form of semantic information, *i.e.,* self-relation in more fine-grained dimensions, for self-supervised ViTs.

2.3 Property of Vision Transformer

Recent works have shown that the remarkable success of ViTs on many vision tasks [32,4,12] relay on their strong ability of modeling relations among pixels. Dosovitskiy *et al.* [14] and Kim *et al.* [26] find that the attention attends to semantically relevant regions of images. Raghu *et al.*[37] reveal the representations of ViTs preserve strong spatial information even in the deep layer. They also observe that pixels in ViTs have strong connections with regions that share similar semantics. Caron *et al.* [7] find that self-supervised ViTs capture more explicit semantic region compared to supervised ViTs. These observations indicate that ViTs have the 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 pixel-level and channel-level feature self-relations for self-supervised learning.

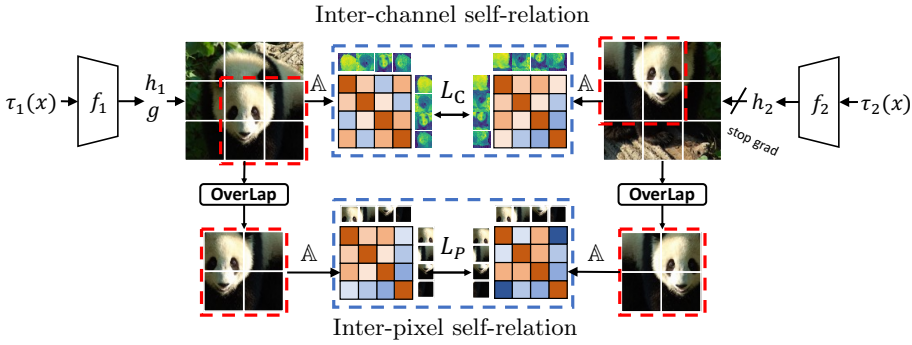

Fig. 2. Overview of our proposed method. The method models the self-relation from pixel and channel dimensions. Given an image x, two views are generated by two random data augmentations, *i.e.,* τ_1 and τ_2. The two views are processed by f_1 and f_2, respectively, where f_2 is the momentum updated network of f_1. We calculate pixel/channel-level relations and force the consistency across different views. The OverLap means extracting the overlapping region between two views, which is indicated by the red dotted box. The $\mathbb{A}$ means the calculating of relations. The h_1 and h_2 are the projection head. The g the is prediction head.

3 Method

3.1 Overview

First, we briefly revisit the framework of common existing self-supervised learning methods [9,21,24,6,7]. 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 self-supervised methods is to maximize the shared information encoded from different views. Firstly, two views are sent to the encoder network to extract feature embeddings $r_1 \in \mathbb{R}^{C\times HW}$ and $r_2 \in \mathbb{R}^{C\times HW}$ with $H \cdot W$ pixels of C channels. According to the needs of self-supervised learning methods, the feature embeddings are then transformed with projection $\mathbb{P}$ to obtain different forms of representations, *e.g.,* image/pixel-level embeddings. Different self-supervised optimization objectives utilize the obtained representations to get the loss as follow:

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

where R means the function that maximizes the consistency across views and can be defined with multiple forms, *e.g.,* contrastive [21], non-contrastive [19], and clustering [6] losses.

Our main focus in this work is to explore new forms of representation transformation $\mathbb{P}$. Motivated by the relation modeling properties in ViTs, instead of directly using feature embeddings, we explore utilizing feature self-relation in multiple dimensions as the representation for self-supervised learning on ViTs.

In the following sections, we introduce two specific forms of self-relation representations for self-supervised ViTs, *i.e.,* pixel/channel-level self-relation.

3.2 Pixel Self-relation

In ViTs, the MHSA module models the relation among pixels with self-attention. Caron *et al.* [7] observe that the self-attention mechanism in ViTs helps locate and segment semantic regions of images. For ViTs with strong representation, pixels that share similar semantic context have stronger relations in feature embedding. Learning representations with accurate pixel-level relation is crucial for many pixel-level dense prediction tasks [44,22], *e.g.,* object detection and semantic segmentation. So we propose to enhance the relation modeling ability of ViTs by utilizing pixel-level self-relations for self-supervised training. Specifically, we transform the feature embeddings encoded by ViTs to feature self-relations in pixel dimension. Then, we apply this self-relation as the representation for self-supervision losses in Equ. (1).

Pixel-level self-relation. Given the feature embeddings $r_1 \in \mathbb{R}^{C\times HW}$ and $r_2 \in \mathbb{R}^{C\times HW}$ from the ViT backbone, we separately calculate their pixel-level self-relations in the spatial dimension. Following common settings [9,7], a pixel-level projection head h_b processes these embeddings to obtain $h_p(r_1)$ and $h_p(r_2)$. Common self-supervised learning methods maximize the representation similarity between overlapping areas of two views. But the random crop in the data augmentation causes miss-alignment between $h_p(r_1)$ and $h_p(r_2)$ in the spatial dimension. Therefore, we apply a region-aligned sampling operation $\mathbb{O}$ to extract the overlapping feature embeddings of two views with the shape of $\mathbb{R}^{C\times H_s W_s}$ where H_s, W_s are the height and width of sampled features. For re-sampled feature of one view, *e.g.,* $\mathbb{O}(h_p(r_1))$, we calculate the pixel-level self-relation matrix $\mathbb{A}_p(\mathbb{O}(h_p(r_1))) \in \mathbb{R}^{H_s W_s \times H_s W_s}$ as follows:

$$\mathbb{A}_p(\mathbb{O}(h_p(r_1))) = \text{Softmax}\left(\frac{\mathbb{O}(h_p(r_1))^T \cdot \mathbb{O}(h_p(r_1))}{\sqrt{C}} \Big/ t_p\right), \tag{2}$$

where the T is the matrix transpose operation, t_p is the temperature parameter that controls the sharpness of the Softmax function. In the pixel-level self-relation matrix, each row represents the relation of one pixel to each other pixels and is normalized by the Softmax function to generate probability distributions.

Self-supervision with pixel-level self-relation. The pixel-level self-relation matrix can be used by arbitrary existing self-supervision methods. For simplicity, we give an example of using self-relation for asymmetric non-contrastive self-supervision loss [19,10] as follows:

$$L_p = \text{R}_e(\mathbb{A}_p(\mathbb{O}(g_p(h_p(r_1)))), \mathbb{A}_p(\mathbb{O}(h_p(r_2)))), \tag{3}$$

where the R_e is the cross entropy loss measuring two self-relation distributions, g_p is the prediction head for asymmetric non-contrastive loss and we empirically

find that it benefits the representation quality. This loss function enforces the consistency among the pixel-level self-relations of different views.

Multi-head self-relation. In ViTs, the MHSA performs multiple parallel self-attention operations by dividing the feature into M parts, where each part has $\frac{C}{M}$ channels. Considering that different heads might focus on different patterns of image [7], we also apply such multi-head scheme when modeling the pixel-level self-relations. Specifically, the feature map of one view is split into M parts across the channel dimension, where each part generates a unique pixel-level self-relation matrix.

3.3 Channel Self-relation

The FFN in ViTs combines features across channels, and implicitly models the relation among channels. The pattern encoded in one channel has different degrees of correlation with the patterns encoded by other channels. This motivates us to explicitly model the channel-level self-relations for self-supervised learning to enhance the modeling ability in the channel dimension. Specifically, the feature embeddings of ViTs are transformed to channel-level the self-relations, which is used by self-supervisions.

Channel-level self-relation. Given the feature embeddings of two views r_1 and r_2, we define a projection head h_c to process these embeddings and obtain $h_c(r_1)$ and $h_c(r_2)$. Then we separately calculate the channel-level self-relations for the features in different views. Taking the feature of one view as an example, *e.g.,* $h_c(r_1)$, we calculate its channel-level self-relation matrix $\mathbb{A}_c(h_c(r_2)) \in \mathbb{R}^{C\times C}$ as follows:

$$\mathbb{A}_c(h_c(r_1)) = \text{Softmax}\left(\frac{h_c(r_1)\cdot h_c(r_1)^T}{H\cdot W}\Big/ t_c\right), \tag{4}$$

where the t_c is the temperature parameter. The channel-level self-relations matrix is normalized by the number of pixels and each row of the matrix is also normalized by the Softmax function.

Self-supervision with channel-level self-relation. The channel-level self-relations, as a representation in a new dimension, can also be applied to arbitrary existing self-supervision losses. Similar to the pixel-level self-relation based loss in Equ. (3), the self-supervision using channel-level self-relation is written as follows:

$$L_c = \text{R}_\text{e}(\mathbb{A}_c(g_c(h_c(r_1))), \mathbb{A}_c(h_c(r_2))), \tag{5}$$

where the R_e is the cross entropy loss and g_d is a prediction head. This loss function enforces the consistent channel-level self-relations across different views and thus enhances the ability to model channel-level self-relations.

3.4 Implementation Details

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

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

where L_I is the loss utilizing image-level embeddings [7]. We show in Section 4.4 that solely using our proposed self-relation as representations achieves competitive and even better performance than exiting image embeddings. Still, combining these three representations results in much better representation quality.

Architecture. Following [21,7], the representations of the two views $\tau_1(x)$ and $\tau_2(x)$ of an image are extracted by an encoder network f_1 and its momentum updated counterpart f_2. During training, the parameters θ_1 of f_1 is updated by gradient descent and the parameters θ_2 of f_2 is updated as $\theta_2 = \lambda\theta_2 + (1-\lambda)\theta_1$ where $\lambda \in [0, 1]$ is the momentum coefficient. Following DINO [7], the λ is set to 0.996 and is increased to 1.0 during training with a cosine schedule. Apart from the ViT backbone, we also use projection heads and prediction heads to process embeddings. The projection heads h_c and h_p consist of a batch normalization [25] layer, and a Relu activation layer. Prediction heads g_c and g_p are composed of a 1×1 conv layer, a batch normalization layer and a Relu activation layer. The prediction heads are only attached to the encoder network to match the encoder outputs and momentum encoder outputs. The self-relations in pixel-level and channel-level share the ViT backbone but the projection and prediction heads are separated for these two dimensions.

Relation in multi-crop. Many existing methods apply multi-crop [7,6] data augmentations to small patches for speeding up the training. The small crops provide richer information with a small amount of extra computation cost. To further facilitate the self-relation modeling ability, we also match the self-relations of small local crops to the self-relations of large global crops. The matching among small crops is ignored due small crops are less likely to have overlapping regions.

4 Experiments

In this section, we verify the effect of using our proposed pixel-level and channel-level feature self-relations as the representation for self-supervised learning. We give the training settings in Section 4.1 and introduce multiple evaluation protocols in Section 4.2. Then, we compare our method with existing methods in Section 4.3, showing stable improvement over multiple baselines. In Section 4.4, we conduct ablations to clarify designing choices.

4.1 Pretraining Settings

To compare with existing self-supervised learning methods for ViTs, we adopt the ViT-S/16 as the backbone network. The DINO [7] is selected as our major baseline method. During pretraining, the model is trained by an AdamW [33] optimizer with the batch size of 512 on 8 GPUs and learning rate of 0.001. Following [7], we apply the multi-crop training scheme where 2 global crops with the resolution of 224×224 and 4 local crops with the resolution of 96×96 are adopted. The global crops are cropped with a ratio between 0.35 and 1.0, and the local crops are cropped with a ratio between 0.05 and 0.35. The number of heads for modeling the pixel-level self-relation is set to 6 by default. The temperature terms for the pixel-level and channel-level self-relation are set to 0.5 and 0.1, respectively. The temperature terms are only applied to the momentum encoder to generate a sharper distribution, speeding up the encoder training. When sampling overlapping regions for pixel-level self-relation, the overlapping regions for global and local crops are set to 13×13 and 6×6, respectively. For performance comparison, we pre-train models on the ImageNet-1k [39] dataset. For ablation, the ImageNet-S_{300} dataset [16] is used to save training cost.

4.2 Evaluation Protocols

We propose new forms of representations for self-supervised learning, *i.e.,* the pixel/channel-level self-relations. The self-relations improve representation in more fine-grained dimensions, so we mainly evaluate our methods in the pixel-level dense prediction tasks. In addition, we also follow the exiting works to provide the comparison on the image-level classification task.

ImageNet-S semi-supervised semantic segmentation. The ImageNet-S dataset [16] extends ImageNet-1k with pixel-level semantic segmentation annotations for training and evaluation images. Evaluating semantic segmentation performance on the ImageNet-S dataset avoids the potential influence of domain shift between pretraining and downstream task datasets. We finetune the pretrained model with the semantic segmentation annotations in the ImageNet-S training set and evaluate the performance on the validation set of ImageNet-S. A randomly initialized 1×1 conv is attached to the ViT-S/16 as the segmentation head. We finetune models for 100 epochs with an AdamW optimizer using the batch size of 256. The learning rate is initially set to 5e-4 with a layer-wise decay of 0.5 and the weight decay is set to 0.05. After a warmup of 5 epochs, the learning rate is decayed to 1e-6 by the cosine decay schedule.

Transfer learning on semantic segmentation. We also evaluate the transfer learning performance on the semantic segmentation task using datasets from other domains, *i.e.,* PASCAL VOC2012 [15] and ADE20K [56] datasets. The UperNet [47] with the VIT-S/16 backbone is used as the segmentation model. Following the training setting in [57], models are trained for 20k and 160k iterations for PASCAL VOC2012 and ADE20K datasets, respectively. We train

	VOC SEG		ADE20K SEG		COCO DET			COCO SEG		
	mIoU	mAcc	mIoU	mAcc	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
DINO[7]	77.1	87.5	42.6	53.4	46.0	64.9	49.7	40.0	62.0	42.8
+Ours	**79.7**	**88.8**	**43.8**	**54.6**	**46.6**	**65.9**	**50.2**	**40.5**	**62.9**	**43.5**

Table 1. Transfer learning on downstream tasks. The models are pretrained on ImageNet-1k for 100 epochs. The AP^b means the bounding box AP for the object detection (DET) and AP^m means the segmentation mask AP for the instance segmentation (SEG).

models with the batch size of 16 on 4 GPUs and the crop size of 512, and we evaluate them on the validation set.

Transfer learning on object detection and instance segmentation. We use the Cascade Mask R-CNN [3] with ViT-S/16 to evaluate the transfer learning performance of pretrained models on object detection and instance segmentation tasks. Models are trained on COCO train2017 set and evaluated on the val2017 set [31]. Following the setting of [57], we train models using the 1× schedule with a batch size of 16.

Finetuning for ImageNet classification. We finetune classification models on ImageNet-1k dataset. During finetuning, a linear layer is attached to the ViTs as the classification head. Following the setting of [20], each layer in ViTs is end-to-end finetuned without any frozen layers. An AdamW optimizer [33] is used with a batch size of 512. The initial learning rate is set to 1e-3 and decayed to 1e-6 with the cosine decay schedule.

4.3 Performance and Analysis

Performance comparison. Table 1 shows the performance comparison between our method and the DINO baseline on downstream tasks. Compared to the DINO, our method improves the semantic segmentation by 2.6% mIoU/1.3% mAcc on the PASCAL VOC2012 dataset and 1.2% mIoU/1.2% mAcc on the ADE20K dataset, respectively. For the instance segmentation on COCO dataset, our method brings the improvement of 0.6% on box AP and 0.5% on mask AP. Table 2 shows the classification and semantic segmentation performance on ImageNet and ImageNet-S datasets, respectively. For semantic segmentation on ImageNet-S, our method brings a considerable gain of 1.8% mIoU. The F-measure (F), which solely measures the quality of segmentation shape and ignores the categories, is improved by 3.5%. Our self-relation enhances the relation modeling ability, enabling ViTs with much stronger shape-related representations. The end-to-end finetuned ViT also achieve the performance gain of 1.2% Top-1 and 0.4% Top-5 image classification accuracy on the ImageNet-1k dataset.

	ImageNet-1k		ImageNet-S	
	Top-1	Top-5	mIoU	F
DINO [7]	79.7	95.1	35.1	70.7
+Ours	**80.9**	**95.5**	**36.9**	**74.2**

Table 2. Top-1 and Top-5 classification accuracy on ImageNet-1k and semi-supervised semantic segmentation mIoU on ImageNet-S.

	VOC SEG		ImageNet-S_{300}	
	mIoU	mAcc	mIoU	F
MoCov3 [11]	65.7	78.7	24.0	62.7
+Ours	**67.5**	**80.6**	**29.1**	**67.0**

Table 3. Comparison with the MoCo v3 [11] framework. The models are pre-trained for 100 epochs on the ImageNet-S_{300} dataset.

	pixel	channel
DINO [7]	0.205	1.18
+Ours	**0.154**	**0.82**

Table 4. The average difference of self-relations between two views on the validation set of ImageNet-S.

	VOC SEG		ImageNet-S_{300}	
	mIoU	mAcc	mIoU	F
DNIO baseline	68.1	81.1	28.8	66.7
+Ours symmetry	72.1	84.4	37.1	71.7
+Ours asymmetric	**73.5**	**84.7**	**41.2**	**72.9**

Table 5. Effect of asymmetric losses in Equ. (3) and Equ. (5).

Apart from DINO, we also apply the proposed self-relation to MoCo v3 [11] to illustrate the generalization ability of our proposed method. Table 3 shows that using self-relation improves the MoCo v3 by 1.8% in mIoU and 2.0% in mAcc for semantic segmentation on the Pascal VOC dataset. Notably, there is a huge improvement of 5.1% in mIoU for the semi-supervised semantic segmentation on the ImageNet-S_{300} dataset.

Analysis and visualization. Intuitively, training self-supervised ViTs with our proposed pixel/channel-level self-relations results in stable and consistent relations among pixels under different data augmentations. To verify this, we measure the difference of pixel/channel-level self-relation matrices between two views in the validation set of ImageNet-S, as shown in Table 4. For the pixel dimension, we report the difference of self-relation matrix on the overlapping regions between two views. All regions are used to calculate the channel-level self-relation matrix. Adding self-relation to the DINO baseline narrows 24% and 31% difference of self-relation between two views in the pixel and channel dimensions, indicating that using self-relation for self-supervised learning makes the ViT learn stronger relation modeling ability. We also visualize the self-relation between one pixel and other pixels of two views. The self-attention of overlapping regions between two views in Figure 3 shows smaller differences when trained using feature self-relations as representation.

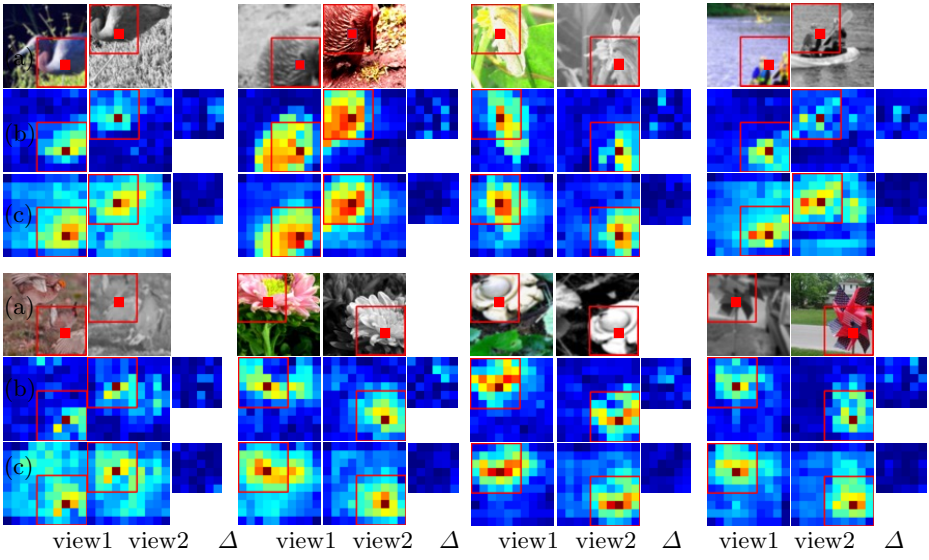

Fig. 3. The difference of pixel-level self-relations across two views. (a) Two views from each image. (b) The pixel-level self-relations generated by DINO. (c) The pixel-level self-relations generated by DINO+Ours. View1 and view2 means the two views generated from an image. The Δ means the difference of self-relation in the overlapping region, which is indicated by red boxes.

4.4 Ablation Studies

To save computational cost, we pretrain all models on the ImageNet-S_{300} [16] dataset with 300 categories for the ablation study. We pretrain the models for 100 epochs with a batch size of 256 on 4 GPUs. There are no local crops and the global crops are cropped with a ratio between 0.14 and 1.0, following the settings in DINO [7]. The semantic segmentation results on both PASCAL VOC and ImageNet-S datasets are evaluated.

Ablation of pixel and channel self-relation. We compare the effectiveness of using different representation forms for self-supervised learning, *i.e.,* our proposed pixel/channel-level self-relations and the feature embeddings used by DINO. As shown in Table 6, compared to the feature embeddings used by DINO, the pixel-level self-relations improves the model performance by 2.4% in mIoU and 1.9% in mAcc on the PASCAL VOC dataset. Similarly, the F-measure on the ImageNet-S_{300} dataset is improved by 2.4%. Training self-supervised ViTs with pixel-level self-relation further enhance the pixel-relation modeling ability of ViT, benefiting the pixel-level dense prediction task, *e.g.,* semantic segmentation. Although inferior to the other two representation forms, channel-level self-relations still help learn reasonable semantics, *i.e.,* the model pretraind with channel-level self-relations performs much better than the randomly initialized

Image embd. L_I	Pixel self-relation L_p	Channel self-relation L_c	VOC SEG mIoU	VOC SEG mAcc	ImageNet-S$_{300}$ mIoU	ImageNet-S$_{300}$ F
✗	✗	✗	25.6	35.7	0.2	0.0
✓			68.1	81.1	28.8	66.7
	✓		71.5	83.0	23.7	69.1
		✓	61.4	75.6	22.5	64.2
✓	✓		70.7	82.6	33.3	69.7
✓		✓	69.8	82.9	36.5	70.3
	✓	✓	71.5	83.3	30.6	71.1
✓	✓	✓	**73.5**	**84.7**	**41.2**	**72.9**

Table 6. Ablation of using different representations for self-supervised training. The L_I, L_p, and L_c denote the loss functions using image-level embeddings [7], pixel-level self-relations, and channel-level self-relations, respectively. The model without these three losses is randomly initialized when finetuning on downstream tasks.

M	VOC SEG mIoU	VOC SEG mAcc	ImageNet-S$_{300}$ mIoU	ImageNet-S$_{300}$ F
1	72.4	84.0	38.7	71.4
3	72.7	84.8	38.9	71.6
6	**73.5**	84.7	**41.2**	**72.9**
12	73.4	**85.1**	40.8	72.9
16	72.5	84.3	39.3	71.7

Table 7. The effect of different number of heads for the pixel-level self-relations.

t_p	t_c	VOC SEG mIoU	VOC SEG mAcc	ImageNet-S$_{300}$ mIoU	ImageNet-S$_{300}$ F
0.50	0.50	72.0	84.2	36.7	70.9
0.50	0.10	**73.5**	**84.7**	**41.2**	**72.9**
0.50	0.01	70.4	82.7	33.6	70.3
1.00	0.10	70.2	83.1	36.7	70.3
0.50	0.10	73.5	84.7	**41.2**	**72.9**
0.10	0.10	**73.7**	**85.0**	39.9	72.9

Table 8. The effect of different temperatures for pixel/channel-level self-relation.

model. Therefore, pixel-level and channel-level self-relations can represent information of input images, supporting the self-supervised training of ViTs.

Then, we verify the orthogonality of these three representation types. Combining with the feature embedding based loss in DINO, the pixel-level self-relation brings the gain of 2.6% in mIoU and 1.6% in mAcc on the PASCAL VOC dataset. Also, the gains on the ImageNet-S$_{300}$ dataset are 4.5% in mIoU and 3.0% in F-measure. Adding channel-level self-relation to DINO also improves 1.7% and 7.7% in mIoU on PASCAL VOC and ImageNet-S datasets. Finally, utilizing three forms of representation further boosts the performance on all benchmarks. We can conclude that our self-relations are orthogonal to existing feature embeddings for self-supervised learning.

Effect of multi-head. We utilize the multi-head pixel-level self-relation following the MHSA module in ViTs. More heads enable diverse pixel-level self-relations, but the channels used for calculating each self-relation are reduced. Table 7 shows the effect of different number of heads M in pixel-level self-relation.

Compared to the single-head version, increasing M to 6 brings the largest performance gain of 1.1 % in mIoU on the PASCAL VOC dataset. $M = 12$ achieves limited extra gains while $M = 16$ suffers a rapid performance drop. Too many heads result in inaccurate estimation of self-relation, hurting the representation quality. So we set the number of heads to 6 by default to balance the diversity and quality of pixel-level self-relation.

Effect of sharpness. The temperature terms in Equ. (2) and Equ. (4) control the sharpness of the self-relation distributions. Small temperature sharpens the distributions while large temperature softens the distributions. In Table 8, we verify the effectiveness of temperatures for both pixel and channel self-relations. For the channel self-relation, decreasing temperature from 0.1 to 0.01 results in rapid performance drops from 73.5% to 70.4% in mIoU on 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 pixel self-relation, the temperature 0.5 performs better than 1.0, but changing temperature from 0.5 to 0.1 has a limited difference. Using a temperature of 0.5 achieves slightly better performance on the large-scale ImageNet-S dataset. Thus, we set the default temperature of pixel self-relation to 0.5. Interestingly, semi-supervised semantic segmentation on ImageNet-S is more sensitive to the sharpness changes in channel-level self-relations than pixel-channel self-relations.

Effect of asymmetric loss. The asymmetric structure is proven effective for non-contrastive loss [19,10] when using image-level embeddings as representations. We verify if the asymmetric loss also benefits our self-relation representations. In Table 5, we compare the effect of self-relation based loss using asymmetric/symmetry structures. The self-relation with asymmetric and symmetry structures both achieve better performance than the DINO baseline. The symmetrical loss outperforms the baseline on PASCAL VOC and ImageNet-S_{300} datasets with 4.0% and 8.3% gains in mIoU. Compared to the symmetry loss, the asymmetric loss brings the improvement of 1.4% in mIoU on PASCAL VOC dataset and 4.1% in mIoU on ImageNet-S_{300}. Therefore, though the asymmetric loss is not indispensable for self-relation representation, it still benefits the self-supervised training of ViTs.

5 Conclusions

In this paper, we propose the feature self-relation based self-supervised learning to enhance the relation modeling ability of self-supervised ViTs. Specifically, instead of directly using feature embeddings as the representation, we propose to use pixel-level and channel-level self-relation of features as representations for self-supervised learning. Self-relation is orthogonal to feature embeddings, which further boosts existing self-supervised methods. We show that feature self-relation helps improve the self-supervised ViTs in a more fine-grained level, benefiting multiple pixel-level dense prediction downstream tasks.

References

1. Atito, S., Awais, M., Kittler, J.: Sit: Self-supervised vision transformer. arXiv preprint arXiv:2104.03602 (2021)
2. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers (2021)
3. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: CVPR (June 2018)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020)
5. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV (2018)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
8. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning (ICML). pp. 1691–1703 (2020)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML). pp. 1597–1607. PMLR (2020)
10. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (June 2021)
11. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: ICCV (October 2021)
12. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: NeurIPS (2021)
13. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
15. Everingham, M., Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**, 303–338 (2009)
16. Gao, S., Li, Z.Y., Yang, M.H., Cheng, M.M., Han, J., Torr, P.: Large-scale unsupervised semantic segmentation. arXiv preprint arXiv:2106.03149 (2021)
17. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
18. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent - a new approach to self-supervised learning. In: NeurIPS (2020)
19. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Ávila Pires, B., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent - a new approach to self-supervised learning. In: NeurIPS (2020)
20. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (2021)

21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (June 2020)
22. Hénaff, O.J., Koppula, S., Alayrac, J.B., van den Oord, A., Vinyals, O., Carreira, J.a.: Efficient visual pretraining with contrastive detection. In: ICCV. pp. 10086–10096 (October 2021)
23. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: ICLR (2019)
24. Hu, Q., Wang, X., Hu, W., Qi, G.J.: Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In: CVPR. pp. 1074–1083 (June 2021)
25. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (ICML). pp. 448–456 (2015)
26. Kim, K., Wu, B., Dai, X., Zhang, P., Yan, Z., Vajda, P., Kim, S.J.: Rethinking the self-attention in vision transformers. In: CVPRW. pp. 3071–3075 (June 2021)
27. Koohpayegani, S.A., Tejankar, A., Pirsiavash, H.: Mean shift for self-supervised learning. In: ICCV. pp. 10326–10335 (October 2021)
28. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: CVPR (July 2017)
29. Li, C., Yang, J., Zhang, P., Gao, M., Xiao, B., Dai, X., Yuan, L., Gao, J.: Efficient self-supervised vision transformers for representation learning. In: ICLR (2022)
30. Li, Z., Chen, Z., Yang, F., Li, W., Zhu, Y., Zhao, C., Deng, R., Wu, L., Zhao, R., Tang, M., Wang, J.: MST: Masked self-supervised transformer for visual representation. In: NeurIPS (2021)
31. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)
32. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
34. Noroozi, M., Favaro, P.: Unsupervised learning of visual representions by solving jigsaw puzzles. In: ECCV (2016)
35. Noroozi, M., Pirsiavash, H., Favaro, P.: Representation learning by learning to count. In: ICCV (Oct 2017)
36. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (June 2016)
37. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) NeurIPS (2021)
38. Roh, B., Shin, W., Kim, I., Kim, S.: Spatilly consistent representation learning. In: CVPR (2021)
39. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
40. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning (ICML). pp. 10347–10357. PMLR (2021)

41. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: International Conference on Machine Learning (ICML) (2008)
42. Wang, L., Liang, F., Li, Y., Zhang, H., Ouyang, W., Shao, J.: Repre: Improving self-supervised vision transformer with reconstructive pre-training. arXiv preprint arXiv:2201.06857 (2022)
43. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: IEEE ICCV (2021)
44. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: CVPR (2021)
45. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. arXiv preprint arXiv:2112.09133 (2021)
46. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (June 2018)
47. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: The European Conference on Computer Vision (ECCV) (September 2018)
48. Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Sun, P., Li, Z., Luo, P.: Detco: Unsupervised contrastive learning for object detection. In: ICCV. pp. 8392–8401 (October 2021)
49. Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., Hu, H.: Self-supervised learning with swin transformers. arXiv preprint arXiv:2105.04553 (2021)
50. Xie, Z., Lin, Y., Zhang, Z., Cao, Y., Lin, S., Hu, H.: Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In: CVPR. pp. 16684–16693 (June 2021)
51. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling (2021)
52. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. CoRR **abs/2103.03230** (2021)
53. Zhan, X., Xie, J., Liu, Z., Ong, Y.S., Loy, C.C.: Online deep clustering for unsupervised representation learning. In: CVPR (June 2020)
54. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV. pp. 649–666. Springer (2016)
55. Zhao, Y., Wang, G., Luo, C., Zeng, W., Zha, Z.J.: Self-supervised visual representations learning by contrastive mask prediction. In: ICCV. pp. 10160–10169 (October 2021)
56. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (July 2017)
57. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)