SELF-SUPERVISED MODELS ARE GOOD TEACHING ASSISTANTS FOR VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers have shown remarkable progress on computer vision tasks in the past year. Compared to their CNN counterparts, transformers usually need the help of distillation to achieve comparable results on middle or small sized datasets. Meanwhile, recent researches discover that when transformers are trained with supervised and self-supervised manner respectively, the captured patterns are quite different both qualitatively and quantitatively. These findings motivate us to introduce an self-supervised teaching assistant (SSTA) besides the commonly used supervised teacher to improve the performance of transformers. Specifically, we propose a head-level knowledge distillation method that selects the most important head of the supervised teacher and self-supervised teaching assistant, and let the student mimic the attention distribution of these two heads, so as to make the student focus on the relationship between tokens deemed by the teacher and the teacher assistant. Extensive experiments verify the effectiveness of SSTA and demonstrate that the proposed SSTA is a good compensation to the supervised teacher. Meanwhile, some analytical experiments towards multiple perspectives (*e.g*. prediction, shape bias, robustness, and transferability to downstream tasks) with supervised teachers, self-supervised teaching assistants and students are inductive and may inspire future researches.

1 INTRODUCTION

Recently, Vision Transformers (ViTs) have been successfully used for computer vision tasks, including image recognition, object detection, semantic segmentation and so on. Remarkably, ViTs are capable to reach superior performance on image classification task when trained with large-scale datasets, *e.g*. JFT-300M (Dosovitskiy et al. (2020)). However, ViTs achieve lower accuracies than Convolutional Neural Network (CNN) networks on medium-scale or small-scale datasets (Dosovitskiy et al. (2020)). To alleviate the demand for data, DeiT (Touvron et al. (2021)) distills the inductive bias from a large CNN teacher by introducing an extra distillation token and shows satisfactory results.

Self-supervised learning (SSL) and supervised learning (SL) are two different paradigms *w.r.t.* the way they construct training objectives. With the development of transformer, self-supervised learning for transformers has also attracted widespread attention from the community, and many approaches have been proposed (Chen et al. (2021); Caron et al. (2021)). Caron et al. (2021) reported an interesting discovery that self-attention visualizations of self-supervised vision transformers and supervised vision transformers represent different tendentiousness. As shown in Figure 1, vision transformers trained with supervised signal pay more attention to texture, while self-supervised counterparts focus on shape. In addition, when the size of the annotated training dataset is small, the supervised transformer is more prone to overfitting. For example, when the training dataset is ImageNet-1K (Russakovsky et al. (2015)), self-supervised ViT can transfer better to downstream tasks than its counterpart (Caron et al. (2021)).

These observations motivate us to explore and exploit the differences between these two learning paradigms (SSL v.s. SL) applied on ViTs. We measured the similarity index between the feature layers of two randomly initialized supervised transformers and of a supervised transformer and a self-supervised transformer through Centered Kernel Alignment (CKA) indicator. The results are shown in Figure 2. It can be seen that the similarity between the layers of two randomly initialized supervised transformers significantly exceeds that between a supervised transformer and a self-supervised transformer, and the feature similarity of the last few layers is relatively low.

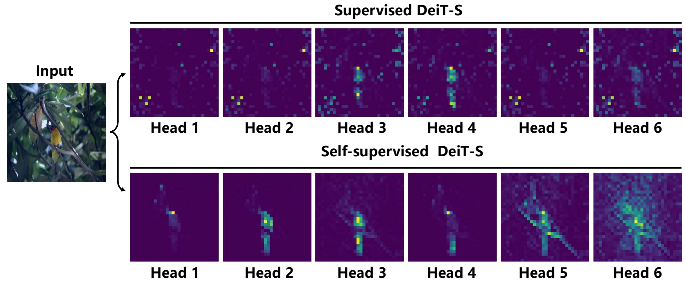


Figure 1: Visualizations of self-attention maps from the last layer of DeiT-S (Touvron et al. (2021)).

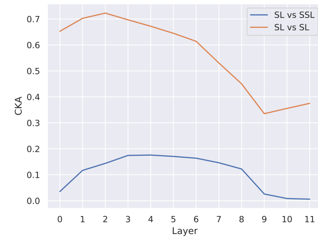


Figure 2: CKA similarities between layers across different learning paradigms.

Since the difference is quantitatively prominent and qualitatively compensating, we propose to introduce an self-supervised teaching assistant (termed as SSTA) besides the commonly used supervised teacher to further improve the performance of transformers. Specifically, we propose a head-level knowledge distillation method that selects the most important head of the supervised teacher and the self-supervised teaching assistant, and let the student to mimic the attention distribution of these two heads, so as to make the student focus on the relationship between tokens deemed by the teacher and the teacher assistant. Extensive experiments demonstrate that the proposed SSTA is a good compensation to the supervised teacher. Meanwhile, compared with supervised teaching assistant, SSTA with greater difference can bring more improvements.

The success of SSTA prompted us to further reveal the otherness between the self-supervised ViTs and supervised ViTs. We explore the differences between two different teachers and the students distilled from different teachers on prediction, shape bias, robustness, and transferability to downstream tasks, some of which are counter-intuitive and are studied for the first time.

Our contributions are summarized as follows:

- By observing that self-supervised learning and supervised learning provide information from different perspectives, we exploit adding an self-supervised transformer as a teaching assistant to complement to commonly used supervised teacher, and firstly propose a head-level knowledge distillation approach for data efficient vision transformer learning.
- To effectively transfer the knowledge via heads, a heuristic head selection strategy is designed to choose most informative heads from teacher. Meanwhile, an early stop learning strategy is further derived to facilitate distillation.
- Extensive experiments are conducted to demonstrate the advantage of the self-supervised teaching assistant. Besides, by comprehensive analyzing the variant combination of two teachers, some interesting findings, regarding the prediction, shape bias, robustness, and transferability, are studied for the first time.

2 Related Work

2.1 Vision Transformer

Recently, ViTs have made tremendous development, and various Transformer architectures for computer vision tasks have been proposed (Dosovitskiy et al. (2020); Touvron et al. (2021); El-Nouby et al. (2021)). The Self-Attention mechanism allows transformers to capture long-distance relationships and become content-aware. Compared to CNN, ViTs are more robust to severe occlusions, perturbations, and domain shifts and significantly less biased towards textures (Naseer et al. (2021)). However, ViTs are very hungry for data, when training on medium-scale or small-scale datasets, ViTs can't exceed the results of CNN (Dosovitskiy et al. (2020)). Therefore, a lot of works (Touvron et al. (2021); Graham et al. (2021)) introduce the inductive bias of a supervised pre-trained large CNN teacher through knowledge distillation, thereby alleviating the demand for annotated data.

2.2 KNOWLEDGE DISTILLATION

Knowledge Distillation (KD) was first proposed by Hinton et al. (2015), which aims to transfer the knowledge of a larger teacher model to a smaller student model. Many approaches have achieved great success on CNN, *e.g*. Romero et al. (2014); Zagoruyko & Komodakis (2016); Park et al. (2019), however due to the differences of transformers, few of them can be directly applied to transformers. DeiT (Touvron et al. (2021)) is the first work applying knowledge distillation to transformer, which adds an extra distillation token to transfer the inductive bias of a larger CNN to a relatively small transformer in the form of hard or soft output label. Ren et al. (2021) use different architectural inductive biases to co-advise the student transformer. These methods all rely on the inductive bias of other network structure, the knowledge needed by the student transformer and the effective transmission method are still to be explored. Furthermore, the teachers used in the existing distillation methods are all obtained by supervised training, and as far as we know, we are the first to try to use self-supervised representations to assist supervised training.

2.3 SELF-SUPERVISED LEARNING

Self-supervised Learning (SSL) is a generic framework that gets supervision from the data itself without any tags from human labor. Earlier methods heavily rely on constructing negative samples, *e.g*. SimCLR (Chen et al. (2020a;b)), MoCo (He et al. (2020); Chen et al. (2020c)), while recent works eliminate the need for negative samples, *e.g*. BYOL (Grill et al. (2020)), SimSiam (Chen & He (2021)). With the development of vision transformer, some works (Caron et al. (2021), Chen et al. (2021)) apply contrastive learning to vision transformers. Compared to supervised counterparts, self-supervised vision transformers exhibit some properties. As described in Caron et al. (2021), self-supervised ViT features explicitly contain the scene layout and object boundaries. In this work, we show that the difference between self-supervised ViT representations and supervised ViT representations is far from that.

2.4 KD MEETS SSL

Recently, some works have combined KD and SSL. SSKD (Xu et al. (2020)) adds an SSL branch next to the supervisory branch and regards the information contained in the SSL task as additional dark knowledge. CRD (Tian et al. (2019)) proposes a contrastive-based objective for knowledge distillation, which allows student to capture more information in the teacher's representations of data. SEED (Fang et al. (2021)) employs knowledge distillation as a means to improve the representation capability of small models in self-supervised learning. These methods are for CNN, and there is only one teacher with the same training paradigm as the student. While in our method, the teacher and the student are under different training paradigms and the two teachers are trained by different paradigms with obvious different tendentiousness.

3 METHODOLOGY

In this section, we introduce the proposed *Self-Supervised Teacher Assistant* (SSTA). We first present the overall architecture in Section 3.1, and then introduce the specific head-level distillation in detail in Section 3.2. Finally, the entire training process is described in Section 3.3.

3.1 OVERALL ARCHITECTURE

The framework of the proposed method is shown in Figure 3, consisting of three transformer encoders. The *Student* in the middle is the encoder that we want to improve, the *SL Teacher* on the left is the pre-trained teacher obtained via supervised learning, and the *SSTA* on the right is the pre-trained teaching assistant obtained through self-supervised learning. For each input $X \in \mathbb{R}^{H \times W \times C}$, where H, W and C represents the height, width and channel of the image respectively, it is input to three encoders respectively. After patch embedding, the input image is projected to $X_{PE} \in \mathbb{R}^{N \times D}$ where N is the number of tokens and D is the dimension of each token, and X_{PE} is then fed into stacked layers. As shown in Figure 3, each layer consists of LayerNorm (Ba et al. (2016)), Multi-head Self Attention (MSA), Multi-Layer Perceptron (MLP) and residual connections.

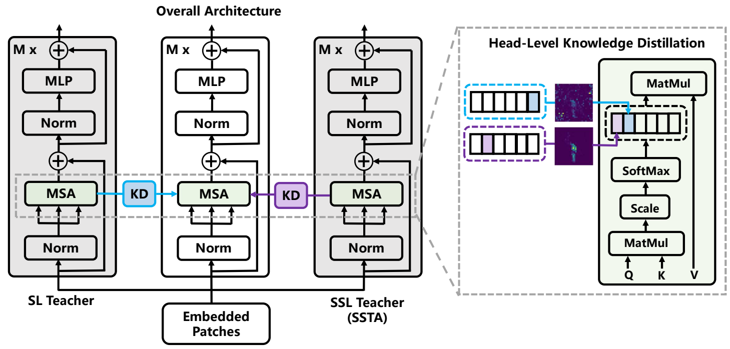


Figure 3: The overall architecture of the proposed method. One image is first projected into tokens, and then input to three transformer encoders, one is the learnable student, one is a frozen pre-trained SL teacher, and the other is fixed pre-trained SSTA. The areas concerned by the head of the students are required to be consistent with the areas focused by the most important head in the SL teacher and SSTA simultaneously via constraining the attention distributions.

For MSA, we first compute $Q = X_{PE} \cdot W_{\mathrm{Q}} \in \mathbb{R}^{N\times h\times d}$, $K = X_{PE} \cdot W_{\mathrm{K}} \in \mathbb{R}^{N\times h\times d}$ and $V = X_{PE} \cdot W_{\mathrm{V}} \in \mathbb{R}^{N\times h\times d}$ via linear transformations W_{Q}, W_{K}, W_{V}, where h is the number of heads, and d is the dimension of each head ($d = D/h$). Figure 3 (right) shows the details of MSA, Q and K produce an attention matrix via inner product and then the matrix is rescaled by $\sqrt{d}$ and normalized with a softmax function. Finally the normalized attention matrix is multiplied by V to get the output of the MSA layer. The entire procedure can be formulated as:

$$AttnMat = Softmax(Q \times K^T/\sqrt{d}), \tag{1}$$

$$Output = AttnMat \times V, \tag{2}$$

note the dimension of $AttnMat$ is $h \times N \times N$. For more details, please kindly refer to (Dosovitskiy et al. (2020)).

$AttnMat$ describes the attention distribution, which is computed based on the similarity between tokens. The higher the value, the more the relevance. The attention distribution reflects the relationship between tokens, and the relationship between [cls] token and other patch tokens can further reflect where the model is focusing on, as shown in Figure 1. Existing work (Zagoruyko & Komodakis (2016)) has demonstrated that the attention maps of a powerful teacher network are effective knowledge in CNN. As Transformers are based on attention mechanism, we consider to adopt attention distribution as knowledge to transfer. The specific knowledge transfer method is the newly proposed head-level distillation, which will be introduced in section 3.2.

3.2 Head-level distillation

As (Caron et al. (2021)) observed that different heads focus on different locations, we consider distilling diverse knowledge from the heads of different teachers. Figure 1 shows that the heads of SL transformers focus more on textures of background, while the heads of SSL transformers have high activation on objects. For the human visual system, both the objects and textures of background are important judgment bases when determining the categories of images, which inspires us to fully utilize these different attention preferences. We propose a head-level knowledge distillation method, which is illustrated in the dashed box of Figure 3. Specifically, two heads of the student imitate the attention distribution of the most important head of the two diverse teachers (*i.e.* SSL teacher (SSTA) and SL teacher) respectively via knowledge distillation loss, so as to pay attention to the most significant relationship deemed by different teachers simultaneously. Next, we will introduce the choice of the most important head from the teachers and the definition of distillation loss.

3.2.1 Head selection strategy

The first critical problem of head-level distillation is the selection of the most important heads. Since different vision transformers have different numbers of heads, aligning the number of teacher and student heads is a thorny problem. To avert this problem, we propose a head selection strategy which only selects the most important head for each layer from the teacher for knowledge distillation.

Considering that the greater the contribution to the accuracy, the more important the head is, we first evaluate the accuracy drop by alternatively setting different head to zero, and then regard the head corresponding to the highest drop as the most important one. Supposing the index of the head to be estimated is $i \in \{1, 2, ...h\}$ for the l-th layer, the reset process can be expressed as:

$$AttnMat_l[i, :, :] = \mathbf{0}, \tag{3}$$

the new $AttnMat$ is remarked as $AttnMat'$. Then, we define the importance of the head as follows:

$$I = Acc(\phi(AttnMat)) - Acc(\phi(AttnMat')), \tag{4}$$

where $\phi(AttnMat)$ is the model with original heads, $\phi(AttnMat')$ is the model that partial heads are reset as zero and $Acc(\cdot)$ is the accuracy of model. The higher the I value, the more important it is. Note that we estimate the importance of heads on the pre-trained model. For the assigned layer set L, we select the most important head for each layer, then we can obtain most important head index set $H' = \{i^l\}$, where $l \in L$. It is worth noting that when conducting distillation on multi-layers (*i.e.* $|\mathrm{L}| > 1$), we evaluate the most important combination of multiple heads over multiple layers. For example, in this paper, $L = \{10, 11, 12\}$, for the head combination $\{1^{10}, 2^{11}, 3^{12}\}$ that to be evaluated, the 1st head of 10th layer, the 2nd head of 11th layer and the 3rd head of 12th layer are reset to 0.

3.2.2 Objective Function

After selecting the most important heads of the SL teacher and SSL teacher (SSTA), we let two heads in each layer of the student mimic the most important head of SL teacher and SSL teacher (SSTA) in the corresponding layer respectively through minimizing Kullback-Leibler divergence between the head-level attention distributions. The objective function of knowledge distillation is as follows:

$$L_{KD}^{SL} = \sum_{i \in H'_{SL}} KL(AttnMat_S[0, :, :], AttnMat_l^{SL}[i, :, :]), \tag{5}$$

$$L_{KD}^{SSL} = \sum_{j \in H'_{SSL}} KL(AttnMat_S[1, :, :], AttnMat_l^{SSL}[j, :, :]), \tag{6}$$

where $KL(\cdot)$ is Kullback-Leibler divergence, H'_{SL} and H'_{SSL} are the most important heads sets of SL teacher and SSL teacher (SSTA) and l is the index of the layer. .

3.3 Training Process

Total loss. The total loss is defined as follows:

$$L_{Total} = \alpha \cdot L_{CE}(f^S(X), y) + \beta \cdot L_{KD}^{SL} + \lambda \cdot L_{KD}^{SSL}, \tag{7}$$

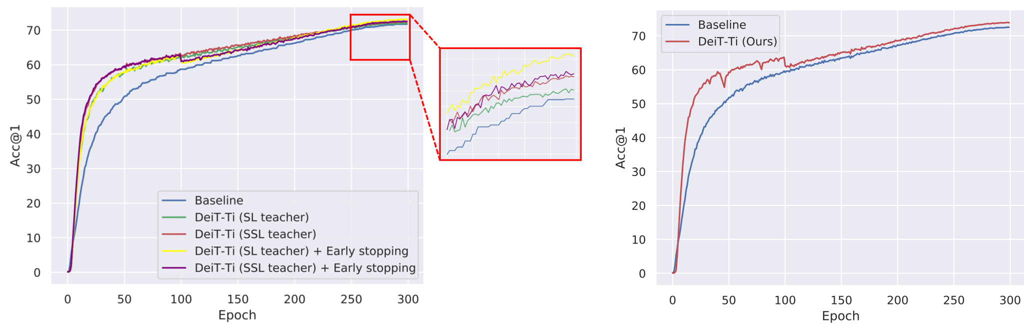
where $L_{CE}(\cdot)$ denotes Cross Entropy, and y is ground truth. α, β and λ are the hyper-parameters that control the weights of CE loss and distillation loss.

Early stop strategy. Figure 4 (a) shows the curve of training accuracy, it can be observed that both the SL teacher and SSL teacher can accelerate the convergence of student in the early stage, and the acceleration of SSL teacher is more significant in particular. However, this property has no benefit for student in the later period, and the performance of student even declined. Based on this observation, we propose the early stop strategy to take advantage of this property and avoid performance degradation. Specifically, the distillation is only conducted in the early stage (*e.g.*100 epochs), when entering the next epoch, β and λ of Eq. 7 are set to 0.

4 Experiments

4.1 Implementation Details

Datasets. ImageNet (Russakovsky et al. (2015)) is used to verify the effectiveness of our method. CIFAR10 and CIFAR100 (Krizhevsky et al. (2009)) are adopted for downstream transfering tasks. ImageNet-C (Hendrycks & Dietterich (2019)) is utilized to analyze the robustness of the representations. SIN dataset (Geirhos et al. (2018)) is used to evaluate the shape bias of models.



(a) DeiT-Ti with only one teacher (SL or SSL). (b) DeiT-Ti with our method.

Figure 4: Accuracy curves during training. (a) exhibits that both the SL teacher and SSL teacher can accelerate the convergence of the student in the early stage, and the acceleration of SSL teacher is more significant. However, this superiority disappear in the later stage. (b) demonstrates that our method can take advantage of the ability of SSL teacher to accelerate convergence in the early stage, allowing students to converge faster in the early stage while stably surpassing the baseline in the later stage. The distillation is stopped at 100 epoch.

Teacher Pre-training Settings. The SSTAs are obtained by DINO (Caron et al. (2021)) and both the pre-training and linear evaluation are conducted on ImageNet-1K. The SL teachers are obtained by DeiT (Touvron et al. (2021)) and XCiT (El-Nouby et al. (2021)) respectively without distillation.

Distillation Training Settings. Following DeiT and XCiT, the total number of distillation epochs are 300 and 400 for DeiT and XCiT respectively, and the corresponding early stop epochs are 100 and 150. The teachers are frozen during the distillation.

Downstream Transfer Training Settings. In order to analyze the generalization of representations, we further conduct linear evaluation on Cifar10 and Cifar100. Since the image resolution of the Cifar dataset is 32×32, all the images are resized to 224×224 with bicubic re-sampling, following Gao et al. (2021). All the training hyper-parameters are consistent with Gao et al. (2021).

4.2 The Effectiveness on ImageNet

We first verify the effectiveness of the proposed method on ImageNet-1K. The results are shown in Table 1, from which we have the following observations:

i. The proposed method outperforms all the baselines significantly. Specifically, our method can bring 1.8% improvement on DeiT-Ti (74% v.s. 72.2%). When applying to DeiT-Ti⚗, which is a strong baseline that enhances the model by introducing the inductive bias from a large pre-trained CNN teacher, our method can still bring a further 0.7% gain.

ii. The proposed method is not limited to transformer architectures, and can also bring considerable improvement on XCiT-T12.

Teacher1	Acc@1	Teacher2	Acc@1	Student	Acc@1
-	-	-	-	DeiT-Ti	72.2
DeiT-S (SSL)	77.0	DeiT-S (SL)	79.9	DeiT-Ti	**74**
-	-	-	-	DeiT-Ti⚗	74.5
DeiT-S (SSL)	77.0	DeiT-S (SL)	79.9	DeiT-Ti⚗	**75.2**
-	-	-	-	DeiT-S	79.9
DeiT-B (SSL)	78.2	DeiT-B(SL)	81.8	DeiT-S	**81.4**
-	-	-	-	XCiT-T12	77.0‡
XCiT-S12 (SSL)	77.8	XCiT-S12 (SL)	82.0	XCiT-T12	**77.5**

Table 1: Results on ImageNet-1K. A(B) stands for the teacher of A structure obtained by B training paradigm, ⚗ denotes the student uses the hard label output by RegNetY-16G (Radosavovic et al. (2020)) for distillation and ‡ means our reproduction. The students without any teacher are baselines.

4.3 Ablation Study

In this section, we testify the effectiveness of each important component in the proposed method, *i.e*. SSTA, head selection strategy, early stop strategy and distillation layers. Note the teachers in all experiments of this part are DeiT-S, and the students are DeiT-Ti.

Model	SL KD	SSL KD	Early Stop	Head Sel.	Acc@1
Baseline	×	×	×	-	72.2
Single Teacher					
SL_KD	✓	×	×	imp.	72.0
SSL_KD	×	✓	×	imp.	72.2
SL_KD_early100	✓	×	100ep	imp.	73.2
SSL_KD_early100	×	✓	100ep	imp.	72.6
Multiple Teachers					
2SL_KD	✓	✓	×	imp.	71.4
SSTA_KD	✓	✓	×	imp.	72.2
2SL_KD_early100	✓	✓	100ep	imp.	73.2
SSTA_KD_avg_early100	✓	✓	100ep	avg.	73.2
SSTA_KD_rand_early100	✓	✓	100ep	rand.	73.5
SSTA_KD_early100 (Ours)	✓	✓	100ep	imp.	**74.0**

Table 2: Ablation study on ImageNet-1K. 100ep denotes 100 epochs, imp. stands for selecting the most important head for distillation, avg. means using the average of multiple heads, and rand. denotes randomly select one head from teacher.

Effectiveness of SSTA. As expected, the distillation results of two teachers will be better than that of a single teacher since more knowledge is transferred to student. However, as shown in Table 2, there is no difference between using single SL teacher (*SL_KD_early100*) and two different SL teachers (*2SL_KD_early100*). On the contrary, our method which adds an SSTA to the SL teacher can significantly improve the performance. In particular, our approach can bring an accuracy improvement of 0.8%, 1.4% and 0.8%, compared to training with single SL teacher (*SL_KD_early100*), single SSL teacher (*SSL_KD_early100*) and two different SL teachers (*2SL_KD_early100*), respectively. The results demonstrate the effectiveness of SSTA and inspire us to further explore the otherness of different teachers and students. We provide detailed analyses in Sec. 5.

Effectiveness of head selection strategy. Besides selecting the most important heads based on the contribution to accuracy, we also tried to use the average of the attention distribution of all heads or randomly choose one head within one layer as the knowledge to transfer. As shown in the bottom three rows of Table 2, choosing the most important head has an improvement of 0.8% or 0.5% compared to taking the average attention distribution of the heads or random selection, which indicates the effectiveness of the proposed head selection strategy.

Effectiveness of early stop strategy. It can be seen from Table 2 that the students do not work well when using head-level distillation in all epochs. Nevertheless, after applying the early stop strategy, our method can significantly boost the performance of students (up to 1.8% accuracy). The experimental results prove that the early stop strategy can make good use of the advantages of the head-level distillation to accelerate the convergence of students in the early stage, so as to achieve better results, the corresponding training accuracy curve is shown in Figure 4 (b). Ablation study of different early stop epochs can be found in appendix.

Multiple layers for distillation. We tried distillation on different layers and the results are shown in appendix.

4.4 Comparison against existing KD methods

In this section, we compare the head-level distillation against two widely-used distillation methods, logits distillation (LKD) (Hinton et al. (2015)) and attention transfer (AT) (Zagoruyko & Komodakis (2016)). We follow the common practice that using SL model as the teacher for logits distillation and attention transfer. Since our method adopts two teachers, to be fair, we add SSTA to the above distillation methods during training. The results are shown in Table 3, it can be observed that SSTA combining head-level knowledge from SL teacher is better than combining the form of AT/logits. We also find that combining AT performs even worse than baseline.

4.5 Transfer learning on downstream tasks

In order to analyze the generalization of representations obtained by our method, we further conduct linear evaluation on Cifar10 and Cifar100, and the results are shown in Table 4. It can be seen that compared to the baseline without any distillation, our method can significantly improve the

SL Teacher KD Method	SSTA KD Method	Student Acc@1
-	-	72.2
LKD	Head-level	73.4
AT	Head-level	70.0
Head-level	Head-level	**74**

Table 3: Comparison against existing distillation methods. All the teachers are Deit-S, and students are Deit-Ti.

Dataset	Teacher1	Acc@1	Teacher2	Acc@1	Student	Acc@1
CIFAR100	-	-	-	-	DeiT-Ti	71.9
	DeiT-S (SL)	78.0	-	-	DeiT-Ti	72.2
	DeiT-S (SSL)	80.9	-	-	DeiT-Ti	72.2
	DeiT-S (SL)	79.6	DeiT-S (SL)	78.0	DeiT-Ti	72.0
	DeiT-S (SSL)	80.9	DeiT-S (SL)	78.0	DeiT-Ti	**72.8**
	-	-	-	-	DeiT-S	78.0
	DeiT-B (SSL)	84.5	DeiT-B(SL)	82.6	DeiT-S	**80.4**
CIFAR10	-	-	-	-	DeiT-Ti	90.4
	DeiT-S (SL)	93.9	-		DeiT-Ti	90.7
	DeiT-S (SSL)	95.0	-		DeiT-Ti	91.1
	DeiT-S (SL)	94.5	DeiT-S (SL)	93.9	DeiT-Ti	91.2
	DeiT-S (SSL)	95.0	DeiT-S (SL)	93.9	DeiT-Ti	**91.6**
	-	-	-	-	DeiT-S	93.9
	DeiT-B (SSL)	96.4	DeiT-B(SL)	95.9	DeiT-S	**95.2**

Table 4: Performance of transfering to downstream classification task on CIFAR.

classification accuracy on both Cifar10 and Cifar100. Furthermore, when using SSL teacher with better generalization as teaching assistant, the student is better than using SL teacher as teaching assistant. The results prove that the introduction of SSL teacher (SSTA) can make the students have better generalization, which further verifies the effectiveness of our method.

5 ANALYSIS

In this section, we did some in-depth analyses towards the otherness between the representations obtained by different learning paradigms. Firstly, we explore the prediction preference of SL teacher and SSL teacher, and then further analyze the shape bias of teachers and students, and the robustness of networks, finally we provide some visualizations. Note the teachers in all experiments of this part are DeiT-S, and the students are DeiT-Ti.

5.1 PREDICTION PREFERENCE

Figure 5 demonstrates the distribution of predictions of the top 10 categories by SL teacher and SSL teacher. It can be seen that these two models have different tendencies for the predicted categories. Furthermore, we counted the number of samples in which one of SL teacher and SSL teacher has a correct prediction but the other has a wrong prediction, which accounted for 11.3% of the validation dataset. In addition, the top 3 categories that SL teacher predicted correctly but SSL teacher predicted incorrectly are *lighter*, *spatula* and *coffee mug*, but the top 3 classes that SL teacher predicted incorrectly but SSL teacher predicted correctly are *cornet*, *sports car* and *drum*. **These data prove that the two models with the same structure obtained with different learning paradigms have different prediction preferences, which is what we are trying to exploit.**

5.2 SHAPE BIAS

Tuli et al. (2021) reported that the errors of vision transformers are more consistent with those of humans, compared to CNN. We are interested in comparison of ViTs with different representations and human vision. Following (Geirhos et al. (2018)), we evaluate shape bias on SIN dataset. We find that our proposed SSTA lets students have a higher shape bias and behave more like human. More detail can be found in appendix.

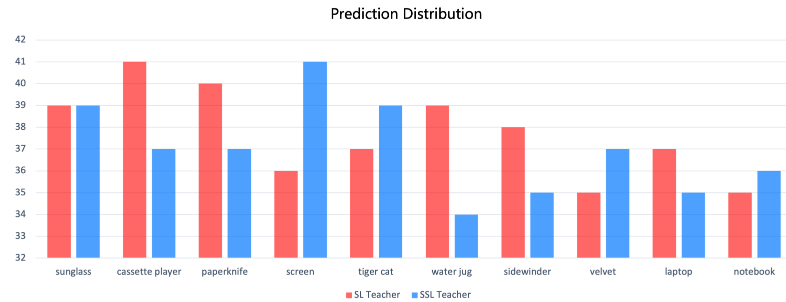


Figure 5: Prediction distribution. The abscissa is the top 10 categories in the validation dataset of ImageNet predicted by SL teacher and SSL teacher, and the ordinate is the specific number.

5.3 ROBUSTNESS

We measure the robustness on ImageNet-C, and the results are shown in Table 5. **It is obvious that our proposed SSTA can improve the robustness of student**, compared to both the student distilled by two different SL teachers (52.1 v.s. 53.0) and without distillation (52.1 v.s. 54.0). Moreover, it is worth noting that the results in Table 5 show that SL teachers have stronger robustness, while the robustness of the student distilled by two SL teachers is worse than the student distilled by a SL teacher together with another SSTA, which further proves the effectiveness of SSTA.

Model	mCE (↓)
Teachers	
DeiT-S (SL)	41.4
DeiT-S (SL) *	40.7
DeiT-S (SSL)	51.5
Students	
DeiT-Ti (Baseline)	54.0
DeiT-Ti (2 SL teachers)	53.0
DeiT-Ti (Ours)	**52.1**

Table 5: Performance on ImageNet-C. * represents the model is obtained by different initialization. The lower the mCE value, the better.

5.4 VISUALIZATIONS

As shown in Figure 6, compared to baseline (right) which is trained without any distillation, **our student pays more attention to objects**, especially the first head since it mimics the most important head of SSL teacher (SSTA). For example, when recognizing the *loggerhead* (the first input), since the key areas are not focused, baseline misjudges it as *pug-dog*, but our student can predict correctly. More visualizations can be seen in appendix, including the most important heads and the attention maps of the last layer.

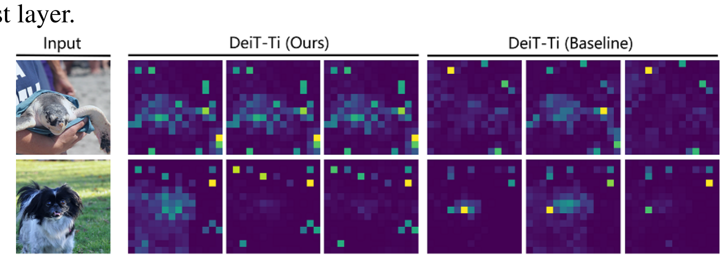


Figure 6: Visualizations of self-attention from the last layer. DeiT-Ti(Ours) consists of 3 heads, and the 1st head and 2nd head are distilled from SSL teacher (SSTA) and SL teacher respectively.

6 CONCLUSION

In this paper, we exploit using a self-supervised transformer as the teaching assistant besides the commonly used supervised teacher, and propose a head-level knowledge distillation approach to achieve this. Experiments demonstrate that self-supervised models are good teaching assistants for transformers. Meanwhile, some analytical experiments towards the difference between the supervised and self-supervised learning paradigms are inductive and may inspire future researches.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised visual transformers. *arXiv e-prints*, pp. arXiv–2104, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *arXiv preprint arXiv:2106.09681*, 2021.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021.

Yuting Gao, Jia-Xin Zhuang, Ke Li, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Xing Sun. Disco: Remedy self-supervised learning on lightweight models with distilled contrastive learning. *arXiv preprint arXiv:2104.09124*, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. *arXiv preprint arXiv:2104.01136*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *arXiv preprint arXiv:2105.10497*, 2021.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10428–10436, 2020.

Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, and Hang Zhao. Co-advise: Cross inductive bias distillation. *arXiv preprint arXiv:2106.12378*, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.

Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*, pp. 588–604. Springer, 2020.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.