Attention Distillation: self-supervised vision transformer students need more guidance

Kai Wang
kwang@cvc.uab.es
Fei Yang (*corresponding author)
fyang@cvc.uab.es
Joost van de Weijer
joost@cvc.uab.es

Computer Vision Center
Universitat Autònoma de Barcelona
Barcelona, Spain

Abstract

Self-supervised learning has been widely applied to train high-quality vision transformers (ViT). Unleashing their excellent performance on memory and compute constraint devices is therefore an important research topic. However, how to distill knowledge from one self-supervised ViT to another has not yet been explored. Moreover, existing self-supervised knowledge distillation (SSKD) methods focus on ConvNet architectures and are suboptimal for ViT knowledge distillation. In this paper, we study knowledge distillation of self-supervised vision transformers (ViT-SSKD). We show that directly distilling information from the crucial attention mechanism from teacher to student can significantly narrow the performance gap between both. In experiments on ImageNet-Subset and ImageNet-1K, we show that our method AttnDistill outperforms existing self-supervised knowledge distillation (SSKD) methods and achieves state-of-the-art k-NN accuracy compared with self-supervised learning (SSL) methods learning from scratch (with the ViT-S model). We are also the first to apply the tiny ViT-T model for self-supervised learning. Moreover, AttnDistill is independent of self-supervised learning algorithms, and it can be adapted to ViT based SSL methods to improve performance in future research.

1 Introduction

Vision transformers [18] have been widely applied in computer vision tasks, including image classification [55, 62, 70], object recognition [4, 13, 19, 52, 76] and semantic segmentation [12, 50, 64, 73]. ViTs contain a self-attention mechanism [59] that allows for information exchange between distant patches and consequently leads to a more holistic understanding of image content. Another important aspect of transformers is that they are often pretrained in a self-supervised manner, followed by a finetuning stage to adapt to the downstream task [15, 37]. ViTs suffer from high memory requirements and substandard optimizability [11, 14, 25, 41, 60, 68], making them unsuitable for applications on memory or computation constraint devices. Consequently, methods that can reduce the memory footprint while maintaining the performance of ViTs are in demand.

One transfer learning technique is knowledge distillation [32]. Initial works focussed on knowledge transfer for networks trained in a supervised manner [7, 57, 66, 69]. Recently, the theory was extended to distill knowledge of self-supervised feature representations generated by large networks [21, 42, 43]. Since these networks do not output a conditional probability over a label set, but rather a feature representation, alternative distillation techniques needed to be developed [1, 21]. With the advent of transformers, supervised knowledge distillation for transformers has recently been investigated [33, 38, 54]. However, methods that can transfer self-supervised ViTs to smaller variants have not yet been explored.

Therefore, we explore knowledge distillation of self-supervised ViTs. We find that existing theory designed to transfer knowledge of ConvNets trained in a self-supervised manner results in a significant performance gap between teacher and student. To address this problem, we explore attention distillation that focuses on transferring the information present in the self-attention mechanism. Rather than just communicating the teacher's conclusion which is the focus of most traditional knowledge distillation methods, attention distillation provides more guidance to the student network by identifying the important regions for understanding the image content. The potential of attention distillation has been explored for ConvNet [24, 71], however, since *for these networks attention is not explicitly computed, additional computation and attention definition are needed.* Since the attention mechanism is an integral and crucial part of transformers and *no additional* computation is required, we argue that attention distillation is a *natural extension* of the existing distillation theory for transformer networks.

In this paper, we focus on *self-supervised knowledge distillation of self-supervised vision transformers* (ViT-SSKD). First, we propose to use a *projector alignment (PA)* module to align the class tokens from teacher and student models. Second, we propose *attention guidance (AG)* with the Kullback–Leibler divergence to guide the student to obtain similar attention maps as the teacher model to further enhance the distillation. With these two modules, we can obtain state-of-the-art performance compared with self-supervised algorithms. Furthermore, we are the *first* to successfully train a small ViT-T model based on self-supervised learning (SSL) with knowledge distillation. More importantly, there might be more complex and outperforming SSL pretrained models in the future. In that case our method can be *applied directly* to obtain a smaller model while keeping competitive performance. Our main contributions are:

- We are the first to study the important ViT-SSKD problem allowing to transfer knowledge to small transformers in a self-supervised fashion.
- We propose an attention distillation loss for improved guidance of the student during knowledge distillation. Our method, AttnDistill, significantly reduces the gap between teacher and student models.
- We are the first to train a self-supervised ViT-T model. It obtains a performance almost (-0.3%) at par with the supervised ViT-T model.

2 Related work

Self-supervised learning. SSL [8, 9, 16, 23, 26, 29, 30, 58, 63] automatically derives a supervisory signal for the training of high-quality feature representations, preventing the need of large labeled datasets. The common paradigm here is to pretrain on ImageNet [49] and then evaluated on downstream tasks, on which it has reached excellent performance, closing the gap with supervised methods. Recent popular SSL methods can be divided into two streams. *Contrastive learning* [6, 8, 27, 29] is the most popular stream. Another stream of representation learning, named *masked image encoding* [3, 17, 30, 74], learns representations from corrupted images. *In this paper, we study knowledge distillation for SSL based on both technical streams.* From the aspect of backbone architectures, the previous methods are all based on ConvNet [5, 8, 26]. Recently, with the appearance of ViT, these are also applied for SSL (DINO [6], MoCo [10], etc). Compared with ConvNet, the attention-based ViTs suffer less from an image-specific inductive bias and have a larger potential when training on large scale datasets. *In this paper we focus on the original ViT design, but the method could also be generalized to Swin Transformer [39] based SSL methods [37, 65].*

Knowledge distillation for self-supervised models. Most knowledge distillation [32] techniques are proposed under the supervised learning scenario [2, 7, 44, 46, 48, 53, 57, 69, 71, 72]. Under the *SSL* settings, CC [43] exploit pseudo labels from clustering teacher embeddings as distillation signals. Then SEED [21] and CompRess [1] maintain memory banks to store a huge number of samples to calculate instance-level similarity score distributions for aligning the teacher and student models. SimReg [42] has similar projector architecture as our method, where they use the projector to align the teacher and student features. However, in some cases when the student ViT architectures become quite different from the teacher model, only projector regression is not sufficient to transfer knowledge from the teacher to the student model. Reg [67] is specified for metric learning, which could also be applied to self-supervised representation distillation. Recently, KDEP [31] propose the power temperature scaling to distill representation from a supervised teacher model.

Except for these examples in computer vision, there are several distillation attempts in NLP [20, 34, 51, 61]. However, these methods are limited to the case that teacher-student models share similar architectures. Also, Pelosin et al. [45] apply attention distillation between similar transformer architectures for continual learning. Our proposal is a more generalizable framework and allows for attention distillation between different ViT architectures.

3 Methodology

3.1 Preliminaries

Vision Transformers Architecture. Here we consider the ViT proposed in [18] but the theory is general and can be extended to other transformer architectures. The ViT consists of a patch embedding part, where the transformer encoder is a stack of L multi-head self-attention blocks (MH-SAB). In each MH-SAB, there are two parts: a multi-head self-attention module (MSA) and a fully connected feedforward module (MLP). Each self-attention module has H *heads*. We will further use d as the output dimension for each head and N as the number of patches. Also considering the class token, we can denote the output of the lth MH-SAB as $\mathbf{z}^l \in \mathbb{R}^{(N+1)\times(dH)}, l \in [1,L]$. $\mathbf{z}^0$ is the encoded patch embedding of image $\mathbf{x}$ (i.e., from $\mathbf{z}^0_1$ to $\mathbf{z}^0_N$) and the initial class token (i.e., $\mathbf{z}^0_0$). For the hth head in the self-attention module, learnable parameters $W_Q^{l,h}, W_K^{l,h}, W_V^{l,h}$ implemented as FC layers, map one slice of the input tokens $\mathbf{z}^{l-1,h}$ into the queries, keys and values ($Q^{l,h}, K^{l,h}, V^{l,h} \in \mathbb{R}^{(N+1)\times(d)}$). We obtain the attention map with Eq. 1, where $\mathbb{A}^{l,h} \in \mathbb{R}^{(N+1)\times(N+1)}$, and the output of this head is obtained by Eq. 2. Combining the multiple head outputs, we obtain the final output of this multi-head self-attention layer (see Eq. 3). Finally, with the MLP and layer normalization (LN), the output tokens produced by the lth MH-SAB are given by Eq. 4. We have shown the process of the last MH-SAB from both the teacher and student ViT separately in Fig. 1 (the design is

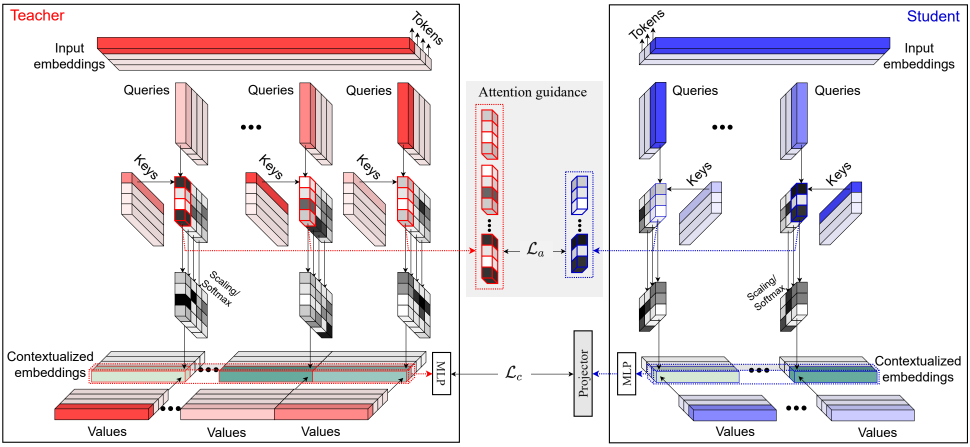


Figure 1: AttnDistill for ViT-SSKD on the last block of the ViT. It is composed of the projector alignment loss $\mathcal{L}_c$ and attention guidance loss $\mathcal{L}_a$. The class tokens are taken from the last layer of the teacher and student ViT. We only consider the attention vectors that are formed by the interaction of the class token query with all keys for distillation.

inspired from [22].) which includes the key components in our approach.

$$\mathbb{A}^{l,h} = \text{Softmax}(Q^{l,h} \cdot (K^{l,h})^T / \sqrt{d}) \tag{1}$$

$$\mathbf{y}^{l,h} = \mathbb{A}^{l,h} \cdot V^{l,h} \tag{2}$$

$$\mathbf{y}^{l} = concat(\mathbf{y}^{l,1}, \mathbf{y}^{l,2}, ..., \mathbf{y}^{l,H}) \tag{3}$$

$$\mathbf{z}^{l} = \text{MLP}\left(\text{LN}\left(\mathbf{y}^{l} + \mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1}. \tag{4}$$

3.2 AttnDistill: Attention distillation

An outline of AttnDistill is given in Fig. 1. It can be divided into two parts: projector alignment (PA) and attention guidance (AG). For our distillation, we do not need to generate multiple views, since we focus on distilling the knowledge of the teacher. So other than contrastive-based SSL methods [6, 74, 75], in this phase *we do not rely on multi-crop augmentations*. Thus, we have less computation costs and *the effective epoch is equal to the training epoch.*

Projector alignment (PA). Suppose we have a teacher-student pair, each with a ViT architecture named V_t and V_s. In self-supervised knowledge distillation, our aim is to distill knowledge from the teacher to the student model in a SSL way while maintaining its transferability. In most cases, we expect a smaller student model compared to the teacher. *The parameter size is highly dependent on the feature dimension in the ViT. Thus, V_t and V_s typically have different feature dimensions.* Therefore, we introduce a linear mapping projector $\mathcal{P}$ to map the student to the teacher feature space for alignment. And since in ViT, *the class token embedding* $\mathbf{E}^c = \mathbf{z}_0^L$ is the most representative embedding for a classification decision, in AttnDistill, we only map the class token from the last layer for aligning the teacher and student model with a MSE loss to *communicate* with the final output from the teacher:

$$\mathcal{L}_c = ||\mathbf{E}_t^c - \mathcal{P}(\mathbf{E}_s^c)||_2 \tag{5}$$

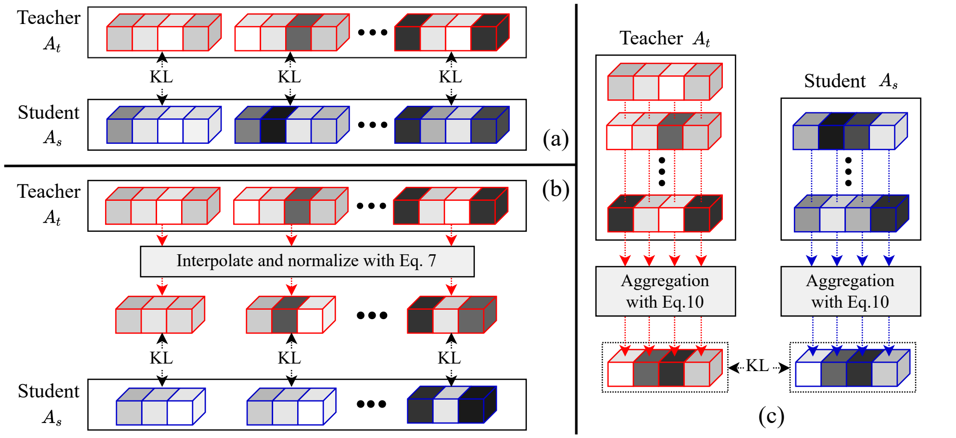


Figure 2: Attention guidance between varying transformer architectures.

where 's/t' subscripts represent 'student/teacher'. In our ablation study, we also explored the influence of aligning the image patch embeddings $\mathbf{z}_i^L, i \in [1,N]$ with the linear mapping $\mathcal{P}$.

Attention guidance (AG). However, aligning the class tokens can only tell the student model about "what" is in the image. More guidance from the teacher explaining "why" it reached this conclusion could be helpful. This extra guidance can be extracted from the multi-head self-attention (MSA) mechanism in ViTs. The MSA module pays attention to the decisive and informative parts in the image. Actually, we can observe that distillation with other methods will lead to *attention drift* where the student attention differs from the teacher (see the attention maps in Fig. 3.)

For a ViT, each $\mathbb{A}^{l,h}$ given by layer l head h is an attention map from all tokens to all tokens. And since $\mathbb{A}^{L,h}_{0,j}, j \in [0,N]$ ('0' represents the first row of the attention map $A^{L,h}$.) contains the attention probabilities for the class token, it represents the importance of each token for the classification prediction of the image. By denoting $A = \mathbb{A}^L_0, A^h = \mathbb{A}^{L,h}_0, a^h_j = \mathbb{A}^{L,h}_{0,j}$, we propose to apply Kullback–Leibler divergence ($\mathcal{KL}$) to make the student model pay attention to the same regions as the teacher model by aligning A_s and A_t. We also study the performance with attentions from all layers in our ablation study section. To address knowledge transfer between ViTs with different designs, several cases need to be addressed. Here, we categorize them into four cases that might occur and discuss the solution below. The illustration of these variations are provided in Fig. 2.

(a) The teacher and student models have the same number of heads $H = H_t = H_s$ and patches $N = N_t = N_s$ (see Fig.2 (a)). This is the simplest case, where we align according to:

$$\mathcal{L}_a = \sum_{h \in [1,H]} \mathcal{KL}(A^h_t || A^h_s) \tag{6}$$

(b) The teacher and student models have the same number of heads H but a different number of patches N_t and N_s (see Fig.2 (b)). In this case, we propose to interpolate (**IP**, by default we apply bicubic function [35] as a smoother interpolation) the teacher model attention map $(a^h_j)_t, j \in [1,N_t]$ into $(a^h_j)'_t, j \in [1,N_s]$ ($N_t = w \times h, N_s = w' \times h'$), and then normalize (**NR**) them into $1-(a_0^h)_t$ by scaling up to have the attentions sum to 1 as in Eq. 7. Then the *attention guidance* loss is given by Eq. 8.

$$(a_j^h)_t' = \mathbf{NR}_{1-(a_0^h)_t}(\mathbf{IP}((a_j^h)_t)) \tag{7}$$

$$\mathcal{L}_a = \sum_{h \in [1,H]} \mathcal{KL}((A^h)_t' || A_s^h) \tag{8}$$

(c) The teacher and student models have the same number of patches N but a different number of heads H (see Fig.2 (c)). Here we merge the attentions from all heads for distillation. We have considered several aggregation functions, including mean, maximum and soft-maximum. We found that using the *log* summation to aggregate the attention probabilities for both teacher and student models (see Eq. 9 and Eq. 10) leads to slightly superior results compared to max-based fusion. Then the *attention guidance* loss $\mathcal{L}_a$ is as Eq. 11.

$$a_j = \frac{1}{T} \cdot \sum_{h \in [1,H]} log(a_j^h) = \frac{1}{T} \cdot log(\prod_{h \in [1,H]} a_j^h) \tag{9}$$

$$A = \texttt{Softmax}([a_0, a_1, ..., a_N]) \tag{10}$$

$$\mathcal{L}_a = \mathcal{KL}(A_t || A_s) \tag{11}$$

This aggregation could effectively highlight the maximum probabilities from the attention maps of the H heads, as can be seen from our ablation study.

(d) The teacher and student models have a different number of heads H and patches N. This case is a combination of the above two, thus we apply interpolation and aggregation sequentially and then apply distillation.

Finally, the self-supervised loss to update the student model V_s is:

$$\mathcal{L} = \mathcal{L}_c + \lambda \cdot \mathcal{L}_a \tag{12}$$

4 Experiments

4.1 Pre-Training setup

Datasets. In our experiments, ImageNet-Subset [49] is used for ablation study and to compare with other self-supervised knowledge distillation methods. This dataset contains 100 classes and ≈130k images in high resolution (resized to 224×224) [47]. For comparison with SSL methods, we employ the ImageNet-1K dataset [49].

Architecture. For the Teacher-Student pairs, we focus on knowledge distillation from a larger ViT teacher model to a smaller ViT student model. Due to the high computation demands of ViT, we select Teacher-Student pairs as below:

- On ImageNet-1K we select the following three pairs (Teacher →Student): (a) Mugs(ViT-S/16) →ViT-T/16; (b) Mugs(ViT-B/16) →ViT-S/16; (c) DINO(ViT-S/8) →ViT-S/16;
- On ImageNet-Subset, we fix the teacher model as MAE(ViT-S/16) with 12-Layer, 6-Head, 16-Patch and vary the design of the student model in the ablation study.

Teacher model	Method	Student Arch.	Par.(M)	Train Epo.	Effect Epo.	k-NN	LP.
✗	Supervised	ViT-T/16	5.7	-	-	72.2	72.2
SwAV (RN-50)	CRD	RN-18	11	240	240	44.7	58.2
SwAV (RN-50)	CC	RN-18	11	100	100	51.0	60.8
SwAV (RN-50)	Reg	RN-18	11	100	100	47.6	60.6
SwAV (RN-50)	CompRess-2q	RN-18	11	130	130	53.7	62.4
SwAV (RN-50)	CompRess-1q	RN-18	11	130	130	56.0	65.6
SwAV (RN-50)	SimReg	RN-18	11	130	130	59.3	65.8
SwAV (RN-50×2)	SEED	RN-18	11	200	200	55.3	63.0
SwAV (RN-50×2)	SEED	EffNet-B1	7.8	200	200	60.3	68.0
SwAV (RN-50×2)	SEED	EffNet-B0	5.3	200	200	57.4	67.6
SwAV (RN-50×2)	SEED	MbNet-v3	5.5	200	200	55.9	68.2
Mugs (ViT-S/16)	AttnDistill	ViT-T/16	5.7	500	500	**71.4**	**71.9**
✗	Supervised	ViT-S/16	22	-	-	79.8	79.8
✗	SimCLR	ViT-S/16	22	300	600	-	69
✗	BYOL	ViT-S/16	22	300	600	-	71
✗	MoCo v3	ViT-S/16	22	600	1200	-	73.4
✗	SwAV	ViT-S/16	22	800	2400	66.3	73.5
✗	DINO	ViT-S/16	22	800	3200	74.5	77.0
✗	iBOT	ViT-S/16	22	800	3200	75.2	77.9
✗	MUGS	ViT-S/16	22	800	3200	75.6	**78.9**
SwAV (RN-50×2)	SEED	RN-34	21	200	200	58.2	65.7
SwAV (RN-50×2)	SEED	RN-50	24	200	200	59.0	74.3
SimCLR (RN-50×4)	CompRess-1q	RN-50	24	130	130	63.3	71.9
SimCLR (RN-50×4)	CompRess-2q	RN-50	24	130	130	63.0	71.0
SimCLR (RN-50×4)	CC	RN-50	24	100	100	55.6	68.9
SimCLR (RN-50×4)	SimReg	RN-50	24	130	130	60.3	74.2
Mugs (ViT-B/16)	AttnDistill	ViT-S/16	22	800	800	76.8	78.6
DINO (ViT-S/8)	AttnDistill	ViT-S/16	22	800	800	**77.4**	78.8
Teacher Models statistics							
Mugs (ViT-S/16)	-	-	22	800	3200	75.6	78.9
DINO (ViT-S/8)	-	-	22	800	3200	78.3	79.7
Mugs (ViT-B/16)	-	-	85	400	1600	78.0	80.6
SwAV (RN-50)	-	-	24	800	2400	64.8	75.6
SwAV (RN-50×2)	-	-	94	800	2400	-	77.3
SimCLR (RN-50×4)	-	-	375	1000	2000	64.5	75.6

Table 1: Comparison with state-of-the-art SSL methods with k-NN and linear probing (LP.) on **ImageNet-1K**. "Effect Epo." is the effective pretraining epochs computed by multiplying number of views processed by the models following iBOT [74].

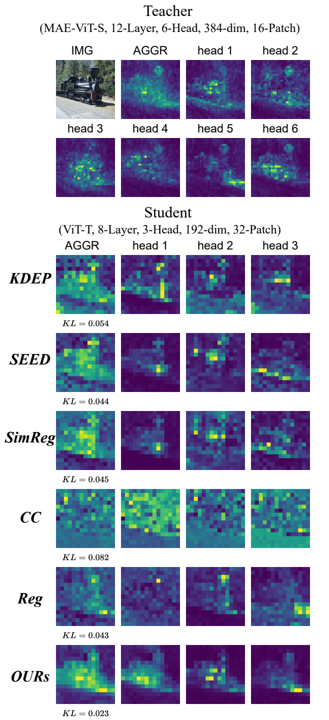


Figure 3: For the teacher model, we show the original image, the aggregated attention map (AGGR) and the attention maps for each head. The KL distances to the teacher AGGR are shown.

Moreover, since MAE/MoCo-v3 and DINO/iBOT/Mugs are following different position embedding strategies (fixed vs. learnable), our teacher-student pairing setups can show the effectiveness of AttnDistill for these two kind of position encodings.

Implementation details. We train our AttnDistill with the AdamW [40] optimizer. The learning rate is linearly ramped up during the first 40 epochs to the base learning rate $lr = (1.5e-4) \times \text{batchsize}/256$. After the warming up epochs, we decay it with a cosine schedule till 800 epochs (except the distillation from Mugs(ViT-S/16) to ViT-T/16, where we train for 500 epochs because of performance saturation). By default, we set $T = 10.0$ and $\lambda = 0.1$, the projector $\mathcal{P}$ is a 4-layer linear mapping. For the evaluations of the student model, we found it is optimal to perform with the features before the $\mathcal{P}$. For the experiments on ImageNet-Subset, we fix the teacher model as a ViT-S/16 pre-trained with MAE [30] method with 3200 epochs. More details are in our supplementary materials.

4.2 Comparison with state-of-the-art

Comparison with SSKD methods on ViT. As shown in Table 2, we compare several methods to distill a MAE(ViT-S/16) teacher to a ViT-T student on *ImageNet-Subset*. Next to the common ViT-T architecture (with 12-layers, 6-heads, 16-patches), we also consider a harder variant (with 8-layers, 3-heads, 32-patches) as the student model. AttnDistill clearly outperforms them all in both cases. The margin is larger for the harder setup. This also indicates the importance of the attention distillation guidance. A comparison of the attention maps from the 8-layer 3-head 32-patch student case is shown in Fig. 3.

Comparison with SSL methods with linear probing, *k*-NN and finetuning accuracies. For linear probing and *k*-NN evaluation on *ImageNet-1K*, we follow the commonly used setting in DINO [6] and iBOT [74]. The comparison is shown in Table 1. We draw the following conclusions:

- Based on ViT-T/16 distilled from Mugs(ViT-S/16), our method AttnDistill gets state-of-the-art *k*-NN and Linear probing performance compared with previous knowledge distillation methods based on ConvNet. AttnDistill (ViT-T/16) is with only 5.7M parameters but outperforms the previous methods by a large margin and gets quite close to the supervised ViT-T/16 learning from scratch. In this case, we only train the student model for 500 epochs since we have observed marginal improvement on linear probing after that.

- Based on ViT-S/16 distilled from DINO(ViT-S/8), our method AttnDistill gets state-of-the-art in *k*-NN and the second in linear probing. This distillation decreases the ViT computational demand since there are 75% less patches in ViT-S/16 than ViT-S/8. Then, based on ViT-S/16 distilled from Mugs(ViT-B/16), AttnDistill gets the second in *k*-NN and the third in linear probing evaluations. In this case, the model size is decreased by 75%.

Moreover, we could further observe the advantage of ViT in *k*-NN evaluations, which means the extracted features from self-supervised pretrained ViT are more beneficial without learning an extra classifier as in linear probing evaluations. The accuracy curves during training are shown in Fig. 4.

For the finetuning comparison on ImageNet-1K shown in Table 3, we compared with existing methods working on ViT-T and ViT-S. AttnDistill (ViT/T) and AttnDistill (ViT/S) are distilled from Mugs (ViT-B) and Mugs (ViT-S) respectively. In both cases, AttnDistill works better than supervised learning methods and just marginally worse than the state-of-the-art with ViT-S. In conclusion, whereas AttnDistill is state-of-the-art for *k*-NN evaluation (Table 1), this is not the case when evaluating by means of finetuning.

Comparison with SSL methods on downstream tasks. For semi-supervised learning, results with SSL methods based on ViT-S/16 are shown in Table 4. In this setting, first, models are trained self-supervised on all *ImageNet-1K* data. Next labels for a small fraction of data (1% or 10%) are used to perform fine-tuning, linear probing or *k*-NN classification. Under all three settings with only 1% of the data, we can observe a considerable advantage of AttnDistill with a 3.9%/2.6%/5.8% improvement compared with Mugs(ViT-S/16). With 10% data, the improvement is less but still notable as 0.4%/2.7%/3.1%.

We also evaluate AttnDistill (ViT-T) for transfer learning. We compare with supervised ViT models, since there are no papers using ViT-T for SSL. Results are summarized in Table 5) for several small datasets. AttnDistill gets a 2.3% improvement compared with the previous best supervised learning method CCT-7/3x1 [28]. Next, we consider transfer learning of AttnDistill (ViT-S) in Table 6. Here we compare with previous SSL methods, we are only

Method	Top-1	Top-5	Top-1	Top-5
Student ViT-T Arch.	12-L, 6-H, 16-P		8-L, 3-H, 32-P	
Teacher MAE-S/16	79.4	93.6	79.4	93.6
SEED [illegible]	71.7	90.7	66.8	88.4
CompRess [1]	71.9	90.8	67.2	88.7
CC [illegible]	63.8	87.6	47.1	74.9
KDEP [illegible]	66.7	87.8	57.8	82.4
Reg [illegible]	76.2	92.5	69.6	89.0
SimReg [illegible]	77.8	93.4	68.6	88.8
AttnDistill	**79.3**	**94.1**	**73.8**	**91.7**

Table 2: Compare with SSKD methods on ImageNet Subset with top-1/top-5 (LP.)

Method	Arch.(M)	Effect Epo.	FT Acc.
Supervised learning			
-	ViT-T/16	-	72.2
-	ViT-S/16	-	79.8
DeiT	ViT-S/16	-	81.3
DeiT-III [illegible]	ViT-S/16	-	81.4
Manifold [illegible]	ViT-S/16	-	81.5
MKD [illegible]	ViT-S/16	-	82.1
Self-Supervised learning			
MoCo v3	ViT-S/16	600	81.4
DINO	ViT-S/16	3200	82.0
iBOT	ViT-S/16	3200	82.3
Mugs	ViT-S/16	3200	**82.6**
AttnDistill	ViT-T/16	500	72.9
AttnDistill	ViT-S/16	800	81.6

Table 3: Finetuning comparison on ImageNet1K.

Method	Arch	FT 1%	FT 10%	LP 1%	LP 10%	*k*-NN 1%	*k*-NN 10%
SimCLR	RN50	57.9	68.1	-	-	-	-
BYOL	RN50	53.2	68.8	-	-	-	-
SwAV	RN50	53.9	70.2	-	-	-	-
SimCLR+SD	RN50	60.0	70.5	-	-	-	-
DINO	ViT-S/16	60.3	74.3	59.1	70.3	61.2	69.0
iBOT	ViT-S/16	61.9	75.1	61.5	71.8	62.5	70.1
Mugs	ViT-S/16	66.8	76.8	64.1	72.2	63.6	70.6
AttnDistill	ViT-S/16	**70.7**	**77.2**	**66.7**	**74.9**	**69.4**	**73.7**

Table 4: Semi-supervised learning on ImageNet1K. AttnDistill (ViT-S) is distilled from the teacher model Mugs (ViT-B/16) for 800 epochs.

Method	Par.(M)	CIFAR100	CIFAR10
Supervised learning			
SL-CaiT [36]	9.2	80.3	95.8
SL-T2T [36]	7.1	77.4	95.6
SL-Swin [36]	10.2	80.0	95.9
CVT-7/4 [28]	3.7	73.0	92.4
CCT-7/3x1 [28]	3.8	82.7	98.0
Self-Supervised + Transfer learning			
AttnDistill (ViT-T)	5.7	**85.0**	**98.1**

Table 5: Transfer learning comparison on CIFAR10/CIFAR100. AttnDistill (ViT-T) is distilled from Mugs(ViT-S/16).

Method	Par.(M)	CIFAR100	CIFAR10	Flowers	Cars
Supervised + Transfer learning					
-	22	89.5	99.0	98.2	92.1
DeiT-III [56]	22	90.6	98.9	96.4	89.9
Self-Supervised + Transfer learning					
BEiT	22	87.4	98.6	96.4	92.1
DINO	22	90.5	99	98.5	93.0
iBOT	22	90.7	99.1	98.6	**94.0**
Mugs	22	**91.8**	**99.2**	**98.8**	93.9
AttnDistill (ViT-S)	22	91.6	99.1	98.6	93.8

Table 6: Compared with SSL methods on four small datasets. AttnDistill (ViT-S) is distilled from the teacher Mugs (ViT-B/16).

marginally worse than the state-of-the-art and much better than the supervised distillation method DEiT [54].

4.3 Ablation study

To prove the generalizability of AttnDistill, we perform an ablation study on *ImageNet-Subset* with a fixed MAE(ViT-S/16) teacher and vary the architecture of the student model. Extended ablation studies are in the supplementary. In Fig. 5, our ablation study contains three parts:

(a) **The architecture of ViT (in Fig. 5-(a))** : To verify the effectiveness of AttnDistill for various architectures of ViT, we modify the number of heads, patch sizes and the number of block layers. In all cases, AttnDistill significantly improves the PA baseline and closes the gap with the teacher performance. Especially, for the smaller student architectures and those with fewer tokens, attention distillation is shown to be crucial leading to improvements of over 5%.

(b) **The various aggregation functions (in Fig. 5-(b))** : Here we fix the design of the student model and vary the strategy to compute the attention guidance loss. To verify the superiority of the used log summation in Eq. 9, we replace it with *MEAN/MIN/MAX* strategies to aggregate attention maps from different heads. However, they are all suboptimal.

(c) **Alternative self-supervised losses (in Fig. 5-(c))** : A recent work for distillation of self-supervised representations of ConvNet is CompRess [1]. Here, we apply the knowledge distillation loss from CompRess [1] to our PA module, we can clearly observe that this

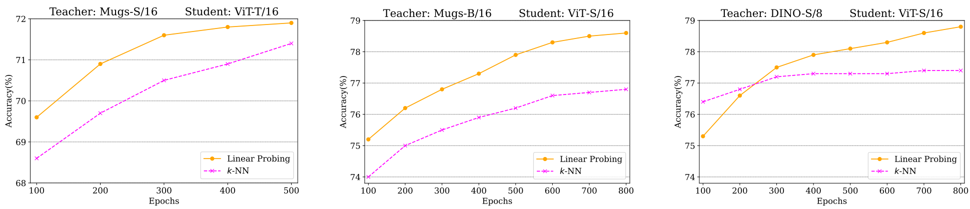
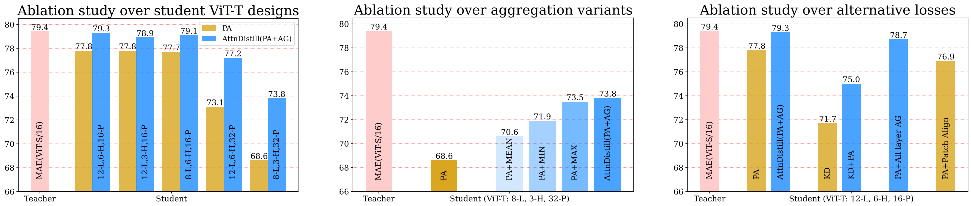


Figure 4: More results of AttnDistill with different training epochs on **ImageNet-1K**.



(a) ViT architectures (b) aggregation functions (c) alternative losses

Figure 5: Ablation study for ViT architectures, aggregation functions and alternative losses.

KD loss is an obstacle in ViT distillation since when combined with our PA module it leads to a performance drop. Except distilling the attention maps from the last layer, we also experiment the distillation over attention maps from all layers. This is 0.6% lower than AttnDistill based on only the last layer. Finally, we also align the patch tokens with our PA module. This is 2.4% worse than without the patch token alignment, thus aligning patch tokens is not necessary.

5 Conclusion

In this paper, we explored the ViT-based self-supervised knowledge distillation problem. Observing that the previous SSKD methods focussed on ConvNet do not work well on ViT, we proposed *AttnDistill* to distill the knowledge from a pretrained teacher model to its student model. The experiments clearly show that AttnDistill outperforms other SSKD methods. Furthermore, our distilled ViT-S gets state-of-the-art in k-NN accuracy and is second in linear probing compared with SSL methods. Also, our method AttnDistill is especially advantageous in semi-supervised learning evaluation and competitive in transfer learning evaluation. To prove the effectiveness of AttnDistill, we also implement various ablation studies on ImageNet-Subset. For future work, we are interested to explore AttnDistill for knowledge distillation between ConvNets and ViT.

Limitations. A drawback of the attention mechanism is that it is tailored for transformer usage and requires additional computation when applied to ConvNets (namely the computation of the attention maps). A further limitation is that the theory only applies to a single teacher-student pair. In case of multiple teacher models, further thought has to be given on how the multiple attention maps can be meaningfully communicated with the student.

Acknowledgement

We acknowledge the support from Huawei Kirin Solution, and the Spanish Government funded project PID2019-104174GB-I00/AEI/10.13039/501100011033.

References

[1] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. *Advances in Neural Information Processing Systems*, 33:12980–12992, 2020.

[2] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019.

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *International Conference on Learning Representations*, 2022.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[7] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021.

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

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

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[11] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[12] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1601–1610, 2021.

[14] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34: 3965–3977, 2021.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[17] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

[18] 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. *International Conference on Learning Representations*, 2020.

[19] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34, 2021.

[20] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1428–1438, 2021.

[21] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. In *International Conference on Learning Representations*, 2021.

[22] Romain Futrzynski. Getting meaning from text: self-attention step-by-step video. https://peltarion.com/blog/data-science/self-attention-video, 2020. Accessed: 2022-05-17.

[23] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations*, 2018.

[24] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[25] Benjamin 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12259–12269, 2021.

[26] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[27] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006.

[28] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021.

[29] 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*, pages 9729–9738, 2020.

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[31] Ruifei He, Shuyang Sun, Jihan Yang, Song Bai, and Xiaojuan Qi. Knowledge distillation as efficient pre-training: Faster convergence, higher data-efficiency, and better transferability. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[32] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Advances in Neural Information Processing Systems*, 2014.

[33] Ding Jia, Kai Han, Yunhe Wang, Yehui Tang, Jianyuan Guo, Chao Zhang, and Dacheng Tao. Efficient vision transformers via fine-grained manifold distillation. *arXiv preprint arXiv:2107.01378*, 2021.

[34] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *Empirical Methods in Natural Language Processing*, 2020.

[35] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981.

[36] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

[37] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=fVu3o-YUGQK.

[38] Jihao Liu, Boxiao Liu, Hongsheng Li, and Yu Liu. Meta knowledge distillation. *arXiv preprint arXiv:2202.07940*, 2022.

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

[41] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *International Conference on Learning Representations*, 2022.

[42] KL Navaneet, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Simreg: Regression as a simple yet effective tool for self-supervised knowledge distillation. *Proceedings of the British Machine Vision Conference*, 2021.

[43] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9359–9367, 2018.

[44] 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*, pages 3967–3976, 2019.

[45] Francesco Pelosin, Saurav Jha, Andrea Torsello, Bogdan Raducanu, and Joost van de Weijer. Towards exemplar-free continual learning in vision transformers: an account of attention, functional and weight regularization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[46] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019.

[47] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[48] 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.

[49] 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.

[50] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[51] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. *Annual Meeting of the Association for Computational Linguistics*, 2020.

[52] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3611–3620, 2021.

[53] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *International Conference on Learning Representations*, 2020.

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021.

[55] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021.

[56] Hugo Touvron, Matthieu Cord, and Herve Jegou. Deit iii: Revenge of the vit. *European Conference on Computer Vision*, 2022.

[57] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[58] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[60] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[61] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

[62] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021.

[63] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[64] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

[65] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021.

[66] Fei Ye and Adrian Bors. Lifelong teacher-student network learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[67] Lu Yu, Vacit Oguz Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, and Arnau Ramisa. Learning metrics from teachers: Compact networks for image embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2907–2916, 2019.

[68] Shixing Yu, Tianlong Chen, Jiayi Shen, Huan Yuan, Jianchao Tan, Sen Yang, Ji Liu, and Zhangyang Wang. Unified visual transformer compression. *International Conference on Learning Representations*, 2022.

[69] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020.

[70] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[71] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *International Conference on Learning Representations*, 2017.

[72] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. *arXiv preprint arXiv:2203.08679*, 2022.

[73] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[74] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022.

[75] Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. Mugs: A multi-granular self-supervised learning framework. *arXiv preprint arXiv:2203.14415*, 2022.

[76] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *International Conference on Learning Representations*, 2021.