Self-slimmed Vision Transformer

Zhuofan Zong[1⋆], Kunchang Li[3,4⋆], Guanglu Song[2], Yali Wang[3,5], Yu Qiao[3,6], Biao Leng[1], Yu Liu[2✉]

[1]School of Computer Science and Engineering, Beihang University
[2]SenseTime Research
[3]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[4]University of Chinese Academy of Sciences
[5]SIAT Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society
[6]Shanghai AI Laboratory

Abstract. Vision transformers (ViTs) have become the popular structures and outperformed convolutional neural networks (CNNs) on various vision tasks. However, such powerful transformers bring a huge computation burden, because of the exhausting token-to-token comparison. The previous works focus on dropping insignificant tokens to reduce the computational cost of ViTs. But when the dropping ratio increases, this hard manner will inevitably discard the vital tokens, which limits its efficiency. To solve the issue, we propose a generic self-slimmed learning approach for vanilla ViTs, namely SiT. Specifically, we first design a novel Token Slimming Module (TSM), which can boost the inference efficiency of ViTs by dynamic token aggregation. As a general method of token hard dropping, our TSM softly integrates redundant tokens into fewer informative ones. It can dynamically zoom visual attention without cutting off discriminative token relations in the images, even with a high slimming ratio. Furthermore, we introduce a concise Feature Recalibration Distillation (FRD) framework, wherein we design a reverse version of TSM (RTSM) to recalibrate the unstructured token in a flexible auto-encoder manner. Due to the similar structure between teacher and student, our FRD can effectively leverage structure knowledge for better convergence. Finally, we conduct extensive experiments to evaluate our SiT. It demonstrates that our method can speed up ViTs by **1.7**× with negligible accuracy drop, and even speed up ViTs by **3.6**× while maintaining **97**% of their performance. Surprisingly, by simply arming LV-ViT with our SiT, we achieve new state-of-the-art performance on ImageNet. Code is available at https://github.com/Sense-X/SiT.

Keywords: Transformer, Token Slimming, Knowledge Distillation

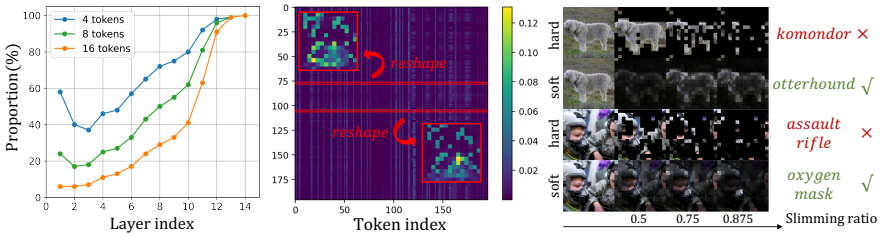
1 Introduction

Since vision transformer (ViT) [10] started the era of transformer structure in the fundamental computer vision tasks [3,36,5], variant transformers have been

⋆ Z. Zong and K. Li contribute equally during their internship at SenseTime. Emails: zongzhuofan@gmail.com and liuyuisanai@gmail.com

Table 1: **Comparison to recent model pruning methods for ViT.** Our SiT surpasses all the other methods based on structure pruning or hard dropping.

Type	Method	Reference	ImageNet Top-1(%)	Throughput (image/s)	(%)
Baseline	DeiT[29]	ICML21	79.8	1637	0
Structure Pruning	S^2ViTE[6]	NeurIPS21	79.2(**−0.6**)	2117	29.3
Token Hard Dropping	PS-ViT[28]	CVPR22	79.4(**−0.5**)	2351	43.6
	IA-RED^2[22]	NeurIPS21	79.1(**−0.7**)	2369	44.7
	Dynamic-ViT[24]	NeurIPS21	79.3(**−0.5**)	2575	57.3
	Evo-ViT[37]	AAAI22	79.4(**−0.4**)	2629	60.6
	EViT[20]	ICLR22	79.1(**−0.7**)	2783	70.0
Token Soft Slimming	Our SiT	ECCV22	79.8(**−0.0**)	2344	43.2
		ECCV22	79.4(**−0.4**)	3308	**102.1**



(a) Token similarity becomes higher in deeper layers.

(b) All tokens tend to focus on the same tokens in deeper layers.

(c) Our *soft slimming* can automatically zoom the attention scope based on the object size.

Fig. 1: **Our motivation.** In Fig(a), we calculate the correlation coefficients among tokens and count the proportion that is at least similar ($\geq$0.7) to 4/8/16 tokens in different layers. As for Fig(b), we randomly select two tokens in the tenth layer to show their attention. Moreover, we compare different token pruning methods in Fig(c). Darker tokens get less attention.

designed to challenge the dominance of convolutional neural networks (CNNs). Different from CNNs that stack convolutions to encode local features progressively, ViTs directly capture the long-term token dependencies. However, because of the exhausting token-to-token comparison, current powerful transformers require huge computation, limiting their wide application in reality [12]. Hence, in this paper, we aim to design a generic learning framework for boosting the efficiency of vanilla vision transformers.

To make ViTs more efficient, we tried to explore the inherent properties of the token-to-token comparison. We conduct a series of experiments based on LV-ViT, which reveals that sparse attention with high token similarity exists in ViTs. Fig. 1a shows that token similarity becomes higher in deeper layers, especially in the last three layers. Besides, the attention tends to focus on the specific tokens in the deeper layers (Fig. 1b), which indicates the number of decision-relevant tokens becomes fewer. These observations demonstrate that only a few token candidates indicate meaningful information. It inspires us a feasible structure-

agnostic dimension, token number, to reduce the computational cost. Intuitively, we can progressively drop the redundant tokens as the network deepens.

Recent studies have tried to compress tokens via data-independent dropping with minimizing reconstruction error [28], or data-dependent dropping with differentiable scoring [24]. However, data-independent dropping requires layer-by-layer optimization, which is hard to generalize. Moreover, the token hard dropping will inevitably discard the vital tokens as the dropping ratio increases, *e.g.*, the shape of the otterhound is destroyed in the deep layer (Fig. 1c), thus limiting its performance as shown in Table 1.

Can we design a flexible method of token slimming, thus decision-relevant information can be dynamically aggregated into a slimmed token set? To answer this question, we propose token soft slimming. It contains a concise Token Slimming Module (TSM), which generates decision-relevant tokens via a data-dependent weight matrix. As shown in Fig. 1c, by simply inserting multiple TSMs in LV-ViT, our network can learn to localize the key object tokens. More importantly, the attention scope can be zoomed automatically without cutting off the discriminative token relations, *e.g.*, our network can adaptively concentrate on the most informative parts of the otterhound in *softer* manner, while the oxygen mask in *harder* manner.

In DynamicViT [24], self-distillation is introduced in the last layer to minimize the performance drop brought by token sparsification. However, it ignores hint knowledge in the intermediate layers, leading to inevitable knowledge loss. To solve this issue, we introduce a concise Feature Recalibration Distillation (FRD) to achieve stable and efficient model slimming optimization. Note that previous hint knowledge distillation methods [25,42,39,17] are designed for spatial structured tokens. Since the neighbor token information is contiguous, they can apply contiguous upsampling (*e.g.*, deconvolution and interpolation) to find the correspondence between tokens of teacher and student. In contrast, our TSM generates *unstructured* token set, which can not be allocated corresponding supervision directly. To align the token relations among unstructured tokens, we design a reverse version of the token slimming module (RTSM) in a flexible auto-encoder manner. Such an effective way helps our FRD densely transfer all the token information block to block. Benefiting from the innate knowledge inheritance (structure knowledge), our FRD is more suitable for teaching itself. As shown in Table 1, our SiT is able to improve the throughput by 43.2% without any performance drop, and accelerate the inference speed by over 100% with negligible accuracy decrease.

Our self-slimmed learning method is flexible and easy to generalize to all vanilla vision transformers (SiT), *e.g.*, DeiT [29], LV-ViT [16] etc. We conduct extensive experiments on ImageNet [8] to verify the effectiveness and efficiency. Interestingly, our method without self-distillation can perform even better than DynamicViT [24] with distillation. Besides, the SiT-XS achieves 81.8% top-1 accuracy with $\mathbf{3.6}\times$ inference speed and SiT-L achieves competitive 85.6% top-1 accuracy while running $\mathbf{1.7}\times$ faster. More importantly, our SiT based on LV-ViT

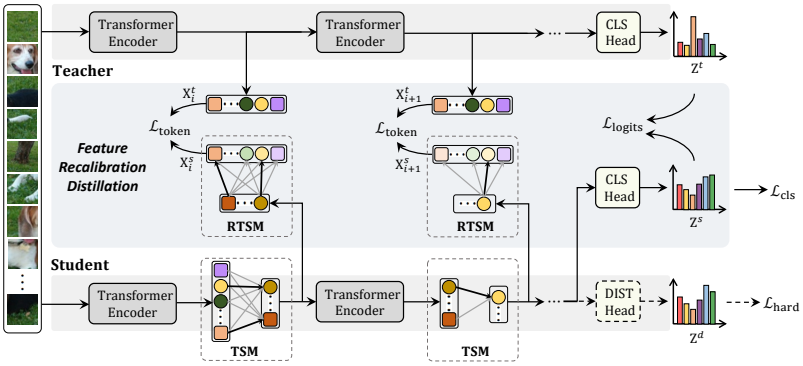


Fig. 2: **The framework of our self-slimmed learning.** We insert our token slimming modules (TSM) into vanilla vision transformer. To reduce decision-relevant information loss, we apply Feature Recalibration Distillation (FRD), wherein the reverse version of TSM (RTSM) is utilized to recalibrate unstructured token. The dash lines indicate the prediction supervision from an extra CNN teacher is optional and complementary to our method.

achieves the new state-of-the-art performance on ImageNet, surpassing recent CNNs and ViTs.

2 Related Works

2.1 Vision Transformers

Transformer architecture [31] was first proposed for machine translation. The success of transformer in NLP inspires the application of transformers in various vision tasks, for example, DETR [3] for object detection and ViT [10] for image recognition. ViT is the first pure transformer that achieves the state-of-the-art performance on ImageNet [8]. Recent ViT variants mainly focus on better optimization and more powerful performance [43,41,40,34,13,4,9,35,11,7,38,15,19,18]. However, few of them explore to improve the efficiency of vision transformers [12]. In this paper, we aim to design a general optimization framework named self-slimmed learning to promote the efficiency of ViTs.

2.2 Transformer Slimming

The large computation of self-attention hinders the wide application of ViTs, such as detection and segmentation with the high-resolution input image. To solve this problem, several prior works concentrate on designing sparse attention [33,21] or structure pruning [6]. S^2ViTE [6] dynamically extracts and trains sparse subnetworks of ViTs, while sticking to a fixed small parameter budget. However, model structure pruning struggles to trim down the inference latency.

Other works try to reduce the token redundancy [24,28,22,37] by entirely dropping the unimportant tokens, which brings more improvements on throughput compared to structure pruning. Different from the above works, our SiT aggregates all tokens into fewer informative tokens in a soft manner by a concise slimming module. It can automatically zoom the attention scope to localize the key object for better recognition.

3 Method

In this section, we describe our self-slimmed learning for vision transformer (SiT) in detail. First, we introduce the overall architecture of SiT. Then, we explain the vital design of our SiT, *i.e.*, token slimming module (TSM) and feature recalibration distillation (FRD). Finally, we thoroughly compare our TSM and FRD with other counterparts.

3.1 Overview of Self-slimmed Learning

The overall framework is illustrated in Fig. 2. We first design a lightweight Token Slimming Module (TSM) for conventional ViTs to perform token slimming, and its reverse version (RTSM) to recalibrate unstructured tokens for hint knowledge distillation. We divide the slimmed vision transformer into multiple stages (*e.g.*, 4 stages as in prior works [12,21]), where different numbers of tokens participate in feed-forward computation. To decrease the information loss, we propose a block-to-block feature recalibration distillation (FRD), wherein the original vision transformer can serve as a teacher to minimize the difference between itself and the slimmed student. Finally, we integrate TSM and FRD to form a general self-slimmed learning method for all vanilla ViTs.

3.2 Token Slimming Module

Given a sequence of N input tokens with C channels $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_N] \in \mathbb{R}^{N \times C}$, (class token is omitted as it will never be pruned), token slimming aims to dynamically aggregate the redundant tokens to generate $\hat{N}$ informative tokens $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1; \hat{\mathbf{x}}_2; \cdots; \hat{\mathbf{x}}_{\hat{N}}]$:

$$\hat{\mathbf{X}} = \hat{\mathbf{A}}\mathbf{X}, \tag{1}$$

where $\hat{\mathbf{A}} \in \mathbb{R}^{\hat{N} \times N}$ is a normalized weight matrix:

$$\sum_{i=1}^{\hat{N}} \hat{\mathbf{A}}_{i,j} = 1, \quad \texttt{where} \quad j = 1 \dots N. \tag{2}$$

Such operation is differentiable and friendly to end-to-end training. We follow the design paradigm of self-attention [32] and propose a lightweight token slimming module (TSM) shown in Fig. 3a:

$$\hat{\mathbf{A}} = \text{Softmax}(\frac{W_q \sigma(\mathbf{X} W_k)^T}{\tau}), \tag{3}$$

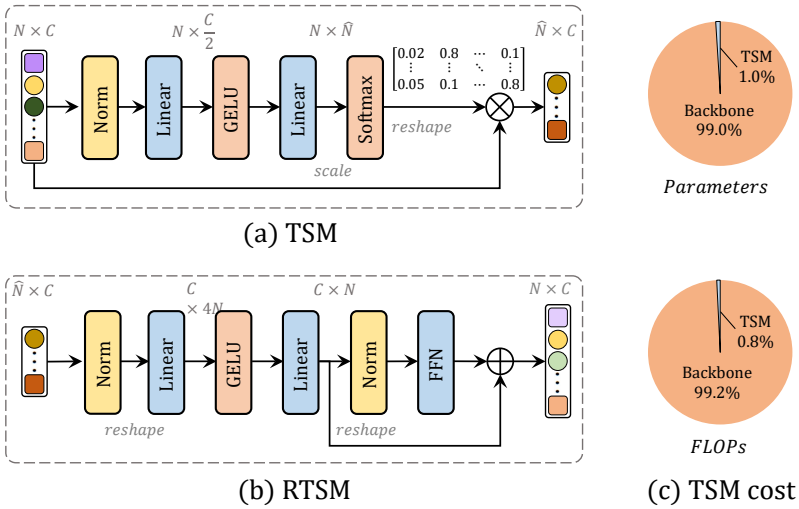


Fig. 3: The token slimming module (TSM) and its reverse version (RTSM).

where $W_k \in \mathbb{R}^{C\times\frac{C}{2}}$ and $W_q \in \mathbb{R}^{\hat{N}\times\frac{C}{2}}$ are both learnable parameters. σ and τ represents the nonlinear function (GELU) and scaling factor respectively. Similar to self-attention, TSM generates a global attention matrix, but it requires much fewer overhead in terms of throughput and memory usage during both training and inference. Fig. 3c shows that TSM blocks only require negligible cost. Thanks to the learnable scaling factor τ, the attention tends to be sparse in our experiments, which means it learns to focus on the most informative tokens.

Hard dropping vs. soft slimming. The prior works have tried to compress tokens via hard dropping [28,24], in which the slimming weight $\hat{\mathbf{A}}_{i,j} \in \{0, 1\}$ is a binary decision matrix, *i.e.*, dropping or keeping the corresponding token. However, this approach with binary decision leads to severe information loss if numerous tokens are discarded. Such weakness limits their high efficiency on ImageNet [8], wherein the objects often occupy a large part in the pictures. On the contrary, we design soft slimming with a learnable normalized weight $\hat{\mathbf{A}}_{i,j} \in (0, 1)$, which is able to discriminate the meaningful tokens in a global view. As shown in Fig. 1c, our soft slimming can dynamically zoom the attention scope to cover the significant regions for classification. It reveals that $\hat{\mathbf{A}}$ can adaptively become a one-hot matrix to help our SiT focus on the most informative part.

3.3 Feature Recalibration Distillation

Feature recalibration. Though token slimming significantly reduces the inference latency, when using a large slimming rate, it inevitably decreases the accuracy because of the information loss. Hint knowledge distillation is the common method to maintain meaningful information in intermediate layers, wherein

the challenge is to align the feature semantics between student and teacher. Previous works [25,39] adopt spatial deconvolution or interpolation to cope with this misalignment between spatial contiguous features. However, it is not suitable for slimmed unstructured tokens with spatially discrete semantics. To solve this problem, we design a reverse version of the token slimming module (RTSM) to recalibrate the unstructured tokens in a flexible auto-encoder manner (Fig. 3b). Therefore, all the token information can be seamlessly transferred from the teacher. Note that we only perform RTSM when training, thus no extra computation is introduced during inference. We first linearly transform the informative tokens into plenty of token candidates, thus utilizing a non-linear function (GELU) to filter the vital representations. Finally, another linear transformation is performed to compress the token candidates:

$$\hat{\mathbf{X}}' = \mathbf{A}_2(\sigma(\mathbf{A}_1\hat{\mathbf{X}})), \tag{4}$$

where $\mathbf{A}_1 \in \mathbb{R}^{4N\times\hat{N}}$ and $\mathbf{A}_2 \in \mathbb{R}^{N\times 4N}$ in our experiments. To further enhance the token representations, we introduce an extra multi-layer perception (MLP) block [32] with residual connection [14]:

$$\mathbf{X}' = \hat{\mathbf{X}}' + \mathrm{MLP}(\hat{\mathbf{X}}'). \tag{5}$$

The recalibrated features $\mathbf{X}'$ will be forced to be consistent with the teacher features in FRD, ensuring the sufficient information of the slimmed tokens $\hat{\mathbf{X}}$.

Knowledge distillation. Due to the invariant model structure, we design a block-to-block knowledge distillation for the recalibrated features:

$$\mathcal{L}_{\text{token}} = \frac{1}{LN}\sum_{i=1}^{L}\sum_{j=1}^{N}(\mathbf{X}^s_{i,j} - \mathbf{X}^t_{i,j})^2, \tag{6}$$

where $\mathbf{X}^s_{i,j}$ and $\mathbf{X}^t_{i,j}$ refer to the j-th token embedding at the i-th layer of the student and teacher, respectively. L means the layer number. Note that $\mathbf{X}^s$ refers to the recalibrated tokens $\mathbf{X}'$ in Eq. 5. With such reconstruction loss, the student model will be forced to maintain as much as knowledge in the informative tokens $\hat{\mathbf{X}}$. Besides, to further alleviate the classification performance deterioration caused by token slimming, we introduce the logits distillation to minimize the predictions difference between the student and teacher:

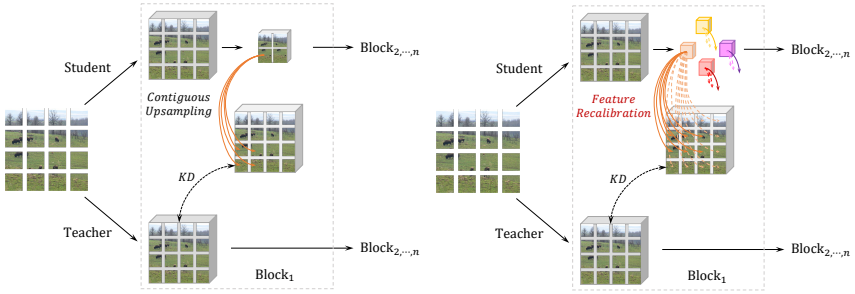
$$\mathcal{L}_{\text{logits}} = \mathrm{KL}(\psi(\mathrm{Z}^s), \psi(\mathrm{Z}^t)), \tag{7}$$

where KL denotes Kullback–Leibler divergence loss and ψ is the softmax function. Z^s and Z^t are respectively the predictions of the student and teacher model. Moreover, the above FRD is complementary to the hard distillation in [29]:

$$\mathcal{L}_{\text{hard}} = \mathrm{CrossEntropy}(\psi(\mathrm{Z}^d), y^c), \tag{8}$$

where Z^d indicates the prediction of distillation head and y^c is a hard decision of the extra CNN teacher. It can further improve the performance with longer training epochs. Our final objective of distillation for self-slimmed learning is:

$$\mathcal{L}_{\text{dist}} = \lambda_{\text{token}}\mathcal{L}_{\text{token}} + \lambda_{\text{logits}}\mathcal{L}_{\text{logits}} + \lambda_{\text{hard}}\mathcal{L}_{\text{hard}}, \tag{9}$$



(a) Structured Tokens. (b) Unstructured Tokens.

Fig. 4: Hint knowledge distillation for structured and unstructured tokens.

where λ is the coefficient balancing the three distillation losses. We set $\lambda_{\text{logits}} = 2, \lambda_{\text{token}} = 2$ by default. λ_{hard} is set to 1 when the CNN teacher is involved. As for the training objective of self-slimmed learning, we treat the classification task and the distillation task equally:

$$\mathcal{L}_{\text{cls}} = \text{CrossEntropy}(\psi(\text{Z}^s), \overline{y}), \tag{10}$$

$$\mathcal{L}_{\text{global}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dist}}, \tag{11}$$

where $\overline{y}$ means the ground truth, *i.e.*, one-hot label.

FRD vs. other knowledge distillation. Firstly, current vision transformers [29,30] simply select a strong teacher network with totally different architectures, such as RegNet for DeiT. Only a few knowledge (*e.g.*, final predictions) can be inherited, thus the semantic information in the intermediate layer is ignored. In FRD, thanks to the consistency between the teacher and student, we naturally conduct densely token-level supervision for each block, which greatly improves the stability of the model mimicking. Besides, the popular hint knowledge distillation method [25,39] are mainly designed for spatial structured tokens. As shown in Fig. 4a, they can simply apply local and contiguous upsampling to reconstruct tokens. However, as shown in Fig. 4b, the token slimming generates an unstructured token set. Each token contains partial information of previous tokens. To recalibrate the unstructured features, we design a concise RTSM in a flexible auto-encoder manner. Thus via reconstruction loss, our FRD can force the student model to maintain sufficient knowledge in the informative tokens.

4 Experiments

4.1 Implementation Details

In this section, we conduct comprehensive experiments to empirically analyze the effectiveness of our proposed self-slimmed learning for vision transformer (SiT). All the models are evaluated on the ImageNet dataset [8]. For our teacher models,

Table 2: **Main results on ImageNet.** We apply our self-slimming learning on the state-of-the-art vanilla vision transformer LV-ViT [16]. ‡ means we adopt an extra CNN teacher. Our SiT can speed up LV-ViT **1.7**× with a slight accuracy drop. For fast inference, our SiT can maintain 97% of the performance while speeding up the original transformers by **3.6**×.

Model	Depth	Stage	Embed Dim	Heads	Resolution	#Params (M)	FLOPs (G)
SiT-Ti	14	{1,1,1,11}	320	5	224^2	15.9	1.0
SiT-XS	16	{1,1,1,13}	384	6	224^2	25.6	1.5
SiT-S	16	{9,3,2,2}	384	6	224^2	25.6	4.0
SiT-M	20	{10,4,3,3}	512	8	224^2	55.6	8.1
SiT-L	24	{10,4,3,7}	768	12	288^2	148.2	34.4

(a) Model architecture settings

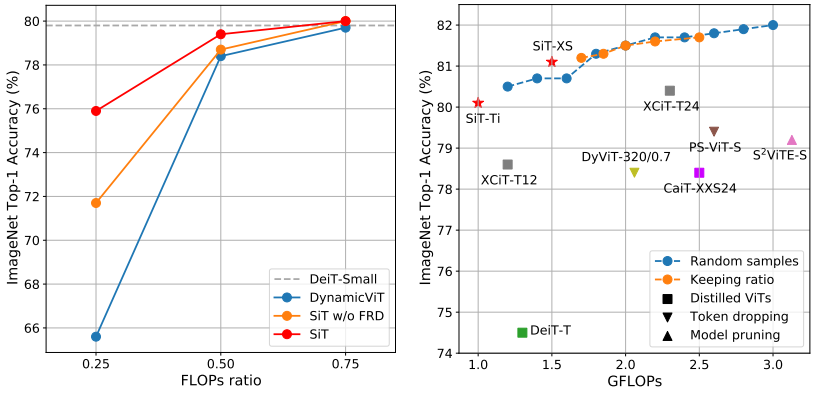
Model	Student			Teacher	
	Throughput (image/s)	Top-1 (%)	Top-1‡ (%)	Throughput (image/s)	Top-1 (%)
SiT-Ti	5896 (**3.2**×)	80.1 (−2.0)	80.6 (−1.5)	1827	82.1
SiT-XS	4839 (**3.6**×)	81.1 (−2.2)	81.8 (−1.5)	1360	83.3
SiT-S	1892 (**1.4**×)	83.2 (**−0.1**)	83.4 (+0.1)	1360	83.3
SiT-M	1197 (**1.5**×)	84.1 (**−0.1**)	84.3 (+0.1)	804	84.2
SiT-L	346 (**1.7**×)	85.6 (**−0.1**)	-	204	85.7

(b) Efficiency comparisons

we train LV-ViTs [16] following the original settings, but we replace the patch embedding module with lightweight stacked convolutions inspired by LeViT [12]. As for student models, all the training hyper-parameters are the same as DeiT [29] by defaults. For initialization, we load all the weights from the corresponding teacher models to accelerate the convergence and train them for 125 epochs. If utilizing an extra CNN teacher, we extend the training time to 300 epochs for better improvements. Moreover, we set different initial learning rates for the backbone and the feature recalibration branch, which are $0.0002 \times \frac{batch\ size}{1024}$ and $0.001 \times \frac{batch\ size}{1024}$ respectively. For token slimming, we insert TSM three times, thus there are four stages in SiT. The default keeping ratio $\hat{N}/N$ is set to 0.5, which means the token number is halved after slimming.

4.2 Main Results

We conduct our self-slimmed learning for LV-ViT [16], which is the state-of-the-art vanilla vision transformer. Table 2 shows our detailed settings for different SiT variants. For SiT-Ti and SiT-XS, we explore their capacity for fast inference, thus we insert TSMs in the early layers. It demonstrates that our self-slimmed method is able to speed up the original vision transformers by **3.6**×, while maintaining at least 97% of their accuracy. Besides, we adopt another CNN teacher to provide the hard label as in DeiT [29]. The results show that complementary prediction supervision can further improve performance. As for other variants,



(a) Our SiT achieves much higher accuracy than DynamicViT even without the distillation.

(b) Our randomly sampled models consistently outperform other distilled and pruned ViTs.

Fig. 5: **Effectiveness and robustness study.** We compare our SiT with the DynamicViT in Fig(a). To verify the robustness of our method, we randomly change the numbers of blocks in each stage (*i.e.*, TSM location) and adjust the keeping ratio of TSM from 0.3 to 0.7 to sample a series of SiT-Ti models in Fig(b). All the models are trained for 125 epochs without a CNN teacher.

we insert TSMs in the deeper layers. Surprisingly, with negligible accuracy drop, our SiTs are up to **1.7**× faster than their teacher models. It is worth mentioning that, extra CNN distillation brings little improvement, mainly because the CNN teacher is inferior to the original transformer teacher (82.9% *vs.* 83.3%/84.2%).

4.3 Effectiveness and Robustness

Comparison to DynamicViT. In Fig. 5a, we compare our method with DynamicViT [24] on DeiT-S [29]. When dropping too many tokens, the performance of DynamicViT deteriorates dramatically. Though it utilizes knowledge distillation to minimize the gap, our SiT without the distillation consistently surpasses it under different FLOPs ratios, especially under the smallest ratio. Besides, when armed with FRD, our SiT can maintain performance better.

TSM locations and keeping ratio. To verify the robustness of our method, we conduct experiments on SiT-Ti as shown in Fig. 5b. It clearly shows that all of the randomly sampled models outperform popular ViTs with knowledge distillation, *e.g.*, DeiT [29] and XCiT [1]. Besides, compared with other counterparts based on token hard dropping [24,28] and structure pruning [6], our models surpass them by a large margin. These results demonstrate our SiT is insensitive to the setting of TSM locations and keeping ratio. To make a fair comparison with the state-of-the-art ViTs, we set these hyper-parameters according to the FLOPs.

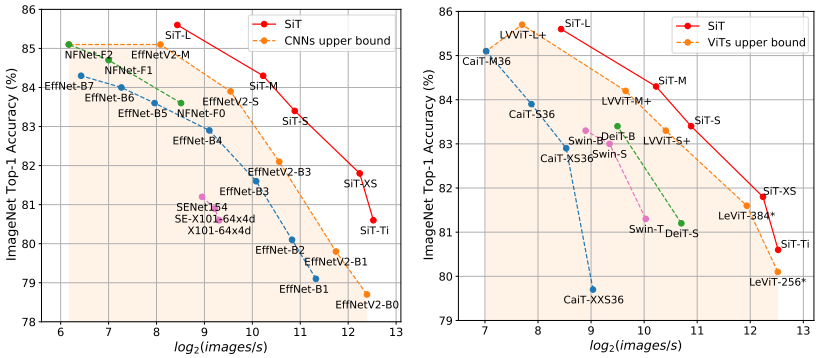
Table 3: **Comparison to the state-of-the-art on ImageNet.** The models marked in *gray* color are trained with distillation supervision from a powerful CNN for 300 epochs. Our SiT achieves the best performance trade-off.

Model	Resolution	#Params (M)	FLOPs (G)	Throughput (image/s)	ImageNet Top-1(%)
EfficientNet-B1 [26]	240^2	7.8	0.7	2559	79.1
EfficientNet-B2 [26]	260^2	9.1	1.1	1808	80.1
DeiT-T [29]	224^2	5.9	1.3	3346	74.5
LeViT-256 [12]	224^2	18.9	1.1	5802	80.1
SiT-Ti	224^2	15.9	1.0	**5896**	80.1
SiT-Ti	224^2	16.2	1.0	5833	**80.6**
EfficientNet-B3 [26]	300^2	12.2	1.9	1062	81.6
Swin-T [21]	224^2	28.3	4.5	1023	81.3
DeiT-S [29]	224^2	22.4	4.6	1598	81.2
LeViT-384 [12]	224^2	39.1	2.4	3876	81.6
SiT-XS	224^2	25.6	1.5	**4839**	81.1
SiT-XS	224^2	26.0	1.5	4798	**81.8**
EfficientNet-B4 [26]	380^2	19.3	4.6	545	82.9
Swin-B [21]	224^2	87.8	15.5	474	83.3
DeiT-B [29]	224^2	87.3	17.7	718	83.4
LV-ViT-S [16]	224^2	26.2	6.6	1270	83.3
SiT-S	224^2	25.6	4.0	**1892**	83.2
SiT-S	224^2	26.0	4.0	1873	**83.4**
EfficientNet-B6 [26]	528^2	43.0	19.9	153	84.0
EfficientNetV2-S [27]	384^2	21.5	8.5	742	83.9
CaiT-S36 [30]	224^2	68.2	13.9	233	83.9
LV-ViT-M [16]	224^2	55.8	12.7	768	84.1
SiT-M	224^2	55.6	8.1	**1197**	84.1
SiT-M	224^2	56.2	8.1	1185	**84.3**
EfficientNetV2-M [27]	480^2	54.1	25.0	271	85.1
NFNet-F1 [2]	320^2	132.6	36.0	128	84.7
CaiT-M36 [30]	224^2	270.1	53.4	130	85.1
LV-ViT-L [16]	288^2	150.1	58.8	208	85.3
SiT-L	288^2	148.2	34.4	**346**	**85.6**

4.4 Comparison to state-of-the-art

In Table 3, we compare SiT with other competitive CNNs and ViTs. For a fair comparison, we group these methods according to their top-1 accuracies. The throughput is measured on a single 16GB V100 GPU under the same setting as LeViT [12]. Our SiT-Ti is competitive with LeViT, while the throughput is **3.2**× than that of EfficientNet [26]. Note that EfficientNet is designed via extensive neural architecture search and LeViT is elaborately designed for fast inference. For our larger model variants, they perform better than EfficientNetV2 [27] with simple training strategies. Compared with the original LV-ViT [16], our SiT is **1.5**× faster than those with similar accuracy.

We further visualize the comparisons to the upper bounds of CNNs and ViTs in Fig. 6a and 6b. It clearly shows that our SiT achieves the best balance between throughput and accuracy, surpassing the recent state-of-the-art CNNs and ViTs.



(a) Comparison to CNNs. (b) Comparison to ViTs.

Fig. 6: **Speed *vs.* Accuracy.** We compare our SiT with the previous state-of-the-art CNNs and ViTs in Fig(a) and Fig(b), respectively. "LV-ViT+" denotes our improved LV-ViT teacher. Our SiT surpasses the SOTA methods by a large margin, even the efficient models EfficientNetV2 [27] and LeViT [12].

Table 4: Efficiency comparison.

Method	Top-1	Throughput
Structure-width	76.3	2947
Structure-depth	69.4	5652
DynamicViT[24]	75.7	5762
SiT w/o FRD	**77.7**	**5896**

Table 5: Inherited knowledge.

Knowledge	Self	CaiT	RegNet
	83.3	83.5	82.9
Scratch	**80.1**	79.9	79.2
Fine-tuning	**80.5**	80.2	80.0
Fine-tuning+Structure	**81.1**	80.6	80.2

4.5 Ablation Studies

If not otherwise specified, all experiments for ablations are conducted on SiT-Ti and run with only 125 training epochs under the supervision of the original teacher model. "Token-MLP" refers double linear layers along the token dimension.

Does token slimming outperform model scaling down? In Table 4, we compare token slimming with model scaling down rules under the same computation limit. For model scaling down, we adapt the channel and depth individually. Note that the above two models are trained from scratch for 300 epochs with token labeling [16]. For token slimming, we simply insert TSMs without FRD. We also drop tokens and train it with extra distillation as in DynamicViT [24]. It shows that scaling along the channel achieves higher accuracy than scaling along the depth but with lower throughput. Besides, token slimming can largely improve the throughput with higher performance. However, DynamicViT performs worse than our SiT without distillation, since token hard dropping loses much discriminative information with a large slimming ratio. Such results demonstrate simply inserting our TSM into vanilla ViT is able to achieve great performance.

Does structure knowledge matter to self-slimmed learning? We further investigate whether the structure knowledge benefits the performance as shown

Table 6: Robustness of slimming ratios.

Ratio	$\mathcal{L}_{\text{logits}}$+$\mathcal{L}_{\text{token}}$	$\mathcal{L}_{\text{hard}}$
1	82.1	82.1
0.75	82.0	82.0
0.5	81.6	81.3
0.25	80.1	78.4

Table 7: Slimming.

Method	GFLOPs	Top-1
Baseline	3.5	82.1
3×3 AvgPool	1.0	77.4
3×3 Conv	1.0	79.3
Token-Mixer	1.1	79.3
Our TSM	**1.0**	**80.1**

Table 8: Recalibration.

Method	Top-1
Baseline	79.0
Interpolation	78.3
Deconvolution	78.4
Token-MLP	79.0
Our RTSM	**80.1**

Table 9: Knowledge distillation.

Method	Top-1
Baseline	77.7
+$\mathcal{L}_{\text{logits}}$	79.0(**+1.3**)
+$\mathcal{L}_{\text{logits}}$+$\mathcal{L}_{\text{token}}$	80.1(**+2.4**)
+$\mathcal{L}_{\text{logits}}$+$\mathcal{L}_{\text{token}}$+$\mathcal{L}_{\text{hard}}$	80.2(**+2.5**)
+$\mathcal{L}_{\text{logits}}$+$\mathcal{L}_{\text{token}}$+$\mathcal{L}_{\text{hard}}$+Longer training	80.6(**+2.9**)

Table 10: Loss weights.

λ_{token}:λ_{logit}:λ_{hard}	Top-1
1:1:1	79.3
1:2:1	79.4
1:2:2	79.5
2:1:1	**79.6**
2:2:1	**79.6**

in Table 5. For the teacher models, we adopt different architectures (LV-ViT-S[16], CaiT-S24[30], and RegNetY-16GF[23]) but similar accuracies for a fair comparison. It shows that training with the pre-trained weights for 125 epochs converges to higher results than those trained from scratch for 300 epochs. Moreover, we utilize structure knowledge via block-to-block mimicking, which can further boost the performance. It also reveals that higher similarity between students and teachers can bring greater improvements.

Is self-slimmed learning robust to different FLOPs ratios? In Table 6, we empirically train models with different FLOPs ratios. When the ratio is large than 0.5, our FRD and CNN distillation are both helpful for maintaining performance. However, when the ratio is small, CNN distillation leads to a higher performance drop, while our FRD only drops the accuracy by 2.0%. These results demonstrate that our method is robust to different FLOPs ratios.

Dynamic *vs.* Static: Which aggregation manner works better for token slimming? To explore whether dynamic aggregation is better for token slimming, we perform ablation experiments as shown in Table 7. For static aggregation, we choose different data-independent operations and maintain similar computation: 3×3 average pooling/convolution with stride 2×2, and double linear layers with GELU function ("Token-MLP"). It shows that learnable parameters are vital for token slimming since average pooling leads to a severe accuracy drop. Besides, the static aggregation methods with data-independent weights yield similar but inferior performance to our TSM (79.3% *vs.* 80.1%). Such comparisons prove that our TSM can generate more informative tokens.

Can contiguous upsampling recalibrate the features? We first recalibrate the original tokens by contiguous upsampling methods, *e.g.*, bilinear interpolation and deconvolution. As shown in Table 8, these two spatial contiguous methods misalign the token relations and hurt the capacity compared with the baseline (without block-to-block mimicking). In contrast, "Token-MLP" does not hurt the token representation, and its accuracy can be further boosted to 80.1% by the insertion of an MLP.

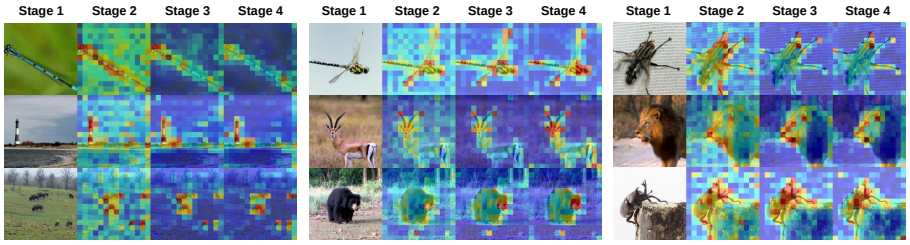


Fig. 7: **Visualizations of our progressive token slimming.** The blue/red tokens contribute less/more to the final informative tokens. Our method can zoom the attention scope to cover the key object, even with only 12.5% of tokens.

Does each distillation supervision help? Table 9 presents that the soft logits supervision $\mathcal{L}_{\text{logits}}$ brings 1.4% accuracy gain. When further introducing block-to-block knowledge supervision, our model improves the accuracy by 1.1%. Finally, combining complementary hard label supervision, the accuracy reaches 80.6% with longer training epochs.
What are the appropriate loss weights? Table 10 shows the settings of loss wights are robust in our SiT (trained for 100 epochs). In fact, we simply choose the weight of 2:2:1 to ensure different loss values are close in the early training.

4.6 Visualization

Qualitative token slimming visualization. Fig. 7 shows the original images and the token slimming procedure of our SiT-Ti. We observe that the tokens of higher scores, *i.e.*, brighter tokens, are concentrated and tend to cover the key objects in the image. It demonstrates our proposed TSM is able to localize the significant regions and predict accurate scores for the most informative tokens.

5 Conclusions

In this paper, we propose a generic self-slimmed learning method for vanilla vision transformers (SiT), which can speed up the ViTs with negligible accuracy drop. Our concise TSM softly integrates redundant tokens into fewer informative ones. For stable and efficient training, we introduce a novel FRD framework to leverage structure knowledge, which can densely transfer token information in a flexible auto-encoder manner. Extensive experiments demonstrate the effectiveness of our SiT. By simply arming LV-ViT with our SiT, we achieve new state-of-the-art performance on ImageNet, surpassing recent CNNs and ViTs.
Acknowledgements. This work is partially supported by National Key R&D Program of China under Grant 2019YFB2102400, National Natural Science Foundation of China (61876176), the Joint Lab of CAS-HK, Shenzhen Institute of Artificial Intelligence and Robotics for Society, the Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

References

1. Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al.: Xcit: Cross-covariance image transformers. Advances in neural information processing systems (2021)
2. Brock, A., De, S., Smith, S.L., Simonyan, K.: High-performance large-scale image recognition without normalization. In: International Conference on Machine Learning (2021)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision (2020)
4. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
5. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
6. Chen, T., Cheng, Y., Gan, Z., Yuan, L., Zhang, L., Wang, Z.: Chasing sparsity in vision transformers: An end-to-end exploration. Advances in Neural Information Processing Systems (2021)
7. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. Advances in Neural Information Processing Systems (2021)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2009)
9. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. ArXiv **abs/2107.00652** (2021)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
11. d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. In: International Conference on Machine Learning (2021)
12. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: a vision transformer in convnet's clothing for faster inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
13. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. Advances in Neural Information Processing Systems (2021)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (2016)
15. Heo, B., Yun, S., Han, D., Chun, S., Choe, J., Oh, S.J.: Rethinking spatial dimensions of vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
16. Jiang, Z.H., Hou, Q., Yuan, L., Zhou, D., Shi, Y., Jin, X., Wang, A., Feng, J.: All tokens matter: Token labeling for training better vision transformers. Advances in Neural Information Processing Systems (2021)

17. Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: Network compression via factor transfer. Advances in neural information processing systems (2018)
18. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.J.: Uniformer: Unifying convolution and self-attention for visual recognition. ArXiv **abs/2201.09450** (2022)
19. Li, Y., Zhang, K., Cao, J., Timofte, R., Gool, L.V.: Localvit: Bringing locality to vision transformers. ArXiv **abs/2104.05707** (2021)
20. Liang, Y., GE, C., Tong, Z., Song, Y., Wang, J., Xie, P.: EVit: Expediting vision transformers via token reorganizations. In: International Conference on Learning Representations (2022)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
22. Pan, B., Panda, R., Jiang, Y., Wang, Z., Feris, R., Oliva, A.: Ia-red2: Interpretability-aware redundancy reduction for vision transformers. Advances in Neural Information Processing Systems (2021)
23. Radosavovic, I., Kosaraju, R.P., Girshick, R.B., He, K., Dollár, P.: Designing network design spaces. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
24. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems (2021)
25. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. CoRR **abs/1412.6550** (2015)
26. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning (2019)
27. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning (2021)
28. Tang, Y., Han, K., Wang, Y., Xu, C., Guo, J., Xu, C., Tao, D.: Patch slimming for efficient vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12165–12174 (2022)
29. 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 (2021)
30. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems (2017)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems (2017)
33. Wang, P., Wang, X., Wang, F., Lin, M., Chang, S., Xie, W., Li, H., Jin, R.: Kvt: k-nn attention for boosting vision transformers. ArXiv **abs/2106.00515** (2021)
34. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

35. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
36. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems (2021)
37. Xu, Y., Zhang, Z., Zhang, M., Sheng, K., Li, K., Dong, W., Zhang, L., Xu, C., Sun, X.: Evo-vit: Slow-fast token evolution for dynamic vision transformer. Proceedings of the AAAI Conference on Artificial Intelligence (2022)
38. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal self-attention for local-global interactions in vision transformers. ArXiv **abs/2107.00641** (2021)
39. Yim, J., Joo, D., Bae, J.H., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
40. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
41. Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S.: Volo: Vision outlooker for visual recognition. ArXiv **abs/2106.13112** (2021)
42. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. ArXiv **abs/1612.03928** (2017)
43. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Hou, Q., Feng, J.: Deepvit: Towards deeper vision transformer. ArXiv **abs/2103.11886** (2021)