Spatial-Channel Token Distillation for Vision MLPs

Yanxi Li [1 2] **Xinghao Chen** [2] **Minjing Dong** [1 2] **Yehui Tang** [2 3] **Yunhe Wang** [2] **Chang Xu** [1]

Abstract

Recently, neural architectures with all Multi-layer Perceptrons (MLPs) have attracted great research interest from the computer vision community. However, the inefficient mixing of spatial-channel information causes MLP-like Vision Models to demand tremendous pre-training on large-scale datasets. This work solves the problem from a novel knowledge distillation perspective. We propose a novel **Spatial-channel Token Distillation** (**STD**) method, which improves the information mixing in the two dimensions by introducing distillation tokens to each of them. A mutual information regularization is further introduced to let distillation tokens focus on their specific dimensions and maximize the performance gain. Extensive experiments on ImageNet for several MLP-like architectures demonstrate that the proposed token distillation mechanism can efficiently improve the accuracy. For example, the proposed STD boosts the top-1 accuracy of Mixer-S16 on ImageNet from 73.8% to 75.7% without any costly pre-training on JFT-300M. When applied to stronger architectures, e.g. CycleMLP-B1 and CycleMLP-B2, STD can still harvest about 1.1% and 0.5% accuracy gains, respectively.

1. Introduction

In the past few decades, convolutional neural networks (CNNs; LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016) have in fact dominated computer vision (CV) tasks. However, the recent advances of Transformers (Vaswani et al., 2017; Devlin et al., 2018; Brown et al., 2020) in natural language processing (NLP) also impact CV researchers to design pixel-level Transformers for vision tasks (Parmar et al., 2018; Hu et al., 2019; Ramachandran et al., 2019; Zhao et al., 2020) and finally lead to patch-based Vision Transformer (ViT; Dosovitskiy et al., 2021). ViT splits an image into small patches and feeds them into several stacked transformer blocks to obtain the final image classification results. ViT and its variants have shown impressive performance on ImageNet (Russakovsky et al., 2015) classification as well as downstream tasks like object detection and segmentation.

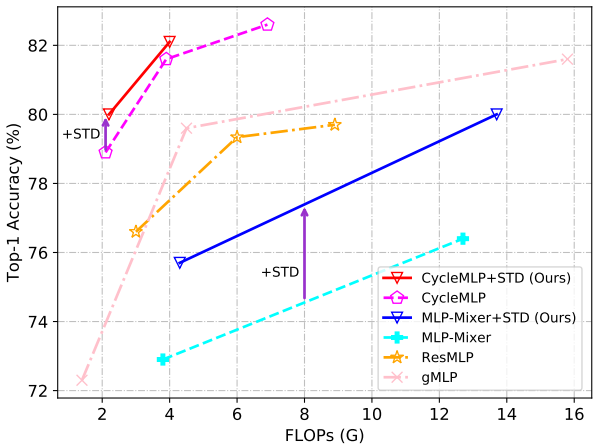

Figure 1. Top-1 accuracy and FLOPs on ImageNet-1K of MLP-like Vision Models distilled by STD (solid lines) compared to state-of-the-art results (dotted lines).

Following this line, MLP-Mixer (Tolstikhin et al., 2021) reconsiders the possibility of using pure multi-layer perceptrons (MLPs) for vision tasks and builds the pure MLP architecture by applying MLPs to both the spatial and channel dimensions of image patches, which are known as token mixing and channel mixing, respectively. Despite the simplicity of its architecture, MLP-Mixer is hard to obtain superior performance without costly pre-training on large-scale datasets, such as ImageNet-21K and JFT-300M (Hinton et al., 2015). For example, Mixer-S16 only obtains a 72.9% top-1 accuracy when trained on ImageNet-1k, which is still slightly lower than that of current state-of-the-art architectures. Therefore, there are many attempts to design novel mixing methods (Hou et al., 2021; Touvron et al., 2021a; Chen et al., 2021; Guo et al., 2022). However, they could

[1]School of Computer Science, University of Sydney, Australia [2]Huawei Noah's Ark Lab [3]School of Artificial Intelligence, Peking University. Correspondence to: Yunhe Wang <yunhe.wang@huawei.com>, Chang Xu <c.xu@sydney.edu.au>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

introduce additional complexity to the models. For example, Hou et al. (2021) permutates 3D matrices along different axes and then flattens them into 2D; and Chen et al. (2021) uses a stair-like operation implemented by deformable convolution, which is much more complex than pure MLPs. Moreover, the mixing of spatial and channel information has not been fully exploited, preventing current MLP-like Vision Models from achieving higher performance.

In this work, we propose to improve the spatial and channel mixing from a novel knowledge distillation (KD) perspective. To be specific, we design a special KD mechanism for MLP-like Vision Models called **Spatial-channel Token Distillation** (**STD**), which improves the information mixing in both the spatial and channel dimensions of MLP blocks. Instead of modifying the mixing operations themselves, STD adds spatial and channel tokens to image patches. After forward propagation, the tokens are concatenated for distillation with the teachers' responses as targets. Each token works as an aggregator of its dimension. The objective of them is to encourage each mixing operation to extract maximal task-related information from their specific dimension. Besides, this manner also allows STD to be very flexible and to be applied to different KD settings. It supports not only single-teacher distillation and distillation of the last layer but also multi-teacher distillation and distillation of intermediate layers. A remaining obstacle is that the spatial information and channel information are highly entangled with each other in the image patches, which is contrary to the goal of our distillation. To force the tokens to focus on their specific dimensions, we further introduce a *mutual information* (MI) regularization.

We perform extensive experiments for various MLP-like architectures to demonstrate that STD can efficiently distill MLP-Mixers and reach better performances than the costly pre-training. For example, Mixer-S16 with STD obtains 75.7% top-1 accuracy on ImageNet-1K compared to 72.9% by training from scratch and 73.8% by pre-training on JFT-300M. Besides, Mixer-B16 with STD can reach 80.0% top-1 accuracy, which is very close to the performance of ResMLP-B24 and CycleMLP-B2. When applied to stronger architectures, e.g. CycleMLP-B1 and CycleMLP-B2, STD can still harvest about 1.1% and 0.5% accuracy gains, respectively. Figure 1 further compares top-1 accuracy and FLOPs on ImageNet-1K of MLP-like Vision Models distilled by STD with state-of-the-art results. Our method reaches superior results with marginal FLOPs raising.

2. Related Work

2.1. Transformer-based Vision Models.

Transformers are a family of models utilizing multi-head self-attention (MSA; Vaswani et al., 2017), which are initially designed for NLP tasks. Transformer (Vaswani et al., 2017) first introduce MSA. BERT (Devlin et al., 2018) introduces a classification token, and pre-training. The GPT series of works (Radford et al., 2018; 2019; Brown et al., 2020) highly focus on pre-training tasks.

Inspired by those language models, there are several early successes to explore Transformer-based architectures for CV tasks. Image Transformer (Parmar et al., 2018) applies Transformer for a sequence modeling formulation of image generation tasks. Hu et al. (2019) and Zhao et al. (2020) use local multi-head dot-product self-attention blocks for image classification. Ramachandran et al. (2019) further expands self-attention for both classification and object detection.

Recently, ViT (Dosovitskiy et al., 2021) demonstrates that applying a pure Transformer to sequences of image patches can reach state-of-the-art performance on ImageNet without any convolution. However, ViT demands costly pre-training on large-scale datasets, such as ImageNet-21K and JFT-300M. To avoid the pre-training, DeiT (Touvron et al., 2021b) replaces it with distillation. A distillation token inspired by the classification token in BERT (Devlin et al., 2018) is introduced to improve its performance.

2.2. MLP-like Vision Models.

The success of using image patches as inputs for Transformers has encouraged researchers to rethink and revive MLPs as vision models. MLP-Mixer (Tolstikhin et al., 2021) applies pure MLPs to both the spatial and channel dimensions of image patches, which are referred as token mixing and channel mixing, respectively. However, those mixing operations are inefficient. Just like its ancestor, ViT (Dosovitskiy et al., 2021), MLP-Mixer also demands costly pre-training on those large-scale datasets.

There are several attempts to improve the efficiency of MLP-like vision models from an architecture view. ViP (Hou et al., 2021) uses various permutation of patches in parallel to improve the mixing. ResMLP (Touvron et al., 2021a) designs a residual MLP layer. CycleMLP (Chen et al., 2021) introduces a Cycle Fully-Connected layer to enlarge the receptive field. Hire-MLP (Guo et al., 2022) proposes a hierarchical rearrangement mechanism to aggregate the local and global spatial information. Wave-MLP (Tang et al., 2022) propose to represent each token as a wave function consisting of an amplitude part and a phase part for dynamical aggregation.

2.3. Knowledge Distillation.

KD is initially inspired by model compression (Bucilua et al., 2006) and introduced by Hinton et al. (2015) to transfer knowledge from a large ensemble of models into a single small model. The main idea is to let the small student network mimic the behavior of its large teachers. There are

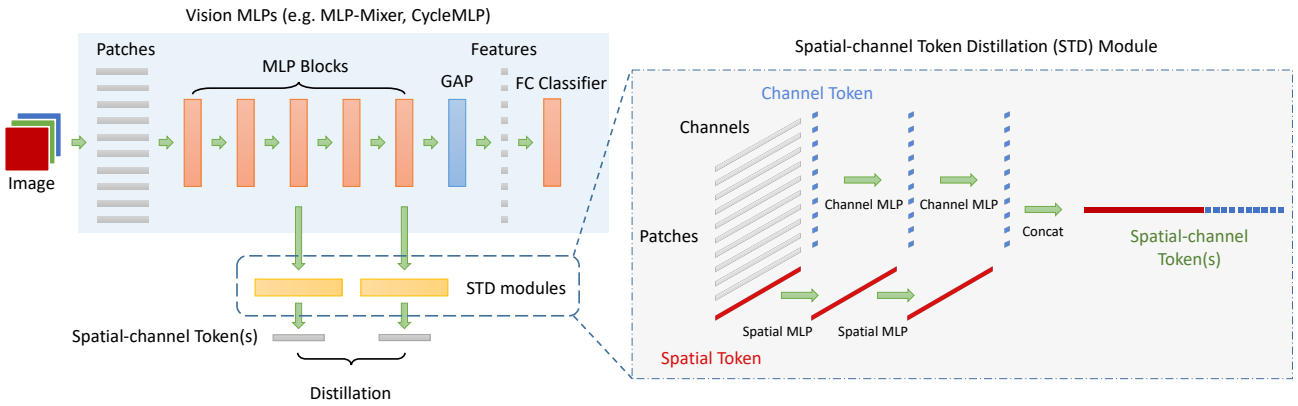

Figure 2. The overall pipeline of STD.

several explanations of why KD works. Yuan et al. (2019) explains it from a label smoothing perspective. Wei et al. (2020) argues KD is equivalent to a data augmentation.

As for multi-teacher distillation, Hinton et al. (2015) simply uses an averaged response from all teachers. In this way, every teacher has the same importance. You et al. (2017) not only uses the responses, but also consider features from the intermediate layers. Chen et al. (2019) uses different teachers for different purposes. They use one teacher for response-based distillation and one teacher for feature-based distillation. Guo et al. (2019) designs distillation objectives regarding prediction scores and gradients of examples to enhance the robustness of student networks. Chen et al. (2020) introduces a locality preserving loss to encourage student networks to generate low-dimensional features inheriting intrinsic properties from corresponding high-dimensional teacher's features. Park & Kwak (2020) and Asif et al. (2019) add additional teacher branches to the student network to mimic the intermediate features of teachers.

3. Methodology

The proposed STD includes two major components, the spatial-channel tokens and a mutual information regularization on those tokens. We first introduce how the spatial-channel tokens are combined with MLP-like vision models and how they participate in the distillation. Then, we describe a method estimating the mutual information between the spatial and channel tokens, which is to regularize them to focus on their specific dimensions. Finally, the overall pipeline of STD is described, including multi-teacher token distillation, the token distillation of both the intermediate layer and last layer, and the overall distillation objective. The overall pipeline of STD is illustrated in Figure 2.

3.1. Spatial-channel Token Distillation

MLP-like vision models typically split an image into small patches and mix features along two dimensions: 1) the **spatial MLP** layers mix feature across different spatial locations and share weights among channels, and 2) the **channel MLP** layers mix features across channels at a given spatial location and share weights among locations. Given the feature $\boldsymbol{Z}^{(l-1)} \in \mathbb{R}^{P\times N}$ of P patches with N channels, a MLP-like block can be represented by

$$\boldsymbol{U}^{(l)} = \mathrm{MLP}_S^{(l)}(\mathrm{LN}(\boldsymbol{Z}^{(l-1)})) + \boldsymbol{Z}^{(l-1)}, \quad (1)$$

$$\boldsymbol{Z}^{(l)} = \mathrm{MLP}_C^{(l)}(\mathrm{LN}(\boldsymbol{U}^{(l)})) + \boldsymbol{U}^{(l)}, \quad (2)$$

where $l = 1, \ldots, L$ are L blocks, $\mathrm{LN}(\cdot)$ is the layer normalization, and $\mathrm{MLP}_S^{(l)}$ and $\mathrm{MLP}_C^{(l)}$ are the spatial and channel MLP layers in block l, respectively. Note, $\mathrm{MLP}_S^{(l)}$ and $\mathrm{MLP}_C^{(l)}$ are flexible and are not limited to be the token-mixing and channel-mixing MLPs in MLP-Mixer, but can also be other complex MLP layers.

The spatial-channel mixing is a common paradigm widely existing in various MLP-like vision models (Tolstikhin et al., 2021; Hou et al., 2021; Touvron et al., 2021a; Chen et al., 2021; Guo et al., 2022). Based on this fundamental characteristic, we design the spatial-channel tokens for distillation of MLP-like vision models. In a previous work of Transformer-based vision models, DeiT (Touvron et al., 2021b) introduces a distillation token by adding an extra patch after all the image patches. It separates the distillation objective from the classification objective and improves the network performance, but it has limitation to be applied to MLP-like vision models.

MLP-like vision models uses spatial and channel MLP layers instead of self-attention. If we add a new patch $T_S \in \mathbb{R}^{1\times N}$ as a spatial token, it interacts with other patches in the spatial MLP layers:

$$\mathrm{MLP}_S^{(l)}(\boldsymbol{x}')_{*,j} = \boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x}'_{*,j}), \quad (3)$$

where σ is an non-linear activation function, $\boldsymbol{x}' \in \mathbb{R}^{P+1\times N}$, $j = 1, \ldots, N$, and $\boldsymbol{x}_{P+1,*} = T_S$. Equation (3) only allows T_S to capture cross-location features per channel, but cross-channel features per location are ignored. To capture the

cross-channel features, we propose a new channel token $T_C \in \mathbb{R}^{P\times 1}$. It interacts with other channels in the channel MLP layers:

$$\mathrm{MLP}_C^{(l)}(\boldsymbol{x}'')_{i,*} = \boldsymbol{W}_4\sigma(\boldsymbol{W}_3\boldsymbol{x}''_{i,*}), \tag{4}$$

where $\boldsymbol{x}'' \in \mathbb{R}^{P\times N+1}$, $i = 1,\dots,P$, and $\boldsymbol{x}''_{*,N+1} = T_C$. Using both Equations (3) and (4) allows us to capture per-channel and per-location features at the same time.

As aforementioned, the classification and distillation objective are independent. Therefore, we design the spatial-channel tokens as information aggregator and do not want them copy information back to the features. This target can be achieved by slightly modifying weights in Equations (3) and (4). Let $\boldsymbol{W}_1 \in \mathbb{R}^{d_S\times P+1}$ $\boldsymbol{W}_2 \in \mathbb{R}^{1\times d_S}$ $\boldsymbol{W}_3 \in \mathbb{R}^{d_C\times N+1}$ $\boldsymbol{W}_4 \in \mathbb{R}^{1\times d_C}$, where d_S and d_C is the hidden dimension of MLP layers. We can add K MLP blocks for distillation in parallel with Equations (1) and (2):

$$T_S^{(k)} = \mathrm{MLP}_S^{(k)}(\mathrm{LN}([\boldsymbol{Z}^{(l)}||T_S^{(k-1)}])) + T_S^{(k-1)} \tag{5}$$

$$T_C^{(k)} = \mathrm{MLP}_C^{(k)}(\mathrm{LN}([\boldsymbol{Z}^{(l)}||T_C^{(k-1)}])) + T_C^{(k-1)} \tag{6}$$

where $\boldsymbol{Z}^{(l)}$ is the output of a MLP layer l from the original network, and $[\cdot||\cdot]$ represents concatenation. Finally, the spatial and channel tokens are concatenated as the final output $[T_S^{(K)}||T_C^{(K)}]$ for distillation. The right part of Figure 2 demonstrates how the spatial-channel token works.

There are several differences between our token distillation for MLPs and the existing token distillation methods for Transformers, e.g. DeiT (Touvron et al., 2021b). First of all, Transformers use self-attention and do not consider spatial and channel differently. Therefore, DeiT uses only spatial tokens for distillation. However, MLP-like vision models utilize independent operations for spatial and channel mixing, so we add tokens to both dimensions. Secondly, the objective of using spatial-channel tokens is to improve the two kinds of mixing operations. To reach this goal, we further design a mutual information regularization to disentangle the spatial and channel information, which encourages the spatial and channel tokens to extract more informative features. Finally, DeiT updates the image patches, classification token, and distillation token together. We update the patches and distillation tokens separately. The distillation tokens aggregate information from patches, but do not copy the information back to them. Patches are passed to tokens by residual connections and won't be updated. This characteristic allows us to insert tokens flexibly and makes multi-teacher distillation and distillation of intermediate layers possible.

3.2. Mutual Information Regularization

Although the spatial and channel tokens are separate, they can share joint information from the features. To make them focus on their own dimension, we design a MI regularization term to disentangle the spatial and channel information. The MI is a measure of dependence between random variables based on the Shannon entropy. It is equivalent to the Kullback-Leibler (KL-) divergence between the joint distribution and the product of the marginal distribution of the random variables. Given two random variable X and Y, the MI can be calculated by

$$I(X;Y) := D_{KL}(\mathbb{P}_{XY}||\mathbb{P}_X \otimes \mathbb{P}_Y), \tag{7}$$

where $D_{KL}(\mathbb{P}||\mathbb{Q}) := \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}}\right]$ is the KL-divergence. The direct calculation of Equation (7) is costly. To efficiently measure the MI, we use a estimation called mutual information neural estimation (MINE) (Belghazi et al., 2018).

Algorithm 1 describes how MINE is used to regularize our spatial-channel token. MINE uses a statistics network $\psi_\theta : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ parameterized by $\theta\in\Theta$ to estimate a *neural information measure* as

$$I_\Theta(X;Y) = \sup_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_{XY}}[\psi_\theta] - \log\left(\mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_Y}\left[e^{\psi_\theta}\right]\right). \tag{8}$$

Equation (8) is a supremum of expectation, and it can be empirically calculated by maximizing

$$\frac{1}{b}\sum_{i=1}^{b}\psi_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)}) - \log(\frac{1}{b}\sum_{i=1}^{b}e^{\psi_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)},\bar{\boldsymbol{y}}^{(i)})}) \tag{9}$$

with gradient ascent on $\theta\in\Theta$, where $\left(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)}\right)\sim\mathbb{P}_{XY}, i=1,\dots,b$ are samples from a joint distribution of X and Y, and $\bar{\boldsymbol{y}}^{(i)}\sim\mathbb{P}_Y, i=1,\dots,b$ are samples from a marginal distribution of Y. In practice, we use paired spatial and channel tokens $(\boldsymbol{T}_S^{(i)},\boldsymbol{T}_C^{(i)})$ from the image $\boldsymbol{x}_i$ as the joint distribution and use unpaired tokens $(\boldsymbol{T}_S^{(i)},\boldsymbol{T}_C^{(j)})$ from random images $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as the marginal distribution. Then, we can minimize Equation (9) by optimizing the spatial-channel token for minimal mutual information.

3.3. The Overall Pipeline of STD

Multi-teacher Distillation with Tokens. Distillation with multiple teachers has turned out to be effective to improve the performance of student than using a single teacher (Hinton et al., 2015; Sau & Balasubramanian, 2016; You et al., 2017), because different teacher networks can provide their unique information. We also consider multi-teacher distillation with our spatial-channel tokens. To achieve multi-teacher distillation, we only need to add extra tokens into Equations (5) and (6). This allows each teacher to correspond to a specific latent representation in the student network. An intuitive way to utilize those teachers is using the averaged response from them (Hinton et al., 2015). However, we argue that different teachers have various importance. Therefore, we further introduce an entropy-based

Algorithm 1 The mutual information regularization on spatial-channel tokens.

Input: the vision model f and parameters $\boldsymbol{\omega}$, the MINE network ψ and parameters $\boldsymbol{\theta}$, input images $\boldsymbol{x}$, the number of samples b, the number of epochs T

for $t = 1$ **to** T **do**
  Get the vision model outputs: $(\hat{y}, \boldsymbol{T}_S, \boldsymbol{T}_C) \leftarrow f_{\boldsymbol{\omega}}(\boldsymbol{x})$;
  Draw random pairs $(\boldsymbol{T}_S^{(1)}, \boldsymbol{T}_C^{(1)}), \ldots, (\boldsymbol{T}_S^{(b)}, \boldsymbol{T}_C^{(b)})$ from $(\boldsymbol{T}_S, \boldsymbol{T}_C)$ as the joint distribution;
  Draw random samples $\bar{\boldsymbol{T}}_C^{(1)}, \ldots, \bar{\boldsymbol{T}}_C^{(b)}$ from $\boldsymbol{T}_C$ as the marginal distribution;
  Calculate Equation (9): $\mathcal{V}(\boldsymbol{\theta}) \leftarrow \frac{1}{b}\sum_{i=1}^{b} \psi_{\boldsymbol{\theta}}(\boldsymbol{T}_S^{(i)}, \boldsymbol{T}_C^{(i)}) - \log(\frac{1}{b}\sum_{i=1}^{b} e^{\psi_{\boldsymbol{\theta}}(\boldsymbol{T}_S^{(i)}, \bar{\boldsymbol{T}}_C^{(i)})})$;
  Calculate $\boldsymbol{g}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}\mathcal{V}(\boldsymbol{\theta})$ and update $\boldsymbol{\theta}$ by gradient ascent: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{g}_{\boldsymbol{\theta}}$;
  Calculate $\boldsymbol{g}_{\boldsymbol{\omega}} = \nabla_{\boldsymbol{\omega}}\mathcal{V}(\boldsymbol{\theta})$ and update $\boldsymbol{\omega}$ by gradient descent: $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \boldsymbol{g}_{\boldsymbol{\omega}}$;
end for

confidence re-weighting term. The losses regarding different teachers are re-weighted by their confidence about a sample. We use the negative entropy (Wan, 1990; Zaragoza & d'Alché Buc, 1998) of the softmax distribution of a teacher network to measure its confidence score:

$$S(g_i) = -\sum_{i=1}^{K} P\left[g_i(\boldsymbol{x})|\boldsymbol{x}\right] \cdot \log P\left[g_i(\boldsymbol{x})|\boldsymbol{x}\right], \tag{10}$$

where K is the number of class, and g_i is the i-th teacher network. Finally, the multi-teacher distillation loss re-weighted by 10 is calculated. The confidence score for each teacher is normalized by softmax to ensure the range of the final loss is proper. The overall multi-teacher distillation loss, therefore, is

$$\mathcal{L}_{\text{dist}} = \sum_{i=1}^{M} \left(\sigma([\mathrm{S}(g_j)]_{j=1,\ldots,M})^{(i)} \cdot \ell_{\text{CE}}(f(\boldsymbol{x}), g_i(\boldsymbol{x}))\right), \tag{11}$$

where M is the number of teachers, σ is the softmax function, and ℓ_{CE} is the cross-entropy loss.

The Spatial-channel Token Distillation Network. An image is first split into patches and embedded to the target dimension. The patches are then passed through MLP blocks. As aforementioned, the proposed method can be used for both last layer distillation and intermediate layer distillation. To distill intermediate layers, the nature of STD allows us to directly add tokens to the target layers. Because STD only back-propagates gradients to the previous layers and does not propagate outputs to the following layers, we can safely use it in the middle of a network without worrying it conflicts with tokens for the distillation of the last layer. In our practice, we find distilling shallow layers with a small teacher, e.g. ResNet-50 (He et al., 2016), and deep layers with a large teacher, e.g. ResNet-101, can outperform distilling the whole network with the same two networks. For the classification head, we use a global average pooling (GAP) before the fully-connected classification head. To get the final prediction, we follow DeiT (Touvron et al., 2021b) and use the average of the distillation head(s) and classification head for the final prediction.

The Overall Distillation Objective. Finally, the concatenated distillation tokens are fed to fully connected classification heads. The predictions are used for *hard-label distillation*, where the hard decisions of teachers are the targets of the student. Letting the teacher $g_i(\cdot)$ in Equations (10) and (11) outputs the argmax of its classification head, we can have its hard decisions. The objective associated with this hard label distillation is:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{class}} + \alpha\mathcal{L}_{\text{dist}}, \tag{12}$$

where $\mathcal{L}_{\text{class}} = \ell_{\text{CE}}(f(\boldsymbol{x}), y)$ is the classification loss regarding the ground truth labels, and $\mathcal{L}_{\text{dist}}$ is the distillation loss calculated by Equation (11).

4. Experiments

In this section, we demonstrate the performance of STD with several popular architectures for MLP-like vision models. We first compare models distilled by STD and trained with the origin schema. Then, we discuss the distillation settings in STD. Finally, we perform extensive ablation studies to evaluate each component in STD.

Table 1. The throughput of various MLP-like architectures tested on ImageNet-1K with 8×NVIDIA V100 GPUs. We use the largest possible batch size to evaluate the maximal speed.

Architecture	Variant	Patch Size	Over-lapping	Throughput (images/s)
Mixer	S16	16	✗	962.45
ResMLP	S24	16	✗	733.59
CycleMLP	B1	7	✓	822.39
Mixer	B16	16	✗	562.80
ResMLP	B24	16	✗	346.09
CycleMLP	B2	7	✓	664.14

Table 2. Comparison between STD and SOTA methods. We compare models distilled by STD with models trained from scratch on ImageNet-1K and models pre-trained on large-scale datasets and then fine-tuned on ImageNet-1K.

Model	Params (M)	FLOPs (G)	Top-1 Acc. (%)
CNN			
ResNet-18 (He et al., 2016)	12.5	1.8	69.8
ResNet-50 (He et al., 2016)	22.0	4.1	78.9
RSB-ResNet-18 (Wightman et al., 2021)	12.5	1.8	71.5
RSB-ResNet-50 (Wightman et al., 2021)	22.0	4.1	80.4
Transformer-based			
ViT-B/16/384 (Dosovitskiy et al., 2021)	86.0	-	77.9
ViT-L/16/384 (Dosovitskiy et al., 2021)	307.0	-	76.5
DeiT-Ti (Touvron et al., 2021b)	6.0	-	74.5
DeiT-S (Touvron et al., 2021b)	22.0	-	81.2
DeiT-B (Touvron et al., 2021b)	87.0	-	83.4
MLP-like			
Mixer-S16 (Tolstikhin et al., 2021)	18.5	3.8	72.9
+ JFT-300M	18.5	3.8	73.8 (+0.9)
+ DeiT Distillation (Touvron et al., 2021b)	20.0	3.8	74.2 (+1.3)
+ **STD (ours)**	22.2	4.3	**75.7** (+2.8)
Mixer-B16 (Tolstikhin et al., 2021)	59.9	12.7	76.4
+ JFT-300M	59.9	12.7	80.0 (+3.6)
+ ImageNet-21K	59.9	12.7	80.6 (+4.2)
+ **STD (ours)**	66.7	13.7	80.0 (+3.6)
ResMLP-S24 (Touvron et al., 2021a)	30.0	6.0	79.4
+ **STD (ours)**	32.5	6.2	**80.0** (+0.6)
ResMLP-B24 (Touvron et al., 2021a)	115.7	23.0	81.0
+ **STD (ours)**	122.6	24.1	**82.4** (+1.4)
CycleMLP-B1 (Chen et al., 2021)	15.2	2.1	78.9
+ **STD (ours)**	18.4	2.2	**80.0** (+1.1)
CycleMLP-B2 (Chen et al., 2021)	26.8	3.9	81.6
+ DeiT Distillation (Touvron et al., 2021b)	28.6	3.9	81.9 (+0.3)
+ **STD (ours)**	30.1	4.0	**82.1** (+0.5)

4.1. Setup

Datasets. We use the ImageNet-1K (Russakovsky et al., 2015) dataset for both distillation and evaluation. It has 1.3 million images covering 1,000 classes. One of our baselines, MLP-Mixer, uses additional datasets, including ImageNet-21K and JFT-300M (Sun et al., 2017), for pre-training. ImageNet-21K is a superset of ImageNet-1K, which contains 14 million images covering 21,000 classes. JFT-300M is a private dataset, which contains 300 million images covering 18,000 classes. We do not use any extra images or labels from them.

Student Networks. We evaluate STD with various MLP-like architectures, including MLP-Mixer (Tolstikhin et al., 2021), ResMLP (Touvron et al., 2021a) and CycleMLP (Chen et al., 2021). Each architecture has multiple variants, we compare the variants with a similar throughput. The throughput is tested on ImageNet-1K with 8×NVIDIA V100 GPUs and is listed in Table 1. The first group includes Mixer-S16, ResMLP-S24, and CycleMLP-B1, and the second group includes Mixer-B16, ResMLP-B24, and CycleMLP-B2.

Teacher Networks. We mainly use CNNs as the teacher networks. We consider two variants of ResNet (He et al., 2016), including ResNet-50 and ResNet-101. They have 79.6% and 80.7% top-1 accuracy on ImageNet-1K, respectively. We also perform an ablation study by distilling with a Transformer-based vision model, Swin-Transformer (Liu et al., 2021), as a teacher. We use Swin-B pre-trained on ImageNet-22K with the identical 224×224 resolution to ResNets. It has 85.2% top-1 accuracy on ImageNet-1K.

4.2. Comparison with State-of-the-Art Methods

We compare with three kinds of SOTA MLP-like vision models, including models trained from scratch on ImageNet-1K, models pre-trained with large-scale datasets (e.g., JFT-300M and ImageNet-21K) and fine-tuned on ImageNet-1K, and models distilled with prior methods, such as DeiT. We

Table 3. The top-1 accuracy on ImageNet-1K of CycleMLP-B2 distilled with various selections of teachers.

Teachers				Student
Archtecture	ResNet-50	ResNet-101	Swin-B/224	Top-1 Acc. (%)
Params (M)	25.58	44.57	87.77	
FLOPs (G)	4.36	8.09	15.14	
Selection	✓	✗	✗	81.47
	✗	✗	✓	81.91
	✓	✓	✗	81.96

Table 4. The top-1 accuracy on ImageNet-1K of CycleMLP-B2 distilled with and without the propsoed spatial-channel tokens.

Teachers		S+C Tokens	Student
ResNet-50	ResNet-101		Top-1 Acc. (%)
✓	✗	✗	81.40
✓	✗	✓	81.47
✓	✓	✗	81.89
✓	✓	✓	81.96

also compare to Transformer-based vision models both with and without distillation. The results are reported in Table 2.

Compared to MLP-Mixers, models distilled with our STD without additional datasets obtain consistently better performance. With the proposed STD, Mixer-S16 can reach 75.7% top-1 accuracy, which is 1.0% higher than the model pre-trained on JFT-300M. As for Mixer-B16 with STD, it can reach 80.0%, which is higher than the models pre-trained on ImageNet-1K and JFT-300M and is competitive to the model pre-trained on ImageNet-21K. Besides, we find that if there is no pre-training, even though Mixer-S16 is 3 times smaller than Mixer-B16 in terms of FLOPs and parameters, Mixer-S16+STD can still reach a competitive performance to Mixer-B16 trained from scratch. It demonstrates the difficulty of optimizing MLP-Mixers from scratch and the effectiveness of our STD.

As for stronger architectures, e.g. ResMLP and CycleMLP, if Mixer-B16 is pre-trained on ImageNet-1K instead of the two large datasets, JFT-300M and ImageNet-21K, the performance of it is consistently lower than ResMLP and CycleMLP regardless of the model size. By distilling Mixer-B16 with STD, it can outperform ResMLP-S24 nad CycleMLP-B1 and reach very close to ResMLP-B24 and CycleMLP-B2. When applying to ResMLP and CycleMLP, STD can further improve their performance. Among them, STD reaches the maximal accuracy gain of 1.1% on CycleMLP-B1.

Compared with DeiT, it can improve the accuracy of Mixer-S16 to 74.2%, which is still lower than DeiT-Ti. In contrast, STD can improve Mixer-S16 to be better than DeiT-Ti by 1.2%. The larger MLP-Mixer, Mixer-B16, performs lower than both ViTs. When applying STD to it, it can reach 2.1% and 3.5% higher accuracy than ViT-B/16/384 and ViT-L/16/384, respectively.

4.3. Distillation Settings

In this subsection, we discuss our distillation setting, including the selection of teachers, the spatial-channel tokens, the distillation of the intermediate layer, and the manner of prediction. We use the CycleMLP-B2 as the student, whose accuracy is the best among our models in Table 2.

Different Teachers. We first consider the selection of teachers for STD. There are three different options in Table 3, including a small teacher (i.e. ResNet-50), a large teacher (i.e. Swin-B/224), and the combination of two small teachers (i.e. ResNet-50 and ResNet-101). Even though it is no surprise that the large Swin-B/224 can distill a better student than the small ResNet-50, we find the combination of two small teachers can reach competitive performance to one large teacher. The Swin-B/224 has 87.77M parameters and 15.14G FLOPs, yet the combination of ResNet-50 and ResNet-101 only has 70.15M parameters and 12.45G FLOPs. Besides, both of the two ResNets have lower accuracy than Swin-B/224 (79.6% and 80.7% vs. 85.2%). Nevertheless, distilling with the two ResNets reaches 81.96% top-1 accuracy on ImageNet-1K, which is slightly higher than the accuracy by distilling with Swin-B/224.

Besides, we find the student network can outperform its teachers by distilling with ResNet-50 or the combination of two ResNets. Although Swin-B/224 has higher accuracy than both of the ResNets, its student has much lower performance than it. This advantage does not maximize the benefit of its student. We argue this could be caused by the huge gap between the model sizes of CycleMLP-B2 and Swin-B/224. The student only has 30.1M parameters and 4.0G FLOPs, which is about three times smaller than the teacher.

Spatial-channel Tokens. Then, we study the impacts of our spatial-channel token, under both single teacher distillation and multi-teacher distillation. ResNet-50 is used for the single teacher distillation, and the combination of ResNet-50 and ResNet-101 is used for the multi-teacher distillation. As can be seen in Table 4, the spatial-channel tokens can obtain consistent performance gains.

Intermediate Layer Distillation. We evaluate the intermediate layer distillation. Instead of distilling different layers with the same teacher, we consider the multi-teacher setting, whose performance is the best in Table 3. We use the

Table 5. The top-1 accuracy on ImageNet-1K of CycleMLP-B2 distilled at the last layer and at multiple positions.

Teachers			Student
Architecture	ResNet-50	ResNet-101	Top-1 Acc. (%)
Position	Last	Last	81.96
	Inter	Last	82.09

Table 6. Ablation study on using different prediction methods, including using the averaged response, using the classification head, and using distillation heads.

Model	No Dist	Mean	Class Head	Dist Head
Mixer-S16	72.9	75.74	75.40 (-0.34)	75.70 (-0.04)
Mixer-B16	76.4	80.05	78.78 (-1.27)	80.05 (-0.00)
CycleMLP-B1	78.9	79.96	79.59 (-0.37)	79.92 (-0.04)
CycleMLP-B2	81.6	82.11	82.01 (-0.10)	81.97 (-0.14)

ResNet-50 to distill the intermediate layer and the ResNet-101 to distill the last layer. The intermediate distillation tokens are inserted into the $2/3$ position of the network. Table 5 demonstrates distilling both the intermediate layer and last layer can improve the accuracy by 0.13%.

Prediction Heads. As aforementioned, we follow DeiT (Touvron et al., 2021b) and use the averaged response from the classification head and distillation heads. We also perform an ablation study to evaluate this prediction manner with both MLP-Mixer and CycleMLP. The results are reported in Table 6. As can be seen, the accuracy always drops whether the classification head or distillation heads are used alone. Although the distillation heads perform better than the classification head in most cases, neither of them is necessarily better. A possible explanation of this phenomenon is that the classification head learns ground-truth labels, and distillation heads learn teachers' outputs. They can learn different hypothesis. Using them together is similar to ensemble learning, which can improve the predictive performance. It is also worth noting that both of the heads in networks distilled by STD can reach a better performance than the networks without distillation, which demonstrates the effectiveness of STD.

4.4. Spatial-channel Token Distillation

We perform ablation studies on different components in STD, including the teachers' confidence, spatial-channel tokens, and mutual information regularization. Models trained from scratch and pre-trained on JFT-300M are considered as baselines. The results are reported in Table 7. In this experiment, we use Mixer-S16 as the student and distill it with the two default CNN teachers in our work, i.e. ResNet-50 and ResNet-101.

Teachers' Confidence Weights. We first consider the teachers' confidence weights, which can always be applied to multi-teacher distillation regardless of the use of tokens. Whether distilling without any token, with spatial tokens, or with spatial-channel tokens, the confidence weights can always increase the accuracy by around 0.12% to 0.14%.

Distillation Tokens. As for the distillation tokens, we consider three different settings: distilling without any token, distilling with spatial tokens only, and distilling with our spatial-channel tokens. Firstly, we find even vanilla distillation can indeed improve the performance of MLP-Mixer. By distilling without any token, the accuracy is increased by 1.11% compared to pre-training on JFT-300M and 2.07% compared to training from scratch. However, distilling with spatial tokens only reduces the performance gain. By adding spatial tokens, the accuracy drops 0.26%. Our spatial-channel token can improve the accuracy by 0.72% and reach 75.66%. This is an increase of 1.83% compared to JFT-300M and 2.79% compared to from scratch.

Besides, we also find the classification token can harm the performance of MLP-like vision models, even though it works well in DeiT. Distilling with spatial tokens in this experiment is similar to DeiT, but we use a GAP before the classification head instead of a classification token like them. Comparing with the accuracy of Mixer-S16 distilled with the DeiT's distillation method reported in Table 2, Mixer-S16 can reach 75.56% top-1 accuracy with spatial distillation token and GAP classification head, which is about 0.4% higher than the former.

Mutual Information Regularization. Because the mutual information regularization is between the spatial and channel tokens, it does not apply to distillation without any token or with the spatial tokens only. Therefore, we mainly study its impacts on STD. As can be seen, the accuracy of STD can further increase to 75.74%, which is 1.91% higher than the pre-training on JFT-300M. The computational cost of MINE is also marginal. We use a three-layer MLP with 512 dimensions as the MINE network. It has only 0.003G FLOPs and 0.84M parameters, which is much smaller than the models.

5. Conclusion

In this work, we propose a distillation mechanism designed for MLP-like vision models, namely Spatial-channel Token Distillation (STD). STD adds distillation tokens into both the spatial and channel dimension of MLP blocks. Those tokens are designed to improve the spatial and channel mix-

Table 7. Ablation studies on components in STD with Mixer-S16 as the student network.

Method	Pre-training	Distillation	Token	Teachers' Confidence	MIR	Top-1 Acc. (%)
None	✗	✗	-	-	-	72.87
Pre-training	JFT-300M	✗	-	-	-	73.83
Other Distillation	✗	✓	✗	✗	-	74.80
	✗	✓	✗	✓	-	74.94
	✗	✓	S	✗	-	74.56
	✗	✓	S	✓	-	74.68
STD (ours)	✗	✓	S + C	✗	✗	75.52
	✗	✓	S + C	✓	✗	75.66
	✗	✓	S + C	✓	✓	75.74

ings. We also introduce a mutual information regularization to disentangle the spatial and channel information. The proposed spatial-channel tokens are not only suitable for last layer distillation but also applicable for the distillation of intermediate layers. By inserting additional pairs of tokens, STD also supports multi-teacher distillation. Extensive experiments demonstrate that STD can improve the performance of MLP-like vision models. Ablation studies show that distilling with the spatial-channel tokens can outperform vanilla distillation without any token and DeiT's distillation with spatial tokens only.

Acknowledgements

The authors would like to thank the area chairs and the reviewers for their constructive comments. This work was supported in part by the Australian Research Council under Project DP210101859 and the University of Sydney SOAR Prize.

References

Asif, U., Tang, J., and Harrer, S. Ensemble knowledge distillation for learning improved and efficient networks. *arXiv preprint arXiv:1909.08097*, 2019.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mine: mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 531–540. PMLR, 2018.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Bucilua, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Chen, H., Wang, Y., Xu, C., Xu, C., and Tao, D. Learning student networks via feature embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1): 25–35, 2020.

Chen, S., Xie, E., Ge, C., Liang, D., and Luo, P. Cyclemlp: A mlp-like architecture for dense prediction. *arXiv preprint arXiv:2107.10224*, 2021.

Chen, X., Su, J., and Zhang, J. A two-teacher framework for knowledge distillation. In *International Symposium on Neural Networks*, pp. 58–66. Springer, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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*, 2021.

Guo, J., Tang, Y., Han, K., Chen, X., Wu, H., Xu, C., Xu, C., and Wang, Y. Hire-mlp: Vision mlp via hierarchical rearrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Guo, T., Xu, C., He, S., Shi, B., Xu, C., and Tao, D. Robust student network learning. *IEEE transactions on neural networks and learning systems*, 31(7):2455–2468, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Hou, Q., Jiang, Z., Yuan, L., Cheng, M.-M., Yan, S., and Feng, J. Vision permutator: A permutable mlp-like architecture for visual recognition. *arXiv preprint arXiv:2106.12368*, 2021.

Hu, H., Zhang, Z., Xie, Z., and Lin, S. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3464–3473, 2019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

Park, S. and Kwak, N. Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks. In *ECAI 2020*, pp. 1411–1418. IOS Press, 2020.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. Image transformer. In *International Conference on Machine Learning*, pp. 4055–4064. PMLR, 2018.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding with unsupervised learning. 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., and Shlens, J. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Sau, B. B. and Balasubramanian, V. N. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*, 2016.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Tang, Y., Han, K., Guo, J., Xu, C., Li, Y., Xu, C., and Wang, Y. An image patch is a wave: Phase-aware vision mlp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021a.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Wan, E. A. Neural network classification: a bayesian interpretation. *IEEE Transactions on Neural Networks*, 1(4): 303–305, 1990.

Wei, L., Xiao, A., Xie, L., Zhang, X., Chen, X., and Tian, Q. Circumventing outliers of autoaugment with knowledge distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 608–625. Springer, 2020.

Wightman, R., Touvron, H., and Jégou, H. Resnet strikes back: An improved training procedure in timm. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, 2021.

You, S., Xu, C., Xu, C., and Tao, D. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1294, 2017.

Yuan, L., Tay, F. E., Li, G., Wang, T., and Feng, J. Revisit knowledge distillation: a teacher-free framework. 2019.

Zaragoza, H. and d'Alché Buc, F. Confidence measures for neural network classifiers. In *Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowlegde Based Systems*, volume 9, 1998.

Zhao, H., Jia, J., and Koltun, V. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10076–10085, 2020.