DISTILL VISION TRANSFORMERS TO CNNS VIA TEACHER COLLABORATION

Sunqi Lin[1], Chong Wang[1,2], Yujie Zheng[1], Chenchen Tao[1], Xinmiao Dai[1] and Yuqi Li[1,2]*

[1] Faculty of Electrical Engineering and Computer Science, Ningbo University, China
[2] Zhejiang Engineering Research Center of Advcanced Mass Spectrometry and Clinical Application, China

ABSTRACT

The vision transformer (ViT) has recently emerged as a leading approach in various domains, outperforming other methods. Therefore, it is logical to explore the possibility of transferring the superior knowledge from ViT to more compact and cost-effective convolutional neural networks (CNNs). However, due to substantial architectural disparities in representation and logits between these models, conventional knowledge distillation methods have proven ineffective in this context. To address this issue, a novel cross-architecture knowledge distillation scheme based on teacher collaboration is proposed to alleviate the architecture gap. Two different teachers, i.e. one ViT and one CNN, are utilized to simultaneously distill the student by feature reaggregation and logit correction. The experiments show that the proposed scheme outperforms conventional methods on CIFAR-100 dataset. The code is available at https://github.com/SunkiLin/RCD.

Index Terms— Knowledge Distillation, Vision Transformer, Convolutional Neural Network, Cross Architecture

1. INTRODUCTION

The attention mechanism brings versatility and powerful modeling capabilities to Vision Transformer (ViT) [1], allowing it to excel in several fields such as image classification [2], object detection [3], and semantic segmentation [4]. However, the high computational complexity that comes along with the attention mechanism also constrains their application on edge devices. On the other hand, Convolutional Neural Network (CNN) already had established a complete family of hardware accelerated libraries and held several mature and effective backbones .Hence it is natural to explore how to transfer the knowledge from ViT to compact CNN.

The existing model compression methods include pruning [5, 6], quantization [7], low-rank decomposition [8] and knowledge distillation [9]. Knowledge distillation (KD) is a simple and efficient technique that typically utilizes heavy models (teachers) to provide guidance to light models (students) and thus achieve performance improvements in students without introducing extra inference costs. The concept of KD was first proposed in [10], which defined prediction logits from teachers as knowledge. Followed by it, different types of knowledge are broadly explored for better guidance, including logits, intermediate layer feature, and feature relationship. FitNet [11] leverages hints from the intermediate layer to facilitate student learning. RKD [12] takes the relative relationships among the outputs within a batch as guidance.

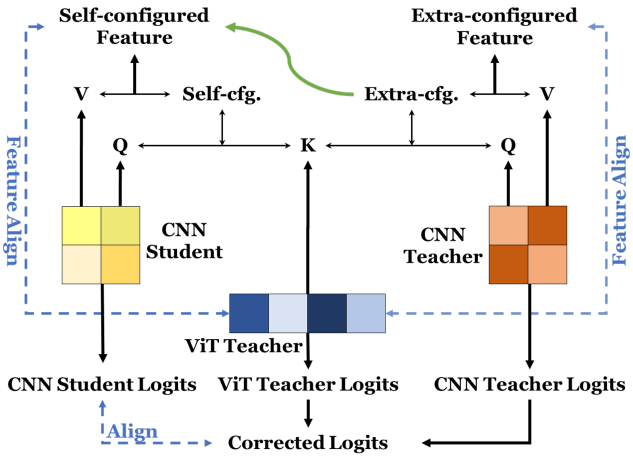

Fig. 1. The illustration of proposed reaggregation and correction distillation (RCD), where **cfg.** indicates the configuration matrix.

However, conventional methods are delicately designed for homo-architecture, e.g., CNN to CNN, Transformer to Transformer, and are not tuned to cross-architecture, because of the representation gap and logits gap between them. For the representation gap, CNN utilizes multiple convolutional layers to progressively expand the receptive field to extract features, whereas ViT feeds the patched images (tokens) into multi-head self-attention layers to receive global receptive fields throughout and obtain features. This also leads to the features from them are completely different not only in the manner of representation but also in the coding region. For logits gap, the ViT teacher has higher accuracy than traditional CNN in most cases, but this also brings over-confidence to the teacher's logits distribution, which is similar to or even worse than, the gap problem in CNN distillation.

Inspired by the cross-attention mechanisms [13], a novel cross-architecture feature reaggregation and logits correc-

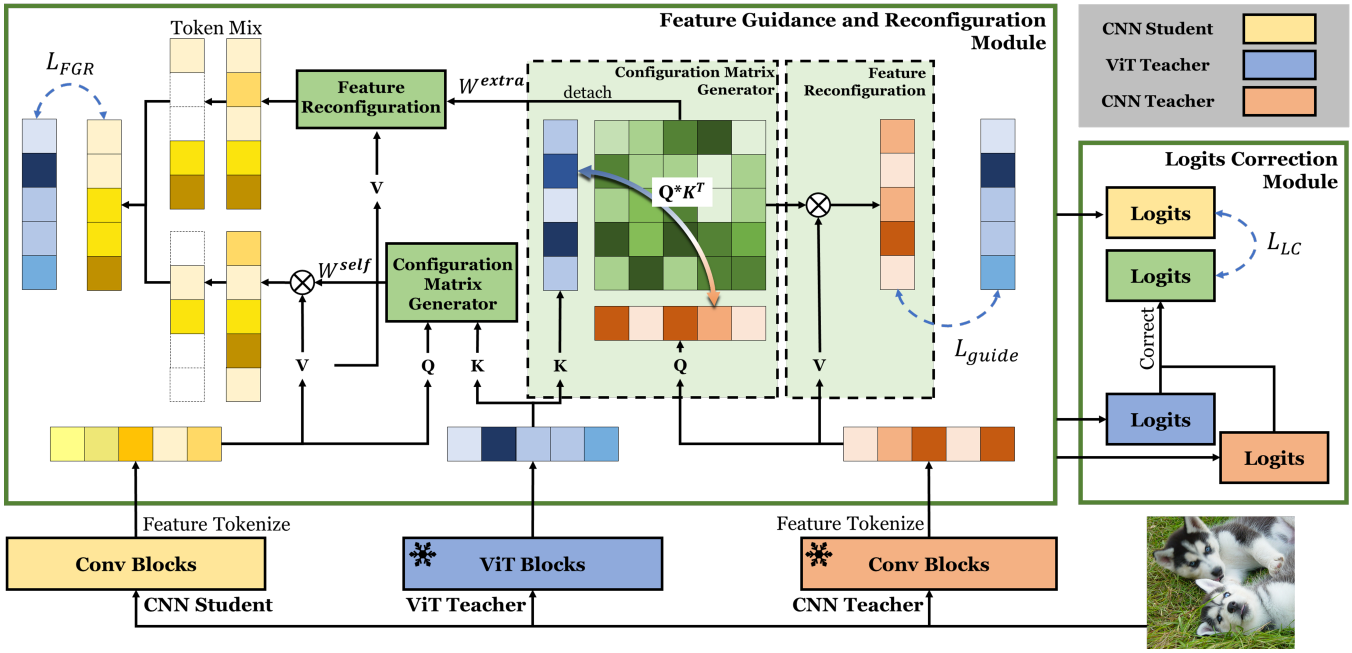

Fig. 2. The overview of the proposed reaggregation and correction distillation network.

tion distillation scheme based on teacher collaboration is proposed, as shown in Fig.1. We loosen the point-to-point feature constraints in the classic method and use a plane-to-point form instead, in which each point in the student's feature autonomously assigns the proportion of learning for each token across the teacher feature. Meanwhile, an auxiliary CNN teacher is introduced to guide the learning process for student representation and enrich the information entropy of the ViT teacher's logits distribution.

2. METHODOLOGY

To leap the representation gap and logits gap in cross-architecture KD, a novel feature reaggregation and logits correction distillation scheme based on teacher collaboration is proposed. As shown in Fig.2 , the student's feature is reconfigured by the self-configuration matrix and guided by the extra-configuration matrix in the feature guidance and reconfiguration module (FGR) before feature aligning. Meanwhile, the teacher's logits are corrected by the auxiliary teacher's logits for better instruction in the logits correction module (LC).

2.1. Preliminary

Denote the ViT teacher, CNN teacher and CNN student by t_v, t_c and s. $F^{t_v} \in \mathbb{R}^{N\times E}$, $F^{t_c} \in \mathbb{R}^{C\times H\times W}$ and $F^{s} \in \mathbb{R}^{C'\times H'\times W'}$ refer to their feature respectively, where H and W are the height and width of the feature map, C and E represent the number of channels and the token embedding dimension, respectively. The token numbers N are determined by the patch size h and w, where $N = (H \times W)/(h \times w)$.

Given a training sample $x \in \mathbb{R}^{3\times H\times W}$ with label $y \in \mathbb{R}^{M}$, where M is the number of classes. The output logits of teacher and student can be expressed as $z^t = t(x)$ and $z^s = s(x)$. Therefore, a classic distillation loss $\mathcal{L}_{\text{KD}}$ can be defined as follows, to measure the difference between z^t and z^s,

$$\mathcal{L}_{\text{KD}} = \sum_{i=1}^{N} \sigma\left(\frac{z_i^t}{\tau}\right) \log\left(\sigma\left(\frac{z_i^s}{\tau}\right)\right) \tag{1}$$

Where $\sigma(\cdot)$ denotes the softmax function, τ represents the distillation temperature, z_i^t is the i-th item of z^t, and so on is z_i^s.

2.2. Feature Guidance and Reconfiguration

In conventional feature distillation, F^{t_c} and F^s are required for pixel-wise matching exactly, which implicitly assumes that the intermediate representations of different networks have the same extent of semantic information. However, due to the ViT teacher's significantly broader receptive field compared to the CNN student, it is not reasonable to expect the CNN student to model the ViT teacher's representation exactly with its limited receptive field. Therefore, our approach aims to represent each ViT token using the entire CNN feature map. Specifically, the conception of the query-key is adopted from cross-attention mechanisms, and a self-configuration matrix is derived by employing the CNN student as the query and the ViT teacher as the key. Additionally, we compute an extra-configuration matrix by replacing the CNN student's query with the CNN teacher's query to guide the student's

learning process.

For generality, the simplest linear interpolation and convolution method is used for aligning F^{t_c} and F^s with F^{t_v}. As a result, the aligned feature of t_c and s can be denoted as $\widehat{F^{t_c}} \in \mathbb{R}^{N\times E}$ and $\widehat{F^s} \in \mathbb{R}^{N\times E}$. Then, the self-configuration matrix can be denoted as W^{self} to map the entire F^s to each token of F^{t_v} and the self-configured $\widehat{F^s_{self}}$ can be obtained by the following,

$$W^{\text{self}} = \sigma\left(Q\left(\widehat{F^s}\right) \times \left(K\left(F^{t_v}\right)\right)^T\right) \in \mathbb{R}^{N\times N} \quad (2)$$

$$\widehat{F^s_{self}} = W^{\text{self}} \times V(\widehat{F^s}) \in \mathbb{R}^{N\times E} \quad (3)$$

where $Q(\cdot)$, $K(\cdot)$, and $V(\cdot)$ contain a linear projection layer and a layer normalization operation based on the attention heads number. However, it is still difficult for CNN students to align the ViT teacher's features directly through the feature self-configuration alone. So we introduce an auxiliary pre-trained CNN teacher t_c to guide the feature reconfiguration process. The auxiliary teacher engages with the ViT teacher in the same manner as the CNN student does, acquiring the extra-configuration matrix denoted as W^{extra} and producing the self-configured feature $\widehat{F^{t_c}_{self}}$. This process can be described as follows,

$$W^{\text{extra}} = \sigma\left(Q\left(\widehat{F^{t_c}}\right) \times \left(K\left(F^{t_v}\right)\right)^T\right) \in \mathbb{R}^{N\times N} \quad (4)$$

$$\widehat{F^{t_c}_{extra}} = W^{\text{extra}} \times V(\widehat{F^{t_c}}) \in \mathbb{R}^{N\times E} \quad (5)$$

The $Q(\cdot)$, $K(\cdot)$, $V(\cdot)$ have the same definition as above. Additionally, it can be observed that the $\widehat{F^{t_c}}$ after the generation layer produces an improved W^{extra} in practice. Subsequently, the loss function of $\widehat{F^{t_c}_{extra}}$ can be denoted as,

$$\mathcal{L}_{\text{guide}} = \|F^{t_v} - \widehat{F^{t_c}_{extra}}\|_2 \quad (6)$$

Notably, the auxiliary CNN teacher only utilizes the detached extra-configuration matrix W^{extra} for guidance, and the loss $\mathcal{L}_{\text{guide}}$ is computed independently.

Therefore, the guided student feature $\widehat{F^s_{extra}}$ is obtained as follows,

$$\widehat{F^s_{extra}} = W^{\text{extra}} \times V(\widehat{F^s}) \in \mathbb{R}^{N\times E} \quad (7)$$

With reference to curriculum learning, a customized training scheme is developed to allow the student to learn progressively from easy to hard. Specifically, during the early phase of training, the feature $\widehat{F^s_{extra}}$, guided and reconfigured by the auxiliary teacher's W^{extra}, is mixed with the student's self-configuration feature $\widehat{F^s_{self}}$ in a token-wise manner. This training scheme can be formulated as follows,

$$\widehat{F^s} = TokenMix(\widehat{F^s_{self}}, \widehat{F^s_{extra}}; \mu) \quad (8)$$

where the $TokenMix(\cdot,\cdot;\mu)$ is a method to mix features in token dimension, and the μ is a dynamic hyper-parameter that controls the intensity of mixing. In addition, the value of μ decreases as the training progresses. After this scheme, the difficulty of the student's training is effectively reduced. Therefore, the total feature guidance and reconfiguration (FGR) loss can be expressed as,

$$\mathcal{L}_{\text{FGR}} = \left\|F^{t_v} - \widehat{F^s}\right\|_2 \quad (9)$$

2.3. Logits Correction

The logits gap is a well-known issue in knowledge distillation, where large CNN teachers often struggle to effectively transfer their knowledge to students. Recent studies [17, 18] have attributed the logits gap to the overconfidence of the large teachers, which results in assigning higher logits value to the target class and subsequently reducing the information entropy in their logits distribution. This phenomenon is further exacerbated by the fact that ViT teachers tend to achieve higher accuracy compared to CNN teachers. Therefore, they have attempted to address this issue by refining temperature adjustment and extracting more information from the teacher's own logits. However, in the extreme state, this becomes trivial since the teacher's logits approach the one-hot state, making them insensitive to temperature adjustments. To overcome this limitation, we introduce the rich knowledge of inter-class relationships from the auxiliary CNN teacher to correct the ViT teacher's logits distribution. Consequently, we define the logits of t_c and t_v as z^{t_c} and z^{t_v} respectively, the corrected ViT teacher logits $\widehat{z^t}$ can be obtained as follows,

$$\widehat{z^t} = \omega z^{t_c} + (1-\omega) z^{t_v} \quad (10)$$

where ω is a hyper-parameter. The logits correction loss $\mathcal{L}_{LC}$ is defined as:

$$\mathcal{L}_{LC} = \sum_{i=1}^{N} \sigma\left(\frac{\widehat{z^t_i}}{\tau}\right) \log\left(\sigma\left(\frac{z^s_i}{\tau}\right)\right) \quad (11)$$

Finally, the total loss of our proposed method is given as,

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{task}} + \beta\mathcal{L}_{\text{LC}} + \gamma\mathcal{L}_{\text{FGR}} \quad (12)$$

It is worth noting that $\mathcal{L}_{\text{task}}$ is the cross-entropy loss for image classification, and is calculated by the student's logits z^s and the ground truth label y, where α, β, and γ are hyper-parameters.

3. EXPERIMENTS

3.1. Implementation Details

CIFAR-100 contains 100 categories of 60K images each with 32×32 pixels, and each category contains 500 training samples and 100 test samples. Our ViT teacher is finetuned 100

Table 1. Results on the CIFAR-100 dataset (%). Res and ViT-T/S denote ResNet and ViT-Tiny/Small respectively.

Student	**Res20** 68.92		**Res20** 68.92		**Res32** 71.14		**Res20** 68.92		**VGG8** 70.36	
Teacher	**Res56** 72.34	**ViT-T** 81.35	**Res110** 74.31	**ViT-T** 81.35	**Res110** 74.31	**ViT-T** 81.35	**Res56** 72.34	**ViT-S** 85.72	**VGG13** 74.64	**ViT-T** 81.35
KD[10]	70.66	64.83	70.67	64.83	73.08	69.34	70.66	64.35	72.98	72.58
FitNet[11]	69.21	64.53	68.99	64.53	71.06	68.14	69.21	63.83	71.02	71.32
SP[14]	69.67	68.54	70.04	68.54	72.69	71.64	69.67	65.48	72.68	72.79
RKD[12]	69.61	69.38	69.25	69.38	71.82	72.45	69.61	68.28	71.48	72.13
PKT[15]	70.34	69.06	70.25	69.06	72.61	71.75	70.34	68.77	72.88	72.97
CRD[16]	71.16	-	71.46	-	73.48	-	71.16	-	**73.94**	-
Ours	**71.69**		**71.49**		**73.75**		**71.20**		73.89	

epochs from ImageNet in CIFAR-100, referring to [19]. The learning rate is initialized as 4e-3 with a 20-epoch linear warmup followed by a cosine decaying schedule. Furthermore, in CNN student training, as the same training settings of [20], all of the models are trained for 240 epochs and optimized by standard stochastic gradient descent (SGD). The initial learning rate is 0.05 with a decay rate of 0.1 at epochs 150, 180, and 210. The α, β, and γ are set as 0.5, 0.6, and 0.1 respectively.

3.2. Comparison with Different Distillation Methods

The proposed method is implemented on diverse network architectures including ViT, ResNet, VGG, and is compared with the previous state-of-the-art methods, as depicted in Table 1. Through the evaluation of five different teacher-student pairs on CIFAR-100, our method demonstrates remarkable performance, surpassing previous works. Furthermore, it has been observed that the logits-based method [10], which originally performed well, shows weak performance in cross-architecture distillation. Whereas the method based on intermediate feature [11] and relationship[12, 14, 15] can obtain a marginal boost when the CNN students possess sufficient capabilities. However, our proposed method, which combines intermediate features and logits, consistently achieves significant enhancements across all teacher-student pairs, thus reaffirming the generality and effectiveness of our method.

3.3. Ablation Study

To validate the effectiveness of feature guidance and reaggregation (FGR) and logits correction (LC), we conducted ablation experiments on the CIFAR-100 dataset. ViT-Tiny, ResNet56, and ResNet20 are employed as ViT teacher, CNN teacher, and CNN student, respectively. The knowledge distillation (KD) and feature reconfiguration (FR) refer to the student models guided by logits or features of single teacher models. As shown in Table. 2, the LC module demonstrates an improvement of 6.02% and 0.19% respectively compared to the CNN student guided solely by the single ViT teacher with KD [10]. The results confirm that our module effectively corrects the logits of the ViT teacher and provides more suitable knowledge to the CNN student. Additionally, we observe that the FGR module also obtained certain improvements compared to the student instructed by the single teacher. Furthermore, while combining them, our method exhibits a significant improvement of 2.77% compared to the baseline. This finding confirms the presence of a synergistic effect between our modules, enabling an effective transfer of knowledge from ViT teachers to CNN students.

Table 2. Ablation study on the CIFAR-100 dataset (%).

ViT teacher	**CNN teacher**	**KD**	**LC**	**FR**	**FGR**	**Accuracy**
						68.92
✓		✓				64.83
	✓	✓				70.66
✓	✓	✓	✓			70.85
✓				✓		68.54
	✓			✓		69.12
✓	✓			✓	✓	69.23
✓	✓	✓	✓	✓	✓	**71.69**

4. CONCLUSION

In this work, a novel feature reaggregation and logits correction distillation scheme based on teacher collaboration has been proposed to leap the architecture gap. The feature reaggregation allows students to align ViT teacher's representation in a plane-to-point manner and be guided by auxiliary teacher. The logits correction module corrects ViT teacher's logits distribution and enriches the information entropy contained in teacher's logits with the help of auxiliary teacher. Experiments on CIFAR-100 dataset with various teacher-student pairs demonstrated the effectiveness of the RCD scheme we proposed in this paper.

5. ACKNOWLEDGES

This work was supported by the Ningbo Municipal Natural Science Foundation of China (No.2022J114), Innovation Challenge Project of China (Ningbo) (No.2022T001).

6. REFERENCES

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

[2] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10347–10357.

[3] 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*. Springer, 2020, pp. 213–229.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[5] Zheng Huo, Chong Wang, Weiwei Chen, Yuqi Li, Jun Wang, and Jiafei Wu, "Balanced stripe-wise pruning in the filter," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4408–4412.

[6] Weiwei Chen, Chong Wang, Zhehao Zhang, Zheng Huo, and Linlin Gao, "Reweighted dynamic group convolution," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3940–3944.

[7] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 525–542.

[8] Wei Wen, Cong Xu, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li, "Coordinating filters for faster deep neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 658–666.

[9] Yujie Zheng, Chong Wang, Yi Chen, Jiangbo Qian, Jun Wang, and Jiafei Wu, "Enlightening the student in knowledge distillation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

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

[11] 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.

[12] 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*, 2019, pp. 3967–3976.

[13] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang, "Knowledge distillation via the target-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10915–10924.

[14] Frederick Tung and Greg Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1365–1374.

[15] Nikolaos Passalis and Anastasios Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 268–284.

[16] Yonglong Tian, Dilip Krishnan, and Phillip Isola, "Contrastive representation distillation," in *International Conference on Learning Representations*, 2019.

[17] Xin-Chun Li, Wen-Shu Fan, Shaoming Song, Yinchuan Li, Shao Yunfeng, De-Chuan Zhan, et al., "Asymmetric temperature scaling makes larger networks teach well again," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3830–3842, 2022.

[18] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang, "Curriculum temperature for knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 1504–1512.

[19] Tianjin Huang, Lu Yin, Zhenyu Zhang, Li Shen, Meng Fang, Mykola Pechenizkiy, Zhangyang Wang, and Shiwei Liu, "Are large kernels better teachers than transformers for convnets?," *arXiv preprint arXiv:2305.19412*, 2023.

[20] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11953–11962.