Integrating Local and Global Frequency Attention for Multi-Teacher Knowledge Distillation

Zhidi Yao[1], Xin Cheng[1], Zhiqiang Zhang[2],Mengxin Du[3] Wenxin Yu[4,5]
[1]Graduate School of Science and Engineering, Hosei University, Tokyo, Japan
[2]School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China
[3]Instrumentation Technology and Economy Institute, Beijing, China
[4]Sichuan Civil-military Integration Institute, Mianyang, China
[5]Fujiang Laboratory, Mianyang, China

***Abstract*—Knowledge distillation, particularly in multi-teacher settings, presents significant challenges in effectively transferring knowledge from multiple complex models to a more compact student model. Traditional approaches often fall short in capturing the full spectrum of useful information. In this paper, we propose a novel method that integrates local and global frequency attention mechanisms to enhance the multi-teacher knowledge distillation process. By simultaneously addressing both fine-grained local details and broad global patterns, our approach improves the student model's ability to assimilate and generalize from the diverse knowledge provided by multiple teachers. Experimental evaluations on standard benchmarks demonstrate that our method consistently outperforms existing multi-teacher distillation techniques, achieving superior accuracy and robustness. Our results suggest that incorporating frequency-based attention mechanisms can significantly advance the effectiveness of knowledge distillation in multi-teacher scenarios, offering new insights and techniques for model compression and transfer learning.**

***Index Terms*—knowledge distillation, frequency attention mechanisms, model compression, deep learning**

I. Introduction

In the last decade, deep neural networks (DNNs) [1] have achieved significant advancements in various vision-related tasks, including image classification [2]–[4], object detection [5], [6], and semantic segmentation [7], [8]. Nevertheless, these high-performance models typically require extensive computational resources and storage capacity, making them challenging to deploy on resource-constrained edge devices. A promising strategy to address this challenge is knowledge distillation (KD) [9], which utilizes the "dark knowledge" from a powerful but complex teacher network to train a more efficient student network. The goal of KD is to enable the student network to replicate the teacher's predictions while using significantly fewer parameters.

Nonetheless, traditional KD methods typically depend on a single pre-trained teacher network. Recently, inspired by human cognitive learning processes, researchers have explored the potential benefits of students learning from multiple teachers. This exploration has given rise to multi-teacher distillation (MKD), which seeks to harness the diverse and valuable insights provided by several teacher networks to enhance the student network's performance. Various MKD approaches have shown that students can indeed gain advantages from multiple teachers [10]–[15]. However, many of these MKD methods often fall short in assigning appropriate importance to each teacher, as they frequently use identical or fixed weights for all teachers [10]–[12]. This can result in suboptimal integration of knowledge from multiple teachers, preventing the student from fully capitalizing on the combined knowledge. Some recent approaches [13]–[15] have introduced various strategies to address the issue of ineffective knowledge integration. However, these methods have certain limitations that hinder their ability to fully capitalize on the potential of knowledge integration. As a result, the performance improvements they offer are often constrained and do not reach their full potential.

Our objective is to ensure that the student model not only captures high-level abstract information but also retains detailed features, such as object parts, from the teacher model. One effective approach to achieve this is by analyzing the student's features in the frequency domain rather than the traditional spatial domain. The frequency domain offers a distinct advantage in interpreting images, particularly those containing repetitive or periodic patterns that may be challenging to detect using conventional spatial domain methods. By transforming features into the frequency domain, the student model can more effectively recognize and preserve intricate details, leading to a more comprehensive understanding of both the finer and broader aspects of the visual data.

In this paper, we propose an adaptive Multi-Teacher Knowledge Distillation with Local and Global Frequency Attention method called **LGMKD**, which introduces a novel approach that enhances the knowledge distillation process by leveraging both local and global frequency attention mechanisms. This method aims to improve the student's ability to capture and integrate detailed and abstract information from multiple teacher networks. By analyzing features in the frequency domain, this technique effectively identifies and preserves critical patterns and structures that might be missed by traditional spatial domain methods. The integration of local and global frequency attention enables the student model to achieve a more com-

*Corresponding author: Mengxin Du(email: dudumengxin@126.com)

prehensive understanding of the knowledge transferred from the teachers, leading to superior performance and more effective utilization of the diverse insights provided by multiple sources. We demonstrate the effectiveness of our method on the CIFAR-100 and ImageNet benchmark datasets. Our main contributions are summarized as follows:

- Novel Dual-Attention Framework: Introduces a novel dual-attention framework that combines both local and global frequency domain analyses. This approach enhances the student's ability to capture detailed and high-level features by leveraging the complementary strengths of both types of attention, leading to more effective knowledge transfer from multiple teacher models.
- Dynamic Knowledge Integration: Proposes an innovative method for dynamically integrating knowledge from multiple teachers through frequency domain attention. This technique overcomes the limitations of fixed-weight methods by adaptively adjusting the influence of each teacher, resulting in improved performance and more robust student models.
- Extensive experiments on image classification datasets CIFAR-100 and ImageNet validate the effectiveness and flexibility of our method.

II. Related Work

Knowledge Distillation. Knowledge distillation has gained significant attention as a promising technique for model compression, utilizing the supervisory signals from complex teacher networks to train lightweight student models. Traditional approaches, such as vanilla KD [9], focused solely on transferring the teacher network's soft labels to the student network. FitNet [16] advanced this by introducing the idea of having the student network replicate the intermediary layer features of the teacher. Building on this, AT proposed aligning the attention maps of teacher and student features, which led to enhanced student performance. CRD [17] further improved distillation effectiveness by employing contrastive learning strategies. SimKD [18] innovated by using the discriminative classifier from the pre-trained teacher model for student inference, aligning features through a single $\mathcal{L}_2$ loss. DKD [19] introduced a decoupling of the original KD loss function into target class and non-target class components. Despite these advancements, previous distillation methods have relied on a single pre-trained network. In contrast, our approach introduces a novel method by extracting knowledge from multiple teacher networks, aiming to leverage the diverse insights provided by multiple teachers.

Multi-Knowledge Distillation. Multi-teacher knowledge distillation (MKD) leverages the principle that collective intelligence can surpass the knowledge of any single individual. By harnessing the diverse insights provided by multiple teacher networks, MKD aims to enhance the performance of the student network. Several MKD approaches have been developed to achieve this goal. For instance, some methods [10]–[12] assign equal weight to each teacher, treating them as equally valuable. While this approach is straightforward, it overlooks the varying significance of different teachers. To address this limitation, RLKD [20] employs reinforcement learning to filter out less suitable teachers and then averages the logits from the remaining, more relevant teachers. Although this method improves the selection process, it still may not fully account for the unique contributions of each teacher. To better integrate the diverse knowledge of multiple teachers, advanced methods have been introduced. EBKD [21] assigns weights based on the entropy of the predicted logits, emphasizing teachers that provide more informative outputs. AMTML-KD [13] calculates teacher importance weights using latent factors, thereby adapting the influence of each teacher according to their contribution to the student model's learning. Additionally, AEKD [14] examines teacher diversity through the gradient space, enhancing the integration of varied teacher insights. CA-MKD [15] further refines this by evaluating the importance of teachers based on their prediction confidence, determined by the cross-entropy between the teacher's logits and the ground-truth labels. These advanced techniques represent significant progress in addressing the limitations of earlier methods by more effectively recognizing and utilizing the unique contributions of each teacher network.

III. Method

In this section, we first explain how to extract local and global frequencies of the teacher network. Secondly, we integrate and design a new multi-teacher knowledge distillation framework based on our newly proposed module.

Notations. We denote the labeled training dataset as $D = x_i, y_i i = 1^N$, where N represents the total number of samples, and K is the number of teachers. Let F denote the feature output from the second network block, represented as a tensor with dimensions $h \times w \times c$. The logit outputs are denoted as $z = [z_1, \ldots, z_C]$, where C indicates the number of categories. The model's final prediction is obtained through a softmax function $\sigma(z^c) = \frac{\exp(z^c/\tau)}{\sum j \exp(z^j/\tau)}$, where τ is the temperature parameter. An illustration of our proposed module is provided in Figure **??**, and the proposed multi-teacher knowledge distillation framework is shown in Figure 2

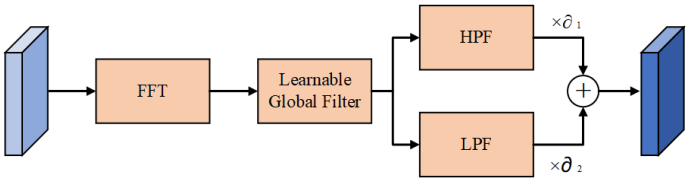

Fig. 1. Local and Global Frequency Attention Module. HPF stands for a high pass filter. LPF stands for a low pass filter. The outputs of the global and local branches are added and the resulting feature map is compared with the teacher's feature map. α_1 and α_2 are the learnable weighting parameters of the global and local frequency, respectively.

A. Local and Global Frequency Attention Module

As illustrated in Figure 2, the proposed module comprises both global and local branches. For a feature map X with

dimensions $C_{in} \times H \times W$, the global branch begins by converting this feature map into the frequency domain using Fast Fourier Transform (FFT). The FFT is applied independently to each channel. For the i-th channel X_i of the feature map X, its 2-D discrete FFT is denoted as Y_i and is expressed as follows:

$$Y_i(u,v) = \sum_{k=0}^{H-1} \sum_{l=0}^{W-1} X_i(k,l) e^{-i2\pi\left(\frac{uk}{H}+\frac{vl}{W}\right)} \tag{1}$$

To adjust the frequencies of Y_i, we apply a learnable global filter K. we design the global filter K with dimensions $C_{out} \times C_{in} \times H \times W$, where C_{out} corresponds to the number of channels in the teacher's feature map. Each kernel in the global filter K has the same spatial dimensions as the 3D input tensor X, which is $C_{in} \times H \times W$. This kernel performs element-wise multiplication with the input tensor X, resulting in a 3D feature map that retains the same dimensions as the input. Subsequently, these 3D frequency feature maps are aggregated through sum-pooling within each $C_{in} \times 1 \times 1$ block, producing a 2D output with dimensions $H \times W$. This process is repeated for all C_{out} kernels in the global filter, yielding a final 3D feature map with dimensions $C_{out} \times H \times W$.

After that, we suppress low frequencies to encourage students to shift their attention away from non-salient areas. To do this, we add a high-pass filter (HPF) after the learnable global filter to remove some of the lowest-frequency components. The HPF is applied to each channel separately. Specifically, for each channel, we use an ideal HPF to suppress 1% of the lowest frequency. In addition, we also use a low-pass filter (LPF) to filter out some high-frequency information, because we found that some high-frequency information interferes with the model's judgment ability. And, we use two parameters α_1 and α_2 to balance the importance of the two.

Next, we convert the frequency domain representation back to the spatial domain using the inverse Fast Fourier Transform (IFFT). For a given frequency feature map $\bar{X}$ that results from the high pass filtering (HPF) and low pass filtering (LPF), the 2-D IFFT of the i-th channel $\bar{X}_i$ of $\bar{X}$ is denoted as $\bar{X}_i$ and is expressed as follows:

$$X_i(k,l) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \bar{X}_i(u,v) e^{i2\pi\left(\frac{uk}{H}+\frac{vl}{W}\right)} \tag{2}$$

The proposed module consists of HPF and LPF. We take the frequency knowledge of the high-pass filter output F_{local} as the local knowledge output and the output of the low-pass filter F_{global} as the global knowledge. The output of the frequency attention module F_{out} is calculated as follows:

$$F_{out} = \alpha_1 * F_{global} + \alpha_2 * F_{local} \tag{3}$$

where α_1 and α_2 are the learnable weighting parameters of the global and local frequency, respectively.

B. Proposed Multi-teacher Knowledge Distillation Framework

Let I represent the indices of the selected layers from the teacher network for intermediate feature-based distillation. The loss for layer-to-layer knowledge distillation is defined as follows:

$$L_{feat} = \sum_{i \in I} D_{FT_i}\left(f(F_{S_j})\right) \tag{4}$$

where F_{S_j} denotes the feature map obtained from the j-th layer of the student network, which is selected to receive knowledge from the feature map F_{T_i} of the i-th layer of the teacher network. The function f represents a transformation applied to the student's feature map. In our approach, f is implemented as the proposed module. The term D refers to a distance function used to measure the discrepancy between the transformed feature maps of the student and the teacher. For our experiments, we utilize the L_2 distance as the chosen distance function. It is important to note that in our framework, the teacher network remains unchanged throughout the process, meaning that no transformations are applied to the feature maps of the teacher network. For simplicity, we only draw one teacher network in the figure, but no matter how many teacher networks there are, similar operations are performed.

C. The Training of Student Network

During the training phase of the student network, we augment the learning process by incorporating not only the ground-truth labels but also the class attention maps generated by multiple teachers as supplementary knowledge to collectively guide the student.

In addition to the aforementioned loss, a regular cross entropy with the ground-truth labels is calculated,

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y^c log(\sigma(z_s^c)). \tag{5}$$

The overall loss function of the proposed **LGMKD** is given as:

$$\mathcal{L}_{overall} = \mathcal{L}_{CE} + \beta \mathcal{L}_{lgmkd}, \tag{6}$$

Here, β is hyperparameters that balance the effects of each loss.

IV. Experiments

A. Datasets and Implementation Details

Datasets. Our experiments are carried out on two prominent datasets: CIFAR-100 [22] and ImageNet [23]. CIFAR-100 is a widely used image classification dataset with 32 $\times$ 32 pixel images categorized into 100 different classes. It comprises 50,000 images for training and 10,000 images for validation. On the other hand, ImageNet is a large-scale dataset designed for image classification across 1,000 categories, featuring 1.2 million training images and 50,000 validation images.

Implementation Details. We utilize a stochastic gradient descent (SGD) optimizer with Nesterov momentum set at 0.9 for all teacher-student model pairs. The training procedure spans a total of 240 epochs, with the learning rate reduced by a factor of 10 at the 150th, 180th, and 210th epochs. For MobileNet [24] and ShuffleNet [4], [25] architectures, the initial learning rate is 0.01, while for other architectures [2], [26], [27], it is set to 0.05. The mini-batch size is maintained

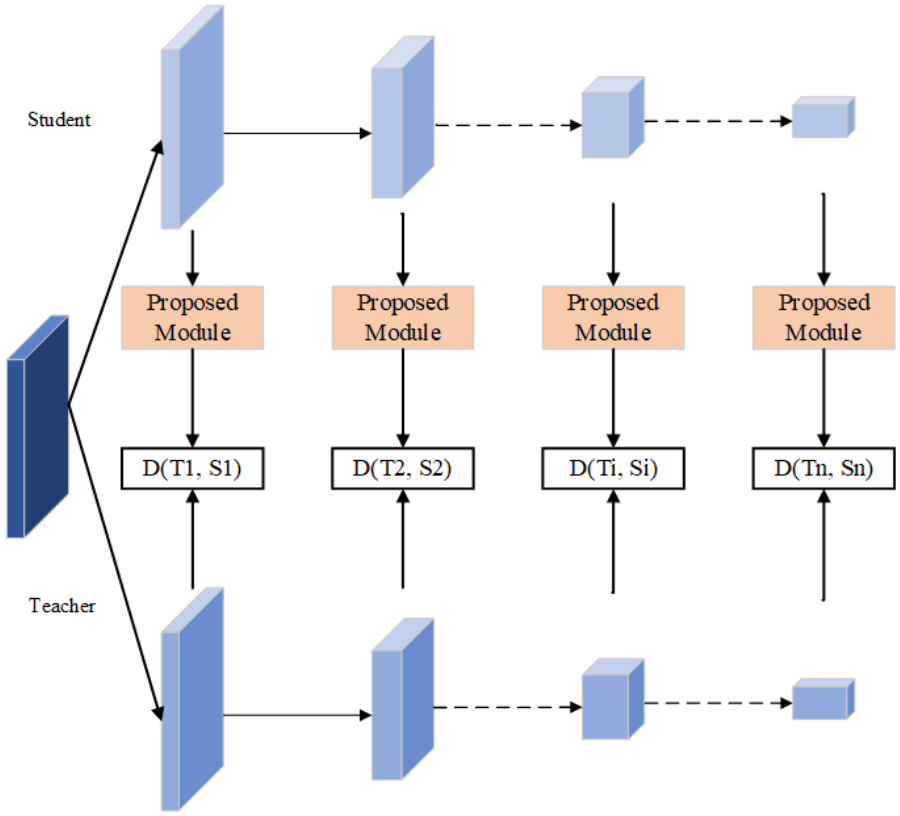

Fig. 2. An overview of our LGMKD. Unlike traditional feature-based multi-teacher knowledge distillation, we do not transfer feature knowledge to the student network but local and global frequency attention knowledge. For simplicity, we only draw one teacher network in the figure.

at 64, and weight decay is configured to 5×10^{-4}. In the knowledge distillation (KD) loss function, the temperature parameter τ is set to 4, and α is set to 10.

To ensure fairness in comparison, the results for the benchmark methods are derived using the authors' released codes and are evaluated under our experimental conditions. The reported results for the CIFAR-100 and ImageNet datasets are averaged over three independent trials.

B. Distillation Performance

In this section, we employ four teacher networks with identical architectures but varied initialization parameters and distillation strategies to guide the student network. We assess the performance improvements of our proposed method relative to baseline approaches using top-1 accuracy as our primary evaluation metric.

The results, as shown in Tables 1, 2 and 3, indicate that our proposed approach surpasses both widely used single-teacher and multi-teacher KD methods across all teacher-student configurations. Notably, even the most basic multi-teacher KD method, AVER, demonstrates superior performance compared to many single-teacher KD approaches, highlighting the benefits of drawing knowledge from multiple teachers. Nevertheless, it is evident that multi-teacher KD methods do not always outperform single-teacher KD methods. Despite the potential for enhanced knowledge diversity with multiple teachers, achieving effective distillation performance requires careful consideration of how well the knowledge from the teacher ensemble aligns with the student network.

Results on teachers with same architectures. Table 1 provides a comparative analysis of top-1 accuracy on CIFAR-100, including results from teacher ensembles employing a majority voting strategy. The data clearly show that LGMKD surpasses all other methods across different architectures. Remarkably, our approach achieves an average accuracy improvement of 0.55% over the second-best method, CA-MKD, and realizes an impressive absolute accuracy gain of 1.05% in the optimal case. In addition, as shown in Table 3, our proposed method also performs well on ImageNet datasets and obtains relative improvement compared with CA-MKD [15]. This demonstrates the effectiveness of our method.

Results on teacher-student pairs have different architecture. The experiment conducted above involved teacher networks in each teacher-student pair with identical architecture. In order to assess the flexibility of our approach, we employed disparate teacher networks across the teacher-student pairs. Specifically, we opted for ResNet8x4, ResNet20x4, and ResNet32x4 as the teacher combination network, while VGG8 served as the student network. The comparative results of top-1 accuracy are presented in Table IV, which further highlights the superiority of our method in relation to other compared methods.

TABLE I
TOP-1 TEST ACCURACY (%) OF VARIOUS MULTI-TEACHER KNOWLEDGE DISTILLATION APPROACHES ON CIFAR-100.

Teacher	WRN40-2	ResNet56	VGG13	ResNet32x4	ResNet32x4
	76.62±0.17	73.19±0.30	74.89±0.18	79.45±0.19	79.45±0.19
Student	ShuffleNetV2	MobileNetV2	MobileNetV2	ShuffleNetV1	VGG-8
	73.07±0.06	65.46±0.10	65.46±0.10	71.58±0.30	70.70±0.26
AVER-KD [9]	76.98±0.19	70.68±0.11	68.89±0.10	75.02±0.25	73.51±0.22
AVER-FitNet [16]	77.29±0.14	70.63±0.23	68.87±0.06	74.75±0.27	73.00±0.16
AEKD [14]	77.02±0.17	70.36±0.19	69.07±0.22	75.11±0.19	73.21±0.04
EBKD [21]	76.75±0.13	69.89±0.14	68.09±0.26	74.95±0.14	73.01±0.01
CA-MKD [15]	77.64±0.19	**71.19±0.28**	69.29±0.09	76.37±0.51	75.02±0.12
LGMKD	**77.76±0.15**	71.06±0.23	**70.01±0.10**	**77.42±0.12**	**75.32±0.10**

TABLE II
TOP-1 TEST ACCURACY (%) OF VARIOUS SINGLE-TEACHER KNOWLEDGE DISTILLATION APPROACHES ON CIFAR-100.

Teacher	ResNet32x4	WRN-40-2	WRN-40-2
	79.31±0.14	76.62±0.26	76.62±0.26
Student	MoblieNetV2	MoblieNetV2	WRN-40-1
	65.64±0.19	65.64±0.19	71.39±0.22
KD [9]	67.57±0.10	69.31±0.20	74.22±0.09
FitNet [16]	67.87±0.08	69.01±0.18	74.28±0.17
AT [28]	67.38±0.21	69.18±0.37	74.83±0.15
VID [29]	67.78±0.13	68.57±0.11	74.37±0.22
CRD [17]	69.04±0.16	70.14±0.06	74.82±0.06
SemCKD [30]	68.86±0.26	69.61±0.05	74.41±0.16
SRRL [31]	68.77±0.06	69.44±0.13	74.60±0.04
DKD [19]	70.07±0.12	70.01±0.13	74.80±0.04
LGMKD	**70.14±0.04**	**70.56±0.17**	**75.56±0.17**

TABLE III
TOP-1 TEST ACCURACY (%) OF VARIOUS MULTI-TEACHER KNOWLEDGE DISTILLATION APPROACHES ON IMAGENET.

Teacher	ResNet32x4	VGG13
	53.21±0.21	49.17±0.11
Student	MobileNetV2	MobileNetV2
	38.46±0.14	38.46±0.14
AVER-KD [9]	39.98±0.19	39.68±0.11
AVER-FitNet [16]	39.29±0.14	39.63±0.23
AEKD [14]	40.02±0.17	40.36±0.19
EBKD [21]	40.75±0.13	40.89±0.14
CA-MKD [15]	40.64±0.19	40.19±0.28
LGMKD	**40.86±0.15**	**41.08±0.23**

TABLE IV
TOP-1 TEST ACCURACY (%) OF VARIOUS MULTI-TEACHER KNOWLEDGE DISTILLATION APPROACHES ON CIFAR-100, WHERE TEACHER HAVE DIFFERENT ARCHITECTURES.

Teacher	ResNet8x4	ResNet20x4	ResNet32x4
	72.69	78.28	79.31
Student		VGG8	
		70.70±0.26	
AVER-KD [9]		74.53±0.17	
AVER-FitNet [16]		74.38±0.23	
AEKD [14]		74.75±0.21	
EBKD [21]		74.27±0.14	
CA-MKD [15]		75.21±0.16	
LGMKD		**75.46±0.12**	

C. Ablation Study

The proposed **LGMKD** method is comprised of three key components: the loss function with the ground-truth label, local frequency, and global frequency. To evaluate the impact of each component on the performance of **LGMKD**, we conducted ablation experiments with four distinct variants: 1) Variant A, which utilizes only $\mathcal{L}_{ce}$, represents standard training from scratch; 2) Variant B, which incorporates only local frequency; and 3) Variant C, which includes both local and global frequency.

The results of these ablation experiments are presented in Table V, showcasing the superior performance of our proposed method compared to all other variants. The observed accuracy improvements for each variant highlight the contribution of each component to the overall effectiveness of our method. Specifically, the comparison between Variant A and the other variants emphasizes the value of knowledge distillation. Additionally, the contrast between Variant A and Variant B underscores the importance of incorporating frequency attention knowledge, while the comparison of Variant B with Variant C demonstrates the benefits of integrating both local and global frequency information.

TABLE V
ABLATION STUDY WITH RESNET32-VGG8 PAIR ON CIFAR-100.

Variants	$\mathcal{L}_{ce}$	local frequency	global frequency	Top-1
A	✔	✘	✘	70.70±0.26
B	✔	✔	✘	73.51±0.22
C	✔	✔	✔	**75.39±0.20**

V. CONCLUSION

In this paper, we introduced an advanced approach for knowledge distillation in multi-teacher settings by integrating local and global frequency attention mechanisms. Our method

addresses the limitations of conventional distillation techniques by enhancing the student model's capability to capture and leverage both fine-grained local details and broader global patterns from multiple teacher models. The integration of these frequency-based attention mechanisms significantly improves the performance of the student model, as demonstrated through comprehensive experiments across various benchmark datasets. These experiments show that our approach not only surpasses traditional distillation methods in terms of accuracy but also enhances the robustness of the student model. By offering a more nuanced and effective means of knowledge transfer, our work advances the state-of-the-art in model compression and transfer learning. The success of our approach highlights the potential of combining local and global attention mechanisms to optimize multi-teacher knowledge distillation, providing valuable insights and promising directions for future research and development in the field.

ACKNOWLEDGMENT

This work are supported by the National Key Research and Development Project of China (Grant No. 2022YFF0605200) and Fujiang Laboratory.

REFERENCES

[1] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

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

[3] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[4] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[7] E. Shelhamer, J. Long, T. Darrell *et al.*, "Fully convolutional networks for semantic segmentation." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

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

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

[11] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers." in *Interspeech*, 2017, pp. 3697–3701.

[12] C. C.-T. Wu, Meng-Chieh and K.-H. Wu, "Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2202–2206.

[13] Y. Liu, W. Zhang, and J. Wang, "Adaptive multi-teacher multi-level knowledge distillation," *Neurocomputing*, vol. 415, pp. 106–113, 2020.

[14] S. Du, S. You, X. Li, J. Wu, F. Wang, C. Qian, and C. Zhang, "Agree to disagree: Adaptive ensemble knowledge distillation in gradient space," *advances in neural information processing systems*, vol. 33, pp. 12 345–12 355, 2020.

[15] H. Zhang, D. Chen, and C. Wang, "Confidence-aware multi-teacher knowledge distillation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4498–4502.

[16] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[17] K. D. Tian, Yonglong and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[18] D. Chen, J.-P. Mei, H. Zhang, C. Wang, Y. Feng, and C. Chen, "Knowledge distillation with the reused teacher classifier," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 933–11 942.

[19] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.

[20] F. Yuan, L. Shou, J. Pei, W. Lin, M. Gong, Y. Fu, and D. Jiang, "Reinforced multi-teacher selection for knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 284–14 291.

[21] K. Kwon, H. Na, H. Lee, and N. S. Kim, "Adaptive knowledge distillation based on entropy," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7409–7413.

[22] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[25] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[28] ——, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[29] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9163–9171.

[30] D. Chen, J.-P. Mei, Y. Zhang, C. Wang, Z. Wang, Y. Feng, and C. Chen, "Cross-layer distillation with semantic calibration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 7028–7036.

[31] J. Yang, B. Martinez, A. Bulat, and G. Tzimiropoulos, "Knowledge distillation via softmax regression representation learning," in *International Conference on Learning Representations*, 2020.