Be an Excellent Student: Review, Preview, and Correction

Qizhi Cao, *Student Member, IEEE*, Kaibing Zhang, Xin He, and Junge Shen

***Abstract*—In the letter, we propose a novel yet effective knowledge distillation scheme which mimics an all-round learning process of an excellent student from the teacher, i.e, knowledge review, knowledge preview, and knowledge correction, to acquire more informative and complementary knowledge. In the newly proposed method, to better leverage comprehensive feature knowledge from the teacher model, we propose Knowledge Review and Knowledge Preview Distillation to amalgamate multi-level features from different intermediate layers in both forward and backward pathways and fully distill them through hierarchical context loss, which greatly improves the student's feature learning efficiency. Moreover, we further present a Response Correction Mechanism to reinforce the prediction of student, which can more fully excavate the student's own knowledge, effectively alleviating the negative influence caused by the knowledge gap between the teacher and the student. We verify the effectiveness of our method with various networks on the CIFAR-100 datasets and the proposed method achieves competitive results compared with other state-of-the-art competitors.**

***Index Terms*—Features fusion, knowledge review and preview distillation, response correction mechanism.**

I. Introduction

KNOWLEDGE distillation (KD) [1], [2], as a promising technology for lightweighting complicated deep networks [3], [4], has great potential applicability in Computer Vision (CV) [5], [6], [7], Natural Language Processing (NLP) [8], Speech Recognition (SR) [9] and so on. The main concept behind knowledge distillation (KD) is to deliver valuable and informative knowledge from a larger, high-performing teacher model to a lightweight student model. Employing the KD strategy enables the simple student model to be readily deployed in resource-constrained environments, such as mobile terminals and embedded devices with limited hardware resources.

Manuscript received 31 July 2023; revised 22 October 2023; accepted 3 November 2023. Date of publication 15 November 2023; date of current version 4 December 2023. This work was supported by the National Natural Science Foundation of China under Grants 61971339 and 61471161. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong Cui. *(Corresponding author: Kaibing Zhang.)*

Qizhi Cao and Xin He are with the School of Electronics and Information, Xi'an Polytechnic University, Xi'an 710048, China (e-mail: caoqizhi@stu.xpu.edu.cn; hexin@stu.xpu.edu.cn).

Kaibing Zhang is with the Shaanxi Key Laboratory of Clothing Intelligence, the School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China (e-mail: zhangkaibing@xpu.edu.cn).

Junge Shen is with the Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: shenjunge@nwpu.edu.cn).

The code will be available at https://github.com/kbzhang0505/RPC.

Digital Object Identifier 10.1109/LSP.2023.3333240

Different knowledge distillation strategies categorize existing approaches into three major groups: Response-based [10], [11], Feature-based [12], [13], and Relation-based [14], [15] knowledge distillation. This letter primarily focuses on the first two methods.

In the process of feature-based KD, previous techniques mainly concentrated on transferring the teacher's knowledge from intermediate layers to the corresponding layers of the student. Representative methods include direct feature match (e.g., Fitnet [16]) and indirect feature match (e.g., Attention Transfer [17]). The kind of these peer-layer feature-based KD approaches strive to capture relevant knowledge across different levels, leading to limited performance improvement. To alter the scenario, Chen et al. [18] proposed a particular knowledge review scheme to explore cross-level feature transfer paths between the teacher and the student. The core idea behind this method is to utilize low-level features learned from the teacher to supervise the learning of high-level features in the student. While this approach demonstrates notable superiority over peer-layer knowledge distillation methods, an innegligible weakness is that the high-level features in the teacher are not fully explored to guide the student's learning, leading to limited performance.

In the process of response-based knowledge distillation, a common approach is to directly employ the KL loss to minimize the output disparity between the teacher and the student (e.g., Tf-KD [19] and Snapshot Distillation [20]). Nevertheless, when there is a substantial gap between the outputs of the teacher and the student, training a satisfactory student model becomes very challenging, primarily due to evident structural differences between the two models. To establish a stable response-based knowledge distillation, Mirzadeh et al. [21] suggested a warm multi-step distillation scheme to progressively train multiple teacher assistants to guide the student, indicating that learning knowledge only from the teacher could result in the student's excessive reliance on the teacher. Therefore, it becomes meaningful to encourage the student to learn from its own experiences and self-discover knowledge.

Taking the above considerations account into, we propose a novel yet effective knowledge distillation scheme which mimics an all-round learning process of an excellent student from a teacher, i.e., knowledge review, knowledge preview, and knowledge correction, which is abbreviated as RPC, to acquire more informative and complementary knowledge for performance improvement. Acknowledging the significance of high-level knowledge for the student, we propose a knowledge preview process to further enhance the proficiency of the student. To emulate this process, we advocate the transfer of the teacher's high-level features to the student's low-level features by employing a novel knowledge preview strategy. The procedure involves a backward feature fusion, moving from deeper to shallower levels, followed

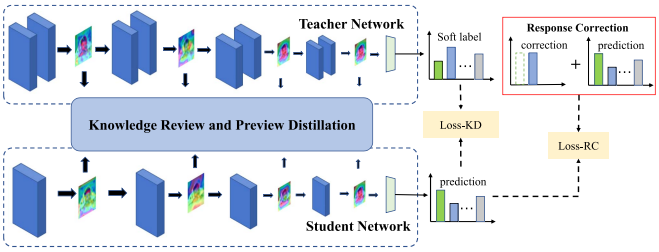

Fig. 1. Overall framework of proposed RPC-based KD. The red box represents the response correction mechanism. Loss-RC and Loss-KD represent response correction loss and original knowledge distillation loss, respectively.

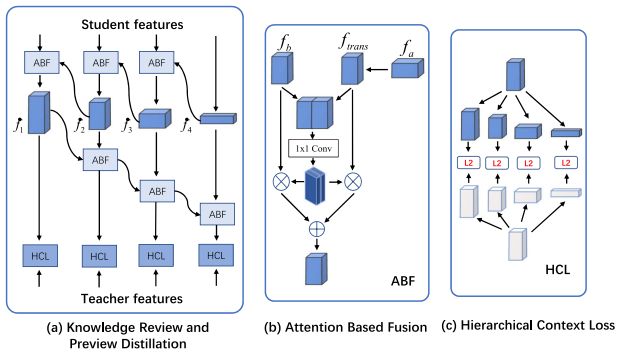

Fig. 2. Details of knowledge review and preview. (a)Architecture of knowledge review and preview distillation. (b) Architecture of ABF. (c) Architecture of HCL.

by a forward fusion of the obtained fused features from shallower to deeper levels. Through this approach, the student's omni-level features can be effectively guided by those of the teacher.

Moreover, to target the philosophy of learning from the student's errors, we further develop a simple yet efficient correction mechanism to reinforce the prediction distribution of the student. To be more specific, we adds the maximum predicted probability value in the prediction distribution to the predicted probability value of the correct class. When the prediction is incorrect, we correct the output of the student by making the student learn from his own mistakes. By contrast, when the prediction is correct, we further strength the confidence of the student by a correction loss, which is beneficial to reducing the negative impact caused by the knowledge gap between the student and teacher. In summary, the contributions of proposed PRC are threefold:

1) We propose a novel knowledge review and knowledge preview strategy to deliver multi-level features in the teacher's intermediate layers to the student, which effectively covers the shortage of insufficient utilization of multi-level features in previous methods.
2) We develop an effective correction mechanism to leverage the student model's own knowledge for self-improvement, which effectively mitigates the negative impact caused by the gap between the teacher and student. Moreover, the proposed response correction mechanism benefits to other self-distillation frameworks.
3) Our method demonstrates excellent performance on both homogenous and heterogeneous scenarios, showing compelling competitiveness compared with other existing KD approaches.

II. Proposed Method

The overall framework of our proposed method is demonstrated in Fig. 1, where two major components, i.e., Knowledge Review and Preview Distillation and the Response Correction Mechanism, are contained. The further elaboration on each component will be explained in the following.

A. Knowledge Review and Preview Distillation

Fig. 2 depicts the overall architecture of proposed Knowledge Review and Preview Distillation. Different from the Review Distillation [18], we elaborately design a bidirectional progressive fusion strategy which considers not only a backward feature fusion from the deeper levels to the shallower levels (called Knowledge Review) but also a forward feature fusion from the shallower levels to the deeper levels (called knowledge preview). In the process of the bidirectional progressive fusion, an attention-based fusion (ABF) scheme is applied to aggregate the feature maps from different layers. Furthermore, a hierarchical context loss (HCL) module is employed to explore more informative details from different layers. In addition, the coordinate attention (CA) [22] is embedded to extract more significant features without obviously increasing computational cost.

1) ABF (Attention Based Fusion): Assuming that the input image is x, the student network is S, and the teacher network is T, we use $F_s = S(x)$ to represent the features of the student $F_s = \{ f_1, f_2, \cdots, f_n \}$. Similarly, we denote F_t as the teacher's features. In the ABF module, $G()$ is the feature transformation function used to convert the feature f_a into the same shape as the feature f_b (seeing Fig. 2(b)), which is expressed as below:

$$f_{trans} = G(f_a), \tag{1}$$

where the f_a and f_b represent the features corresponding to two different layers. f_{trans} denotes the transformed features. The transformed features f_{trans} is first concatenated with f_b, and then a convolution $Con()$ is applied to obtain the attention maps of the two features. After that, the obtained attention maps are performed a dot-wise product operation with their corresponding feature maps and summed together to obtain the fused features, which is formulated as below:

$$f_{fuse} = Con(f_{trans}) \times f_{trans} + Con(f_b) \times f_b, \tag{2}$$

where f_{fuse} represents the fused features and the '$\times$' represents the dot-wise product operation. Finally, the ABF function can be expressed as:

$$F_{ABF}(f_b, f_a) = Con(G(f_a)) \times G(f_a) + Con(f_b) \times f_b. \tag{3}$$

2) Knowledge Review and Preview: In the process of knowledge review for the student's features, we employ a backward progressive fusion through the ABF module to obtain the first-order fused features. The fused features contain all the deeper features than the i-th layer features. This process is similar to that of the review mechanism [18]. Mathematically, it can be expressed as:

$$\dot{f}_i = f_i \circ f_{i+1} \cdots \circ f_n, \tag{4}$$

where $f_i \circ f_{i+1}{=}F_{ABF}(f_i, f_{i+1})$ and the $\dot{f}_i$ denotes the i-th first-order feature. From (4), we can obtain the first-order fused features $\dot{F}_s = \{\dot{f}_1, \dot{f}_2, \ldots, \dot{f}_n\}$.

In the knowledge preview process, we apply a forward progressive fusion pathway to the first-order features to gain the second-order fusion features. Similar to (4), the process of Knowledge Preview can be represented as:

$$\ddot{f}_i = \dot{f}_1 \diamond \dot{f}_2 \cdots \diamond \dot{f}_i, \tag{5}$$

where $\dot{f}_i \diamond \dot{f}_{i+1}{=}F_{ABF}(\dot{f}_{i+1}, \dot{f}_i)$ and $\ddot{f}_i$ denotes the i-th second -order features. The second-order fused features $\ddot{F}_s = \{\ddot{f}_1, \ddot{f}_2, \ldots, \ddot{f}_n\}$ obtained through feature knowledge review and preview contain abundant and omni-level knowledge from each layer, which is conducive to transferring richer cross-layer feature knowledge from the teacher to the student.

3) HCL (Hierarchical Context Loss): Since the second-order fused features from different layers involve multiple scales and different dimensional sizes, directly distilling of the fused features will result in inadequate feature transfer. The problem is solved by an elaborately designed HCL module which utilizes a pyramid pooling to dispose the features across multiple scales and dimensions in both the teacher and the student. It maximally remains the details of initial features from different layers. The process is represented as:

$$L_{RP} = D(P(\ddot{F}_s), P(F_t)), \tag{6}$$

where $P()$ represents the pyramid pooling transformation and $D()$ corresponds to the L_2 distance function.

B. Response Correction Mechanism

In knowledge distillation, the network's output undergoes a transformation through the Softmax function as follows:

$$P_i^\tau = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)}, \tag{7}$$

where z_i is the logits for the i-th sample and τ is a temperature coefficient. The loss of the original knowledge distillation consists of two parts: one is the cross-entropy loss between the classification prediction of the student and the hard label Y and the other is the KL divergence that measures the discrepancy between the classification probability of the teacher and the student. The KD loss is expressed as:

$$L_{KD} = \alpha L_{CE}(X_S, Y) + (1-\alpha)\tau^2 H(X_T^\tau, X_S^\tau), \tag{8}$$

where $X_T^\tau = \text{softmax}(\frac{P_T}{\tau})$, $X_S^\tau = \text{softmax}(\frac{P_S}{\tau})$, $L_{CE}()$ represents the cross-entropy loss function, $H()$ is the KL divergence function, and α is a hyperparameter to adjust the weight between two loss terms in the KD loss.

Assuming that the student's prediction for the i-th sample is $P = \{\, p_1, p_2, ..., p_{cls}, ..., p_n \,\}$. We add the maximal predicted value in the predicted distribution to the prediction value for the correct class, which is expressed as follows:

$$p_{rc} = p_{cls} + Max(P), \tag{9}$$

where the p_{cls} is the probability value of prediction for the correct class. We can get $P_{rc} = \{\, p_1, p_2, ..., p_{rc}, ..., p_n \,\}$ and let the student learn towards the corrected label. When the prediction of the student is incorrect, we correct its prediction, which serves as a learning label for the student. Inversely, we reinforce its prediction by imposing a correction loss to make the student more confident on its own judgments. The correction loss is represented as follows:

$$L_{RC} = H(X_{rc}^\tau, X^\tau), \tag{10}$$

where $X_{rc}^\tau = \text{softmax}(\frac{P_{rc}}{\tau})$ and $X^\tau = \text{softmax}(\frac{P}{\tau})$.

C. Loss Function

The overall loss is comprised of the Knowledge Distillation loss, the Knowledge Review and Preview Distillation loss, and the Response Correction loss as below:

$$L_{ALL} = L_{KD} + \lambda L_{RP} + \mu L_{RC}, \tag{11}$$

where λ and μ are the hyperparameters to balance the loss functions.

III. Experimental Results and Analysis

A. Implementation Details

In our experiments, all the models are trained for 240 epochs with a batch size of 128 on the CIFAR-100 datasets [23]. The temperature τ is set to 4 and the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 is used to optimize the model. The learning rate is initialized to 0.02 for ShuffleNet and 0.1 for other models. The weight decay factor of 5.0×10^{-4} is used and the learning rate is multiplied by 0.1 at the 150th, the 180th and the 210th epoches, respectively.

B. Comparisons With State-of-the-Art Methods

To verify the effectiveness of our method, we employ on various networks, including WideResNet [24], ResNet [25], Vgg [26], ShuffleNetV1 [27], ShuffleNetV2 [28], and MobileNetV2 [29] and compare with the state-of-the-art methods, including vanilla KD [30], Attention Transfer [31], FT [32], CRD [33], ReviewKD [18], Spot-adaptive Distillation [34], ICKD-C [35], and Decouple KD [36].

Comparisons of Experimental Results: Table I tabulates the Top-1 accuracy of various methods on Peer-Architecture and Cross-Architecture networks. Based on the presented results of Peer-Architecture portion, our proposed method gains the best performance apart from the VggNet. In terms of the contrastive results obtained from Cross-Architecture models, although our method may not achieve the top performance on certain networks, it still exhibits strong competitiveness. Furthermore, we visualize the feature maps and the correlation matrix differences between the logits of teacher network(ResNet32×4) and those of the student network(RseNet8×4) corresponding to different schemes in Fig. 3. As demonstrated, in contrast to the feature maps obtained by the other two schemes, our RPC can more obviously enhance the distinction between the background and the the foreground. Furthermore, we also find that the logits difference of our RPC is much smaller than that of other two methods. This is due to that the proposed RPC is capable of providing more informative knowledge to the student networks, which benefits to narrowing the logits difference between the teacher and student networks.

Robustness Evaluation: In the domain of KD, the robustness against the architecture gap between the teacher and the student

TABLE I
TOP-1 ACCURACY RATE (%) OF VARIOUS METHODS ON BOTH PEER-ARCHITECTURE AND CROSS-ARCHITECTURE NETWORKS

Architecture		Peer-Architecture							Cross-Architecture				
Method	Year	T: ResNet56 S: ResNet20	ResNet110 ResNet20	ResNet110 ResNet32	ResNet32x4 ResNet8x4	WRN40-2 WRN16-2	WRN40-2 WRN40-1	Vgg13 Vgg8	ResNet32x4 ShuffleNetV1	ResNet32x4 ShuffleNetV2	WRN40-2 ShuffleNetV1	Vgg13 MobileNetV2	ResNet50 MobileNetV2
Teacher	-	73.08	74.42	74.42	79.42	76.48	76.48	74.65	79.42	79.42	76.48	74.65	79.34
Student	-	69.14	69.14	71.14	72.32	73.42	71.98	70.63	70.47	71.75	70.47	64.62	64.62
Vanilla KD [30]	2015	71.08	70.84	73.08	73.25	74.77	73.54	72.73	74.11	74.36	74.69	67.41	67.35
AT [31]	2017	71.57	70.96	73.11	73.36	75.23	72.77	73.07	73.62	73.57	73.80	60.80	59.72
FT [32]	2019	71.52	70.44	73.37	73.64	75.10	73.04	73.14	72.58	72.73	72.21	62.03	61.84
CRD [33]	2020	71.62	71.50	73.57	75.44	75.51	74.14	74.11	75.14	75.60	75.67	69.71	69.11
Review KD [18]	2021	71.89	71.72	73.89	75.53	76.01	75.09	74.57	77.04	**77.50**	77.14	**69.91**	69.82
CRD + SAKD [34]	2022	71.79	71.53	73.42	75.60	75.85	74.67	74.29	75.33	75.82	75.92	69.88	69.75
ICKD-C [35]	2022	71.75	71.72	73.72	74.80	75.58	74.32	73.88	74.69	74.56	74.63	67.89	67.80
Decouple KD [36]	2022	71.94	71.49	74.11	76.32	76.24	74.81	**74.71**	76.39	76.94	76.49	69.75	**70.32**
Review-Preview	2023	72.21	72.47	73.99	75.74	76.05	75.11	74.59	77.09	77.01	77.23	69.42	69.17
Ours	2023	**72.54**	**72.72**	**74.21**	**76.91**	**76.24**	**75.19**	74.67	**77.15**	77.13	**77.43**	69.59	69.19

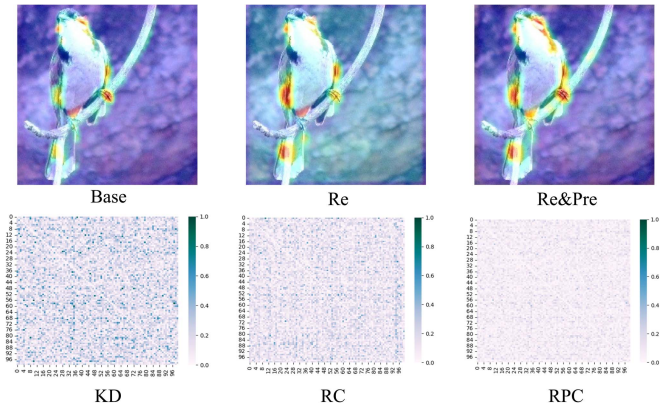

Fig. 3. Visualization comparisons of feature maps and correlation matrices for different schemes. The "Base," "Re," and "Re & Pre" represent the original features, features from knowledge review, and features from knowledge review and preview, respectively. The "KD," "RC," and "RPC" denote the correlation matrices between the logits of the teacher network and the student network of original knowledge distillation, response correction, and the proposed method, respectively.

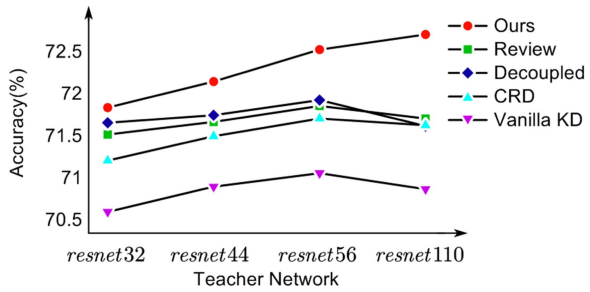

Fig. 4. Robustness evaluation among different KD strategies.

is another indictor to evaluate the effectiveness of applied KD scheme. To this end, we prefix resnet20 as the student and use resnet32, resnet44, resnet56, and resnet 110 respectively as the teacher to compare the performance of Review KD, Decoupled KD, CRD, Vanilla KD, and our method. The contrastive results are displayed in Fig. 4.

As shown, as the difference between the teacher and the student increases, our method steadily improves the performance varying with the complexity of the teacher. Actually, to reduce the gap between the teacher and the student, the knowledge source other than the teacher that are closer to the student's own knowledge should be introduced. The combination of self-distillation and single-teacher distillation may be a promising research direction to alleviate the negative impact caused by the "gap", just as our proposed response correction mechanism can seamlessly integrate with self-distillation.

TABLE II
ABLATION STUDY ON INTEGRATING DIFFERENT KD SCHEMES

T: ResNet110 (74.42%)			S: ResNet20(69.14%)		
Method	CA	Review	Preview	Correction	Accuracy(%)
Vanilla KD			-		70.92
Ours				√	71.58
	√			√	72.02
		√			71.63
	√	√			72.11
	√	√	√		72.47
	√	√	√	√	**72.72**

C. Ablation Study

To verify the contribution of each component, we conduct a group of ablation experiments by integrating different KD strategies and the CA mechanism with different combination manners. In the ablation study, the ResNet110 is used as the teacher and the ResNet20 us used the student. Table II presents the evaluating results, where Review, Preview, and Correction denote the knowledge review distillation, knowledge preview distillation, and response correction mechanism, respectively. In terms of the compared results, the joint RPC strategy yields the best result among the compared schemes.

IV. CONCLUSION

In this letter, we have presented a novel and effective knowledge distillation method inspired by the comprehensive learning process that an outstanding student would undertake. Our approach involves knowledge preview, knowledge review, and knowledge correction. The knowledge review and preview distillation facilitates the transmission of multi-level features from various intermediate layers, enabling the enhancement of information flows in both forward and backward pathways from the teacher to the student. While the knowledge correction mechanism leverages the student model's own knowledge to address the negative impact of the knowledge gap between the student and the teacher. Extensive experiments on both Peer-Architecture and Cross-Architecture networks have indicated impressive superiority of our newly proposed method over other state-of-the-art competitors on CIFAR-100 dataset.

REFERENCES

[1] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 8, pp. 1789–1819, Jun. 2021.

[2] C.-Y. Low, A. B.-J. Teoh, and J. Park, "MIND-Net: A deep mutual information distillation network for realistic low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 28, pp. 354–358, 2021.

[3] G. Tian, J. Chen, X. Zeng, and Y. Liu, "Pruning by training: A novel deep neural network compression framework for image processing," *IEEE Signal Process. Lett.*, vol. 28, pp. 344–348, 2021.

[4] J. Wu, Y. Wang, and X. Zhang, "Lightweight asymmetric convolutional distillation network for single image super-resolution," *IEEE Signal Process. Lett.*, vol. 30, pp. 733–737, 2023.

[5] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D convolution for video super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 1500–1504, 2020.

[6] Y. Shi, H. Zhong, Z. Yang, X. Yang, and L. Lin, "DDet: Dual-path dynamic enhancement network for real-world image super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 481–485, 2020.

[7] W. Zhou, F. Sun, Q. Jiang, R. Cong, and J.-N. Hwang, "WaveNet: Wavelet network with knowledge distillation for RGB-T salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 3027–3039, 2023.

[8] N. Nakashole and R. Flauger, "Knowledge distillation for bilingual dictionary induction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2497–2506.

[9] S. Tian et al., "Knowledge distillation for CTC-based speech recognition via consistent acoustic representation learning," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 2633–2637.

[10] C. Yang, L. Xie, C. Su, and A. L. Yuille, "Snapshot distillation: Teacher-student optimization in one generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2859–2868.

[11] T. Wen, S. Lai, and X. Qian, "Preparing lessons: Improve knowledge distillation with better supervision," *Neurocomputing*, vol. 454, pp. 25–33, Sep. 2021.

[12] H. Qu, X. Su, Y. Wang, X. Hao, and G. Gao, "Noise-separated adaptive feature distillation for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 30, pp. 763–767, 2023.

[13] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2018, pp. 268–284.

[14] C. Lassance, M. Bontonou, G. B. Hacene, V. Gripon, J. Tang, and A. Ortega, "Deep geometric knowledge distillation with graphs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 8484–8488.

[15] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3967–3976.

[16] R. Adriana, B. Nicolas, K. S. Ebrahimi, C. Antoine, G. Carlo, and B. Yoshua, "FitNets: Hints for thin deep nets," *Proc. 3rd Int. Conf. Learn. Representations*, vol. 2, 2015, p. 3.

[17] M. Ji, B. Heo, and S. Park, "Show, attend and distill: Knowledge distillation via attention-based feature matching," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 7945–7952.

[18] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5008–5017.

[19] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3903–3911.

[20] C. Yang, L. Xie, C. Su, and A. L. Yuille, "Snapshot distillation: Teacher-student optimization in one generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2859–2868.

[21] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5191–5198.

[22] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13708–13717.

[23] A. Krizhevsky et al., "Learning multiple layers of features from tiny images," in *Handbook of Systemic Autoimmune Diseases*, Amsterdam, The Netherlands: Elsevier, vol. 1, no. 4, pp. 54–57, Apr. 2009.

[24] E. Zerhouni, D. Lányi, M. Viana, and M. Gabrani, "Wide residual networks for mitosis detection," in *Proc. IEEE 14th Int. Symp. Biomed. Imag.*, 2017, pp. 924–928.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 7–9.

[27] M. G. Hluchyj and M. J. Karol, "ShuffleNet: An application of generalized perfect shuffles to multihop lightwave networks," *J. Lightw. Technol.*, vol. 9, no. 10, pp. 1386–1397, Oct. 1991.

[28] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[30] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Comput. Sci.*, vol. 14, no. 7, pp. 38–39, Mar. 2015.

[31] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 24–26.

[32] J. Kim, S. Park, and N. Kwak, "Paraphrasing complex network: Network compression via factor transfer," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2765–2774.

[33] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *Proc. 8th Int. Conf. Learn. Representations*, 2020, pp. 26–30.

[34] J. Song, Y. Chen, J. Ye, and M. Song, "Spot-adaptive knowledge distillation," *IEEE Trans. Image Process.*, vol. 31, pp. 3359–3370, 2022.

[35] L. Liu et al., "Exploring inter-channel correlation for diversity-preserved knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8251–8260.

[36] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11943–11952.