PRIME KNOWLEDGE WITH LOCAL PATTERN CONSISTENCY FOR KNOWLEDGE DISTILLATION

Qiankun Tang[1]* *Xiaogang Xu*[1, 2] *Jun Wang*[1]

[1]Institute of Artificial Intelligence, Zhejiang Lab
[2] School of Computer and Information Engineering, Zhejiang Gongshang University

ABSTRACT

Intermediate feature maps of teacher model can produce enriched knowledge to improve the performance of student model. Existing works mainly focus on formulating beneficial knowledge for transferring, but ignore the contribution discrepancy of the knowledge to promote performance. To tackle this issue, we propose a simple Importance-based Knowledge Reweighting mechanism, which dynamically measure the importance of knowledge spatially and channel-wisely for teacher-student pairs. This reweighting scheme enables the student model to focus more on the prime knowledge. Furthermore, a local pattern consistency loss based on Structural Similarity Index Measure (SSIM) is presented to narrow the local pattern discrepancy between teacher and student features. Extensive experiments on CIFAR-100 with various combinations of network architectures for teacher and student well demonstrate the effectiveness and superiority of our proposed approach.

Index Terms— Prime knowledge, local pattern consistency, knowledge distillation

1. INTRODUCTION

Model compression techniques gradually become popular research topics in recent years. Typical attempts include network pruning [1], quantization [2], lightweight architecture design [3, 4, 5, 6] and knowledge distillation (KD) [7]. Among them, KD serves as an effective recipe that takes the predictions of a deep teacher model as soft targets, which act as a great regularization [8], to guide the learning of a small student model for improving the generalization ability of student model. The intermediate feature maps of teacher model, which contain richer information, are also formulated as various types of knowledge for distillation [9, 10, 11, 12].

The success of feature-based distillation methods suggests that the intermediate representations are significant value for student. Nevertheless, existing works have two drawbacks. First, most methods mainly focus on studying appropriate representation forms for distillation and treat all the knowledge equally during transferring, as shown in Fig. 1(a). Since the student and teacher model own different abstraction capability, the position with large semantic between teacher-student layer pairs would dominate the transferring procedure. Secondly, only pixel-wise distillation loss is adopted in most works, which assume each pixel of feature maps is independent. In fact, the information contained in each channel of feature maps generally exhibit particular pattern of input [13]. Independent knowledge distillation may cause pattern mismatch between the feature of teacher and student.

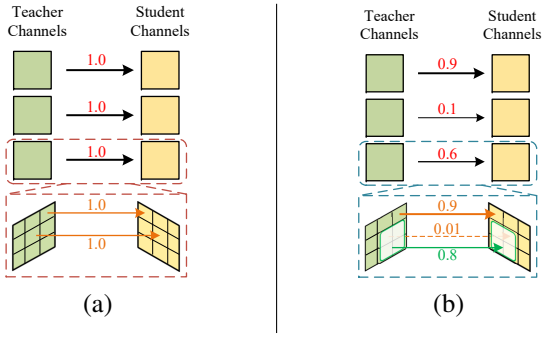

Fig. 1: (a) Existing works treat the knowledge equally and independently during distillation. (b) We suggest to dynamically emphasize on the prime knowledge, and enhance the similarity of local pattern between teacher-student pairs.

To mitigate these issues, we first propose an Importance-based Knowledge Reweighting (IKR) mechanism, as shown in Fig. 1(b). It dynamically models the importance of knowledge by calculating the similarity of spatial and channel-wise features between teacher-student layer pairs. We term the knowledge with large value of importance as prime knowledge, which plays a crucial role in achieving high performance, and emphasize more on it during learning. In addition, inspired by [14], we present a local pattern consistency loss calculated by SSIM to measure the local pattern discrepancy between the student feature map and corresponding teacher feature map. We minimize this loss to encourage the local regions to learn synchronously. Experimental results

*Corresponding author: Qiankun Tang, tangqiankun@zhejianglab.com.
The work is supported in part by the Research Program of Zhejiang Lab (2019KD0AC02), Soybean Intelligent Computing Breeding and Application (2021PE0AC04), Intelligent Technology and Platform Development for Rice Breeding (2021PE0AC05).

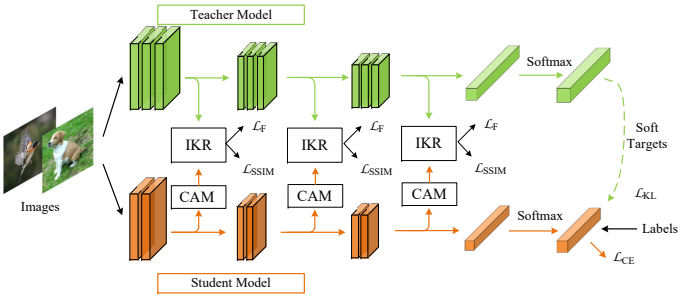

Fig. 2: Overview of the proposed distillation method. CAM means the channel adaption module ϕ.

on classification task validate the effectiveness of our method by adopting various teacher-student combinations based on several popular network architectures. We achieve more attractive performance than existing methods.

2. RELATED WORK

Hinton *et al.* validate that the knowledge from a deep teacher network can well guide the training of a small student model by minimizing the Kullback-Leibler (KL) divergence of their predicted category distributions [7]. Some works observe that the intermediate feature representations in teacher model can provide enriched knowledge to student model. Therefore, they either directly propagate the representations to corresponding student layers as *hints* [9] or enforce the student to align the attentive regions of teacher model [10]. Variational information distillation [11] encourages student to incorporate more certain attributes of input from teacher. Some methods expect the student to preserve the pairwise similarities [15] or structural information [16, 17, 18]. Contrastive learning is also adopted to capture higher-order dependencies in the representation space [12]. In order to narrow the semantic gaps caused by handcrafted layer association, recent works propose automatic layer assignment via a meta-network [19] or exhaustive attention-based layer matching [20].

However, most of the above approaches only consider the particular representation forms but ignore the contribution discrepancy of transferred knowledge. [19] proposes to learn the channel weights, which decide the amount of knowledge transferred to student channel, by a meta-network given only the feature-map of student model. Adaptive sample weighting [21] is introduced to emphasize samples with least uncertainty. In contrast to these solutions, we dynamically formulate the importance of knowledge containing in spatial and channels of teacher model by incorporating information from teacher-student layer pairs and our method is parameter-free.

3. APPROACH

Given a pre-trained powerful teacher model $\boldsymbol{T}$ and a lightweight student model $\boldsymbol{S}$, we denote the output of intermediate convolutional layers as $\boldsymbol{F}_t \in \mathbb{R}^{b\times c_t\times h_t\times w_t}$ and $\boldsymbol{F}_s \in \mathbb{R}^{b\times c_s\times h_s\times w_s}$ (where b is mini-batch size, c means channel number, h and w are spatial dimensions), respectively, as illustrated in Fig. 2.

The feature-based knowledge distillation is formulated as:

$$\mathcal{L}_{\mathrm{F}} = \frac{1}{N}\sum_{b=1}^{N} dist(TF(\boldsymbol{F}_t)[b], TF(\phi(\boldsymbol{F}_s))[b]) \tag{1}$$

where $dist$ means distance function. TF is the feature transformation function, which can be attention [10] or similarity calculation [15]. ϕ is channel adaption module which changes teacher-student layer pairs to the same channel number.

3.1. Prime Knowledge

From the Equ. (1), we can see that the knowledge from positions across spatial and channel contributes equally, without considering the semantic difference among them. Treating all knowledge equally would make the distilled model suffer from negative contributions from low-quality or noisy representations. Thus, we focus on the knowledge which plays a crucial role in promoting the performance of student model.

We propose a simple yet effective weighting scheme, named Importance-based Knowledge Reweighting (IKR), to capture the relative importance of knowledge to the student model. IKR consists of spatial and channel knowledge reweighting. The spatial importance is formulated as:

$$\widetilde{\boldsymbol{F}_s} = \mathcal{R}(\phi(\boldsymbol{F}_s)),\ \widetilde{\boldsymbol{F}_t} = \mathcal{R}(\boldsymbol{F}_t) \tag{2}$$

$$\boldsymbol{\alpha} = \frac{\widetilde{\boldsymbol{F}_t}^T \widetilde{\boldsymbol{F}_s}}{||\widetilde{\boldsymbol{F}_t}^T||_2 ||\widetilde{\boldsymbol{F}_s}||_2} \tag{3}$$

$$\boldsymbol{\alpha}_{\mathrm{sp}} = \frac{1}{2}(\mathsf{diag}(\boldsymbol{\alpha}) + 1) \in [0, 1] \tag{4}$$

where $\mathcal{R}(\cdot)$: $\mathbb{R}^{b\times c\times h\times w} \rightarrow \mathbb{R}^{b\times c\times hw}$ is a reshaping function, and $\boldsymbol{\alpha}$ is a matrix in shape of $b\times hw\times hw$. Finally, we obtain the spatial importance $\boldsymbol{\alpha}_{\mathrm{sp}}$ with shape of $b\times hw$. Since the activations at a position across channels describe the abstract representation of a part of input, the element of $\boldsymbol{\alpha}_{\mathrm{sp}}$ means the similarity of part abstract between teacher and student.

We further formulate the channel importance as:

$$\widehat{\boldsymbol{\alpha}} = \frac{\widetilde{\boldsymbol{F}_t}\widetilde{\boldsymbol{F}_s}^T}{||\widetilde{\boldsymbol{F}_t}||_2 ||\widetilde{\boldsymbol{F}_s}^T||_2} \tag{5}$$

$$\boldsymbol{\alpha}_{\mathrm{ch}} = \frac{1}{2}(\mathsf{diag}(\widehat{\boldsymbol{\alpha}}) + 1) \in [0, 1] \tag{6}$$

where $\widehat{\boldsymbol{\alpha}}$ is in shape of $b\times c\times c$, the channel importance $\boldsymbol{\alpha}_{\mathrm{ch}}$ is in shape of $b\times c$. The features contained in each channel usually exhibit different activation patterns [13], so each entry of $\boldsymbol{\alpha}_{\mathrm{ch}}$ denotes the pattern similarity for teacher-student channel pairs. At last, we re-formulate the loss in Equ.1 by IKR as:

$$\begin{aligned}\mathcal{L}_{\mathrm{F}} &= \frac{1}{N}\sum_{b=1}^{N}\frac{1}{c_t}\sum_{c=1}^{c_t}\boldsymbol{\alpha}^b_{\mathrm{ch}}[c]\boldsymbol{A}[c] \\ \boldsymbol{A} &= \frac{1}{hw}\sum_{i=1}^{hw}\boldsymbol{\alpha}^b_{\mathrm{sp}}[i]dist(TF(\boldsymbol{F}_t)[i], TF(\phi(\boldsymbol{F}_s))[i])\end{aligned} \tag{7}$$

where $\boldsymbol{A} \in \mathbb{R}^{b\times c_t}$, $dist$ is the squared Euclidean distance in our experiments. Based on the above formulation, the reweighted distillation loss has the following merits: (1) **Heterogeneous**: The calculation of $\boldsymbol{\alpha}_{\text{sp/ch}}$ incorporates the features from teacher-student layer pairs, which greatly reveal the learning state of student model. The entry of importance matrix $\boldsymbol{\alpha}_{\text{sp/ch}}$ modulates the amount of knowledge transferred to guide the student model. A higher value means supplying more knowledge from that position or channel. (2) **Dynamic**: The features of student model change during training, which adaptively update each entry of both $\boldsymbol{\alpha}_{\text{sp}}$ and $\boldsymbol{\alpha}_{\text{ch}}$ in iterations. (3) **Diversity**: The dynamical weight entries allow each student layer to learn from the most related spatial positions and channels of its associated teacher layer. Meanwhile, each layer of student model owns its individual reweighting matrices of $\boldsymbol{\alpha}_{\text{sp}}$ and $\boldsymbol{\alpha}_{\text{ch}}$.

3.2. Local Pattern Consistency

The above formulation considers only pixel-wise knowledge transfer. Beyond pixel-wise loss function, we propose a loss to enforce the local pattern consistency. We utilize the SSIM, which is widely used for image quality assessment [22, 14, 23], to measure the local pattern discrepancy centered at each pixel on the feature map of student model and the corresponding region on the teacher feature map. SSIM calculates similarity of two images by three local statistics, *i.e. mean, variance* and *covariance*. The value of SSIM is in the range of -1 to 1 and equals to 1 when the two images are identical.

Following [14], we employ a normalized Gaussian kernel with a size of 3×3 and a standard deviation of 1.0 to estimate the local statistics. The normalized kernel is denoted as $\boldsymbol{\omega} = \{\boldsymbol{\omega}(o)|o \in \boldsymbol{O}, \boldsymbol{O} = (-1,-1),\ \cdots,\ (1,1)\}$, where o is an offset from the center. For each location p on the student feature map $\boldsymbol{F}_s$ and the corresponding teacher feature map $\boldsymbol{F}_t$, the local statistics are computed by:

$$\mu_s(p) = \sum_{o\in\boldsymbol{O}} \boldsymbol{\omega}(o)\cdot \boldsymbol{F}_s(p+o) \tag{8}$$

$$\sigma_s^2(p) = \sum_{o\in\boldsymbol{O}} \boldsymbol{\omega}(o)\cdot [\boldsymbol{F}_s(p+o) - \mu_s(p)]^2 \tag{9}$$

$$\sigma_{st}^2(p) = \sum_{o\in\boldsymbol{O}} \boldsymbol{\omega}(o)\cdot [\boldsymbol{F}_s(p+o) - \mu_s(p)]\cdot [\boldsymbol{F}_t(p+o) - \mu_t(p)] \tag{10}$$

where μ_s and σ_s^2 are the local *mean* and *variance* estimation of $\boldsymbol{F}_s$, μ_t and σ_t^2 of $\boldsymbol{F}_t$ are estimated by the same formulation. $\sigma_{st}^2(p)$ is the local *covariance* estimation. The SSIM between $\boldsymbol{F}_s$ and $\boldsymbol{F}_t$ is calculated as [22, 14]:

$$\text{SSIM}(p) = \frac{(2\mu_s\mu_t + c_1)(2\sigma_{st}^2 + c_2)}{(\mu_s^2 + \mu_t^2 + c_1)(\sigma_s^2 + \sigma_t^2 + c_2)} \tag{11}$$

where c_1 and c_2 are small constant to avoid division by zero [14]. The above estimations are easily implemented with a convolutional layer by setting the weights to $\boldsymbol{\omega}$ and without updating it during back-propagation.

Table 1: Top-1 accuracy of similar network architectures. **Bold** and underline denote the best and second best results, respectively.

Teacher Student	ResNet56 ResNet20	ResNet32×4 ResNet8×4	VGG13 VGG8
Student	69.06	72.51	70.36
KD	70.66	74.40	73.31
FitNet	71.60	74.31	73.54
AT	71.67	74.96	73.52
SP	71.44	74.24	73.44
VID	71.67	74.82	73.66
RKD	71.48	74.47	73.72
HKD	71.20	74.87	73.10
CRD*	71.60	75.80	74.06
SemCKD*	<u>71.91</u>	<u>76.23</u>	<u>74.46</u>
Ours	**72.34**	**76.71**	**74.72**
Teacher	72.41	79.42	74.64

Finally, we define the local pattern consistency loss by considering the IKR as:

$$\mathcal{L}_{\text{SSIM}} = 1 - \frac{1}{N}\sum_{b=1}^{N}\frac{1}{c_t}\sum_{c=1}^{c_t}\boldsymbol{\alpha}_{\text{ch}}^b[c]\cdot\frac{1}{hw}\sum_{i=1}^{hw}\boldsymbol{\alpha}_{\text{sp}}^b[i]\text{SSIM}(i) \tag{12}$$

The total objective of our proposed method is:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{KL}} + \gamma\mathcal{L}_{\text{F}} + \beta\mathcal{L}_{\text{SSIM}} \tag{13}$$

where γ and β are the balancing hyper-parameters.

4. EXPERIMENTS

In this section, we evaluate the effectiveness of the proposed approach for feature-based knowledge distillation on CIFAR-100. We select several popular network architectures, including ResNet [24], VGGNet [25], MobileNet [6], ShuffleNet [4, 5], and construct 2 groups of teacher-student pairs based on whether they have similar network architectures, following [12]. We compare the experimental results with existing works, such as KD [7], FitNet [9], AT [10], SP [15], VID [11], RKD [17], HKD [26], CRD [12], SemCKD [20].

4.1. Evaluation on CIFAR-100

The training details follow the settings in [12]. The channel adaption module ϕ is implemented simply by a stack of three convolutional layers with kernel sizes of 1×1, 3×3 and 1×1 and initialized randomly. We report the average accuracy (Table 1 and Table 2) over 3 runs based on the open-source database [12] or the author-provided codes (marked with "*").

From the reported results, we have the following observations. First, our method consistently surpasses the existing feature-based distillation methods (such as FitNet, AT, VID, SP, HKD) by a large margin. They formalize the knowledge as *hint*, attention or information flow with auxiliary supervision, while our method mainly highlights the prime knowledge in a compact reweighting scheme. Second, comparing

Table 2: Top-1 accuracy of different network architectures. "ResNet" means ResNet32×4.

Teacher Student	VGG13 MobileV2	ResNet ShuffleV1	ResNet ShuffleV2	VGG13 ShuffleV2	ResNet VGG13	ResNet VGG8
Student	64.60	70.50	71.82	71.82	74.64	70.36
KD	67.37	74.73	75.81	75.84	77.23	72.82
FitNet	68.45	74.91	75.81	75.50	77.10	72.89
AT	68.43	74.90	75.80	75.42	77.26	71.93
SP	66.89	73.80	75.95	75.44	77.81	73.10
VID	66.91	74.19	75.98	75.22	77.45	73.22
RKD	68.50	74.49	75.74	75.78	77.54	73.03
HKD	68.12	72.63	76.54	76.28	76.94	72.74
CRD*	68.49	75.54	76.12	75.89	78.12	73.65
SemCKD*	68.49	76.04	77.64	76.27	**79.13**	**75.10**
Ours	**69.78**	**76.86**	**77.91**	**76.62**	78.93	74.88
Teacher	74.64	79.42	79.42	74.64	79.42	79.42

Table 3: Effectiveness of the proposed components.

KD	IKR-sp	IKR-ch	SSIM	ResNet8×4	ShuffleV2
✓				74.40	75.81
✓	✓			75.64	76.51
✓		✓		75.89	76.91
✓			✓	74.60	75.85
✓	✓	✓		76.49	77.21
✓	✓	✓	✓	**76.71**	**77.91**

to the feature-embedding methods (RKD, CRD) which distill knowledge only from the penultimate layer of the teacher model, our method can learn more general and discriminative knowledge especially the prime knowledge from intermediate layers and achieve better accuracy. Third, we obtain slightly worse performance than SemCKD on two network combinations (the last two columns in Table 2). We find this is due to the handcrafted layer matching strategy we adopted for distillation, which cannot associate appropriate teacher layer for student, while the exhaustive matching mechanism [20] is helpful to search the most semantic-related teacher layers. However, an efficient and effective association strategy deserves better research in the future.

4.2. Ablation Study

To verify the benefit of each component, we conduct a series of ablation studies on CIFAR-100 with ResNet32×4 as teacher model and ResNet8×4, ShuffleV2 as student models.

Component analysis. Table 3 reports the effects of each component in our method. We take the vanilla KD as baseline. By progressively applying each component of IKR to baseline, the results are stably improved. It indicates that the IKR is helpful to distill prime knowledge for student model. The introduced local pattern consistency loss based on SSIM can further improve the performance, though its individual only achieves negligible improvement.

Impact of hyper-parameters. In Equ. (13), we adopt γ and β to balance the loss terms. We tune their values to evaluate the impact on final performance, as shown in Table 4. We find that with the increasing of γ, the performance is gradually improved and reaches the peak at γ=20. And β=1 achieves the best results. We adopt these two values in all our experiments.

Table 4: Impact of γ and β.

γ	β	ResNet8×4
10	1	76.03
15	1	76.14
20	1	**76.71**
25	1	76.28
20	2	76.55
20	5	76.18

Table 5: Top-1 accuracy of different reweighting strategies.

	ResNet8×4	ShuffleV2
Baseline	74.40	75.81
Equal	75.22	76.46
Hard_$\boldsymbol{\alpha}'$	74.75	75.75
Hard_$dist$	75.03	76.15
Prime	**76.71**	**77.91**

4.3. Knowledge Reweighting Analysis

Hard example mining is a popular topic in object detection [27], where the "hard" samples are assigned larger weights as they are more important to train the model. Therefore, we want to examine whether hard example mining strategy is applicable to knowledge distillation.

We define the "hard knowledge" for knowledge distillation in two ways: (a) similarity-based. We term the knowledge with lower importance value as "hard knowledge". We reformulate the spatial and channel importance as: $\boldsymbol{\alpha}'_{\text{sp}} = 1 - \boldsymbol{\alpha}_{\text{sp}}$, $\boldsymbol{\alpha}'_{\text{ch}} = 1 - \boldsymbol{\alpha}_{\text{ch}}$ to highlight these "hard knowledge". (b) distance-based. We term the knowledge contained in positions with largest $dist$ values as hard and formulate their weights as: $\boldsymbol{\alpha} = e^d$, where $d = \mathsf{ReLU6}(dist(\cdot,\cdot)/\tau)$, τ is relaxation parameter and set to 4. We utilize ReLU6 to avoid the gradient explosion. We also evaluate the equal weight scheme, as shown in Table 5.

The similarity-based method puts more attention to the most dissimilar knowledge, which achieves negligible or even worse performance. This further indicates that the knowledge with larger $\boldsymbol{\alpha}_{\text{sp/ch}}$ are valuable for distillation. The distance-based method following the loss definition (Equ. (7)) well reveals the hardness of knowledge. However, it only gets limited improvement, while we need carefully define the weight and tune extra parameters. Assigning equal weights achieve much better performance than the above two methods.

5. CONCLUSION

Knowledge distillation has been validated as an effective approach for model compression and acceleration. In this work, we intend to study on what is the most important knowledge for training the student model. We present an Importance-based Knowledge Reweighting mechanism to dynamically emphasize on the knowledge which has larger importance value computed spatially and channel-wise between teacher-student layer pairs. A local pattern consistency loss is further proposed to enforce the local region similarity between feature maps. Extensive experiments on classification task manifest the effectiveness of our method.

6. REFERENCES

[1] Tuanhui Li, Baoyuan Wu, Yujiu Yang, Yanbo Fan, Yong Zhang, and Wei Liu, "Compressing convolutional neural networks via factorized convolutional filters," in *CVPR*, 2019, pp. 3977–3986.

[2] Song Han, Huizi Mao, and William J Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[3] Andrew G Howard, Menglong Zhu, et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[4] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *CVPR*, 2018, pp. 6848–6856.

[5] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *ECCV*, 2018, pp. 116–131.

[6] Mark Sandler, Andrew Howard, et al., "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.

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

[8] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng, "Revisiting knowledge distillation via label smoothing regularization," in *CVPR*, 2020, pp. 3903–3911.

[9] Romero Adriana, Ballas Nicolas, et al., "Fitnets: Hints for thin deep nets," in *ICLR*, 2015, pp. 1–13.

[10] Sergey Zagoruyko and Nikos Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *ICLR*, 2017.

[11] Sungsoo Ahn, Shell Xu Hu, et al., "Variational information distillation for knowledge transfer," in *CVPR*, 2019, pp. 9163–9171.

[12] Yonglong Tian, Dilip Krishnan, and Phillip Isola, "Contrastive representation distillation," in *ICLR*, 2020.

[13] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014, pp. 818–833.

[14] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[15] Frederick Tung and Greg Mori, "Similarity-preserving knowledge distillation," in *ICCV*, 2019, pp. 1365–1374.

[16] Yufan Liu, Jiajiong Cao, et al., "Knowledge distillation via instance relationship graph," in *CVPR*, 2019, pp. 7096–7104.

[17] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho, "Relational knowledge distillation," in *CVPR*, 2019, pp. 3967–3976.

[18] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, et al., "Correlation congruence for knowledge distillation," in *ICCV*, 2019, pp. 5007–5016.

[19] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin, "Learning what and where to transfer," in *ICML*, 2019, pp. 3030–3039.

[20] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen, "Cross-layer distillation with semantic calibration," in *AAAI*, 2021.

[21] Youcai Zhang, Zhonghao Lan, Yuchen Dai, Fangao Zeng, Yan Bai, Jie Chang, and Yichen Wei, "Prime-aware adaptive distillation," in *ECCV*, 2020, pp. 658–674.

[22] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar SSC*, 2003, pp. 1398–1402.

[23] Beibei Jin, Yu Hu, Qiankun Tang, Jingyu Niu, Zhiping Shi, Yinhe Han, and Xiaowei Li, "Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction," in *CVPR*, 2020, pp. 4554–4563.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

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

[26] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas, "Heterogeneous knowledge distillation using information flow modeling," in *CVPR*, 2020, pp. 2339–2348.

[27] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick, "Training region-based object detectors with online hard example mining," in *CVPR*, 2016, pp. 761–769.