FPD: Feature Pyramid Knowledge Distillation

Qi Wang[1], Lu Liu[1], Wenxin Yu[1](✉), Zhiqiang Zhang[2], Yuxin Liu[1], Shiyu Cheng[1], Xuewen Zhang[1], and Jun Gong[3]

[1] Southwest University of Science and Technology, Sichuan, China
[2] Hosei Univeristy, Tokyo, Japan
[3] Southwest Automation Research Institute, China
yuwenxin@swust.edu.cn

Abstract. Knowledge distillation is a commonly used method for model compression, aims to compress a powerful yet cumbersome model into a lightweight model without much sacrifice of performance, giving the accuracy of a lightweight model close to that of the cumbersome model. Commonly, the efficient but bulky model is called the teacher model and the lightweight model is called the student model. For this purpose, various approaches have been proposed over the past few years. Some classical distillation methods are mainly based on distilling deep features from the intermediate layer or the logits layer, and some methods combine knowledge distillation with contrastive learning. However, classical distillation methods have a significant gap in feature representation between teacher and student, and contrastive learning distillation methods also need massive diversified data for training. For above these issues, our study aims to narrow the gap in feature representation between teacher and student and obtain more feature representation from images in limited datasets to achieve better performance. In addition, the superiority of our method is all validated on a generalized dataset (CIFAR-100) and a small-scale dataset (CIFAR-10). On CIFAR-100, we achieve 19.21%, 20.01% of top-1 error with Resnet50 and Resnet18, respectively. Especially, Resnet50 and Resnet18 as student model achieves better performance than the pre-trained Resnet152 and Resnet34 teacher model. On CIFAR-10, we perform 4.22% of top-1 error with Resnet-18. Whether on CIFAR-10 or CIFAR-100, we all achieve better performance, and even the student model performs better than the teacher.

Keywords: Knowledge distillation · Feature pyramid network · Feature pyramid distillation.

1 Introduction

Knowledge distillation is a commonly method for compressing models. In 2015, Geoffrey Hinton et al. added "Temperature (T)" in softmax as the cross-entropy loss function. The initial distillation methods are based on the probability of training samples in the logits layer, such as KD [5]. And then, some classical features representation distillation methods such as Finets [10] added "hints

layer", AT [17] used the attention map of feature maps to transfer knowledge, FSP [16] defined the process of image processing between teacher model and student model as an additional matrix, VID [1] maximized the mutual information between teacher model and student model, and other methods such as AB [4], SP [13]. These methods mainly study the features representation among intermediate layer between the teacher model and student model to improve accuracy of the student model and make the student model as similar as possible to the teacher model. Later, to further improve the accuracy of the student model, some methods combine knowledge distillation with contrastive learning, such as CRD [12]. However, classical distillation methods have a significant gap in feature representation and model ability between teacher model and student model, and contrastive learning distillation method also need massive diversified data for training. The three knowledge distillation methods described above are shown in Figure 1.

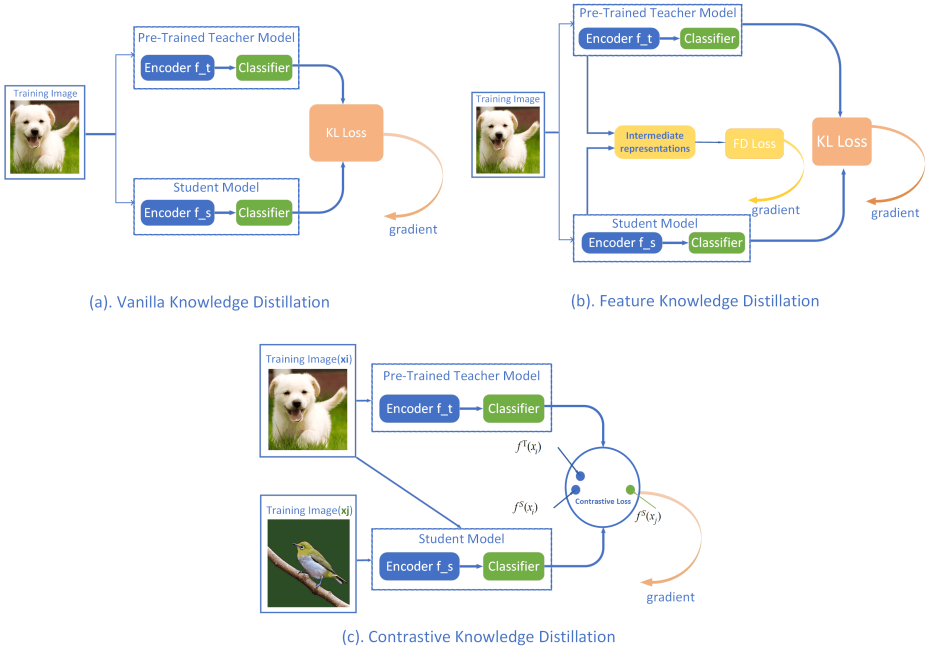

Fig. 1. Three kinds of knowledge distillation methods. (a) Vanilla KD calculates the gradient using final class predictions by pre-trained teacher and student, such as KD [5]. (b) Feature knowledge distillation gathers more gradient information from the intermediate layers through various knowledge representations such as Finets [10], AT [17], FSP [16]. (c) Contrastive knowledge distillation methods such as CRD [12].

Among the previously proposed knowledge distillation methods, the teacher's feature maps differ significantly from the student's due to differences of the net

architecture. Using the feature maps of the teacher model directly to learn will lead the student model to learn insufficiently. In order to solve these existing issues, we propose feature pyramid knowledge distillation (named FPD) to fuse the teacher's and student's feature maps at each stage, which can narrow the gap in feature maps between teacher model and student model during training to help the student model get more knowledge from the teacher model to get better performance. As we expected, our propose method is superior to these methods mentioned above, since, our method can be approximated as combining the logits distillation with the feature representation distillation rather than simply adding another loss function. Although the feature pyramid knowledge distillation reduces the gap of feature representation between teachers and students, but the error will be increased by feature pyramid when the teacher's prediction is wrong, so we use guided knowledge distillation [19] to avoid these errors. As expected, our experiments show that our method outperforms even the use of contrastive distillation methods on small-scale dataset (CIFAR-10), where the accuracy of using contrastive distillation methods significantly decreases in limited sample data. From experimental studies, it is shown that our method is more suitable in some special domains, such as medical and military, where massive and diverse samples are unavailable easily.

Overall, our contributions are summarized as follows:

- We provide an novel view to study feature representation distillation by feature pyramid to help student model have a better performance.
- We reveal limitations of the contrastive distillation methods in some special domain cause by the insufficient of datasets.
- We propose an efficient feature representation distillation method (named FPD) to overcome these issues and limitations. Our method achieves better performance and empirically demonstrates its superiority on CIFAR-100 and CIFAR-10.

2 Related Work

2.1 Knowledge Distillation

KD [5] transferred the teacher's output distribution as a kind of knowledge to the students, prompting student model to approach the teacher's output distribution, it used soft target to learn more about the teacher's predictions rather than learn about ground truth directly. FitNet [10] added some additional monitoring signals in the middle of the network to help the output of the intermediate student layer to be as close to the teacher as possible. AT [17] used the attention map of the teacher model and the student model to distillation. FSP [16] defined a solution process matrix for feature mapping between intermediate layers, representing how the teacher processes information between the two layers and allowing students to learn how the teacher processes information.

2.2 Feature Pyramid

Targets of different sizes all undergo the same downsampling ratio. They will have a sizeable semantic gap, and the most common performance is the relatively low accuracy of small target detection. The feature pyramid has different resolutions at different scales, and targets of different sizes can have suitable feature representations at the corresponding scales. The model's performance can be improved by fusing multi-scale information to predict targets of different sizes at different scales. ASPP [2] proposes to build by multiple branches with different dilation rates of dilated convolution, and after multiple branches concate and then 1x1 convolution, FPN [8] proposes to make the shallow feature map with better semantic information by top-down stacking.

2.3 Attention Mechanisms

The study of attention mechanisms originates from human observation habits and is widely used in natural language processing and computer vision tasks. In SENet [6], global average pool and fully connected is used to obtain weights on feature map channels to obtain channel attention. Attention is paid to channels and spatial locations in the feature map considering global average pool and maximum global pool information in CBAM [14]. The detailed structure of the core module of SENet [6] (denote as **SEBlock**) is shown in Figure 2.

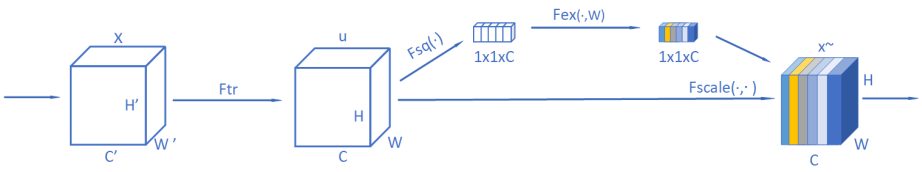

Fig. 2. A core Squeeze-and-Excitation block in SENet [6].

3 Method

In this section, we will introduce the theory behind feature pyramid distillation (named **FPD**), then explain why FPD is performed, and why we use guided knowledge distillation [19], and finally introduce the design of our loss function.

3.1 Feature Pyramid Knowledge Distillation

The FPN [8] consists of two parts: The first part is a bottom-up process, and the second part is a top-down and lateral connection fusion process. The bottom-up process is ordinary convolutional process. The top-down process scales up the

small feature map of the top layer to the same size as the feature map of the previous stage by upsampling. This method takes advantage of the more robust semantic features of the top layer and the detailed information of the bottom layer. The lateral connection fusion process takes the features of the previous layer that have been upsampled to the exact resolution as the current layer and fuses them by summation. The detailed architecture of FPN [8] is shown in Figure 3.

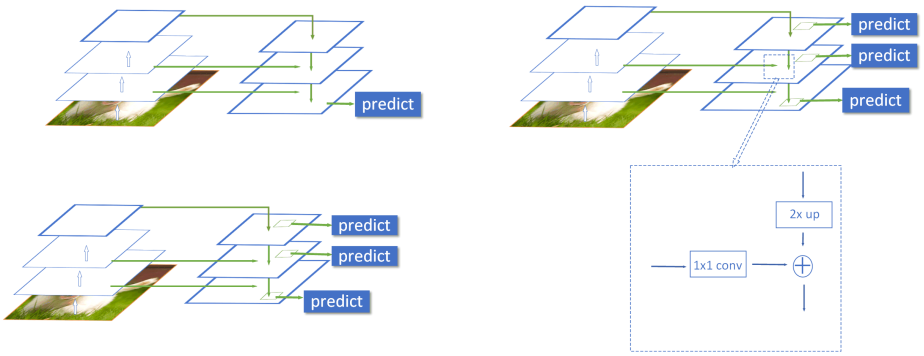

Fig. 3. Left-Top: a top-down architecture with skip connections, where predictions are made on the first level. Left-Bottom: our model that has a similar structure but leverages it as a feature pyramid, with predictions made independently at all levels. Right: A building block illustrating the lateral connection and the top-down pathway, merged by addition.

Our method FPD is derived from the FPN [8]. We believe that feature maps carried by the top layer can be approximated to the logits layer. Meanwhile, the model's bottom and intermediate feature maps carry more details of the original images, such as texture information. Therefore, we think of using the feature pyramid method, which will allow students to learn semantic and texture information at the same time during the training process, this way will narrow the gap of feature maps between teacher and student. In our paper, we take ResNet [3] as the base model, and select the feature map which is derived by residual block of the $conv^1$, $conv^2$, ..., $conv^i$ layers as the distillation feature representation of our FPD, using $\{f_1, f_2, ..., f_i\}$ denotes correspondingly. We put these feature maps into the feature pyramid block to fuse the feature information of each residual block, and feature maps after fusion are represented by $\{p_1, p_2, ..., p_i\}$. And then, we calculate correspondingly the loss value of each pair of feature maps after the feature pyramid between teacher and student. The detailed architecture of our FPD is shown in Figure 4.

We use the mean squared error (called **MSE**) as the loss function to calculate the gap of feature maps between the teacher and student for each pair after the feature pyramid, divide the obtained loss values for each pair by the sum of the

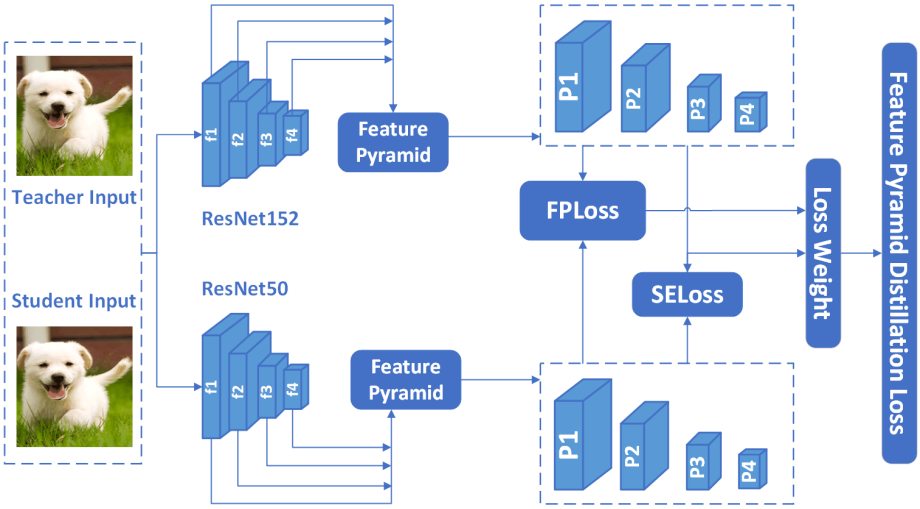

Fig. 4. The structure of our FPD. (1) FPLoss: Mean squared error loss function to calculate the loss value of each pair of feature maps after the feature pyramid. (2) SELoss: feature pyramid loss value of each pair feature maps after SEBlock (the structure of SEBlock is shown in Figure 2).

loss values for all pair, then we use the softmax function to calculate the **Weight** for the corresponding pair, and multiply the loss value by the corresponding **Weight** to obtain the total loss, called **FPLoss**.

$$Weight(t,s) = SoftMax(\frac{MSE(p_t^i, p_s^i)}{\sum_{i=1}^{N} MSE(p_t^i, p_s^i)}) = [w_1 ..., w_N] \tag{1}$$

$$FPLoss(t,s) = \sum_{i=1}^{N} Weight[i] * MSE(p_t^i, p_s^i) \tag{2}$$

The **t**, **s** in all equation denotes feature maps from teacher model and student model respectively. The $\boldsymbol{p_t^i}$, $\boldsymbol{p_s^i}$ denotes i-th feature map of all feature maps after feature pyramid block from teacher and student respectively. The **Weight[i]** denotes the i-th weight of loss each pair of the $\boldsymbol{p_t^i}$, $\boldsymbol{p_s^i}$ calculate by softmax. To gain more knowledge from the teacher model, we put feature maps for each pair after the feature pyramid into **SEBlock** (shown in Figure 2) to get feature maps channel-wise attention, use $\boldsymbol{p_{se_t}^i}$, $\boldsymbol{p_{se_s}^i}$ to denote respectively, and use the same method as **FPLoss** to calculate the **AT_Weight**, and calculate the gap of channel-wise attention between the teacher and student, called **SELoss**.

$$SE(t,s) = SEBlock(p_t^i, p_s^i) = (p_{se_t}^i, p_{se_s}^i) \tag{3}$$

$$SELoss(t,s) = \sum_{i=1}^{N} MSE(p_{se_t}^i, p_{se_s}^i) * AT\_Weight[i] \tag{4}$$

Finally to calculate the weight of **FPLoss** and **SELoss** use above-mentioned same method. The **FP_weight**, **SE_weight** denotes the weight of FPLoss and SELoss respectively. We call the final Loss as **FPDLoss**, the equation is as follows:

$$FPDLoss(t,s) = FPLoss * FP\_weight + SELoss * SE\_weight \tag{5}$$

3.2 Guided Knowledge Distillation

During the experiment, we observe that the feature maps after the feature pyramid help students get more information from the teacher during the training process. Nevertheless, we observe the feature pyramid method will extend the gap of the feature map when the teacher's prediction is wrong. Because the prediction of the teacher's is not always correct, if the prediction of the teacher is wrong, the feature maps of the teacher at the same time are not correct, and error feature maps will lead the student to learn error information from the teacher and lead students to learn insufficiently. For this issue, we try to use guided knowledge distillation [19] to rectify the error feature maps from the teacher to help student perform better. The equation of guided knowledge distillation is defined as follows:

$$GKD(t,s) = \frac{\sum_{i=1}^{N} I(p_t^i, y_i) * KL(p_t^i, p_s^i)}{\sum_{i=1}^{N} I(p_t^i, y_i)} \tag{6}$$

The p_t^i, p_s^i denotes final probability distribution of teacher and student respectively, and y_i denotes true label .The I is an indicator function and $I(p_t^i,y_i)$ is **1** when the output of the teacher equals true label, or else $I(p_t^i,y_i)$ is **0**. The n is the batch size and the $KL(p_t^i, p_s^i)$ is the mean of Kullback-Leibler divergence (KL) between p_t^i and p_s^i.

3.3 Loss Function

We follow the common approach to designing our final loss function, using our **FPDLoss** as the primary loss function, the guided knowledge distillation loss function as the primary auxiliary to rectify the error outputs of teachers, the cross-entropy loss function as a secondary auxiliary. Then, three kinds of loss function is multiplied by the corresponding weight factors, and add them to obtain the final loss function. Our final loss function is as following equation:

$$Loss(t,s) = \alpha * CE(s,y) + \beta * GKD(t,s) + \gamma * FPD(t,s) \tag{7}$$

In above equation, the t, s denote the final probability distribution of teacher and student, and the y denotes the true label in $CE(s,y), GKD(t,s)$, but the t, s denote feature maps after feature pyramid block in $FPD(t,s)$. Besides, the α, β, γ weight value is set based on the ratio of teacher model's error rate to the correct rate.

4 Experiments

We perform experiments on CIFAR-100 [7] and CIFAR-10 [7] datasets and compare with other networks. We use the pre-trained resnet32x4 [15], vgg13 [11], wrn-40-2 [18], resnet152 [3], resnet34 [3] as our teacher models, the pre-trained resnet32x4, vgg13, wrn-40-2 teacher model are publicly available from CRD [12] and resnet152, resnet34 teacher model is trained by ourselves.

4.1 Datasets and baselines

We validated our distillation method on CIFAR-100 [7] and CIFAR-10 [7]. Besides the vanilla KD [5], various approaches are reproduced for comparison, including FitNet [10], AT [17], VID [1], PKT [9], SP [13], CRD [12].To ensure the fairness of the experiments, we use the same data sampling and data augmentation methods as CRD [12], and such as learning rate, epoch and other hyperparameters, are also the same as CRD [12].

4.2 Implement Details

For all experiments on CIFAR-100 and CIFAR-10 datasets, since the image size is 32×32, in order to make the model learn better feature maps on low-resolution images, we use a random crop with padding=4, random horizontal flip and the standard data augmentation and normalize all images by channel means and standard deviations. We still use the SGD optimizer; the weight decay is set to 5e-4, and momentum is set to 0.9. The initial learning rate is set to 0.05, and the decay rate for the learning rate is 0.1, and when epoch = 150, 180, 210, the learning rate decays. The entire model is trained with 240 epochs. All initial settings above are the same with CRD [12]. In addition, for ourselves design, the weight of $CE(s, y)$, α=1, the weight of $GKD(t, s)$, β=5, the weight of $FPD(t, s)$, γ=20. In order to accommodate small networks with few layers, we use the feature maps of the first four stages of the model to compute uniformly. For the feature maps of the first four stages passing through the feature pyramid, we first use the convolution operation with padding=1, kernel_size=3 to shift the number of feature map channels to 32, 64, 128, 256. If the input is smaller than the set number of channels, and if it is larger than the set number of channels, we use the original number of channels and uniformly expand the channels to 256 in the upsampling process of the feature pyramid. According to the previous upsampling experience, we use the "bilinear" interpolation method in the upsampling process.

4.3 Comparison of Test Accuracy

From the Table 1, we can see that our method consistently outperforms all other knowledge distillation methods on CIFAR-100 and the improvements are pretty significant, and the student model performe better than the teacher such

as "ResNet50 & ResNet18". Especially, the performance of ResNet18 [3] even better than ResNet152 [3]. It is shown that our method can help the student to learn more knowledge from the teacher, and can rectify wrong information, making the student more robust. In Table 2, we can see that performance of the contrastive distillation is poor in small-scale datasets. However, our method outperforms in small-scale datasets.

Table 1. Results on the **CIFAR-100** validation. All results are the average over 3 trials.

Distillation Mechanism	Teacher / Student	wrn-40-2 / wrn-16-2	resnet32x4 / resnet8x4	vgg13 / vgg8	resnet152 / resnet50	resnet34 / resnet18
	Acc	75.61 / 73.26	79.42 / 72.50	74.64 / 70.36	79.90 / 79.26	79.12 / 78.36
Logits	KD [5]	74.92	73.33	72.98	80.32	79.53
Intermediate Layer	FitNet [10]	73.58	73.50	71.02	79.85	78.22
	AT [17]	74.08	73.44	71.43	80.46	78.69
	VID [1]	74.11	73.09	71.23	79.12	79.09
	PKT [9]	74.54	73.64	72.88	80.56	79.81
	SP [13]	73.83	72.94	72.68	80.42	79.90
Contrastive Learning	**CRD** [12]	**75.48**	**75.51**	**73.94**	**80.62**	**79.72**
Ours	**FPD**	**75.67** (↑)	**75.71** (↑)	**74.23** (↑)	**80.79** (↑)	**79.99** (↑)

Table 2. Results on the **CIFAR-10** validation. All results are the average over 3 trials.

Distillation Mechanism	Teacher / Student	wrn-40-2 / wrn-16-2	resnet32x4 / resnet8x4	vgg13 / vgg8	resnet34 / resnet18
	Acc	94.73 / 93.83	95.52 / 92.61	94.05 / 91.95	95.47 / 95.13
Logits	KD [5]	94.66	93.80	92.95	95.36
Intermediate Layer	VID [1]	94.02	93.10	92.49	95.26
Contrastive Learning	**CRD** [12]	**88.41** (↓)	**88.89** (↓)	**84.75** (↓)	**89.56** (↓)
Ours	**FPD**	**94.70** (↑)	**94.16** (↑)	**93.13** (↑)	**95.78** (↑)

4.4 Ablation Study

In this section, we use two ablation experiments to verify that student performance will be better during knowledge distillation by using our methods. Meanwhile, verifying the feature map of different levels fuse with feature pyramids will help the student to learn more knowledge from the teacher during training. The feature pyramid magnifies the error when prediction of the teacher is wrong and can affect students performance. Using our methods can narrow the gap in feature maps between teachers and students, reinforce the correct information learned by students, and help students rectify the wrong information learned from the teacher. When we only using FPD, the performance of all student network perform better. It shows that our proposed FPD makes the student learn more knowledge from the teacher. Next, we add guided knowledge distillation as an auxiliary loss function for our FPD, helping students learn more positive knowledge and improve performance. Results of ablation study see in Table 3.

Table 3. Results of ablation study on the **CIFAR-100** validation. All results are the average over 3 trials.

Distillation Mechanism	Teacher Acc	wrn-40-2 75.61	resnet32x4 79.42	vgg13 74.64	resnet34 79.12
	Student Acc	wrn-16-2 73.26	resnet8x4 72.50	vgg8 70.36	resnet18 78.36
Logits	KD [5]	74.92	73.33	72.98	79.53
Intermediate Layer	VID [1]	74.11	73.09	71.23	79.09
	SP [13]	73.83	72.94	72.68	79.90
Contrastive Learning	**CRD** [12]	**75.48**	**75.51**	**73.94**	**79.72**
Ours	**FPD**	74.62	73.94	72.65	78.95
	FPD+GKD	**75.76** (↑)	**75.71** (↑)	**74.23** (↑)	**79.99** (↑)

5 Conclusion

We propose feature pyramid knowledge distillation (named FPD): a novel knowledge distillation method to narrow the gap of feature maps between teachers and students to help get more information. And, the effectiveness and superiority of our method was demonstrated through experiments. Moreover, by using our FPD method allows students to learn both semantic information and texture details during training, which seems to serve as a novel method to learn both the teacher's predicted probability distribution and the teacher's intermediate feature map representations for students during training.

6 Acknowledgements

This research is supported by the Sichuan Science and Technology Program(No.20 20YFS0307), Mianyang Science and Technology Program(2020YFZJ016), Sichuan Provincial M. C. Integration Office Program, and IEDA laboratory of SWUST.

References

1. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9163–9171 (2019)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3779–3787 (2019)
5. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)
6. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
7. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
8. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
9. Passalis, N., Tefas, A.: Probabilistic knowledge transfer for deep representation learning. CoRR, abs/1803.10837 **1**(2), 5 (2018)
10. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
12. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
13. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1365–1374 (2019)
14. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
15. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
16. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4133–4141 (2017)

17. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
18. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
19. Zhou, Z., Zhuge, C., Guan, X., Liu, W.: Channel distillation: Channel-wise attention for knowledge distillation. arXiv preprint arXiv:2006.01683 (2020)