Improved Feature Distillation via Projector Ensemble

Anonymous Author(s)
Affiliation
Address
email

Abstract

Knowledge Distillation has been widely used to improve the performance of the lightweight network (student) by introducing the large network (teacher) to guide training. Among the existing methods, feature matching-based distillation has shown superior performance by minimizing the discrepancy between student and teacher features. Due to the dimension mismatch between student and teacher features, feature distillation methods usually impose a projector on the student or teacher networks to map features into a common space during training. Previous feature distillation methods mainly focus on the design of loss functions and the selection of the distilled layers, while the effect of the feature projector between the student and teacher remains under-explored. To better understand the impact of projectors in distillation, we conduct comprehensive experiments in this paper and observe that the student network benefits from a projector even if the feature dimensions of the student and teacher are the same. One plausible reason is that the projector is optimised towards a "global alignment" that cannot be achieved by just optimising independent feature pairs. Motivated by this, we propose an ensemble of projectors to further improve the distillation performance. Empirical results on a series of teacher-student pairs illustrate the effectiveness of the proposed method.

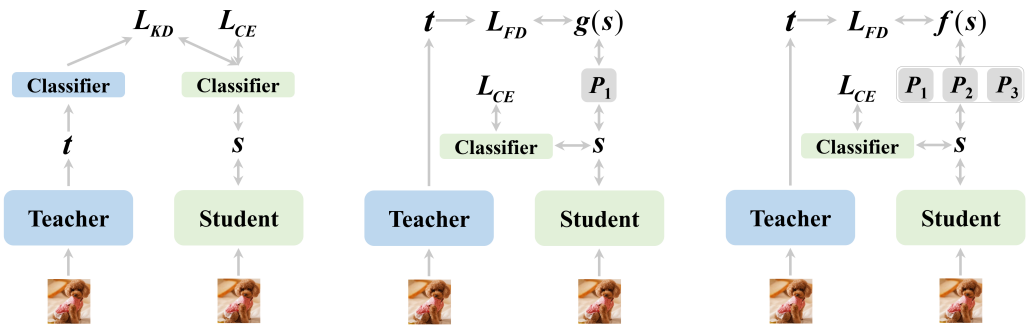
1 Introduction

The last decade has witnessed the rapid development of Convolutional Neural Networks (CNNs) [19, 29, 9, 20] in the field of computer vision. The size of networks has leapt forward along with their performance. This largely limits the applications of CNNs on edge devices [13]. To alleviate this problem, knowledge distillation has been proposed for network compression. The key idea of distillation is to use the knowledge obtained by the large network (teacher) to guide the optimization of the lightweight network (student) [12, 24, 31].

Existing methods can be roughly categorized into logit-based, feature-based and similarity-based distillation [7]. Recent research shows that feature-based methods generally distill a better student network compared to the other two groups [30, 4]. We conjecture that the process of mimicking the teacher's features provides a clearer optimization direction for the training of the student network. Despite the promising performance of feature distillation, it is still challenging to narrow the distribution gap between the student and teacher's feature spaces. To improve the feature learning ability of the student, various feature distillation methods have been developed by designing more powerful objective functions [30, 36, 33, 4] and determining more effective links between the layers of the student and teacher [2, 15, 1].

We found that the feature projected from the student to the teacher's feature space plays a key part in feature distillation and can be redesigned to improve the performance. Since the feature dimensions of student networks are not always consistent with that of teacher networks, a projector is required to map features into a common space for matching. As discussed in [33], imposing a projector on the student network can improve the distillation performance even if the feature dimensions of the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) Logit-based Distillation (b) Single Projector (c) Multiple Projectors

Figure 1: Illustration of (a) the traditional logit-based distillation [12], (b) the general feature-based distillation with single projector [4, 33] and (c) the proposed method with multiple projectors, where $\mathcal{L}_{CE}$, $\mathcal{L}_{KD}$ and $\mathcal{L}_{FD}$ are the cross-entropy loss, logit distillation loss and feature distillation loss, respectively.

student and teacher are the same. By comparing the gradients *w.r.t.* student features between networks with and without projectors, we hypothesize that the network with a projector can better capture the global feature distribution of the teacher, instead of learning from feature pairs only. Empirical study also verifies that the network with a projector obtains lower between-class cosine similarity in the student feature space, which is beneficial to the subsequent classification task. Motivated by this, we propose an ensemble of projectors for further improvement. It is evident that projectors with different initialization would generate diverse transformed features. Therefore, it is intuitive to improve the generalization of the student by using multiple projectors according to the theory behind ensemble learning [37, 32]. Since the projectors will be removed after distillation, the proposed ensemble method does not change the original structure of the student, or its complexity of inference. Figure 1 shows the comparisons of existing distillation methods and our method.

Our contributions are three-fold:

- We investigate the phenomenon that the student benefits from introducing a projector during feature distillation when the feature dimensions of the student and teacher are the same.
- Technically, we propose an ensemble of feature projectors to further improve the performance of feature distillation. The proposed method is extremely simple and easy to implement.
- Experimentally, we conduct comprehensive comparisons between different methods on benchmark datasets with different teacher-student combinations. It is shown that the proposed method consistently outperforms state-of-the-art methods.

2 Related Work

Since this paper mainly focuses on the design of the projector, we divided the existing methods into two categories in term of the usage of the projector as follows:

Projector-free methods. As the most representative distillation method, Knowledge Distillation (KD) [12] proposes to utilize the logits generated by the pre-trained teacher to be the additional targets of the student. The intuition of KD is that the generated logits are able to provide more useful information than the general binary labels for optimization. Motivated by the success of KD, various logit-based methods have been proposed for further improvement. For example, Deep Mutual Learning (DML) [35] proposes to replace the pre-trained teacher with an ensemble of students so that the distillation mechanism does not need to train a large network in advance. Teacher Assistant Knowledge Distillation (TAKD) [21] observes that a better teacher may distill a worse student due to the large performance gap between them. Therefore, a teacher assistant network is introduced to alleviate this problem. Another technical route of projector-free methods is the similarity-based distillation. Unlike the logit-based methods that aim to exploit the category information hidden

in the predictions of the teacher, similarity-based methods try to explore the latent relationships between samples in feature space. For example, Similarity-Preserving (SP) [31] distillation first constructs the similarity matrices of the student and teacher by computing the inner products between features and then minimises the discrepancy between the obtained similarity matrices. Similarly, Correlation Congruence (CC) [23] forms the similarity matrices with the kernel function. Although the logit-based and similarity-based methods do not require an extra projector during training, they are relatively less effective than the feature-based methods as shown in the recent research [4, 33].

Projector-dependent methods. The feature distillation methods aim to make the student and teacher features as similar as possible. Therefore, a projector is essential to map features into a common space. The first feature distillation method FitNets [24] minimizes the L2 distance between the student and teacher feature maps produced by the intermediate layer of networks. Furthermore, Contrastive Representation Distillation (CRD) [30], Softmax Regression Representation Learning (SRRL) [33] and Comprehensive Interventional Distillation (CID) [4] show that the last feature representations of networks are more suitable for distillation. One potential reason is that the last feature representations are closer to the classifier and will directly affect the classification performance [33]. The aforementioned feature distillation methods mainly focus on the design of loss functions such as introducing contrastive learning [30] and imposing causal intervention [4]. A simple 1x1 convolutional kernel or a linear projection is adopted to transform features in these methods. We notice that the effect of projectors is largely ignored. Previous works such as Factor Transfer (FT) [16] and Overhaul of Feature Distillation (OFD) [11] try to improve the architecture of projectors by introducing auto-encoder and modifying the activation function. However, their performance is not competitive enough compared to the state-of-the-art methods [33, 4]. Instead, this paper proposes a simple distillation framework by combining the ideas of distilling the last features and projector ensemble.

3 The Proposed Method

We first define the notations used in the following sections. In line with the observation in recent research [30, 4], we apply the feature distillation loss to the layer before the classifier. $S = \{s_1, s_2, ..., s_i, ..., s_b\} \in \mathbb{R}^{d \times b}$ denotes the last student features, where d and b are the feature dimension and the batch size, respectively. The corresponding teacher features are represented by $T = \{t_1, t_2, ..., t_i, ..., t_b\} \in \mathbb{R}^{m \times b}$, where m is the feature dimension. To match the dimensions of S and T, a projector $g(\cdot)$ is required to transform the student or teacher features. We experimentally find that imposing the projector on the teacher is less effective since the original feature distribution would be destroyed. Therefore, in the proposed distillation framework, a projector will be added to the student as $g(s_i) = \sigma(W s_i)$ during training and be removed after training, where $\sigma(\cdot)$ is the ReLU function and $W \in \mathbb{R}^{m \times d}$ is a projection matrix.

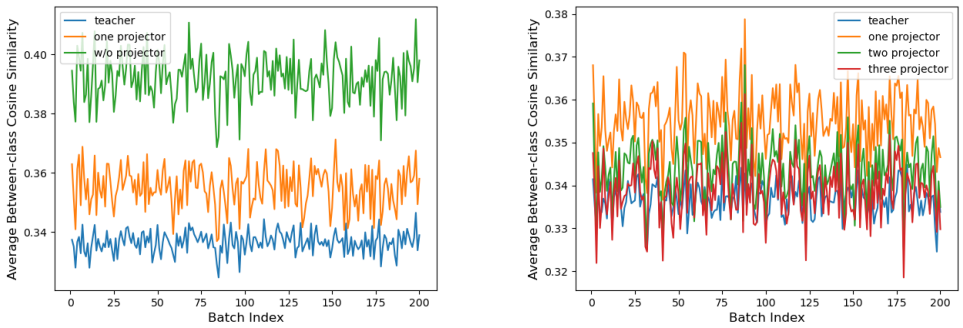
3.1 Feature Distillation

In recent works, SRRL and CID combine the feature-based loss with the logit-based loss to improve the performance. Since distillation methods are sensitive to hyper-parameters and changes of teacher-student combinations, the additional objectives will increase the training cost for coefficients adjustment. To alleviate this problem, we simply use the following Direction Alignment (DA) loss [17, 3, 8] for feature distillation:

$$\mathcal{L}_{DA} = \frac{1}{2b} \sum_{i=1}^{b} || \frac{g(s_i)}{||g(s_i)||_2} - \frac{t_i}{||t_i||_2} ||_2^2 = 1 - \frac{1}{b} \sum_{i=1}^{b} \frac{\langle g(s_i), t_i \rangle}{||g(s_i)||_2 ||t_i||_2}, \tag{1}$$

where $|| \cdot ||_2$ indicates the L2-norm and $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors. As mentioned in the Introduction, we observe that introducing a projector helps to improve the distillation performance even if the feature dimensions of the student and teacher are the same. We attempt to uncover the reason by examining the gradients when $d = m$. In the case of removing the projector during training, the gradients *w.r.t.* student features are as follows:

$$\sum_{i=1}^{b} \frac{\partial \mathcal{L}_{DA}}{\partial s_i^p}, \tag{2}$$



(a) (b)

Figure 2: Average between-class cosine similarities obtained by different methods on CIFAR-100 with teacher-student pair ResNet32x4-ResNet8x4.

where s_i^p is the student feature at the p-th iteration and $g(s_i) = s_i$. By imposing a projector on the student network, the gradients *w.r.t.* student features are as follows:

$$\sum_{i=1}^{b} \frac{\partial \mathcal{L}_{DA}}{\partial s_i^p} = \sum_{i=1}^{b} (D_i^p W^p)^T \frac{\partial \mathcal{L}_{DA}}{\partial g(s_i^p)}, \tag{3}$$

where $D_i \in \mathbb{R}^{m \times m}$ is a diagonal matrix with elements 0 or 1, which indicates the activation of the ReLU function for the i-th sample. On the other hand, the projector is updated as follows:

$$W^p \leftarrow W^{p-1} - \eta \sum_{i=1}^{b} D_i^{p-1} \frac{\partial \mathcal{L}_{DA}}{\partial g(s_i^{p-1})} (s_i^{p-1})^T, \tag{4}$$

where η is the learning rate. The main difference between gradients (2) and (3) is the nonlinear transformation $D_i^p W^p$. Equation (4) shows that W^p is updated according to the previous mini-batch data s_i^{p-1}. We hypothesize that the student network may better capture the global feature distribution by introducing a projector to carry data information during back propagation. Figure 2(a) compares the average between-class similarities between different methods on CIFAR-100 dataset. The between-class cosine similarities of the student features are larger in the case of removing the projector. However, by adding a projector on the student to enhance feature learning, the between-class similarities will be reduced, which makes the features more distinguishable for the subsequent classification task.

3.2 Ensemble of Projectors

The above analysis suggests that the projector can boost the distillation performance of the student. Motivated by this, we propose an ensemble of projectors for further improvement. There are two advantages of using multiple projectors. Firstly, projectors with different initialization would provide different transformed features, which is beneficial to the generalizability of the student [37, 32]. Secondly, it is evident that the capacity of one projector is limited. Using ensemble learning is a natural way to achieve a good trade-off between training error and generalizability. By introducing multiple projectors, the Modified Direction Alignment (MDA) loss is as follows:

$$\mathcal{L}_{MDA} = 1 - \frac{1}{b} \sum_{i=1}^{b} \frac{\langle f(s_i), t_i \rangle}{||f(s_i)||_2 ||t_i||_2}, \tag{5}$$

where $f(s_i) = \frac{1}{q} \sum_{k=1}^{q} g_k(s_i)$, q is the number of projectors and $g_k(\cdot)$ indicates the k-th projector. After imposing multiple projectors, the gradients *w.r.t.* student features are as follows:

$$\sum_{i=1}^{b} \frac{\partial \mathcal{L}_{MDA}}{\partial s_i^p} = \frac{1}{q} \sum_{i=1}^{b} \sum_{k=1}^{q} (D_{k,i}^p W_k^p)^T \frac{\partial \mathcal{L}_{MDA}}{\partial f(s_i^p)}, \tag{6}$$

Algorithm 1 Improved Feature Distillation via Projector Ensemble.

Input: The pre-trained teacher, the structure of the student, training data X and labels.
Parameter: Total iterations N, α and the number of projectors q.
Initialization: Initialize different projectors and the student.
Training:

1: **for** $i = 1 \rightarrow N$ **do**
2: Sample a mini-batch data from X.
3: Generate S, T and the student's prediction by forward propagation.
4: Update projectors and the student network by minimizing objective (8).
5: **end for**

Output: The distilled student.

where $D^p_{k,i}$ denotes the k-th binary diagonal matrix corresponding to the i-th feature at the p-th iteration. The k-th projector is updated as follows:

$$W^p_k \leftarrow W^{p-1}_k - \eta \sum_{i=1}^{b} D^{p-1}_{k,i} \frac{\partial \mathcal{L}_{MDA}}{\partial f(s^{p-1}_i)} (s^{p-1}_i)^T, \tag{7}$$

Similarly, we analyze the feature distribution of the student after introducing more projectors. As shown in Figure 2(b), the global feature distribution is closer to that of the teacher by using more projectors in term of the results of between-class similarities. By combining the distillation loss (5) and the classification loss together, we obtain the following objective function to train the student:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{MDA}, \tag{8}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss and α is a hyper-parameter. The details of our method are shown in Algorithm 1.

4 Experiments

We conduct comprehensive experiments to evaluate the performance of different methods and the effectiveness of the proposed projector ensemble-based feature distillation, on image classification task. The codes of our method are available in Supplementary Material. Implementation details are as follows:

Baselines. We select representative distillation methods in various categories for comparisons, including logit-based method KD [12], similarity-based methods CC [23], SP [31] and Relational Knowledge Distillation (RKD) [22], feature-based methods FitNets [24], FT [16], CRD [30], SRRL [33] and CID [4]. The logit-based and similarity-based methods are projector-free and the feature-based methods require additional projectors. FitNets and SRRL use convolutional kernels to transform the student features. FT adopts an auto-encoder to extract the latent feature representations of the student and teacher. CRD maps the student and teacher features into a low-dimensional space while CID maps the student features into teacher space with linear projections. For simplicity, the proposed method constructs the projector by combining a linear projection and the ReLU function.

Datasets. Two benchmark datasets are used for evaluation in our experiments. ImageNet [25] contains approximately 1.28 million training images and 50,000 validation images from 1,000 classes. The validation images are used for testing. Each image is resized to 224x224. CIFAR-100 [18] dataset includes 50,000 training images and 10,000 testing images from 100 classes. Each image is resized to 32x32. On ImageNet and CIFAR-100, we adopt the commonly used random crop and horizontal flip techniques for data augmentation.

Teacher-student pairs. To validate the generalizability of different distillation methods, we select a group of popular network architectures to form different teacher-student pairs. The teacher networks include ResNet34 [9], DenseNet201 [14], WRN-40-2 [34], VGG13 [27], ResNet32x4 [9] and ResNet50 [9]. The student networks comprise of ResNet18 [9], MobileNet [13], WRN-16-2 [34], VGG8 [27], ResNet8x4 [9] and MobileNetV2 [26]. By combining different teacher and student networks, we can perform distillation between similar architectures (e.g. ResNet34-ResNet18) and different architectures (e.g., DenseNet201-ResNet18).

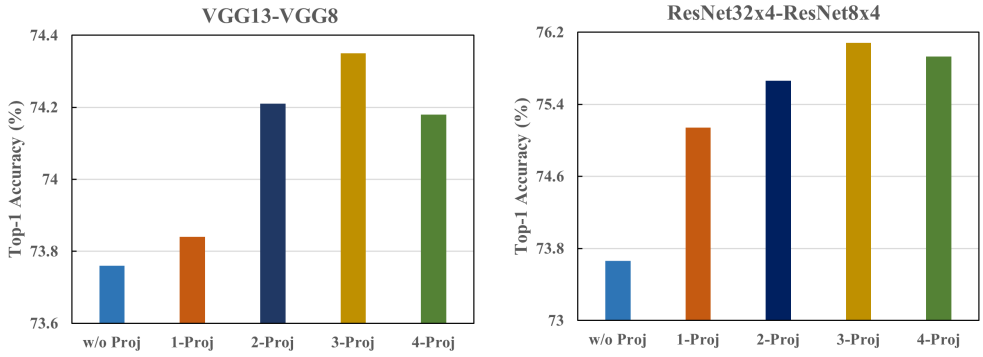


Figure 3: Horizontal ensembles on CIFAR-100 with different teacher-student pairs. 2-Proj, 3-Proj and 4-Proj indicate the number of projectors in the ensemble.

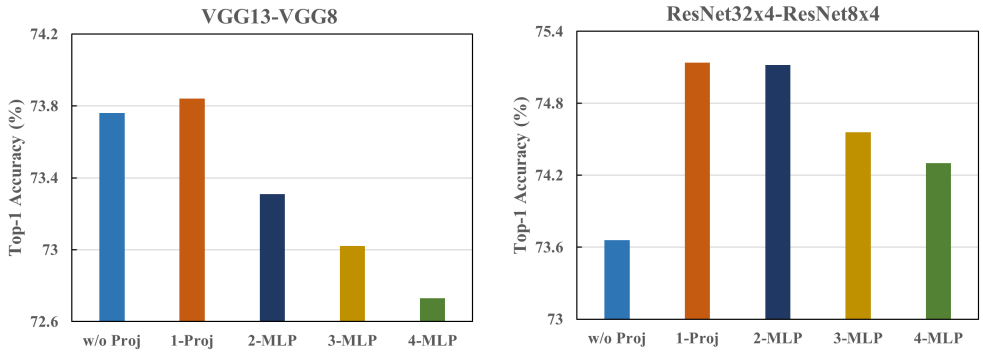


Figure 4: Deep projectors on CIFAR-100 with different teacher-student pairs. 2-MLP, 3-MLP and 4-MLP indicate the depth of the projectors with different number of layers.

Training. Following the settings of previous methods [1], the batch size, epochs, learning rate decay rate and weight decay rate are 256/64, 100/240, 0.1/0.1, and 0.0001/0.0005, respectively on ImageNet/CIFAR-100. The initial learning rate is 0.1 on ImageNet, and 0.01 for MobileNetV2, 0.05 for the other students on CIFAR-100. Besides, the learning rate drops at every 30 epochs on ImageNet and drops at 150, 180, 210 epochs on CIFAR-100. The optimizer is Stochastic Gradient Descent (SGD) with momentum 0.9. All the experiments are performed on an NVIDIA V100 GPU.

Hyper-parameters. By following the conventions in CRD [30], we use the same settings for the hyper-parameters of KD, CC, SP, RKD, FitNets, FT and CRD. For SRRL and CID, the settings of hyper-parameters are provided by the corresponding authors. For the proposed method, we set $\alpha = 25$ and $q = 3$ by tuning with teacher-student pair ResNet34-ResNet18 on ImageNet. For a fair comparison, the hyper-parameters of different methods are fixed in all experiments.

4.1 Ablation Studies

This section studies the effectiveness of the proposed projector ensemble method, and how different ensemble strategies affect the performance. In this experiment, two different network architectures, i.e. VGG-style and ResNet-style networks are used for illustration in Figures 3 and 4.

Horizontal Ensemble of projectors. Figure 3 shows the top-1 classification accuracy of the proposed projector ensemble with different number of projectors. It verifies that imposing a projector improves the distillation performance when the feature dimensions of the student and teacher are the same. A potential reason is that the projector helps to capture the global data distribution across difference

[1] https://github.com/HobbitLong/RepDistiller

Table 1: Classification accuracy (%) on ImageNet with different teacher-student pairs (a) ResNet34-ResNet18, (b) ResNet50-MobileNet, (c) DenseNet201-ResNet18 and (d) DenseNet201-MobileNet.

Pair	Accuracy	Student	KD	SP	CRD	SRRL	CID	Ours	Teacher
(a)	Top-1	69.75	70.83	70.94	70.85	71.71	71.86	**71.94**	73.31
	Top-5	89.07	90.15	89.83	90.12	90.58	90.63	**90.68**	91.41
(b)	Top-1	69.06	70.65	70.14	71.03	72.58	72.25	**73.16**	76.13
	Top-5	88.84	90.26	89.64	90.16	91.05	90.98	**91.24**	92.86
(c)	Top-1	69.75	70.38	70.75	70.87	71.76	71.99	**72.29**	76.89
	Top-5	89.07	90.12	90.01	89.86	90.80	90.64	**90.99**	93.37
(d)	Top-1	69.06	69.98	70.34	70.82	72.28	71.90	**73.24**	76.89
	Top-5	88.84	89.93	89.63	90.09	90.90	90.97	**91.47**	93.37

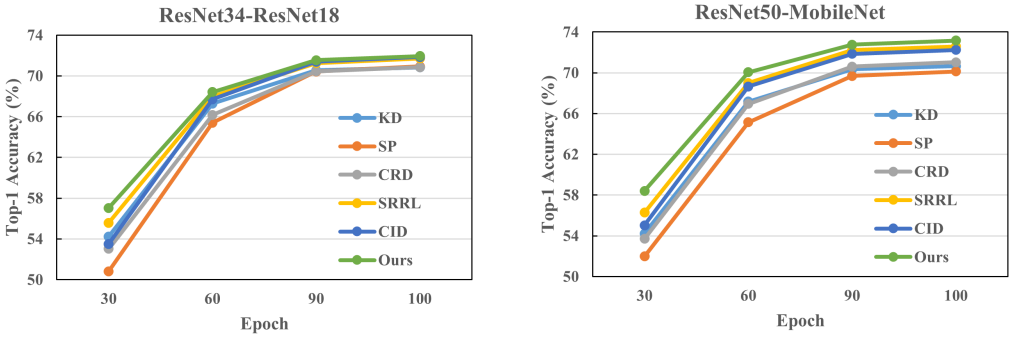


Figure 5: Top-1 accuracy of different methods on ImageNet with different number of epochs and different teacher-student pairs.

samples during the process of back propagation (Equations 4 and 7). Besides, by integrating multiple projectors, the proposed method further increases the classification accuracy by a clear margin with various numbers of projectors.

Deep Cascade of projectors. Another common way to modify the architecture is to increase the depth of the projector. Figure 4 demonstrates the changes of distillation performance by gradually stacking non-linear projections. In this figure, 2-MLP, 3-MLP and 4-MLP are multilayer perceptrons and each layer outputs m-dimensional features followed by a ReLU activation. For instance, the output of 2-MLP is $g(s_i) = \sigma(W_2\sigma(W_1 s_i))$, where $W_1 \in \mathbb{R}^{m\times d}$ and $W_2 \in \mathbb{R}^{m\times m}$ are projection matrices. It is shown that simply increasing the depth of the projector does not improve the performance of the student and tends to degrade the effectiveness of the projector. We hypothesize that with the increase of depth, the teacher's features can be over-fitted by the projector.

4.2 Results on ImageNet

The performance of the students distilled by different methods are listed in Table 1. Compared to the settings in previous methods [30, 33, 4], we introduce more teacher-student pairs for evaluation in this experiment so the generalizability of different methods can be better evaluated. As presented in the table, feature distillation methods (CRD, SRRL, CID and our method) outperform both the logit-based method (KD) and the similarity-based method (SP) in most cases.

One major difference between CRD and the other feature distillation methods is the way of feature transformation. CRD transforms the teacher and student features simultaneously while the other methods only transform the student features. By solely mapping the student features into the teacher space, the original teacher feature distribution can be preserved without losing discriminative information. Therefore, SRRL, CID and our method obtain better performance than CRD. Besides, our method consistently outperforms the state-of-the-art methods SRRL and CID with different teacher-student pairs. With pair DenseNet201-MobileNet, the proposed method obtains 0.96% and 0.57% improvements compared to the second best method in terms of top-1 and top-5 accuracy, respectively. MobileNet (4.2M parameters) distilled by our method can obtain similar performance

Table 3: Training times (in second) of one epoch on ImageNet with teacher-student pair DenseNet201-ResNet18.

Method	KD	SP	CRD	SRRL	CID	Ours
Time	2969	2989	3158	3026	3587	2995

Table 4: Top-1 classification accuracy and standard deviation (%) on CIFAR-100 with different teacher-student pairs.

Teacher	WRN-40-2	VGG13	ResNet32x4	ResNet50	ResNet50
Student	WRN-16-2	VGG8	ResNet8x4	VGG8	MobileNetV2
Teacher	75.61	74.64	79.42	79.34	79.34
Student	73.22±0.13	70.74±0.31	72.93±0.28	70.74±0.31	65.03±0.09
KD	74.92±0.28	72.98±0.19	73.33±0.25	73.81±0.13	67.35±0.32
CC	73.56±0.26	70.71±0.24	72.97±0.17	70.25±0.12	65.43±0.15
SP	73.83±0.12	72.68±0.19	72.94±0.23	73.34±0.34	68.08±0.38
RKD	73.35±0.09	72.21±0.16	71.90±0.11	71.50±0.07	64.43±0.42
FitNets	73.58±0.32	71.02±0.31	73.50±0.28	70.69±0.22	63.16±0.47
FT	73.25±0.20	70.58±0.08	72.86±0.12	70.29±0.19	60.99±0.37
CRD	75.48±0.09	73.94±0.22	75.51±0.18	74.30±0.14	69.11±0.28
SRRL	75.59±0.17	73.44±0.07	75.33±0.04	74.23±0.08	68.41±0.54
Ours	**76.02±0.10**	**74.35±0.12**	**76.08±0.33**	**74.58±0.22**	**69.81±0.42**

and reduce about 80% of the parameters compared to the ResNet34 (21.8M parameters). Figure 5 reports the Top-1 accuracy of different methods with different training epochs. It is shown that the proposed method converges faster than the other distillation methods.

Two recently proposed methods, namely Attention-based Feature Distillation (AFD) [15] and Knowledge Review (KR) [2] are also introduced for comparisons as reported in Table 2. Unlike methods that utilize the last layer of features for distillation (CRD, SRRL, CID and ours), AFD and KR propose to extract information from multiple layers of features. Table 2 shows that the proposed method performs better than AFD and KR with different pairs, which indicates that the last layer of features is sufficient to obtain good distillation performance on ImageNet.

Table 2: Comparisons of the proposed method and distillation methods using multiple layers of features.

Teacher	ResNet34		ResNet50	
Student	ResNet18		MobileNet	
Acc	Top-1	Top-5	Top-1	Top-5
Teacher	73.31	91.41	76.13	92.86
Student	69.75	89.07	69.06	88.84
AFD	71.38	–	–	–
KR	71.61	90.51	72.56	91.00
Ours	**71.94**	**90.68**	**73.16**	**91.24**

We compare the training costs of different methods in Table 3. Since KD and SP are projector-free methods, their training costs are lower than that of the feature distillation methods. The training cost of our method is slightly higher than KD and SP because we use multiple projectors to improve the optimization of the student. On the other hand, the proposed method only uses a naive direction alignment loss to distill the knowledge. Therefore, the computation complexity is lower compared to the other feature-based methods.

In [21], the authors observe that a better teacher may fail to distill a better student. Such phenomenon also exists in Table 1. For example, compared to the pair ResNet50-MobileNet, most of the methods distill a worse student by using a better network DenseNet201 as the teacher. One plausible explanation for this phenomenon is that the knowledge of a better teacher is more complex and is more difficult to learn. To alleviate this problem, TAKD [21] introduces some smaller assistant networks to facilitate training. Densely Guided Knowledge Distillation (DGKD) [28] further extends TAKD with dense connections between different assistants. However, the training costs of these methods are greatly increased by using the assistant networks. As shown in the table, the proposed method has the potential to alleviate this problem without introducing the additional networks.

4.3 Results on CIFAR-100

Table 4 reports the experimental results on CIFAR-100 with five teacher-student pairs. We run our method for three times with different seeds and obtain the average accuracy. Since CID requires different hyper-parameters for different pairs to achieve good performance, we omit it for comparisons on CIFAR-100. Among the projector-free distillation methods, the logit-based method KD shows better performance compared to the similarity-based methods CC, SP and RKD. Furthermore, KD outperforms the projector-based methods FitNets and FT in most cases. Since FitNets is designed to distill the intermediate features, its performance is unstable for teacher-student pairs using different architectures. FT uses an auto-encoder as the projector to extract latent representations of the teacher, which may disturb the discriminative information to some extent and consequently degrade the performance. The recently proposed feature distillation methods CRD and SRRL show competitive performance compared to the previous methods by distilling the last layer of features. By harnessing the power of both distilling the last features and projector ensemble, the proposed method consistently achieves the highest accuracy on CIFAR-100.

We further investigate the effect of the activation function. Table 5 shows that the addition of the ReLU activation function has a significant positive impact on the performance of the proposed method. The reason is that the lack of an activation function and the non-linearity introduced by it limits the diversity of the projectors, as a group of linear projections can be mathematically reduced to a single linear projection through sum-pooling, which will degrade the distillation performance. Recently, Gaussian Error Linear Units (GELU) [10] has been well discussed because of its effectiveness on Transformer [5, 6]. we replace ReLU with GELU in the proposed method to test the performance. It is shown that the performance of the proposed method can be further improved by using GELU.

Table 5: Comparisons of top-1 accuracy with different activation functions on CIFAR-100.

Teacher Student	VGG13 VGG8	ResNet32x4 ResNet8x4
w/ ReLU (Ours)	74.35±0.12	76.08±0.33
w/ GELU	74.39±0.18	76.32±0.27
w/o activation	73.46±0.42	75.04±0.37

5 Conclusion

This paper studies the positive effect of the projector in feature distillation and proposes a projector ensemble-based architecture to improve feature distillation. We first investigate the phenomenon that the addition of a projector improves the distillation performance even when the feature dimensions of the student and the teacher are the same. From the perspective of back propagation, we find that the projector is able to preserve information from data samples across different batches and hence enhance the global feature learning ability of the student. Based on this observation, we propose an ensemble of projectors from the student feature to the teacher's feature space to further improve the distillation performance. Empirical results on ImageNet and CIFAR-100 show that our method consistently achieves competitive performance with different teacher-student combinations, compared to other state-of-the-art methods.

Limitations and future work. In addition to the image classification task, the proposed method can be further applied in other downstream tasks (e.g., object detection and semantic segmentation), which can be explored in future work. Besides, the proposed method focuses on using the direction alignment loss for distillation. How to effectively and efficiently integrate logits and similarity information into the proposed framework is a potential research direction.

References

[1] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *AAAI*, pages 7028–7036, 2021.

[2] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, pages 5008–5017, 2021.

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021.

[4] Xiang Deng and Zhongfei Zhang. Comprehensive knowledge distillation with causal intervention. In *NeurIPS*, 2021.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[7] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 129(6):1789–1819, 2021.

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[11] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, pages 1921–1930, 2019.

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

[13] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

[15] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *AAAI*, pages 7945–7952, 2021.

[16] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *NeurIPS*, 2018.

[17] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical report*, 2009.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

[21] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, pages 5191–5198, 2020.

[22] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pages 3967–3976, 2019.

[23] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *ICCV*, pages 5007–5016, 2019.

[24] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.

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

[28] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *ICCV*, pages 9395–9404, 2021.

[29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[30] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020.

[31] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, pages 1365–1374, 2019.

[32] Xiaofang Wang, Dan Kondratyuk, Eric Christiansen, Kris M Kitani, Yair Alon, and Elad Eban. Wisdom of committees: An overlooked approach to faster and more accurate models. *arXiv preprint arXiv:2012.01988*, 2020.

[33] Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *ICLR*, 2021.

[34] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[35] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018.

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

[37] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. Ensembling neural networks: many could be better than all. *Artificial intelligence*, 137(1-2):239–263, 2002.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]