Local Correlation Consistency for Knowledge Distillation

Xiaojie Li[1[0000−0001−6449−2727]], Jianlong Wu[2(✉)[0000−0003−0247−5221]], Hongyu Fang[3[0000−0002−9945−9385]], Yue Liao[4[0000−0002−2671−0655]], Fei Wang[1[0000−0002−1024−5867]], and Chen Qian[1[0000−0002−8761−5563]]

[1]SenseTime Research
[2]School of Computer Science and Technology, Shandong University
[3]School of Electronics Engineering and Computer Science, Peking University
[4]School of Computer Science and Engineering, Beihang University
lixiaojie@sensetime.com, jlwu1992@sdu.edu.cn, fanghongyu@pku.edu.cn, liaoyue.ai@gmail.com, {wangfei,qianchen}@sensetime.com

Abstract. Sufficient knowledge extraction from the teacher network plays a critical role in the knowledge distillation task to improve the performance of the student network. Existing methods mainly focus on the consistency of instance-level features and their relationships, but neglect the local features and their correlation, which also contain many details and discriminative patterns. In this paper, we propose the local correlation exploration framework for knowledge distillation. It models three kinds of local knowledge, including intra-instance local relationship, inter-instance relationship on the same local position, and the inter-instance relationship across different local positions. Moreover, to make the student focus on those informative local regions of the teacher's feature maps, we propose a novel class-aware attention module to highlight the class-relevant regions and remove the confusing class-irrelevant regions, which makes the local correlation knowledge more accurate and valuable. We conduct extensive experiments and ablation studies on challenging datasets, including CIFAR100 and ImageNet, to show our superiority over the state-of-the-art methods.

Keywords: Knowledge Distillation, Local Correlation Consistency, Class-Aware Attention

1 Introduction

Convolutional Neural Networks have achieved great successes in the vision community, significantly facilitating the development of many practical tasks, such as image classification [14, 23, 7] and face recognition [24, 19, 28]. Currently, many complicated neural networks with deeper and wider architectures have been proposed to pursuit high performance [25, 34]. However, these networks cost plenty of parameters and computations, which limits their deployments on computationally limited platforms, such as mobile devices, embedded systems. Towards this issue, model compression and acceleration become popular research topics

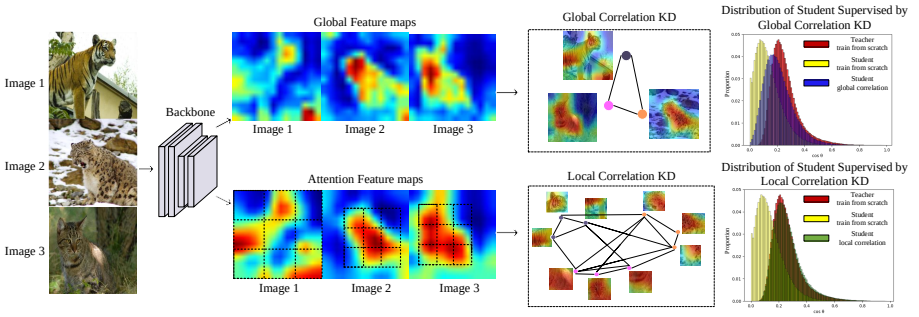

Fig. 1. Comparison between traditional global correlation methods and our proposed method. Instead of directly using the global feature maps to construct the relationship matrix, we first select the class-relevant regions by a class-aware attention module and then construct the local correlation matrix based on the selected local parts to guide the learning of the student network. Feature distributions comparison in the rightmost figure shows that the proposed method can help the student model to better mimic the teacher model than the global correlation method

recently. Typical methods include network pruning [6, 17, 4], compact architecture design [9, 40], network quantization [5, 31, 11], knowledge distillation [8, 41, 35, 36, 1, 26], and so on. Among them, knowledge distillation has been validated as a very effective approach to improving the performance of a light-weight network, *i.e.* student, with the guidance of a pre-trained large deep network, *i.e.* teacher. It encourages the student to learn the teacher's knowledge by applying some consistency based regularization between teacher and student.

The essential point of knowledge distillation is to extract sufficient knowledge from a teacher network to guide a student network. Conventional methods mostly focused on instance-level feature learning, which aims to mimic output activations [8, 41, 2, 1] or transfer the correlation in feature space [16, 18, 27, 20]. The instance-level based methods have achieved good performance, but they still suffer from the following limitations. Firstly, it is hard for a student to thoroughly understand the transferred knowledge from the teacher only based on global supervision. We observe that the local features are also important for the network to understand and recognize an object. As can be seen in Figure 1, the teacher network can make the right predictions for different categories of objects with similar appearance based on those distinguishing local regions, such as the head, the streaks of the body, or the foot appearance, but the student network may fail. We consider that the teacher network with more learn-able parameters can generate more discriminative local features, while the student is hard to achieve that with its limited capacity. Therefore, learning local knowledge from the teacher should be considered as an important factor to improve the discriminative ability of the student network. Secondly, the images may contain regions that are irrelevant to the category information, *e.g.* background. Directly making the student mimic the global features or their relationships without se-

lection is not an optimal way. Besides, each pixel of the class-aware region also has different contributions to the final classification. This property requires the knowledge distillation methods to transfer knowledge selectively according to its importance.

To resolve the above limitations, a novel local correlation exploration framework is proposed for knowledge distillation, which models sufficient relationships of those class-aware local regions. For the first limitation, we greatly enrich the family of network knowledge by proposing three different kinds of local relationships: (1) the local intra-instance relationship across different positions; (2) the local inter-instance relationship in the same position; (3) the local inter-instance relationship across different positions. Based on the above local relationships, we represent the intermediate feature maps using a more concise and structural form. Further, we hope the correlations computed by the teacher network could be well preserved by the student network. Therefore, we define the consistency regularization to minimize their difference between the teacher and student models. For the second limitation, to transfer the knowledge of those valuable class-aware regions and reduce the influence of invalid class-irrelevant information, we propose a novel class-aware attention module to generate the attention maps before the construction of the local correlation matrices.

We conduct extensive experiments on typical datasets to validate the effectiveness of the proposed framework as well as the local relationships. As shown in the rightmost figure of Figure 1, we allocated a set of feature maps from the middle layer of a set of models and draw the $cos(\theta)$ similarity distributions between the local patches of those feature maps. The red one is the distribution of the teacher. The yellow one is from the student trained from scratch. The blue one is from the student supervised by global correlation. And the green one is from the student supervised by our local correlation. The higher the coincidence between the histograms of the student and teacher, the more knowledge the student learns from the teacher. This graph shows that the student supervised by our local correlation achieves higher distribution coincidence with the teacher as well as higher accuracy than the student supervised by global correlation.

Our main contributions are summarized as follows:

1) We make the first attempt to explore local relationships in knowledge distillation and propose a novel local correlation consistency based framework. Instead of the traditional global feature based relationship, we mainly focus on the local correlation knowledge, which contains more details and discriminative patterns. By thoroughly investigating three kinds of local relationships, the student network in our framework can sufficiently preserve the important knowledge of the large teacher network. nt
2) To make the local correlation knowledge more accurate and valuable, we propose a novel class-aware attention module to generate attention masks for valuable class-relevant regions, which can reduce the influence of invalid class-irreleva regions, highlight the contribution of important pixels, and improve the performance as well.

3) Extensive experiments and ablation studies conducted on CIFAR100 [13] and ImageNet [3] show the superiority of the proposed method and effectiveness of each proposed module.

2 Related Works

The concept of knowledge distillation (KD) with neural networks is first presented by Hinton et al. in 2015 [8], where they come up with the teacher-student framework. Since then, many works have been proposed to improve its applicability and generalization ability. According to the types of knowledge to transfer, existing KD methods can be divided into three categories, including the feature representation learning based methods, attention based methods, and graph learning based methods. We briefly introduce them in this section.

Feature learning based methods mainly aim to train the student to mimic output activations of individual data examples represented by the teacher. Zhang et al. [41] learn a projection matrix to project the teacher-level knowledge and its visual representations from an intermediate layer of teacher network to an intermediate layer of student network. Yim et al. [33] construct the flow of solution procedure matrix across two different layers and minimize the difference between that matrix of teacher and student. Aguilar et al. [1] adopt both the activations and internal representations of the teacher network to guide the learning of the student network and achieve good performance on text classification. Chung et al. [2] try to capture the consistent feature map of intermediate layers by the adversarial learning. Similarly, Shu et al. [22] incorporate the intermediate supervision under the adversarial training framework. To better learn discriminative feature representation, Tian et al. [26] come up with the contrastive learning framework. Lan et al. [15] construct a multi-branch network, whose ensemble predictions are taken as supervision for the training of every single branch.

Attention mechanisms have been widely used in computer vision [29, 32, 10] and have been successfully applied in the field of KD [38, 39, 12]. Zagoruyko et al. [38] first show that attention transfer can significantly improve the performance of convolutional neural networks. Zhang et al. [39] present the self distillation framework to distill knowledge within the network itself. Kim et al. [12] make use of the output errors for self-attention based KD models.

Correlation learning [30] receives much attention for KD recently. Instead of directly teach the student to fit the instance features of the teacher, it transfers the correlation among training samples from the teacher network to the student network. Liu et al. [16] construct the instance relationship matrix, which takes the instance features, instance relationships, and feature space transformation into consideration to transfer sufficient knowledge. Park et al. [18] propose distance-wise and angle-wise distillation losses to penalize structural differences in relations. Both Tung et al. [27] and Peng et al. [20] hope to preserve the pair-wise similarity based on the correlation consistency.

Our method focuses on correlation learning and introduces a class-aware attention module. Compared with existing work, our differences mainly lie in

two aspects. First, we are the first to explore local correlation during knowledge transfer, while previous methods mainly use the global features to compute the correlations among instances. Second, our class-aware attention module learns the soft attention mask under the supervision of the ground-truth label, which can strengthen the class-aware regions and weaken the class-irrelevant regions during knowledge transfer. To our knowledge, the above attention mechanism is new in the knowledge distillation area.

3 Methods

In this section, we first summarize the basic framework of traditional global embedding based feature learning and correlation learning KD methods. Then we describe our local relationship based KD framework, and introduce the class-aware attention module to filter the semantic-irrelevant knowledge from the feature maps before the correlation construction. Finally, we come up with the overall loss function to supervise the training of the student network.

3.1 Problem Formulation

Given a teacher model T, a student model S and N training samples $\mathcal{X} = \{x_i\}_{i=1}^{N}$, we denote $f^T(x_i)$ and $f^S(x_i)$ as the outputs of teacher and student network for sample x_i, which can be the final outputs after softmax or intermediate feature maps from the middle layers. In the preliminary stage, the conventional KD methods mainly focus on transferring individual outputs from teacher to student. For example, the milestone of KD proposed by Hinton et al. [8] makes the student mimic the teacher's behavior by minimizing the Kullback-Leibler divergence between predictions of student and teacher:

$$\mathcal{L}_{KD} = \frac{1}{n} \sum_{x_i \in \mathcal{X}} \mathrm{KL}(\mathrm{softmax}(\frac{f^T(x_i)}{\tau}), \mathrm{softmax}(\frac{f^S(x_i)}{\tau})), \tag{1}$$

where τ is a relaxation hyperparameter referred to as temperature in [8]. Recently, many methods have started to take the relationships among instances as a new kind of knowledge for transfer. Based on the outputs of the network, they construct the instance correlations and minimizes the following objective function:

$$\mathcal{L}_{GKD} = \frac{1}{n^2} \mathcal{D}(G(f^T(x_1), f^T(x_2), ...f^T(x_n)), G(f^S(x_1), f^S(x_2), ...f^S(x_n))), \tag{2}$$

where $\mathcal{D}(\cdot)$ is a loss function that penalizes the difference between correlations of teacher and student, and $G(\cdot)$ is the function to construct the similarity correlation, which in this paper is represented by a correlation matrix. Given feature vectors of n samples, the (i, j) element of similarity correlation G is computed as follows:

$$G_{ij} = \varphi(f(x_i), f(x_j)),\ G \in \mathbb{R}^{n \times n}, \tag{3}$$

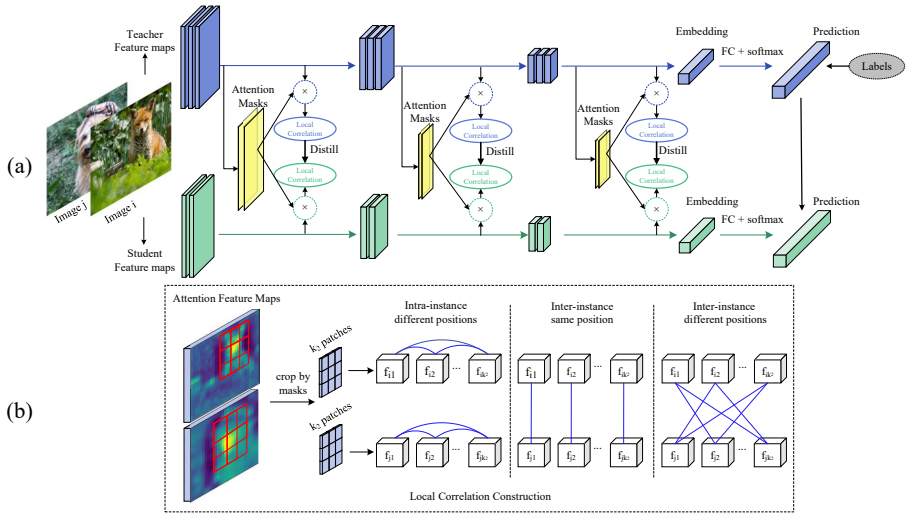

Fig. 2. Overall framework of the proposed method. (a) We supervise the training of the student network by the local correlation consistency losses. Class-aware attention module is trained using the teacher's feature maps to extract those class-specific foreground information before constructing the local correlation matrices of the teacher or student network. (b) We crop the main region of the attention feature maps based on the thresholding attention mask and split it into k^2 patches to investigate the proposed three kinds of local relationships

where φ can be any function that calculates the similarity between two examples, such as the cosine similarity [18] and the Gaussian kernel based similarity [20].

The above similarity correlation among instances has been validated as an effective knowledge for transfer. However, existing methods usually utilize the global features to construct the relationship correlation and neglect the discrimination power implied in local image regions. To make full use of the discriminative local information, we propose our local correlation based knowledge distillation framework.

3.2 Local Correlation Construction

The overall architecture of the proposed method is shown in Figure 2(a). Specifically, we divide both the teacher and the student networks into several stages according to the resolution of the feature maps. For each stage, based on its corresponding feature maps, we investigate three different kinds of local relationships and construct the similarity matrix, after which we minimize the difference of local similarities between the teacher and student models. Before we construct the local correlation, we propose a novel class-aware attention module to extract the semantic foreground area of the image, which will be introduced in the next subsection.

As local information contains more details, we hope to take advantage of this information to learn discriminative correlations and improve the performance of the student model. For the local information, it is a simple way to construct it by dividing the original image or intermediate feature maps into several patches, based on which we can further investigate various correlations. In Figure 2(b), we present the overall procedure of local correlation construction. It models the distilled knowledge of one network stage in a more detailed way, which mainly contains three different kinds of relationships:

(1) Local correlation based intra-instance relationship across different local positions: it corresponds to the relationship between different spatial regions in one image, which can be regarded as a more relaxed way to represent the intermediate features of one image.
(2) Local correlation based inter-instance relationship on the same position: it corresponds to the relationship between regions at the same position among images in one mini-batch, which is a more strict way than the global correlation method to achieve the correlation consistency.
(3) Local correlation based inter-instance relationship across different positions: it corresponds to the relationship between regions at different positions among images in one mini-batch, which contains more abundant knowledge compared with the second relationship and explores more knowledge between local regions without the limitation of position.

For each mini-batch with n images, we compute the correlation matrix of the local regions based on the output feature maps of the teacher and student network. We denote the activation maps produced by the teacher network at l-th stage as $f_l^T \in \mathbb{R}^{n\times c\times h\times w}$, where c, h, w are the size of the channel, height and width, respectively. The corresponding activation maps of the student network can be represented by $f_{l'}^S \in \mathbb{R}^{n\times c'\times h'\times w'}$. Note that c does not necessarily have to equal c' in our method since our correlation-based knowledge transfer method only needs to compute the correlation among features of the same model. For the feature maps f_l^T or $f_{l'}^S$ of each stage, we split it into $k \times k$ patches for each sample and get nk^2 patches for the whole mini-batch, where each patch has the shape of $c \times \frac{h}{k} \times \frac{w}{k}$ or $c' \times \frac{h'}{k} \times \frac{w'}{k}$ (to simplify, here we suppose that h, h', w and w' can be fully divided by k). For the j-th patch from image x_i, we denote $f_l^T(x_{i,j})$ and $f_{l'}^S(x_{i,j})$ as the corresponding local patch features of the teacher and student networks, respectively. After reshaping the features of each patch to a vector, we compute the local correlations we introduced before.

For the first kind of local relationship, it models the intra-instance relationship across different local positions. For the l-th stage, we denote $F_{l,intra}(x_i) = \{f_l(x_{i,1}), ..., f_l(x_{i,k^2})\}$ as the collection of k^2 local features for sample x_i. Then we can define the corresponding loss function in a mini-batch with n samples as:

$$\mathcal{L}_{intra} = \sum_{i=1}^{n}\sum_{l=1}^{L} \|G(F_{l,intra}^T(x_i)) - G(F_{l',intra}^S(x_i))\|_F^2, \tag{4}$$

where $G(\cdot)$ is the function defined in Eq. (3) to construct the similarity matrix, and L is the total number of stages. The permutations of k^2 local features in

$F^T_{l,intra}(x_i)$ and $F^S_{l,intra}(x_i)$ are the same. We adopt the Frobenius norm $\|.\|_F$ to penalize the distance between local correlation matrices computed by student and teacher. For the similarity matrix construction, we use cosine similarity to compute the correlation between the embeddings of two local patches to penalize angular differences.

The second one is the inter-instance relationship on the same local position. Similarly, we denote $F_{l,inter-s}(i) = \{f_l(x_{1,i}), ..., f_l(x_{n,i})\}$ as the collection of local features of the l-th stage, corresponding to the i-th local patch ($i \in [1, 2, \cdots, k^2]$) for n samples of the mini-batch. Then we can define the corresponding loss function as:

$$\mathcal{L}_{inter-s} = \sum_{i=1}^{k^2} \sum_{l=1}^{L} \|G(F^T_{l,inter-s}(i)) - G(F^S_{l',inter-s}(i))\|^2_F. \tag{5}$$

Similarly, the loss function for the third relationship that explores inter-instance relationship across different positions can be defined by:

$$\mathcal{L}_{inter-d} = \sum_{p,q=1,p\neq q}^{k^2} \sum_{i,j=1,i\neq j}^{n} \sum_{l=1}^{L} \left(\varphi\left(f^T_l(x_{i,p}), f^T_l(x_{j,q})\right) - \varphi\left(f^S_{l'}(x_{i,p}), f^S_{l'}(x_{j,q})\right)\right)^2, \tag{6}$$

where $\varphi(\cdot)$ is the function to compute cosine similarity between two feature vectors.

Based on the above loss functions for the above three local relationships, we combine them to get the following overall loss function:

$$\mathcal{L}_{LKD} = \mathcal{L}_{intra} + \mathcal{L}_{inter-s} + \mathcal{L}_{inter-d}. \tag{7}$$

The local correlation based relationships we explored mainly have two advantages. On the one hand, the local features contain more detailed information about this category, which can introduce some discriminative knowledge to facilitate the distillation. For example, many classes in ImageNet belong to a large category. The difference only lies in small local regions, while other regions are very similar. Our local feature based method can well capture and transfer these local patterns, while previous global feature based methods may ignore it. On the other hand, our method investigates various kinds of correlations, which are much more sufficient than previous methods. While the key challenge of knowledge distillation lies in extracting moderate and sufficient knowledge for guidance [16], our method can better guide the learning of the student network.

3.3 Class-Aware Attention

In the previous subsection, we divide the feature map into several non-overlapped patches as the local information. However, the original images also contain a part of unrelated information, which contributes less to the final prediction and may even have a negative influence on the quality of local patches as well as the local

correlation. To solve this issue and extract these high related semantic regions, we introduce a class-aware attention module (CAAT) to filter out the invalid information.

The module consists of two parts: a mask generator and an auxiliary classifier. Supervised by the ground-truth label, CAAT can generate the pixel-level attention mask, which can identify the importance of each pixel and its correlation with the final prediction of the teacher. Given the feature maps of the teacher model $f_l^T \in \mathbb{R}^{n\times c\times h\times w}$, the generated spatial masks $M \in \mathbb{R}^{n\times h\times w}$ can be computed by:

$$M = \mathcal{G}(f_l^T), \tag{8}$$

where $\mathcal{G}(\cdot)$ denotes the mask generator network, which is constructed by a stack of conv-bn-relu blocks followed by the Sigmoid thresholding layer so that each value in the mask is a continuous value between 0 and 1. $M(i,:,:)$ $(i \in [1,n])$ corresponds to the mask for the feature maps of i-th image in the mini-batch. Each value in $M(i,:,:)$ reflects the contribution of the corresponding location to the final prediction of the teacher network. For the same position of different channels, we assign the same mask information. By repeating the mask M along the channel dimension, we can make the mask have the same shape as the feature map f_l^T and $f_{l'}^S$. Then we can get the class-aware attention feature map $\tilde{f}_l^T$ and $\tilde{f}_{l'}^S$ by the following element-wise product:

$$\tilde{f}_l^T = \mathcal{O}_{repeat}(M) \otimes f_l^T, \tilde{f}_{l'}^S = \mathcal{O}_{repeat}(M) \otimes f_{l'}^S, \tag{9}$$

where $\mathcal{O}_{repeat}(.)$ denotes the repeat operation.

To guide the training of network $\mathcal{G}$, we further introduce an auxiliary classifier network $\mathcal{C}$, which takes $\tilde{f}^T$ as input and is supervised by the ground truth label. This sub-network consists of a sequence of bottleneck blocks and utilizes a fully-connected layer for final classification. By minimizing the softmax loss, the auxiliary classifier $\mathcal{C}$ forces the generated mask to pay more attention to informative regions and ignore helpless information like background.

We get the attention feature maps of teacher and student by applying the class-aware attention mask to the original feature maps to highlight those important pixels and weaken those class-irrelevant pixels. Furthermore, we generate a bounding box of the main part of the feature maps based on the thresholding attention mask (the value that larger than threshold $\mathcal{H}$ will be set to 1. The opposite will be set to 0). The top-left point and the right-down point of the bounding box are decided by the boundaries of the thresholding attention mask. We crop the main part of the attention feature maps based on the generated bounding box and divide it into several patches like the way we introduced in the last subsection. Finally, we resize the patches to the same size as the original patch by bilinear interpolation and calculated the local correlation we introduced in the last section. In this part, we modify the proposed losses $\mathcal{L}_{LKD}$ in Eq. (7) by replacing the original local features with the cropped masked local features and then get $\widetilde{\mathcal{L}}_{LKD}$, which is formulated as follows:

$$\widetilde{\mathcal{L}}_{LKD} = \widetilde{\mathcal{L}}_{intra} + \widetilde{\mathcal{L}}_{inter-s} + \widetilde{\mathcal{L}}_{inter-d}. \tag{10}$$

3.4 The Overall Model and Optimization

By combining the cross-entropy loss $\mathcal{L}_{CE}$ supervised by the ground truth labels, the classic KD loss $\mathcal{L}_{KD}$, and the proposed local correlation based consistency loss $\widetilde{\mathcal{L}}_{LKD}$, we come up with the final overall loss function:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{CE} + \alpha\mathcal{L}_{KD} + \beta\widetilde{\mathcal{L}}_{LKD}, \tag{11}$$

where α, β are hyper-parameters to balance contributions of different terms.

During training, we first optimize network $\mathcal{G}$ and $\mathcal{C}$ by minimizing the softmax loss. Then we fix the parameters of mask generator $\mathcal{G}$, and train the student network by minimizing the overall loss function in Eq. (11).

3.5 Complexity Analysis

We present the computational complexity in training a mini-batch. The computational complexities of Eqs. (4), (5) and (6) for l-th stage of teacher are $\mathcal{O}(nk^2chw)$, $\mathcal{O}(nchw)$ and $\mathcal{O}(n^2k^2chw)$, respectively. Therefore, the total computational complexity of our method is $\mathcal{O}(n^2k^2chw)$. For comparison, SP [27] has $\mathcal{O}(n^2chw)$ complexity and CCKD [20] has $\mathcal{O}(n^2pd)$ complexity, where p and d correspond to the p-order Taylor-series and dimension of feature embedding. In fact, k is very small in our method. For example, we set k to 4 on CIFAR100, and 3 on ImageNet. In this case, the complexity of our method is comparable and in the same order with these conventional KD methods. Besides, the complexity of Eq. (4) is much smaller than SP and CCKD. With only Eq. (4) as the loss function, the accuracy of our method is also better than SP and CCKD, which will be proved by the ablation study. Therefore, when the computation resources are limited, you can only use this term as the loss function.

4 Experiments

In this section, we conduct several experiments to demonstrate the effectiveness of our proposed local graph supervision as well as the class-aware attention module. We first compare the results on CIFAR100 [13] and ImageNet [3] with four knowledge distillation methods, including Hinton's traditional knowledge distillation (KD) [8], attention transfer (AT) [38], similarity-preserving knowledge distillation (SP) [27], and correlation congruence knowledge distillation (CC) [20]. Besides, cross-entropy (CE) loss is also chosen as a baseline. Then we perform ablation studies to evaluate the effect of different modules.

4.1 Evaluation on CIFAR100

The CIFAR100 dataset contains 100 classes. For each class, there are 500 images in the training set and 100 images in the testing set. Similar to the settings in [20], we randomly crop 32×32 image from zero-padded 40×40 image, and apply random horizontal flipping for data augmentation. SGD is used to optimize the

Table 1. Comparison of classification accuracy on CIFAR100. The best results of the student network are highlighted in bold

Teacher Net.	Student Net.	CE	KD	AT	SP	CC	LKD	Teacher
ResNet110	ResNet14	67.45	69.78	69.51	69.59	69.77	**70.48**	75.76
ResNet110	ResNet20	69.47	71.47	71.8	71.42	71.78	**72.63**	75.76
WRN-40-2	WRN-16-1	66.79	66.74	66.75	66.4	66.76	**67.72**	75.61
WRN-40-2	WRN-16-2	73.1	74.89	75.15	74.69	75.05	**75.44**	75.61

model with batch size 64, momentum 0.9, and weight decay $5e^{-4}$. For the class-aware attention module, we train the mask generators and auxiliary classifiers for 60 epochs with learning rate starting from 0.05 and multiplied by 0.1 at 30, 40, 50 epochs. The threshold $\mathcal{H}$ is set to 0 because most of the images in CIFAR100 are occupied by the main object. For the extraction of the local features, we set $k = 4$ for all the stages to split the feature maps to 16 patches. Then we train the student network for 200 epochs with the learning rate starting from 0.1 and multiplied by 0.1 at 80, 120, 160 epochs. For CE, we set $\alpha = 0$ in Eq. (11). For traditional KD, AT, CC, SP and our methods, we set $\alpha = 1$ and $\tau = 4$ following the CIFAR100 experiments in [38]. For a fair comparison, we carefully tune the loss weight of all the methods by grid-search for each teacher-student pair and report the average accuracy over 3 runs with the chosen loss weight. $\beta \in [0.001, 0.1]$ works reasonably well for our methods.

We also test the performance under four combinations of teacher and student networks using ResNet [7] and Wide ResNet (WRN) [37]. For the teacher network of ResNet110, the accuracy is 75.76%, and we adopt ResNet14 and ResNet20 as two different student networks. For the teacher network of WRN-40-2, the accuracy is 75.61%, and we adopt WRN-16-1 and WRN-16-2 as two different student networks.

In Table 1, we show the results of different methods on CIFAR100. We can see that our proposed LKD method achieves the best performance under all these four different settings of the teacher and student networks, which can demonstrate the effectiveness and robustness of our method. Based on the results, we also have the following observations. First, our method substantially surpasses the baseline methods KD and AT by a large margin. While these two methods mainly minimize the distance between instance features of the teacher and student models, our improvement can verify that mimicking the correlation between local regions of the feature maps is a more effective way. Second, we find that compared with these methods with global feature based correlation, including SP and CC, our local features based correlation consistency shows the superiority, which can be attributed to the sufficient details and discriminative patterns that local features contain.

Table 2. Comparison of classification accuracy on ImageNet. The best results of the student network are highlighted in bold

Accuracy	CE	KD	AT	SP	CC	LKD	Teacher
Top-1	70.58	71.34	71.33	71.38	71.45	**71.54**	73.27
Top-5	89.45	90.27	90.26	90.28	90.26	**90.30**	91.27

4.2 Evaluation on ImageNet

After successfully demonstrating our method's superiority on the relatively small CIFAR100 dataset, we move to validate its effectiveness on the large-scale ImageNet dataset, which contains 128k training images and 50k testing images. The resolution of input images after pre-processing in ImageNet is 224×224, which is much larger than that in CIFAR100. With more images and larger resolution, classification on ImageNet is more challenging than that on CIFAR100.

Following the setting in AT [38], we adopt ResNet34 as the teacher network and ResNet18 as the student network. Mask generators and auxiliary classifiers are trained for 48 epochs with learning rate starting from 0.8 and multiplied by 0.1 at 36, 44 epochs. The threshold $\mathcal{H}$ is set to 0.1 for the cropping of the attention feature maps. The local relationships based loss function $\widetilde{\mathcal{L}}_{LKD}$ is added on the last stage of the network following the implementation of SP [27] with the loss scale $\beta = 0.5$. The patch number k is set to 3. The student network is trained for 120 epochs with mini-batch size 1024 (on 16 GPUs, each with batch size 64 and weight decay $4e{-}4$). The learning rate starts from 0.4 and is multiplied by 0.1 at 40, 72, and 96 epochs. The α is set to 1 with temperature $\tau = 2$.

In Table 2, we compare the classification accuracy with other methods on ImageNet. We can see that our method continuously outperforms the competing methods on both Top-1 and Top-5 accuracy. Because the ImageNet dataset is very challenging, our small improvement is also very hard. The above result further demonstrates the effectiveness of our LKD on the large-scale and high-resolution dataset.

4.3 Ablation Study

To verify the effectiveness of each of the three kinds of local relationships based knowledge and the class-aware attention module in our method, we conduct ablation studies on CIFAR100 with ResNet110 as the teacher network and ResNet14, ResNet20 as the student networks. Results are shown in Table 3. By adding each of these three local relationship based loss functions into the baseline KD method, the result can be stably improved. By combining these three loss functions, it can achieve a much better result. Based on $\widetilde{\mathcal{L}}_{LKD}$, our class-aware attention module can further improve the performance. The above results can sufficiently show the effectiveness of each local correlation based knowledge as well as the attention module. Besides, we can observe similar results with both ResNet14 and ResNet20, which also demonstrates the robustness and generalization ability of our contributions.

Table 3. Ablation study on CIFAR100. *intra*, *inter-same* and *inter-diff* denote three local relationships introduced in Section 3. CAAT is the class-aware attention module

Methods	Local Relationships			CAAT	Top1 accuracy	
	intra	*inter-same*	*inter-diff*		ResNet14	ResNet20
$\mathcal{L}_{KD}$					69.78	71.47
$\mathcal{L}_{KD} + \mathcal{L}_{intra}$	✓				70.00	72.04
$\mathcal{L}_{KD} + \mathcal{L}_{inter-s}$		✓			70.20	71.96
$\mathcal{L}_{KD} + \mathcal{L}_{inter-d}$			✓		70.03	72.10
$\mathcal{L}_{KD} + \mathcal{L}_{LKD}$	✓	✓	✓		70.37	72.31
$\mathcal{L}_{KD} + \widetilde{\mathcal{L}}_{LKD}$	✓	✓	✓	✓	**70.48**	**72.63**

Table 4. Results on CIFAR100 with different number of k, which denotes how many patches that we divide the feature map into along each axis

Student	LKD(k=1)	LKD(k=2)	LKD(k=4)
ResNet14	69.98	70.09	70.37
ResNet20	71.82	71.86	72.31

4.4 Sensitivity Analysis

Influence of the Parameter k. To extract local features, recall that we split the foreground feature map of each image into $k \times k$ patches. In the above experiments on CIFAR100, we simply set $k = 4$ on CIFAR100. In this part, we purely evaluate the performance of the student network with different k. For simplification, we only add the local correlation based loss on ResNet14 and ResNet20 and do not add the class-aware attention module. The results are presented in Table 4. We can observe that with the increase of k, the performance is improved gradually. The results with $k = 4$ obviously surpasses that of $k = 1$ and $k = 2$. The reason is that the larger k we use to extract the local features, the more sufficient knowledge we will extract from the teacher to transfer, which can bring the performance improvement in return.

Effect of Class-Aware Attention. In this part, we evaluate the effect of our class-aware attention module and show whether it can filter out the invalid information. We conduct experiments on CIFAR100 with several different attention methods, including the activation-based attention in AT [38] and Grad-CAM [21]. In all experiments, the grid number k is set to 4, and the baseline experiment is conducted without the attention module. For a fair comparison, we utilize a sigmoid function to normalize the attention masks obtained by all the attention methods mentioned above.

The results are summarized in Table 5. We can see that our proposed CAAT module works much better than other attention methods as well as the baseline.

We also visualize the attention masks of some sample images in Figure 3. We can find that the informative regions are assigned relatively high value while the confusing background regions are on the contrary. The mask generated by CAAT

Table 5. Top-1 accuracy on CIFAR100 for LKD with different attention methods

Student	LKD	LKD+AT	LKD+Grad-CAM	LKD+CAAT
ResNet14	70.37	69.88	70.41	**70.48**
ResNet20	72.31	72.34	72.18	**72.63**

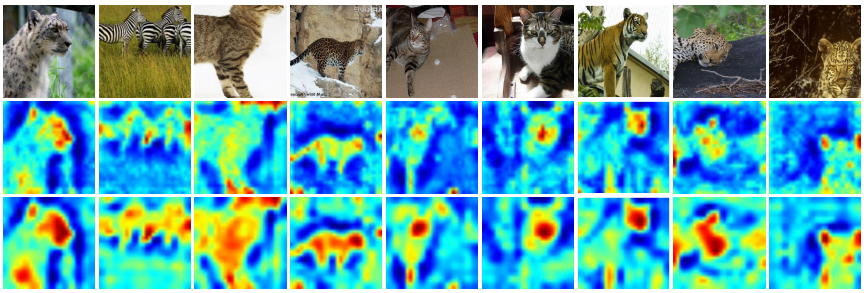

Fig. 3. Visualization of the attention maps. First row: images sampled from ImageNet. Second row: original feature maps generated by the teacher model. Third row: corresponding attention masks generated by CAAT module at the third stage of teacher network. High value is shown in red and low value in blue

can well filter out the background that has less contribution to the classification task. And more importantly, it helps the student network to focus on those class-relevant regions and ignore these confusing regions in images, such as the messy background of all the images in Figure 3.

5 Conclusions

In this paper, we proposed the local correlation consistency: a novel form of knowledge distillation that aims to represent the relationships of local regions in the feature space. By minimizing the local correlation matrices of teacher and student, we could make the student generate more discriminative local features. Furthermore, we applied a class-aware attention mask to both the teacher and the student's feature maps before constructing the local correlation matrices. We trained the class-aware attention module using teacher's feature maps to highlight those informative and class-relevant regions and weaken the effect of those confusing regions. Our Experiments on CIFAR100 and ImageNet demonstrate the effectiveness of the proposed local correlation consistency knowledge distillation and the class-aware attention module.

Acknowledgment.
Jianlong Wu is the corresponding author, who is supported by the Fundamental Research Funds and the Future Talents Research Funds of Shandong University.

References

1. Aguilar, G., Ling, Y., Zhang, Y., Yao, B., Fan, X., Guo, E.: Knowledge distillation from internal representations. In: AAAI (2020)
2. Chung, I., Park, S., Kim, J., Kwak, N.: Feature-map-level online adversarial knowledge distillation. In: ICML (2020)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE CVPR. pp. 248–255 (2009)
4. Dong, X., Yang, Y.: Network pruning via transformable architecture search. In: NeurIPS. pp. 759–770 (2019)
5. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In: ICLR (2016)
6. Han, S., Pool, J., Tran, J., Dally, W.: Learning both weights and connections for efficient neural network. In: NeurIPS. pp. 1135–1143 (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR. pp. 770–778 (2016)
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NeurIPS Deep Learning Workshop (2014)
9. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE CVPR. pp. 7132–7141 (2018)
11. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y.: Quantized neural networks: Training neural networks with low precision weights and activations. Journal of Machine Learning Research **18**(1), 6869–6898 (2017)
12. Kim, H.G., Na, H., Lee, H., Lee, J., Kang, T.G., Lee, M.J., Choi, Y.S.: Knowledge distillation using output errors for self-attention end-to-end models. In: ICASSP. pp. 6181–6185 (2019)
13. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech Report (2009)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS. pp. 1097–1105 (2012)
15. Lan, X., Zhu, X., Gong, S.: Knowledge distillation by on-the-fly native ensemble. In: NeurIPS. pp. 7528–7538 (2018)
16. Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. In: IEEE CVPR. pp. 7096–7104 (2019)
17. Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J.: Pruning convolutional neural networks for resource efficient inference. In: ICLR (2017)
18. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: IEEE CVPR. pp. 3967–3976 (2019)
19. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC (2015)
20. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: IEEE ICCV. pp. 5007–5016 (2019)
21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE ICCV. pp. 618–626 (2017)
22. Shu, C., Li, P., Xie, Y., Qu, Y., Dai, L., Ma, L.: Knowledge squeezed adversarial network compression. arXiv preprint arXiv:1904.05100 (2019)

23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
24. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: NeurIPS. pp. 1988–1996 (2014)
25. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
26. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: ICLR (2020)
27. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: IEEE ICCV. pp. 1365–1374 (2019)
28. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Change Loy, C.: The devil of face recognition is in the noise. In: ECCV. pp. 765–780 (2018)
29. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: IEEE CVPR. pp. 3156–3164 (2017)
30. Wu, J., Long, K., Wang, F., Qian, C., Li, C., Lin, Z., Zha, H.: Deep comprehensive correlation mining for image clustering. In: IEEE ICCV. pp. 8150–8159 (2019)
31. Wu, J., Leng, C., Wang, Y., Hu, Q., Cheng, J.: Quantized convolutional neural networks for mobile devices. In: IEEE CVPR. pp. 4820–4828 (2016)
32. Yang, L., Song, Q., Wu, Y., Hu, M.: Attention inspiring receptive-fields network for learning invariant representations. IEEE TNNLS **30**(6), 1744–1755 (2018)
33. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: IEEE CVPR. pp. 7130–7138 (2017)
34. You, S., Huang, T., Yang, M., Wang, F., Qian, C., Zhang, C.: Greedynas: Towards fast one-shot nas with greedy supernet. In: IEEE CVPR. pp. 1999–2008 (2020)
35. You, S., Xu, C., Xu, C., Tao, D.: Learning from multiple teacher networks. In: KDD. pp. 1285–1294 (2017)
36. You, S., Xu, C., Xu, C., Tao, D.: Learning with single-teacher multi-student. In: AAAI. pp. 4390–4397 (2018)
37. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)
38. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)
39. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: IEEE ICCV. pp. 3713–3722 (2019)
40. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: IEEE CVPR. pp. 6848–6856 (2018)
41. Zhang, Z., Ning, G., He, Z.: Knowledge projection for deep neural networks. arXiv preprint arXiv:1710.09505 (2017)