Exploring Inter-Channel Correlation for Diversity-preserved Knowledge Distillation

Li Liu[1†] Qingle Huang[1†] Sihao Lin[2†‡] Hongwei Xie[1] Bing Wang[1] Xiaojun Chang[3*] Xiaodan Liang[4]

[1]Alibaba Group [2]Monash University [3]RMIT University [4]Sun Yat-sen University

{liuli.ll9412, qingle.hql, linsihao6, hongwei.xie.90, xdliang328}@gmail.com

fengquan.wb@alibaba-inc.com, xiaojun.chang@rmit.edu.au

Abstract

Knowledge Distillation has shown very promising ability in transferring learned representation from the larger model (teacher) to the smaller one (student). Despite many efforts, prior methods ignore the important role of retaining inter-channel correlation of features, leading to the lack of capturing intrinsic distribution of the feature space and sufficient diversity properties of features in the teacher network. To solve the issue, we propose the novel Inter-Channel Correlation for Knowledge Distillation (ICKD), with which the diversity and homology of the feature space of the student network can align with that of the teacher network. The correlation between these two channels is interpreted as diversity if they are irrelevant to each other, otherwise homology. Then the student is required to mimic the correlation within its own embedding space. In addition, we introduce the grid-level inter-channel correlation, making it capable of dense prediction tasks. Extensive experiments on two vision tasks, including ImageNet classification and Pascal VOC segmentation, demonstrate the superiority of our ICKD, which consistently outperforms many existing methods, advancing the state-of-the-art in the fields of Knowledge Distillation. To our knowledge, we are the first method based on knowledge distillation boosts ResNet18 beyond 72% Top-1 accuracy on ImageNet classification. Code is available at: https://github.com/ADLab-AutoDrive/ICKD.

1. Introduction

It is widely witnessed that larger networks are superior in learning capacity compared to smaller ones. Nevertheless, due to the great amount of energy consumption and computation costs, a large network (*e.g.*, ResNet-50 [9]), though powerful, is difficult to deploy on mobile systems. Hence, there is a growing interest in reducing the model size while
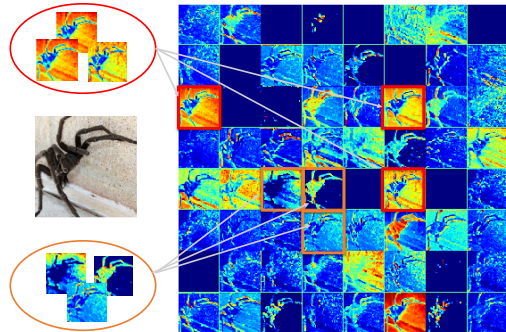

Figure 1: **Illustration of inter-channel correlation**. The channels orderly extracted from the second layer of ResNet18 have been visualized. The channels denoted by red boxes are homologous both perceptually and mathematically (*e.g.*, inner-product), while the channels denoted by orange boxes are diverse. We show the inter-channel correlation can effectively measure that each channel is homologous or diverse to others, which further reflects the *richness* of the feature spaces. Based on this insightful finding, our ICKD can enforce the student to *mimic* this property from the teacher.

preserving comparable performance, which bridges the gap between small networks and large networks.

Knowledge distillation is one of the promising methods to this problem. It is acknowledged that Bucila *et al*. [1] introduced the idea of knowledge distillation and Hinton *et al*. [12] further popularized this concept. The key idea of knowledge distillation is to let the student network *mimic* the teacher model. The underlying principle is that teachers can provide the knowledge that ground truth labels can not tell. Despite its success, this technique, devoted to instance-level classification, may lead the student to mainly learn the instance-level information but not structural information, which limits its application. Prior works [19, 20, 21, 27, 28] have been proposed to help the student network learn the structural representation for better generalization ability. These methods generally utilize the cor-

---

relation of the instances to describe the geometry, similarity, or dissimilarity in the feature space. We call this fashion layer-wise relational knowledge distillation since they mainly focus on exploring the correlation between feature maps in the level of layer. Conversely, we pay more attention to the inter-channel correlations.

Previous works [16, 25] make use of knowledge distillation to reduce the homology (*i.e.*, redundancy) of the student's feature space. Nevertheless, the success of Ghost-Net [7] suggests that small neural networks benefit from increased feature homology. The rich representation can empower the downstream tasks and both the diversity and homology can reflect the richness. Existing literature neglects the importance of feature diversity and homology, yielding an issue that the proportion of feature diversity versus homology may be unbalanced against our expectation that students can learn the representation as rich as the teacher is for better generalization. In Fig. 1, the visualized feature maps show that feature diversity and homology co-exist in the networks. This property can be disclosed by the correlation between channels, where high relevance represents homology and low relevance represents diversity. In this paper, we adopt the Inter-Channel Correlation (**ICC**) as the indicator of the diversity and homology of the feature distribution. However, figuring out the optimal inter-channel correlations manually is impractical. An intuitive solution is to let the student learn better inter-channel correlation from the teacher, as shown in Fig. 2. Due to the discrepancy of learning capacities [4], it is not viable to force the student to mimic the whole feature map of the teacher. Instead, we let the student model learn the inter-channel correlation from the teacher, namely inter-channel correlation knowledge distillation (**ICKD**).

The correlation between the two channels is evaluated by the inner product in this paper. As the inner product collapses the spatial dimension, it naturally does not need to constrain the feature map spatial size of the teacher network and student network to be the same. On the other hand, when it comes to a large feature map, *e.g.* semantic segmentation models, the mapping between the inter-channel correlation measured by the inner product and the original feature space is of high freedom. Thus it will be more difficult to distillate the inter-channel correlation distribution to anchor the teacher's feature space distribution. To alleviate this problem, we propose a grid-level inter-channel correlation distillation method. By dividing the feature map of size $h \times w \times c$ by a pre-defined grid into $n \times m$ patches of size $h_G \times w_G \times c$. Distillation on patch-level is more controllable and we can perform the distillation on the entire feature map by aggregating the inter-channel correlation distillation across these patches. In addition, the local spatial information can be preserved since each patch can keep the knowledge in specific region.

In our experiments, we have evaluated our proposed method in different tasks including classification (Cifar-100 and ImageNet) and semantic segmentation (Pascal VOC). The proposed method shows a performance superior to the existing the state-of-the-arts methods. To our knowledge, we are the first knowledge distillation method that boosts ResNet18 beyond 72% Top-1 accuracy on ImageNet classification. And on Pascal VOC, we achieved 3% mIOU improvement compared to the baseline model.

To summarize, our contributions are:

- We introduce the inter-channel correlation, with the characteristics of being invariant to the spatial dimension, to explore and measure both the feature diversity and homology to help the student for better representation learning.

- We further introduce the grid-level inter-channel correlation to make our framework capable for dense prediction task, like semantic segmentation.

- To validate the effectiveness of the proposed framework, extensive experiments have been conducted on different (a) network architectures, (b) downstream tasks and (c) datasets. Our method consistently outperforms the state-of-the-arts methods a large margin across a wide range of knowledge transfer tasks.

2. Related Works

Knowledge Distillation. Given by [12], the student network is required to minimize the KL-divergence between the logits (before softmax) output by the student and teacher, where a temperature τ is applied to soften the logits. This procedure, making it different from the ground truth label, will increase the low probability in the logits, which is referred to *dark knowledge*.

To learn more generic representation, recent works [19, 20, 21, 27, 28] explored the structural information within the feature space of the teacher and transferred it to the student. Tung and Mori [28] measured the similarity among the given instances in the teacher's feature space and asked the student to match the same similarity. Peng *et al.* [21] presented the kernel-based correlation congruence. A kernel function was employed to measure the correlation metric of each paired instances in the feature space. Similarly, the student is required to share the same correlation metric with the teacher. RKD [19] further introduced the angle-wise relation given a triplet of instances. More recently, Tian *et al.* [27] introduced contrastive learning to maximized the mutual information between the representation of the student and teacher.

Romero *et al.* [22] proposed the distillation in the intermediate layers between the student (*i.e.*, *guided* layer) and the teacher (*i.e.*, *hints* layer). The student is taught to minimize the Euclidean distance of the feature maps from
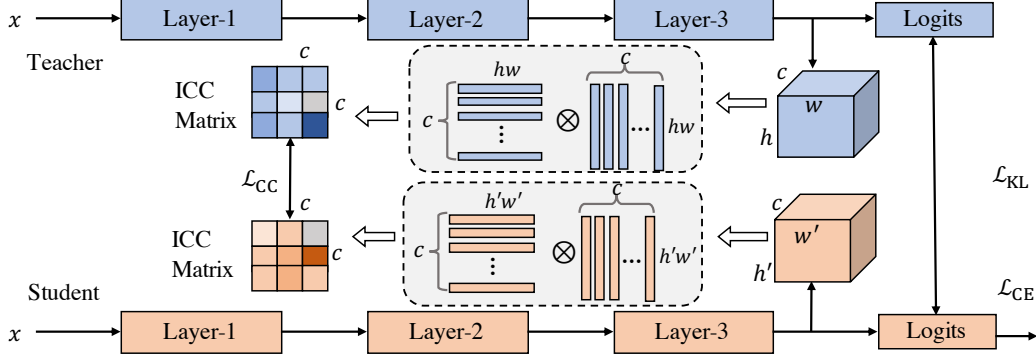
Figure 2: **Illustration of the proposed ICKD**. We measure the inter-channel correlation of the teacher feature and ask the student to share with the same property. The cubes represent the 3D feature tensors extracted from the teacher and student. They are flattened to the corresponding 2D matrices which are used to compute the ICC matrices. We minimize the MSE between the ICC matrices associated with the feature tensors. The student is also asked to minimize the KL divergence between the logits of the teacher and student. Finally the cross-entropy loss is applied on the student.

guided layer and hints layer. Because the semantic information contained within the feature varies from layer to layer according to depth and width, existing works [15, 34, 2] has shown that layer-wisely match a pair of guided layer and hints layer may not be an optimal choice. AT [34] proposed a statistical method to highlight the attention, compressing the 3D feature tensor to a 2D feature map. Chen *et al.* [2] proposed semantic calibration to assign the target teacher layer to the student layer across layers depending on the inner-products of the teacher layers and the student layers. Ji *et al.* [15] measured the similarities, bounded to 1 with a softmax function, between the teacher and student features, which was used as the weights to balance the feature matching.

Semantic Segmentation. In spite of great challenge, some approaches based on knowledge distillation had been proposed in semantic segmentation. He *et al.* [10] pre-trained an auto-encoder to match the features between the student and teacher, which also measured the affinity matrix of the paired instances in the teacher network and transferred it to the student network. Liu *et al.* [17] proposed the structured knowledge distillation consisting of the pair-wise similarity transfer and pixel-wise distillation like [12]. Liu *et al.* also transferred the holistic knowledge via adversarial learning. Wang *et al.* [30] proposed the Intra-class Feature variation Distillation that also measured the pair-wise similarity between the features of each pixel and that of the corresponding class-wise prototype. Heo *et al.* [11] proposed a distillation loss with a designed margin ReLU to boost the performance of a student on semantic segmentation.

We further extend the framework with the grid-level inter-channel correlation for stabilizing the distillation process and preserving the spatial information. Perhaps our work is most closed to Huang and Wang [13] which utilizes the Gram Matrix [6]. Yim *et al.* [32] proposed the flow of

solution procedure that computed the Gram matrix across layers. The difference is that [13, 32] measure the relation between pixel-wise positions and we explore the correlation between two channels.

3. Method

In this section, we first briefly introduce the preliminary of knowledge distillation. Then we formulate the proposed method to show how we can compute the ICC matrix. Finally, we extend the framework with the grid-level inter-channel correlation.

3.1. Preliminary

Let $\mathcal{X}^N$ denote a set of distinct examples with cardinality N. Suppose that we have a teacher model T and a student model S, which are denoted by f^T and f^S, respectively. In practice, f^T and f^S can be any differential function and we parameterize them as convolutional neural network (CNN) here. $F^T \in \mathbb{R}^{c \times h \times w}$ represents the embedding in the teacher network, where c is the number of output channels, h and w represents the height and width of the feature map. Similarly, let $F^S \in \mathbb{R}^{c' \times h' \times w'}$ denote the embedding in the student network. In general, traditional knowledge distillation attempts to minimize the divergence between the embedding of the student and the teacher, in [12] the formulation can be described as:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{N} \sum_{i=1}^{N} D_{\mathrm{KL}}(\sigma(\frac{f^T(x_i)}{\tau}), \sigma(\frac{f^S(x_i)}{\tau})), \quad (1)$$

where $D_{\mathrm{KL}}(\cdot, \cdot)$ measures the Kullback-Leibler divergence, $\sigma(\cdot)$ is the softmax function, τ is the temperature factor, $f^T(x)$ and $f^S(x)$ represent the outputs of the penultimate layer (before softmax) in the teacher network and the student network, respectively.

3.2. Formulation

In this section we introduce the formulation of inter-channel correlation. Given two channels, the correlation metric should return a value reflecting their relevance. A high value indicates homologous otherwise diverse. Ultimately all the correlation metrics are gathered sequentially to represent the holistic diversity of the channels. The inter-channel correlation can be defined by

$$\mathcal{G}^{F^T}_{m,n} = K(\mathrm{v}(F^T_m), \mathrm{v}(F^T_n)), \quad (2)$$

where $F^T_m \in \mathbb{R}^{h \times w}$ denotes the m-th channel of the feature F^T, $\mathrm{v}(\cdot)$ vectorizes a 2D feature map into a vector with length hw, and $K(\cdot)$ is a function that measure the correlation of a input pair, where inner product is employed. Note that Eq. 2 returns a scalar in spite of the spatial dimensions of the channel. This can be rewritten in a manner of matrix multiplication, forming our ICC matrix:

$$\mathcal{G}^{F^T} = \mathrm{f}(F^T) \cdot \mathrm{f}(F^T)^\top, \quad (3)$$

where $\mathrm{f}(F^T) \in \mathbb{R}^{c \times hw}$ flattens the spatial dimensions. The resulting ICC matrix has a size of $c \times c$ regardless of the spatial dimensions h and w. Following the empirical setting [29, 10], we add a linear transformation layer C_l on top of the feature of the student, which consists of a convolution layer with 1×1 kernels and a BN layer without activation function. In case that the output dimension c' of the student mismatches with that of the teacher, C_l can adapt F^S to match the output dimension c of F^T. This procedure would not change the spatial dimensions. We penalize the L2 distance between the ICC matrices of the student and the teacher, allowing the student to obtain similar feature diversity.

$$\mathcal{L}_{\mathrm{CC}} = \frac{1}{c^2} ||\mathcal{G}^{C_l(F^S)} - \mathcal{G}^{F^T}||^2_2. \quad (4)$$

We refer the method described above as ***ICKD-C***, which is mainly developed for image classification. Finally, the objective of our method is given by

$$\mathcal{L}_{\mathrm{ICKD-C}} = \mathcal{L}_{\mathrm{CE}} + \beta_1 \mathcal{L}_{\mathrm{KL}} + \beta_2 \mathcal{L}_{\mathrm{CC}}, \quad (5)$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross entropy loss, β_1 and β_2 are the weight factors.

3.3. Grid-Level Inter-Channel Correlation

In Eq. 3, we simply flatten the entire 3D feature map into the corresponding 2D matrix and then calculate the ICC matrix. When coming into semantic segmentation, the final feature map can be very large, *e.g.* $256 \times 128 \times 128$. The correlation of two channels is generated by the inner-product of two vectors of length 16,384. Generally, these two vectors can be seen as sampled from an independent distribution, the correlation value will be of a very small order of magnitude, that means the correlation result is vulnerable to noise. In this situation, the inter-channel correlation of the student
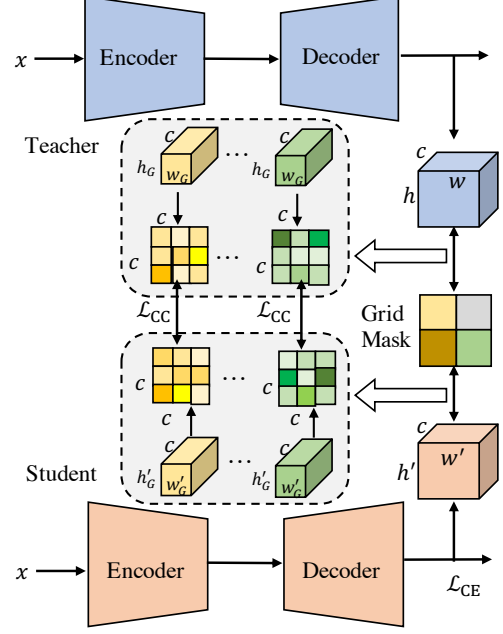

Figure 3: **Grid-level Inter-Channel Correlation**. We evenly divide the original feature into $n \times m$ parts and compute their ICC matrix individually. We then minimize the MSE on each paired ICC matrices.

model in training process may be unstable. Motivated by the divide-and-conquer, we seek to split the feature map and then perform knowledge distillation individually.

Based on this idea, we introduce the grid-level inter-channel correlation. We evenly partition the feature F^T into $n \times m$ parts along the pixel position, denoted by $F^T_{(i,j)}$, $i = 1, 2, ..., n$, $j = 1, 2, ..., m$. Each part is of size $c \times h_G \times w_G$ where $h_G = h/n$ and $w_G = w/m$. Each part presents the semantic on a patch level. The ICC matrix of each part is computed individually as described in Sec. 3.2. Then all the ICC matrices are aggregated.

$$\mathcal{G}^{F^T_{(i,j)}} = \mathrm{f}(F^T_{(i,j)}) \cdot \mathrm{f}(F^T_{(i,j)})^\top, \quad (6)$$

$$\mathcal{G}^{F^S_{(i,j)}} = \mathrm{f}(F^S_{(i,j)}) \cdot \mathrm{f}(F^S_{(i,j)})^\top, \quad (7)$$

$$\mathcal{L}^{n \times m}_{\mathrm{CC}} = \frac{1}{n \times m \times c^2} \sum_i^n \sum_j^n ||\mathcal{G}^{F^T_{(i,j)}} - \mathcal{G}^{F^S_{(i,j)}}||^2_2. \quad (8)$$

As depicted in Fig. 3, we use a Grid Mask to evenly divide the whole feature into different group. Despite the change of spatial dimensions, the size of the resulting ICC matrix always depend on the numbers of channels, *i.e.*, c. In addition, the grid division also helps to extract more spatial or local information, which is beneficial in correctly classifying pixels for semantic segmentation [31]. This variant is referred to ***ICKD-S***. Finally, the objective for semantic segmentation is formulated as:

$$\mathcal{L}_{\mathrm{ICKD-S}} = \mathcal{L}_{\mathrm{Seg}} + \alpha \mathcal{L}^{n \times m}_{\mathrm{CC}}, \quad (9)$$

Table 1: Top-1 accuracy (%) in Cifar-100 testing set. Methods are divided into two groups. The performance of each method against traditional KD [12] are reported. For fair comparison, we also report the performance of our method without $\mathcal{L}_{\mathrm{KL}}$. We find that our ICKD-C outperforms all the other methods.

Method	Network Architecture						
	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	ResNet56 ResNet20	ResNet110 ResNet20	ResNet110 ResNet32	ResNet32×4 ResNet8×4	VGG13 VGG8
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Vanilla	73.26	71.98	69.06	69.06	71.14	72.50	70.36
KD [12]	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet [22]	$73.58_{-1.34}$	$72.24_{-1.30}$	$69.21_{-1.45}$	$68.99_{-1.68}$	$71.06_{-2.02}$	$73.50_{+0.17}$	$71.02_{-1.96}$
AT [34]	$74.08_{-0.84}$	$72.77_{-0.77}$	$70.55_{-0.11}$	$70.22_{-0.45}$	$72.31_{-0.77}$	$73.44_{+0.11}$	$71.43_{-1.55}$
SP [28]	$73.83_{-1.09}$	$72.43_{-1.11}$	$69.67_{-0.99}$	$70.04_{-0.63}$	$72.69_{-0.39}$	$72.94_{-0.39}$	$72.68_{-0.20}$
CC [21]	$73.56_{-1.36}$	$72.21_{-1.33}$	$69.63_{-1.03}$	$69.48_{-1.19}$	$71.48_{-1.6}$	$72.97_{-0.36}$	$70.71_{-2.27}$
RKD [19]	$73.35_{-1.57}$	$72.22_{-1.32}$	$69.61_{-1.05}$	$69.25_{-1.42}$	$71.82_{-1.26}$	$71.90_{-1.43}$	$71.48_{-1.5}$
PKT [20]	$74.54_{-0.38}$	$73.45_{-0.09}$	$70.34_{-0.32}$	$70.25_{-0.42}$	$72.61_{-0.47}$	$73.64_{+0.31}$	$72.88_{-0.10}$
FSP [32]	$72.91_{-2.01}$	NA	$69.95_{-0.71}$	$70.11_{-0.56}$	$71.89_{-1.19}$	$72.62_{-0.71}$	$70.20_{-2.78}$
NST [13]	$73.68_{-1.24}$	$72.24_{-1.3}$	$69.60_{-1.06}$	$69.53_{-1.14}$	$71.96_{-1.12}$	$73.30_{-0.03}$	$71.53_{-1.45}$
ICKD-C (w/o $\mathcal{L}_{\mathrm{KL}}$)	$\mathbf{75.64}_{+0.72}$	$74.33_{+0.79}$	$\mathbf{71.76}_{+1.1}$	$71.68_{+1.01}$	$73.89_{+0.81}$	$75.25_{+1.92}$	$73.42_{+0.44}$
ICKD-C (Ours)	$75.57_{+0.65}$	$\mathbf{74.63}_{+1.09}$	$71.69_{+1.03}$	$\mathbf{71.91}_{+1.24}$	$\mathbf{74.11}_{+1.03}$	$\mathbf{75.48}_{+2.15}$	$\mathbf{73.88}_{+0.9}$

where α is the weight factor and $\mathcal{L}_{\mathrm{Seg}}$ is the supervised segmentation loss. $\mathcal{L}_{\mathrm{Seg}}$, though, can be replaced with other loss for different downstream tasks, this is not the focus of this paper.

4. Experiments

We evaluate the effectiveness of the proposed model on two vision tasks: image classification and semantic segmentation. For image classification, we conduct the experiments on Cifar-100 and ImageNet. To verify the generalization of our framework, we further conduct experiments of semantic segmentation on the large-scale benchmark Pascal VOC.

4.1. Datasets

ImageNet. This dataset has about 1.2**M** training samples labeled into 1,000 categories. The images are resized to 224×224 for both training and testing. Usually, the performance of a model is measured by Top-1 and Top-5 classification accuracy.

Cifar-100. This dataset contains 50,000 training images and 10,000 testing images, labeled into 100 categories. Each image is of size $32 \times 32 \times 3$. Top-1 classification accuracy is adopted to measure the model.

Pascal VOC. This dataset contains 20 foreground object classes plus an extra background class. It has 1,464 images for training, 1,499 images for validation and 1,456 images for testing. We also include the coarse annotated training images from [8], resulting in 10,582 training images in total. We employ mean Intersection over Union (mIoU) to evaluate the effectiveness of the proposed model.

4.2. Implementation Details

For image classification, the feature map before the global average pooling layer is used for distillation. We em-

pirically set the weight factors of β_1 and β_2 in Eq. 5 to 1 and 2.5, respectively. On Cifar-100, the SGD optimizer [26] is applied to train the student model with Nesterov momentum and a batch size of 64. The initial learning rate is 5e-2 and decayed by 0.1 at epochs 150, 180, and 210 with 240 epochs in total. In terms of ImageNet, we use the AdamW optimizer [18] to train the network for 100 epochs with a total batch size of 256. The initial learning rate is 2e-4 reduced by 0.1 at epochs 30, 60, and 90.

As to semantic segmentation, we distillate the knowledge on the last BN [14] layer of DeeplabV3+, whose feature map size is $256\times129\times129$. The weight α in Eq. 9 is set to 20. All students are trained for 100 epochs with a batch size of 12. We use the SGD optimizer with an initial learning rate of 0.007. And the learning rate decays according to the cosine annealing scheduler.

4.3. Image Classification

Results on Cifar-100. We evaluate the proposed method in a variety of network architectures, including VGG [24], ResNet [9] and its variants [33]. As shown in Table 1, our method outperforms other methods by a large margin. In the setting of distillation from WRN-40-2 to WRN-16-2, we achieve 75.57% Top-1 accuracy which is close to the teacher's performance 75.61%. We also compare to the methods that are more relevant to us, including layer-wise relation [20, 28, 21, 19] and those based on Gram matrix [32, 13] which measures the relation between pixel-wise position. In terms of layer-wise relational knowledge distillation methods, we outperform all the state-of-the-arts consistently. For example, in the distillation setting from ResNet56 to ResNet20, our method achieves an accuracy of 71.69% which greatly exceeds the second best method. This consistency proves the important role of feature diversity in knowledge distillation. As can be observed, the
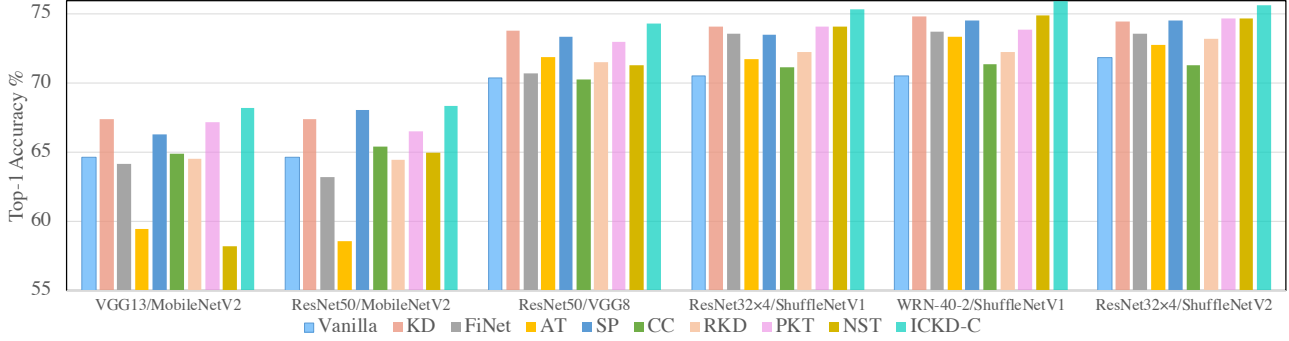
Figure 4: **Knowledge distillation across different architecture on Cifar-100.** Using teacher networks that completely different from that of students for knowledge distillation. The model before the slash is the teacher and the one after is the student. Our method can enable the students to learn more general knowledge regardless of the specific architecture.

Table 2: Top-1 and Top-5 Accuracy (%) on ImageNet validation set. The teacher network is ResNet34 and the student network is ResNet18. Our method outperforms other state-of-the-arts by a significant margin. Methods denoted by * do not release Top-5 accuracy.

	Vanilla	KD [12]	AT [34]	RKD [19]	SCKD* [2]	CRD [27]	CRD+KD	SAD* [15]	CC* [21]	ICKD-C (Ours)	Teacher
w/ $\mathcal{L}_{KL}$		✓		✓	✓		✓		✓	✓	
Top-1	70.04	70.68	70.59	71.34	70.87	71.17	71.38	71.38	70.74	**72.19**	73.31
Top-5	89.48	90.16	89.73	90.37	NA	90.13	90.49	NA	NA	**90.72**	91.42

other state-of-the-arts ranked inconsistently and the traditional KD [12] ranked the second place at most time. We can say that the contribution of mining the relationship of the layer-wise features is less than the guarantee of feature diversity.

We further explore the potential of inter-channel correlation by distillation across different network architectures, including MobileNetV2 [23], ShuffleNetV1 [35], and ShuffleNetV2 [23] (See Fig. 4). The characteristic of an ideal knowledge distillation method is that it can transfer the general knowledge regardless of the specific architecture. We find that some methods even deteriorated the performance of the students. Cho *et al.* [4] has pointed out that the students may fail to catch up with the teachers if their learning capacities mismatch. In the case that VGG13 is adopted as the teacher of MobileNetV2, many methods fail to improve the performance of the student. The situation even became worse when trying unilaterally to lead the student to learn the high-response area of the teacher's features given that AT [34] dropped 5% compared to the vanilla student. On the contrary, due to the characteristic of being invariant to the spatial dimension, ICKD-C can be adopted to transfer knowledge across different architectures, and its performance is always better than other methods. For instance, the transferred layer of VGG13 has a different feature map size from that of MobileNetV2, our method surpasses the second-best about 1% accuracy.

Results on ImageNet. We evaluate our method on the larger scale dataset ImageNet [5]. Note that [19] additionally applied the rotation, horizontal flipping, and color jittering for data augmentation. To compare with other works

more fairly, we choose ResNet34 as the teacher network and ResNet18 as the student network. The result is presented in Table 2. Again, our method consistently outperforms all methods by a significant margin. Our result is remarkable in that it achieves an accuracy rate of more than 72% in the existing literature for the first time.

We also visualize the ICC matrix of the student network and teacher network (See Fig. 5). At first, the feature map of student and teacher show great differences both in the inter-channel correlation and the response on a single channel. However, after distillation, they have become similar in addition to the ICC matrix, and the response on a single channel is also closer. According to the visualization of the feature channels, we can say the student can effectively preserve the feature diversity and has a similar feature pattern with the teacher. More results are displayed in Appendix.

4.4. Semantic Segmentation

Semantic segmentation is a promising but computation-consuming application. Yet methods based on knowledge distillation are rarely successfully applied to semantic segmentation. In this section, we present the experiments on the Pascal VOC semantic segmentation in the setting of knowledge distillation. Specifically, we deploy the ResNet101 as teacher backbone and transfer to student backbone ResNet18 and MoobileNetV2. The DeepLabV3+ [3] is chosen as the baseline model. Semantic segmentation aims at pixel-level classification, which is more challenging than image classification. The result is displayed in Table 3. It shows that we can prompt the student by a large margin (from 72.07% to 75.01%), which demonstrates
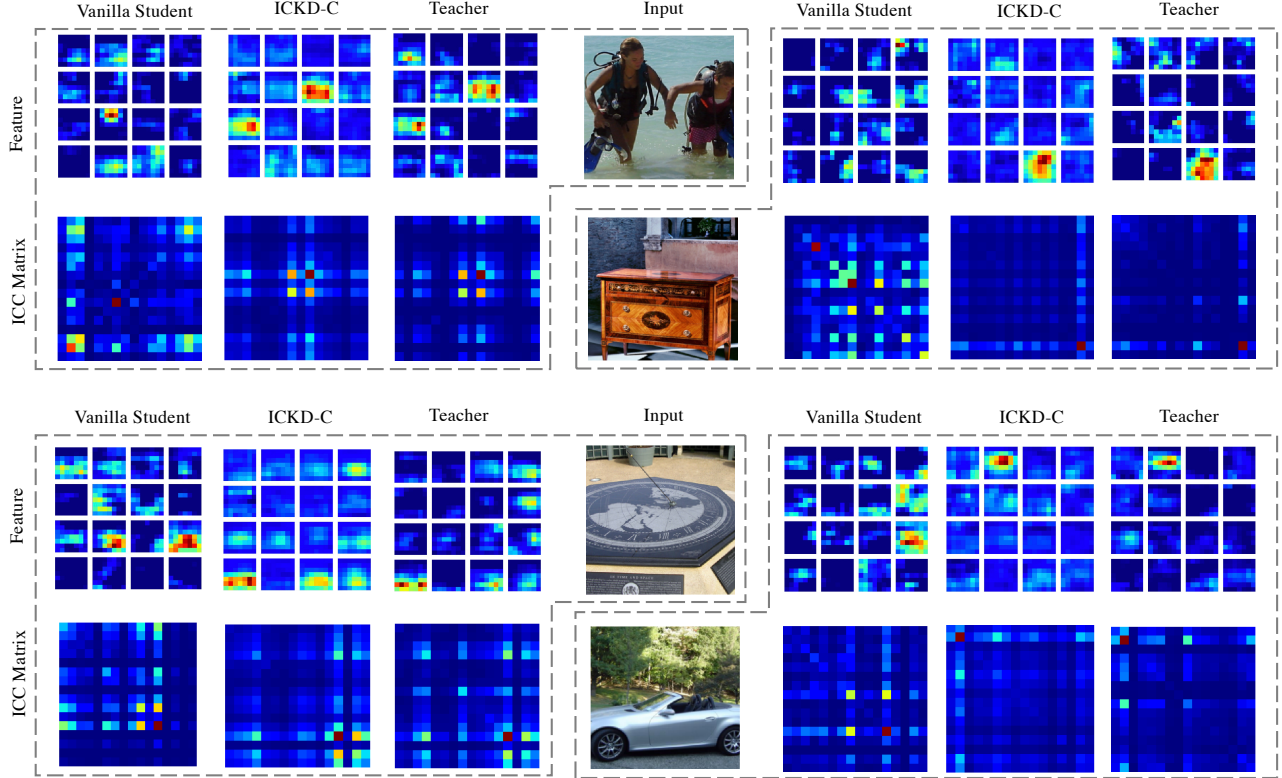
Figure 5: **Visualization of the features and the ICC matrices**. We have visualized the feature maps and the corresponding ICC matrices of the vanilla student, our model (ICKD-C) and the teacher, respectively. The four input images are sampled from ImageNet testing set. The teacher architecture is ResNet34 and the student architecture is ResNet18. Without loss of generality, we orderly select 16 feature maps extracted from the 4-th block (*i.e.*, the distillation layer) of the network. The results show that our model possesses the similar feature diversity and pattern with the teacher, demonstrating that learning inter-channel correlation can effectively preserve feature diversity.

that our method can learn rich representation for different downstream tasks. Particularly, our method bridges the gap between the cumbersome teacher and the inferior student, making it feasible to deploy segmentation models on edge devices.

4.5. The potential of a better teacher

An assumption is that the better the teachers are, the better the students we would have. This assumption seems plausible yet has been demonstrated unpractical because the students may not be able to catch up with the teachers [4]. We use several teacher networks individually to train the same student network (ResNet18) to see the possible improvements. As shown in Fig. 6, although all of the teachers can bring considerable performance gain to the student, the heavier teachers could not consistently prompt the Top-1 accuracy than the lightweight one. The student can achieve the best performance (Top-1 72.31%) when ResNet101 is used as the teacher and the second-best performance (Top-1 72.19%) when ResNet34 is adopted. Except for ResNet101, teachers better than ResNet34 could not bring further improvement. Interestingly, the best teacher (ResNet152)

couldn't obtain a considerable student model compared with others, which may be caused by the huge difference between their channel numbers (2048 for ResNet152 and 512 for ResNet18). We may say that it is unnecessary to employ a very cumbersome teacher network for knowledge distillation since it cannot bring further improvement consistently and spends more cost on pre-training.

Table 3: Performance on semantic segmentation in terms of mIoU (%) on the validation set of Pascal VOC.

Model	ResNet18	MobileNetV2
Vanilla	72.07	68.46
KD [12]	73.74	71.73
FitNet [22]	73.31	69.23
AT [34]	73.01	71.39
Overhaul [11]	73.98	71.19
ICKD-S (Ours)	**75.01**	**72.79**

4.6. Ablation Study

Firstly, we study the impact of the linear transformation layer C_l on Cifar-100. Intuitively, the linear transformation module may hinder the process of inter-channel correlation

knowledge distillation. However, the results presented in Table 6 show that the 1×1 linear transformation leads to a minor gain in most cases. This phenomenon is also observed by Wang *et al*. [29] in which linear transformation acts as an adaptor between the teacher and student.

Secondly, we study the impact of the weight factor β_2 in Eq. 5. In order to exclude the influence of $\mathcal{L}_{KL}$ (Eq. 1) and verify the effectiveness of $\mathcal{L}_{CC}$ separately, we set β_1 to zero. The results in Table 7 illustrate that our method is still impressive without joining $\mathcal{L}_{KL}$ on ImageNet (71.59% Top-1 accuracy, which also surpass the methods in Table 2). And it is also very robust to β_2. We perform the ICC matrix transferring (ResNet34→ResNet18) at different stages. The stage numbers are indicated by the subscript. When a single layer is used, our strategy is the best. When multiple stages are get involved with training, $S_{3,4}$ achieves the best Top-1 accuracy (See Table 4). In addition, we also conduct the experiments under different loss functions and kernel functions (See Table 5).

Lastly, the grid-level inter-channel correlation proposed in Sec. 3.3 is able to bring more performance improvements. Recall that we divide the whole feature map into $n \times m$ parts and if it is set to 1×1, this variant degrades to the ICKD-C without $\mathcal{L}_{KL}$. We conduct some experiments under different settings of $n \times m$ to see its effect. As can be observed in Table 8, our proposed ICKD-C without $\mathcal{L}_{KL}$ still improve the student (ResNet18) by 2.07% (from 72.07 to 74.14) and it can further boost the student to 75.01 after dividing the feature map into 32×32 patches. Generally, meshing the feature map can consistently improve the performance, but it is not the finer the better. Besides, the finer grid means more training cost. Table 9 illustrated the GPU hours cost of training the segmentation model 100 epochs with 2×NVIDIA 2080Ti.

Table 4: ICC Transferring at different places on ImageNet.

	S_1	S_2	S_3	S_4 (ours)	$S_{1,4}$	$S_{2,4}$	$S_{3,4}$	$S_{1,2,3,4}$
Top-1	70.49	70.50	70.87	**72.16**	72.31	72.20	72.33	72.26
Top-5	89.47	89.53	89.59	**90.75**	90.67	90.71	90.55	90.64

Table 5: Different loss functions and kernel functions.

	Loss Functions		Kernel Functions	
	L2 (ours)	Smooth L1	Gaussian kernel	Polynomial kernel
Top-1	72.16	72.29	70.63	72.25
Top-5	90.75	90.75	89.73	90.80

5. Conclusion

This work presents a method for knowledge distillation that explores the inter-channel correlation to mimic the feature diversity of the teacher network. In addition to image classification, we introduce the grid-level inter-channel correlation for semantic segmentation that most prior works do not pay attention to. We empirically demonstrate the effectiveness of the proposed method on a variety of network
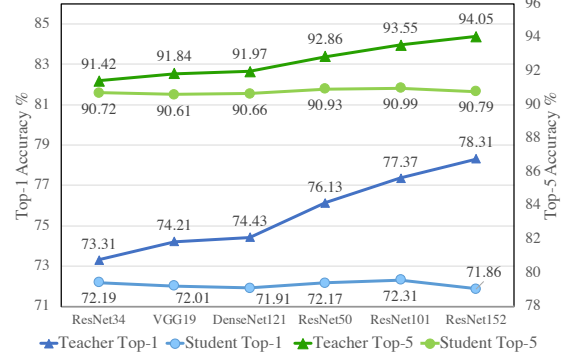

Figure 6: Accuracy (%) of the same student (ResNet18) guided by different teacher on ImageNet.

Table 6: Ablation on Cifar-100.

Teacher	Student	w/o Linear	w/ Linear
WRN-40-2	WRN-16-2	75.10	**75.64**
WRN-40-2	WRN-40-1	73.87	**74.33**
ResNet56	ResNet20	71.72	**71.76**
ResNet110	ResNet20	70.96	**71.68**
ResNet110	ResNet32	**73.90**	73.89
ResNet32×4	ResNet8×4	74.40	**75.25**
VGG13	VGG8	**73.85**	73.42

Table 7: Top-1 accuracy(%) under different β_2 on ImageNet.

β_2	0.2	1.0	2.0	4.0
ACC.	71.09	71.17	**71.59**	71.34

Table 8: mIoU(%) under different setting of $n \times m$ on Pascal VOC.

$n \times m$	1×1	4×4	16×16	32×32
ResNet18	74.14	74.97	74.74	**75.01**
MobileNetV2	72.10	72.26	**72.79**	72.58

Table 9: Training cost (GPU Hours) under different setting of $n \times m$ on Pascal VOC.

$n \times m$	1×1	4×4	16×16	32×32
ResNet18	25.8	25.9	31.8	157.4
MobileNetV2	24.5	24.7	30.4	155.9

architectures and achieve the state-of-the-art in two vision tasks (image classification and semantic segmentation). Besides, the computation of the proposed ICC matrix is invariant to feature spatial dimensions and able to distillate generic knowledge across different network architectures.

References

[1] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 1

[2] Defang Chen, Jian-Ping Mei, Yeliang Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. *ArXiv*, abs/2012.03236, 2020. 3, 6

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 6

[4] J. H. Cho and B. Hariharan. On the efficacy of knowledge distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4793–4801, 2019. 2, 6, 7

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[6] Leon A. Gatys, Alexander S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 3

[7] Kai Han, Yunhe Wang, Q. Tian, Jianyuan Guo, Chunjing Xu, and C. Xu. Ghostnet: More features from cheap operations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1577–1586, 2020. 2

[8] Bharath Hariharan, Pablo Arbeláez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. *2011 International Conference on Computer Vision*, pages 991–998, 2011. 5

[9] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 5

[10] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–587, 2019. 3, 4

[11] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, H. Park, N. Kwak, and J. Choi. A comprehensive overhaul of feature distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1921–1930, 2019. 3, 7

[12] Geoffrey E. Hinton, Oriol Vinyals, and J. Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. 1, 2, 3, 5, 6, 7

[13] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017. 3, 5

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 5

[15] Mingi Ji, Byeongho Heo, and S. Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. *ArXiv*, abs/2102.02973, 2021. 3, 6

[16] Seung Hyun Lee, Dae Ha Kim, and Byung Cheol Song. Self-supervised knowledge distillation using singular value decomposition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–350, 2018. 2

[17] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 3

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[19] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3962–3971, 2019. 1, 2, 5, 6

[20] N. Passalis and A. Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, 2018. 1, 2, 5

[21] Baoyun Peng, Xiao Jin, Jiaheng Liu, Shunfeng Zhou, Y. Wu, Y. Liu, Dong sheng Li, and Z. Zhang. Correlation congruence for knowledge distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5006–5015, 2019. 1, 2, 5, 6

[22] A. Romero, Nicolas Ballas, S. Kahou, Antoine Chassang, C. Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2015. 2, 5, 7

[23] Mark Sandler, A. Howard, Menglong Zhu, A. Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 6

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[25] S. Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In *EMNLP/IJCNLP*, 2019. 2

[26] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013. 5

[27] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2020. 1, 2, 6

[28] F. Tung and G. Mori. Similarity-preserving knowledge distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1365–1374, 2019. 1, 2, 5

[29] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019. 4, 8

[30] Yukang Wang, W. Zhou, T. Jiang, X. Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *ECCV*, 2020. 3

[31] Zhen Wei, Jingyi Zhang, Li Liu, Fan Zhu, Fumin Shen, Yi Zhou, Si Liu, Yao Sun, and Ling Shao. Building detail-sensitive semantic segmentation networks with polynomial pooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7115–7123, 2019. 4

[32] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7130–7138, 2017. 3, 5

[33] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *ArXiv*, abs/1605.07146, 2016. 5

[34] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *ArXiv*, abs/1612.03928, 2017. 3, 5, 6, 7

[35] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 6