A Simple and Generic Framework for Feature Distillation via Channel-wise Transformation

Ziwei Liu[1] Yongtao Wang[1†] Xiaojie Chu[1] Nan Dong[2] Shengxiang Qi[2] Haibin Ling[3]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Chong Qing Changan Automobile Co., Ltd
[3]Stony Brook University
liuziwei@stu.pku.edu.cn, wyt@pku.edu.cn, chuxiaojie@stu.pku.edu.cn
dongnan1@changan.com.cn, shengxiang.qi@gmail.com, hling@cs.stonybrook.edu

Abstract

Knowledge distillation is a popular technique for transferring knowledge from a large teacher model to a smaller student model by mimicking. However, distillation by directly aligning the feature maps between teacher and student may enforce overly strict constraints on the student thus degrading the performance of the student model. To alleviate the above feature misalignment issue, existing works mainly focus on spatially aligning the feature maps of the teacher and the student, with pixel-wise transformation. In this paper, we newly find that aligning the feature maps between teacher and student along the channel-wise dimension is also effective for addressing the feature misalignment issue. Specifically, we propose a learnable nonlinear channel-wise transformation to align the features of the student and the teacher model. Based on this idea, we propose a simple and generic framework for feature distillation, with only one hyper-parameter to balance the distillation loss and the task-specific loss. Extensive experimental results show that our method achieves significant performance improvements in various computer vision tasks including image classification (+3.28% top-1 accuracy for MobileNetV1 on ImageNet-1K), object detection (+3.9% bbox mAP for ResNet50-based Faster-RCNN on MS COCO), instance segmentation (+2.8% Mask mAP for ResNet50-based Mask-RCNN), and semantic segmentation (+4.66% mIoU for ResNet18-based PSPNet in semantic segmentation on Cityscapes), which demonstrates the effectiveness and the versatility of the proposed method.

1. Introduction

Nowadays, the development of deep neural network (DNN) architectures, such as ResNet [10], ResNeXt [28], Faster R-CNN [20], and PSPNet [34], has led to significant performance improvements for various computer vision tasks, such as image classification, object detection, and semantic segmentation. However, the high performance of these DNN models comes at the cost of large size and high computational requirements for these architectures, which poses challenges for their deployment in resource-constrained environments. To address this problem, knowledge distillation [12] has been proposed to achieve high performance with reduced computational cost by transferring the knowledge from a large model (teacher) to a smaller model (student).

Specifically, feature-based knowledge distillation methods, which transfer knowledge from the intermediate layer features of the teacher model to the student model, have been intensively studied and demonstrated as a more effective and generic approach for improving the performance of the student model.

As pointed out in [15], due to the feature misalignment

Task Metric	Cls Top-1 Acc	Det BBox mAP	Seg mIoU
Student	69.9	36.5	69.9
Teacher	73.6	41.0	75.9
Identity	70.3 (+0.4)	38.8 (+2.3)	46.2 (-23.7)
Linear	71.0 (+1.1)	39.3 (+2.8)	71.4 (+1.5)
Task-Specific*	70.9 (+1.0)	39.3 (+2.8)	72.4 (+2.5)
MLP(ours)	71.5 (+1.6)	39.5 (+3.0)	73.5 (+3.6)

* We use TaT [15] in Cls and Seg, FGD [30] in Det

Table 1. Comparison of various transformation methods in knowledge distillation for classification(Cls), Segmentation(Seg) and Detection(Det) tasks. Teacher and student feature maps have the same number of channels. Distillation with the help of the transformation module can improve student performance compared to direct mimics.

†Corresponding author.

of the teacher and student model, directly mimicking the intermediate features of the teacher model via vanilla L_2 distances may enforce overly strict constraints on the student, leading to sub-optimal performance.

To alleviate this problem, existing works design novel distillation loss functions [11, 22] or feature transformation modules [2, 13, 15, 30, 32] to mimic the teacher's features indirectly. Specifically, the latter kind of approach often focuses on the feature transformations along the spatial dimension, such as guiding the student's attention towards the key regions of the feature map [13] or the relationship between different pixels [15, 30, 32].

In this paper, we focus on feature-based knowledge distillation and try to address the feature misalignment problem along the channel dimension rather than spatial dimensions. We have observed that channel-wise transformations (e.g., 1x1 convolution) have been widely used to align the features of different channel sizes in many tasks including feature-based knowledge distillation. Moreover, for the feature-based knowledge distillation task, these channel-wise transformation modules are discarded when the channel sizes of the teacher's feature and the student's feature are already the same. However, we empirically find that a linear channel-wise transformation, *i.e.*, 1x1 convolution, can result in consistent performance improvements for feature-based knowledge distillation, even when the channel sizes of teacher's feature and student's feature are already the same, the results are shown in table 1.

Inspired by our empirical findings about the importance of channel-wise transformations for feature-based distillation, we propose a simple and generic approach that focuses on channel-wise feature alignment. Specifically, without careful selection or design of transformation modules, we implement the channel-wise transformation as a non-linear MLP with one hidden layer, which has been demonstrated to have universal approximation capabilities [5]. With this simple channel-wise transformation module and the conventional $L2$-distance loss, we propose a very simple and generic method for feature-based distillation. With only one tunable hyper-parameter, our method is easy to apply to different tasks.

Our extensive evaluation, as shown in Table 2, reveals that our method consistently outperforms existing feature-based distillation methods on dense prediction tasks. In object detection, we observed consistent performance gains over two-stage, anchor-based, and anchor-free single-stage detectors, with an average improvement of +3.5% in bbox mAP across these settings. For semantic segmentation, our method delivered an average improvement of +4.0% in mIoU over heterogeneous and homogeneous distillation settings on the ResNet-18-based PSPNet. Our method also achieves strong performance on the classification task, with an average increase of +2.4% in Top-1 accuracy, regardless of whether the number of channels in the student and teacher feature maps is the same or not.

	Cls	Det	Ins Seg	Seg	#Hyper
KR [2]	**+2.5**	-	-	-	2
FGD [30]	-	+3.1	+2.4	-	5
CWD [22]	-	-	-	+3.2	2
MGD [31]	+2.4	+3.3	+2.7	+3.3	2
Ours	+2.4	**+3.5**	**+2.8**	**+4.0**	**1**

Table 2. Comparisons of the state-of-the-art methods on image classification (Cls), object detection (Det), instance segmentation (Ins Seg), and semantic segmentation (Seg). The metrics reported are Top-1 accuracy, BBox mAP, Mask AP, and mIoU, improvement relative to students, respectively. Hyper denotes hyperparameters. Our method achieves state-of-the-art results with only 1 hyperparameter.

To sum up, our main contributions are three-fold:

- We reinstate the importance of channel-wise transformation for aligning the student's and teacher's features in feature-based knowledge distillation.
- We propose a simple and generic framework for feature-based knowledge distillation which uses MLP as the channel-wise transformation module to help students learn more powerful features.
- We achieve state-of-the-art distillation results for multiple dense prediction tasks and comparable state-of-the-art results for classification tasks.

2. Related Work

The concept of knowledge distillation was first proposed by Hinton et al. [12], with the goal of transferring dark knowledge from a cumbersome teacher model to a smaller student model to improve the student's performance. Based on the types of dark knowledge, mainstream knowledge distillation methods can be divided into two categories: Logits-based knowledge distillation and feature-based knowledge distillation.

2.1. Logits-based knowledge distillation

Classical logits-based knowledge distillation methods [12] minimize the KL divergence between the output logits of teacher and student models. One recent line of research focuses on refining the vanilla knowledge distillation loss function to better leverage the logits information. WSLD [36] rethinks the knowledge distillation process from a bias-variance trade-off perspective and proposes weighted soft labels for knowledge distillation. DKD [33], reformulates the classical knowledge distillation loss into the target

and non-target part and calculates the distillation loss separately. While these works have improved the performance of logits-based knowledge distillation methods on classification tasks, they have often not achieved significant results on other tasks, such as dense prediction tasks. Another line of work involves modeling other tasks into a classification task and adopting the logits-based knowledge distillation on other tasks. LD [35], reformulates the output form of the regression head to a probability distribution and applies classical knowledge distillation to the regression task. However, it is only for object detection tasks and requires changes to the detection head. RMKD [14] reformulates the ordering between anchors into the form of the probability distribution for knowledge transfer and applies classical knowledge distillation to the regression task. However, it is only limited to anchor-based detectors.

2.2. Feature-based knowledge distillation

For classification. The teacher model feature is a kind of dark knowledge, first used in [21]. Later works focused on better ways to utilize it. AT [13] uses attention maps to help students focus on important regions but does not use channel information. OFD [11] designs a new loss function and uses marginal ReLU to extract key information. CRD [23] uses contrastive learning in knowledge distillation, achieving good performance but with high training costs. KR [2] conducts knowledge distillation on multi-level features in a review manner, resulting in good performance. TaT [15] uses a one-to-all spatial matching approach for knowledge distillation based on similarity generated from a target-aware transformer.

For objection detection. Object detection is a challenging task due to foreground-background pixel imbalance. Knowledge distillation methods attempt to have the student model imitate the key regions of the teacher model. FGFI [25] forces students to focus on foreground regions by using masks. GID [6] identifies regions where the student and teacher models perform differently as key regions for distillation. Defeat [9] distills foreground and background regions separately, and FKD [32] uses attention masks to direct the student model's focus while non-local modules capture relationships between pixels. Additionally, FGD [30] proposes focal and global distillation mechanisms through attention masks and global context blocks [1]. The goal of these methods is to enhance the student model's performance in object detection tasks. This is achieved through knowledge distillation from the teacher model, as well as a focus on crucial regions.

For semantic segmentation. Semantic segmentation is a per-pixel prediction problem, and strictly aligning the feature maps between the student and teacher models may impose overly strict constraints and lead to sub-optimal results [22]. Recent works [17, 27] try to force the student to learn the correlations among different spatial regions. IFVD [27] focuses on the intra-class feature variation among pixels with the same label and designs an IFV module to transfer the structural knowledge. SKDS [17] combines pixel-wise distillation, pair-wise distillation, and holistic distillation using a GAN-based approach to align the output maps of teacher and student models. CIRKD [29] aims to model the pixel-to-pixel and pixel-to-region relationships as supervisory signals for knowledge distillation in the semantic segmentation task. CWD [22] is a method that normalizes the activation maps of each channel and minimizes the KL divergence between these probability maps. This helps to improve the accuracy of predictions in dense tasks like object detection and semantic segmentation in computer vision applications. The effectiveness of this method has been widely acknowledged and used in the field. By utilizing advanced AI technology, CWD is capable of producing highly precise and dependable probability maps. This greatly enhances the efficiency and accuracy of image analysis and processing.

For general tasks. MGD [31] employs a generative approach that involves the use of random masks that randomly to erases a portion of the student's feature map and then force it to generate features similar to the teacher's through an adversarial generator and applies it to classification, detection, and segmentation tasks.

In this paper, we focus on channel-wise transformations and propose a simple and generic method for feature-based knowledge distillation.

3. Method

In this section, we first briefly introduce the basic form of intermediate feature-based knowledge distillation and then present the details of our proposed method.

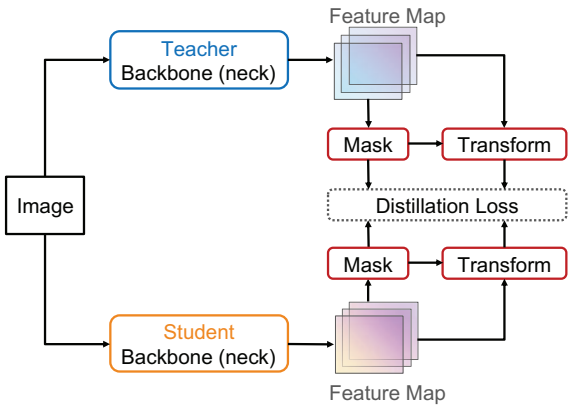
3.1. Revisiting Feature-based Knowledge Distillation

In feature-based knowledge distillation, a student model is generally supervised by a teacher model as [8] :

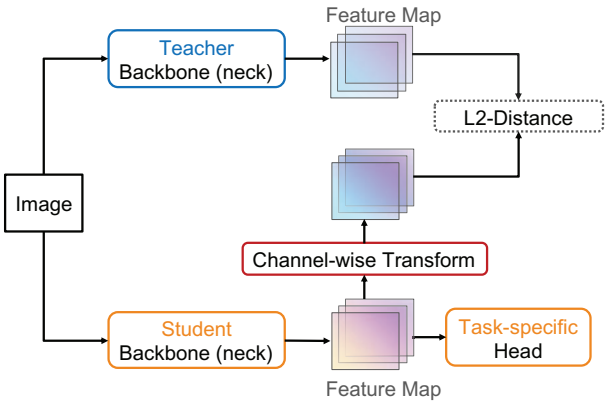
$$L_{feat} = \mathcal{L}_{KD}\left(\mathcal{T}_t\left(\boldsymbol{F}_T\right), \mathcal{T}_s\left(\boldsymbol{F}_S\right)\right), \tag{1}$$

where $\mathcal{L}_{KD}$ represents the similarity function used to match the feature maps of the teacher model, $\mathbf{F_T}$, and the student model, $\mathbf{F_S}$. In addition, the transformation functions $\mathcal{T}_t$ and $\mathcal{T}_s$ will be applied only if the feature maps of the teacher and student models have different dimensions. (e.g., a linear projection layer to align the number of channels in $\mathbf{F_S}$ with those in $\mathbf{F_T}$).

Recently, several works have employed complex transformations ($\mathcal{T}_t$ and $\mathcal{T}_s$) to facilitate the acquisition of knowledge (feature alignment) by student networks from teacher networks. For example, as depicted in Figure 1a, both FKD [32]

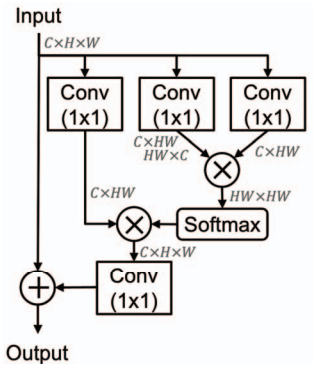


(a) Some previous methods [30, 32] use both sophisticated designed mask and spatial-wise transformation for both teacher and student.

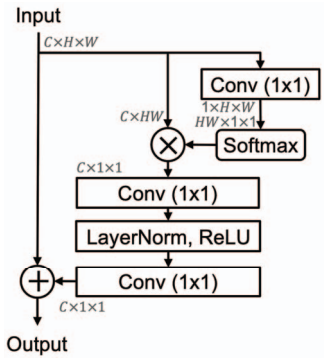


(b) Our proposed method uses learn-able channel-wise transformation only for the student model.

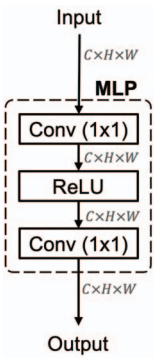
Figure 1. Difference between our method and existing feature-based knowledge distillation methods.



(a) Non-Local Block used in FKD [32]



(b) Global Context Block used in FGD [30]



(c) MLP (Ours)

Figure 2. Comparison of different transformation modules in knowledge distillation. FKD [32] uses a non-local module (a), while FGD [30] employs GCBlock (b) to model the relationships between pixels in an image. Our method utilizes a simple yet effective channel-wise transformation through an MLP (c), consisting of two 1×1 convolution layers and a ReLU activation layer.

and FGD [30] utilize (1) specific modules, such as the Non-local module [26] or GCBlock [1], and (2) channel-wise and spatial-wise attention masks to align the features of the teacher and student networks. This raises the question of whether the use of well-designed modules is necessary for student networks to learn more effective features from the teacher.

In this paper, we perform empirical studies to address the question raised above and find that student models can enhance their representations through a straightforward non-linear channel-wise transformation. Based on this finding, as illustrated in Figure 1b, we introduce a simple method that incorporates a Multi-Layer Perceptron (MLP) into the student features and aligns the transformed student features with the teacher features using a conventional L_2 distance. The specifics of our proposed method are outlined in the subsequent subsection.

3.2. Learnable channel-wise transformation

Instead of using complex transformations on both spatial and channel dimensions, we propose to use a learnable nonlinear channel-wise transformation to align the feature maps of the student and the teacher model. In detail, we use a non-linear MLP with one hidden layer for the student (*i.e.*, $\mathcal{T}_t = identity$ and $\mathcal{T}_s = MLP$ in Equ. 3.1):

$$\mathrm{MLP}(F) = W_2\left(\sigma\left(W_1(F)\right)\right), \tag{2}$$

where $W1$ and $W2$ are learnable parameters implemented as 1×1 convolutions, and σ represents ReLU activation. As illustrated in Figure 2, our transformation module (Figure 2c) is much simpler than the methods proposed in FKD [32] and FGD [30].

Without bells and whistles, we choose L_2 distance for supervising transformed student features and teacher features.

Algorithm 1 Pseudo code of our method in a PyTorch-like style.

```
# f_mlp: 2-layer MLP with ReLU activation
# x_s: student feature [N, C, H, W]
# x_t: teacher feature [N, C, H, W]

def forward(x_s, x_t):
   n = x_s.shape[0] # size of mini-batch

   # channel-wise non-linear transformation
   x_mlp = f_mlp.forward(x_s)

   # calculate l2 distance
   diff = (x_mlp - x_t).pow(2)

   # distillation loss (averaged by batch)
   loss = diff.sum() / n

   return loss
```

Specifically, the feature distillation loss is formulated as:

$$L_{feat} = \sum_{i}^{N} (MLP(\boldsymbol{F}_{S_i}) - \boldsymbol{F}_{T_i})^2. \tag{3}$$

Our approach is straightforward and can be efficiently executed using prevalent machine learning libraries, such as PyTorch, as shown in Algorithm 1. The ease of implementation enables us to leverage existing infrastructure, facilitating the training process of our model.

3.3. Overall loss

Our method can be easily used in various tasks. Combined with task-specific losses, the overall loss can be formulated as:

$$L_{total} = L_{task} + \alpha L_{feat} \tag{4}$$

where α is a hyper-parameter to balance the weight of knowledge distillation loss.

4. Experiment

Our approach, which is a feature-based method, is easy to implement on various models and tasks. In this paper, we conduct experiments on image classification, object detection, instance segmentation, and semantic segmentation, to demonstrate the simplicity, effectiveness, and generality of our method.

4.1. Image Classification on ImageNet

Settings To test our image classification method, we use the Imagenet dataset [7]. We train the model using 1.2 million images from the Imagenet training set and test it using 50,000 images from the validation set. The evaluation metric used is Top-1 accuracy. We use a standard training procedure with the model trained for 100 epochs, learning rate decay at the 30th, 60th, and 90th epochs, SGD optimizer, and an initial learning rate of 0.1. The training is done on 8 GPUs with a batch size of 32 images per GPU. Our method is tested on both homogeneous and heterogeneous distillation settings using two model configurations: ResNet34 as the teacher model and ResNet18 as the student model, and ResNet50 as the teacher model and MobileNet as the student model. We use feature maps from the last stage of the backbone to calculate the distillation loss and set the distillation loss weight α to 7×10^{-5}. Our method is compared with single-layer feature-based methods [19, 23, 31], state-of-the-art logits-based methods [12, 33], and multi-layer feature-based methods [2, 11, 13].

Comparison to baseline. As presented in Table 3, our method demonstrates its effectiveness on the image classification task. Specifically, under the homogeneous setting, the Top-1 accuracy of ResNet-18 is improved by +3.28%. Similarly, under the heterogeneous setting, the Top-1 accuracy of MobileNet is improved by +1.61%. These results highlight the superiority of our method in comparison to the baseline models.

Comparison to single feature distillation methods Our method outperforms all single-feature distillation methods [19,23,31] in the heterogeneous setting and is on par with the state-of-the-art method MGD [31]. These results highlight the effectiveness of our method in comparison to other single-feature distillation techniques.

Comparison to previous state-of-the-art Compared to the previous state-of-the-art method KR [2], which uses multi-stage features, our simple method achieves comparable performance with a difference of less than 0.1% in terms of Top-1 accuracy on both homogeneous and heterogeneous settings.

4.2. Object Detection on COCO

Settings For the object detection task, we evaluate our method on the COCO [16] dataset. Specifically, we use 120,000 images from the COCO training set for model training and 5,000 images from the validation set for model testing, with mAP as the evaluation metric. Our training procedure follows a standard 2x schedule, consisting of 24 training epochs, with the reduction of the learning rate at epochs 16 and 22. The optimization process is performed using Stochastic Gradient Descent (SGD) and the model is trained on 8 GPUs, each with a batch size of 2.

We experiment on multiple detector architectures, including two-stage, single-stage anchor-based, and single-stage anchor-free detectors. The distillation loss is computed on all feature maps output from the neck, and the distillation

Mechanism	Method	Top-1 acc	Top-5 acc	Method	Top-1 acc	Top-5 acc
-	ResNet-50(T)	76.55	93.06	ResNet-34(T)	73.62	91.59
-	MobileNet(S)	69.21	89.02	ResNet-18(S)	69.90	89.43
Logits	KD [12]	70.68	90.30	KD [12]	70.68	90.16
	DKD [33]	72.05	91.0	DKD* [33]	71.37	90.26
Multi Feature	AT [13]	70.72	90.03	AT [13]	70.59	89.73
	OFD [11]	71.25	90.34	OFD [11]	71.08	90.07
	KR [2]	72.56	91.00	KR [2]	71.61	90.51
Single Feature	RKD [18]	71.32	90.62	RKD [18]	71.34	90.37
	CRD [24]	71.40	90.42	CRD [24]	71.17	90.13
	MGD [31]	72.35	90.71	MGD [31]	71.58	90.35
	Ours	72.49	90.81	Ours	71.51	90.32

Table 3. Results of different distillation methods on ImageNet dataset for the image classification task. **T** and **S** mean the teacher and student, respectively. * We report the result inplemented in MMRazor [3].

Method	Input Size	mIoU
PspNet-Res101(T)	512 × 1024	78.34
PspNet-Res18(S)	512 × 512	69.85
SKDS [17]	512 × 512	72.70
CWD [22]	512 × 512	73.53
MGD [31]	512 × 512	73.63
Ours	512 × 512	74.51
PspNet-Res101(T)	512 × 1024	78.34
DeepLabV3-Res18(S)	512 × 512	73.20
SKDS [17]	512 × 512	73.87
CWD [22]	512 × 512	75.93
MGD [31]	512 × 512	76.02
Ours	512 × 512	76.55

Table 4. Results of the semantic segmentation task on CityScapes dataset. **T** and **S** mean teacher and student, respectively.

loss weight α is set to 5×10^{-7} for the two-stage detector and 2×10^{-5} for the one-stage detector. For the instance segmentation task, we use a ResNext-101-based Cascade Mask R-CNN as the teacher model and a ResNet-50-based Mask R-CNN as the student model. The experimental configuration follows that of the two-stage detector distillation.

Object Detection We compare our method with previous state-of-the-art methods designed for object detection [30, 32] and a recent generic distillation method [31]. As shown in Table 5, our simple method can achieve competitive results. For example, on the two-stage detector Faster RCNN-ResNet50, we get the mAP of the student model to rise from 38.4 to 42.3, surpassing the previous state-of-the-art method. On the anchor-based single-stage detector RetinaNet-ResNet50 and the anchor-free single-stage detector Reppoints-ResNet50, we also achieve mAP increases of 3.6 and 3.4, respectively, which are comparable to the results of the state-of-the-art method.

Instance Segmentation Our method demonstrates its effectiveness on the instance segmentation task, as shown in Table 5. The results show that our simple approach leads to +3.2% improvement in bounding box AP and +2.4% improvement in mask AP, respectively, outperforming state-of-the-art methods.

4.3. Semantic segmentation on CityScapes

Settings For the semantic segmentation task, we evaluate our method with the CityScapes dataset [4]. Specifically, our experiments are conducted on 2975 training images and 500 validation images, and the evaluation metric is mIoU. The models are trained for 40,000 iterations using the SGD optimizer on 8 GPUs with a batch size of 2.

We conduct experiments on two model configurations: a) a homogeneous setting with PSPNet-Res101 as the teacher model and PSPNet-Res18 as the student model, and b) a heterogeneous setting with PSPNet-Res101 as the teacher model and DeepLabv3-Res18 as the student model. The input size for both configurations is set to 512×512, and the distillation loss is computed from the features of the last stage. The distillation loss weight α for the homogeneous setting and the heterogeneous setting is set to 2×10^{-5} and 1×10^{-5} respectively.

Results As shown in Table 4, our method achieves remarkable results in both homogeneous and heterogeneous

Teacher	Student	mAP	AP_S	AP_M	AP_L	mAR	AR_S	AR_M	AR_L
RetinaNet ResNeXt101	RetinaNet-ResNet50	37.4	20.6	40.7	49.7	53.9	33.1	57.7	70.2
	FKD [32]	39.6(+2.2)	22.7	43.3	52.5	56.1(+2.2)	36.8	60.0	72.1
	FGD [30]	40.4(+3.0)	23.4	44.7	54.1	56.7(+2.8)	37.6	61.5	72.4
	MGD [31]	40.6(+3.2)	23.4	45.1	54.0	56.7(+2.8)	37.1	61.0	72.5
	PKD [31]	40.8(+3.2)	23.4	45.1	54.0	56.7(+2.8)	37.1	61.0	72.5
	Ours	41.0(+3.6)	23.1	45.5	55.0	56.8(+2.9)	37.2	60.8	72.4
Cascade Mask RCNN ResNeXt101	Faster RCNN-ResNet50	38.4	21.5	42.1	50.3	52.0	32.6	55.8	66.1
	FKD [32]	41.5(+3.1)	23.5	45.0	55.3	54.4(+2.4)	34.0	58.2	69.9
	FGD [30]	42.0(+3.6)	23.8	46.4	55.5	55.4(+3.4)	35.5	60.0	70.0
	MGD [31]	42.1(+3.7)	23.7	46.4	56.1	55.5(+3.5)	35.4	60.0	70.5
	PKD [31]	41.7(+3.7)	23.7	46.4	56.1	55.5(+3.5)	35.4	60.0	70.5
	Ours	42.3(+3.9)	24.2	46.4	56.1	55.3(+3.3)	34.9	59.8	70.4
RepPoints ResNeXt101	RepPoints-ResNet50	38.6	22.5	42.2	50.4	55.1	34.9	59.4	70.3
	FKD [32]	40.6(+2.0)	23.4	44.6	53.0	56.9(+1.8)	37.3	60.9	71.4
	FGD [30]	41.3(+2.7)	24.5	45.2	54.0	58.4(+3.3)	39.1	62.9	74.2
	MGD [31]	41.7(+3.1)	24.1	45.8	55.3	57.9(+2.8)	39.0	62.0	73.6
	PKD [31]	42.3(+3.7)	23.7	46.4	56.1	55.5(+3.5)	35.4	60.0	70.5
	Ours	42.0(+3.4)	24.8	46.0	55.4	57.9(+2.8)	38.9	62.0	73.7

Teacher	Student	Boundingbox AP				Mask AP			
		mAP	AP_S	AP_M	AP_L	mAP	AP_S	AP_M	AP_L
Cascade Mask RCNN ResNeXt101	Mask RCNN-ResNet50	39.2	22.9	42.6	51.2	35.4	19.1	38.6	48.4
	FKD [32]	41.7(+2.5)	23.4	45.3	55.8	37.4(+2.0)	19.7	40.5	52.1
	FGD [30]	42.1(+2.9)	23.7	46.2	55.7	37.8(+2.4)	19.7	41.3	52.3
	MGD [31]	42.3(+3.1)	23.9	46.3	56.2	38.1(+2.7)	17.1	41.1	56.3
	Ours	42.4(+3.2)	23.8	46.3	56.6	38.2(+2.8)	17.3	41.2	56.6

Table 5. Results of detectors on COCO dataset.

configurations. Specifically, the ResNet-18-based PspNet model obtains a mIoU increase of +4.66% under the homogeneous setting, and the ResNet-18-based deeplabv3 model obtains a mIoU increase of +3.28% under the heterogeneous setting. These results demonstrate the effectiveness of our method on the semantic segmentation task.

4.4. Ablation Studies and Analysis

4.4.1 Benefits of Channel-wise Transformation

As shown in Table 1, directly using the features from the teacher and the student without channel-wise transformation can result in significant distillation performance drops in the semantic segmentation task.

To better understand this phenomenon, we calculate the L_2-distance between the student feature map and the teacher feature map on the validation dataset.

The results in Table 6 show that directly mimicking the teacher feature (corresponding to 'Identity' transformation) can achieve a lower L_2-distance to the teacher, but obtain significantly poorer performance compared to those using channel-wise transformations. Compared with it, channel-wise transformation methods can obtain an even lower L_2-distance after the channel-wise transformation, but the L_2-distance before the channel-wise transformation is much larger. Moreover, the distillation performance of the channel-wise transformation methods is much better than directly mimicking.

In the process of distillation, the student model is supervised by two signals: distillation losses and task-specific losses. We conjecture that the limited capacity of the student model makes it difficult to fully capture the knowledge of the teacher, and applying strict distillation constraints (*i.e.*, directly mimicking the teacher feature) may over-optimize the student feature with the distillation supervision and prevent them from being trained with the task-specific supervision, leading to performance degradation. On the contrary, our method exploits the channel-wise transformation module to achieve a better balance between task-specific supervision and distillation supervision.

Transformation	L_2-distance		mIoU
	Before	After	
Identity	0.217	0.217	46.2
Linear	0.4269	0.037	71.4
MLP(ours)	0.7691	0.032	73.5

Table 6. L_2 distances with teacher feature and mIoU scores for different transformations in the semantic segmentation task.

4.4.2 Ablation of Transformation Modules

In this section, we further demonstrate the importance of the channel-wise transformation module in cases where the size of the teacher's feature and the student's feature are not equivalent, *i.e.*, when the number of channels is unequal. We conduct experiments on the heterogeneous segmentation setting.

Table 7 demonstrates that the student model's performance only slightly improves with a single linear layer and no non-linear activation. This indicates that non-linear transformation is crucial for the student's representation ability. However, adding local spatial transformation with non-linear activation results in worse performance compared to an MLP. The Global spatial transformation with Non-Local block yields the lowest mIoU. We conducted experiments on designing more complex transformation modules involving stacking and deforming. Unfortunately, for MLP-deeper, adding an extra hidden layer to the original MLP module or doubling the hidden dimension of the MLP module for MLP-wider did not lead to any further improvement. For the Deformable 3x3 module, we implemented Deformable Conv3×3-ReLU-Deformable Conv3×3 to allow for deformation in the convolution operation. However, the results showed that these modifications did not lead to significant improvements. These modifications were aimed at improving the performance of the student model by modifying the transformation process.

These results show that the transformation of spatial dimensions and more complex designs do not bring additional gain to our method. We conjecture that an overly complex and powerful learnable transformation will make the distillation process concentrate on optimizing the transformation module rather than the student network itself.

4.4.3 Location of MLP

Previous research on feature-based distillation techniques, such as FKD [32] and FGD [30], have employed complex masks in their transformation modules to transform both the student and teacher features. In contrast, our method utilizes a learnable Multi-Layer Perceptron (MLP) for feature transformation.

Module	Transform			mIoU
	Channel	Spatial	Non-Linear	
Stu-Baseline	-	-	-	73.20
Linear	✓	✗	✗	73.40
Conv3×3	✓	Local	✓	75.92
Deformable 3×3	✓	Local	✓	75.69
Non-Local [26]	✓	Global	✓	72.05
MLP-base	✓	✗	✓	76.55
MLP-deeper	✓	✗	✓	76.53
MLP-wider	✓	✗	✓	76.10

Table 7. Performance comparison of different transform modules on semantic segmentation task. The results indicate that our proposed channel-wise non-linear transformation module (MLP) outperforms other methods.

When a learnable MLP is used, it can help students learn better representations from teachers. However, if the learnable transformation is applied to both the student and teacher features and the $L2$ distance is used as the loss function, it can result in a trivial solution where the learnable transformation simply takes the input of the feature and gives zero outputs. This renders feature distillation ineffective. To address this issue, our method only applies the transformation to the student feature. This ensures that the student model can learn from the teacher model without compromising the effectiveness of feature distillation.

5. Conclusion

In this paper, we first present a novel discovery that aligning the feature maps between teacher and student along the channel-wise dimension is also effective for addressing the feature misalignment issue in feature-based knowledge distillation. Then, we exploit a Multi-Layer Perceptron (MLP) as the channel-wise transformation module to align the features of the student and the teacher model. Further, we propose a simple and generic framework for feature distillation based on it, with only one hyper-parameter to balance the distillation loss and the task-specific loss. Extensive experiments are conducted and the results demonstrate that the proposed method can achieve significant performance improvements in various computer vision tasks including image classification, object detection, instance segmentation, and semantic segmentation, even outperforming the state-of-the-art feature-based knowledge distillation methods in some tasks.

6. Acknowledgemet

This work was supported by National Key R&D Program of China (Grant No. 2022ZD0160305). This work was also a research achievement of Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

[1] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 3, 4

[2] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021. 2, 3, 5, 6

[3] MMRazor Contributors. Openmmlab model compression toolbox and benchmark. https://github.com/open-mmlab/mmrazor, 2021. 6

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[5] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989. 2

[6] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2021. 3

[7] 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. 5

[8] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 3

[9] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[11] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019. 2, 3, 5, 6

[12] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 1, 2, 5, 6

[13] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 2, 3, 5, 6

[14] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1306–1313, 2022. 3

[15] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge distillation via the target-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10915–10924, 2022. 1, 2, 3

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[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, 6

[18] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 6

[19] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 5

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[21] 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. 3

[22] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 2, 3, 6

[23] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2019. 3, 5

[24] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 6

[25] 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. 3

[26] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 4, 8

[27] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *European Conference on Computer Vision*, pages 346–362. Springer, 2020. 3

[28] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 1492–1500, 2017. 1

[29] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022. 3

[30] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022. 1, 2, 3, 4, 6, 7, 8

[31] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. *ECCV*, 2022. 2, 3, 5, 6, 7

[32] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020. 2, 3, 4, 6, 7, 8

[33] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022. 2, 5, 6

[34] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1

[35] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9407–9416, 2022. 3

[36] Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias–variance tradeoff perspective. In *International Conference on Learning Representations*, 2020. 2