

Channel-wise Knowledge Distillation for Dense Prediction*

Changyong Shu^{1,4}, Yifan Liu², Jianfei Gao¹, Zheng Yan^{1†}, Chunhua Shen³

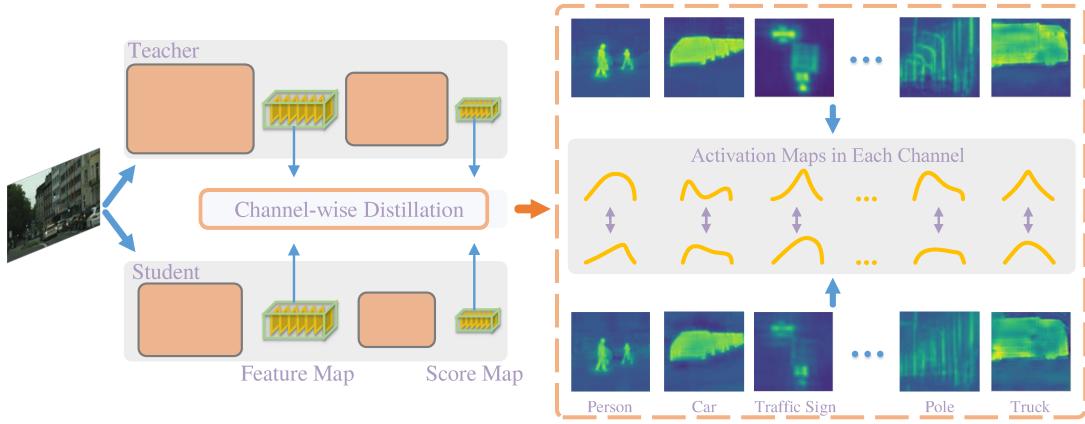


Figure 1 – The overall architecture of our proposed method. The plot on the left shows our teacher-student strategy, where the feature map and the logits map can be used for channel-wsie knowledge distillation. The plot on the right shows an intuitive illustration: activated regions correspond to scene categories.

Abstract

Knowledge distillation (KD) has been proven a simple and effective tool for training compact dense prediction models. Lightweight student networks are trained by extra supervision transferred from large teacher networks. Most previous KD variants for dense prediction tasks align the activation maps from the student and teacher network in the spatial domain, typically by normalizing the activation values on each spatial location and minimizing point-wise and/or pair-wise discrepancy. Different from the previous methods, here we propose to normalize the activation map of each channel to obtain a soft probability map. By simply minimizing the Kullback–Leibler (KL) divergence between the channel-wise probability map of the two networks, the distillation process pays more attention to the most salient regions of each channel, which are valuable for dense prediction tasks.

We conduct experiments on a few dense prediction tasks,

including semantic segmentation and object detection. Experiments demonstrate that our proposed method outperforms state-of-the-art distillation methods considerably, and can require less computational cost during training. In particular, we improve the RetinaNet detector (ResNet50 backbone) by 3.4% in mAP on the COCO dataset, and PSPNet (ResNet18 backbone) by 5.81% in mIoU on the Cityscapes dataset. Code is available at:

<https://git.io/Distiller>

1. Introduction

Dense prediction tasks are a group of fundamental tasks in computer vision, including semantic segmentation [45, 5] and object detection [20, 29]. These tasks require learning strong feature representations for complex scene understanding at the pixel level. Thus, state-of-the-art models usually need high computational costs, making them cumbersome to be deployed to mobile devices. As a result, compact networks designed for dense prediction tasks have drawn much attention. Moreover, effectively training lightweight networks has been studied in previous works using knowledge distillation (KD). A compact network is trained with the supervision of a large teacher network, and

*C. Shu and Y. Liu contributed equally.

¹Shanghai Em-Data Technology Co.

²The University of Adelaide, Australia

³Monash University, Australia

⁴Baidu Inc.

[†]Z Yan is the corresponding author. yanzheng@em-data.com.cn

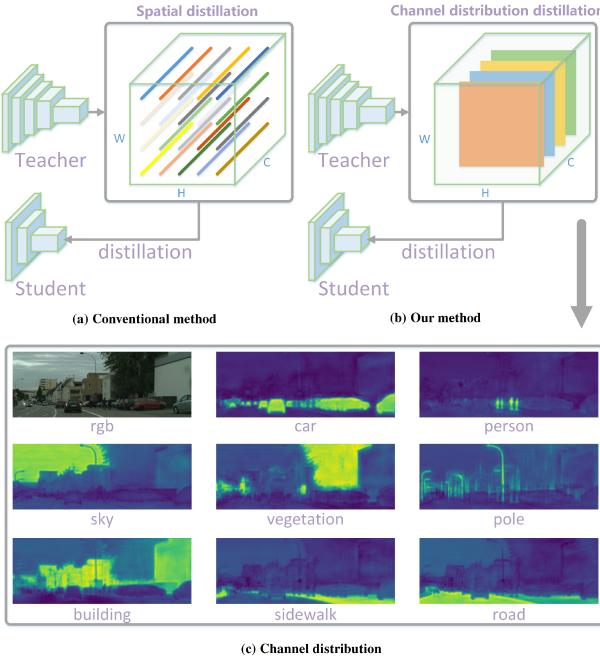


Figure 2 – Spatial knowledge distillation (top-left) works by aligning feature maps in the spatial domain. **Our channel distribution distillation** (top-right) instead aligns each channel of the student’s feature maps to that of the teacher network by minimizing the KL divergence. The bottom plot shows that the activation values of each channel tend to encode saliency of scene categories.

can achieve better performance. Pioneering works [15, 2] are proposed and well studied, mostly for image classification tasks.

Dense prediction tasks are per-pixel prediction problems, which are more challenging than image-level classification. Previous research [24, 19] found that directly transferring the KD methods [15, 2] in classification to semantic segmentation may not lead to satisfactory results. *Strictly aligning the point-wise classification scores or the feature maps between the teacher and student network may enforce overly strict constraints and lead to sub-optimal solutions.*

Recent works [24, 23, 17] pay attention to enforce the correlations among different spatial locations. As shown in Figure 2(a), the activation values¹ on each spatial location are normalized. Then, some tasks specific relationships are conducted by aggregating a sub-set of different spatial locations, such as pair-wise relations [24, 34] and inter-class relations [17]. Such methods may work better than the point-wise alignment in capturing spatial structure information and improve the performance of the student network.

¹The activation values in this work include the final logits and the inner feature maps.

However, every spatial location in the activation map contributes equally to the knowledge transferring, which may bring redundant information from the teacher network.

In this work, we propose a novel channel-wise knowledge distillation by normalizing the activation maps in each channel for dense prediction tasks, as shown in Figure 2(b). Then, we minimize the asymmetry Kullback–Leibler (KL) divergence of the normalized channel activation maps—which is converted into a distribution for each channel—between the teacher and the student networks. We show an example of the channel-wise distribution in Figure 2(c). The activations of each channel tend to encode saliency of scene categories. For each channel, the student network is guided to pay more attention to mimic the regions with significant activation values, leading to a more accurate localization in dense prediction tasks. For example, in object detection, the student network pays more attention to learn the activations of the foreground objects.

Some recent works exploit knowledge contained in channels. Channel distillation [47] proposes to transfer the activation in each channel into one aggregated scalar, which may be helpful for image-level classification, but the spatial aggregation loses all spatial information and thus is not suitable for dense prediction. Other works, such as MGD [39], Channel exchanging [32] and CSC [25] show the importance of channel-wise information. MGD matches the teacher channels with students’ and solves it as an assignment problem. Channel exchanging [32] uses a fusion module to dynamically exchange channels between sub-networks of various modalities.

We show that the simple normalizing operations for each channel can improve the baseline spatial distillation by a large margin. The proposed channel-wise distillation is simple and easy to apply to various tasks and network structures. We summarize our main contributions as follows.

- Unlike those existing spatial distillation approaches, we propose a novel channel-wise distillation paradigm for dense prediction tasks. Our method is simple yet effective.
- The proposed channel-wise distillation significantly outperforms state-of-the-art KD methods for semantic segmentation and object detection.
- We show consistent improvements on four benchmark datasets with various network structures on semantic segmentation and object detection tasks, demonstrating that our method is general. Given its simplicity and effectiveness, we believe that our method can serve as a strong baseline KD method for dense prediction tasks.

2. Related Work

Most works on knowledge distillation focus on classification tasks [10, 11, 15, 26, 35, 37, 43]. Our work here aims to study efficient and effective distillation methods for dense prediction, beyond naively applying pixel-wise distillation as done in classification.

Knowledge distillation for semantic segmentation. In [34], a local similarity map is constructed to minimize the discrepancy of segmented boundary information between the teacher and student network, where the Euclidean distance between the center pixel and the 8-neighborhood pixels is used as knowledge for transferring. Liu *et al.* [23, 24] propose two approaches to capture the structured information among pixels, including pair-wise similarity between pixels and holistic correlations captured by a discriminator. The work in [33] focuses on the intra-class feature variation among the pixels with the same label, where the set of cosine distance between each pixel’s feature and its corresponding class-wise prototype is constructed to transfer the structural knowledge. He *et al.* [13] use a feature adaptor is employed to mitigate the feature mismatching between the teacher and student networks.

Knowledge distillation for object detection. Many methods find that it is important to distinguish the foreground and the background regions in the distillation for object detection. MIMIC [19] forces the feature map inside the RPN of the student network to be similar to that of the teacher network via the L_2 loss, and finds that directly applying pixel-wise loss may harm the performance of object detection. Wang *et al.* [31] propose to distill the fine-grained feature near object anchor locations. Zhang and Ma [41] generate the mask with attention to distinguish the foreground and the background, achieving promising results. Instead, we softly align the channel-wise activations to distinguish the foreground and the background regions.

Channel-wise knowledge. Several recent works [47] also pay attention to the knowledge contained in each channel. Zhou *et al.* calculate the mean of the activation in each channel and align a weighted difference for each channel in classification. CSC [25] calculates the pair-wise relations among all spatial locations and all channels for transferring the knowledge. Channel exchanging [32] proposes that the information contained in each channel is general and can be shared across different modalities.

3. Our Method

We first review relevant spatial knowledge distillation methods in the literature.

3.1. Spatial Distillation

Existing KD methods often employ a point-wise alignment or align structured information among spatial loca-

tions, which can be formulated as:

$$\ell(y, y^S) + \alpha \cdot \varphi(\phi(y^T), \phi(y^S)). \quad (1)$$

Here the task loss $\ell(\cdot)$ is still applied with y being the ground-truth labels. For example, the cross-entropy loss is usually employed in semantic segmentation. By slightly abusing the notation, here y^S and y^T represent either the logits or inner activations of the student and teacher network, respectively. Here α is a hyper-parameter to balance the loss terms. Subscripts $.^T$ and $.^S$ denote teacher and student networks. We list representative spatial distillation methods in Table 1.

A brief overview of these methods is as follows. Attention Transfer (AT) [40] uses an attention mask to squeeze the feature maps into a single channel for distillation. The pixel-wise loss [16] directly aligns the point-wise class probabilities. The local affinity [34] is computed by the distance between the center pixel and its 8 neighborhood pixels. The pairwise affinity [24, 13, 23] is employed to transfer the similarity between pixel pairs. The similarity between each pixel’s feature and its corresponding class-wise prototype is computed to transfer the structural knowledge [33]. The holistic loss in [24, 23] uses the adversarial scheme to align the high-order relations between feature maps from the two networks. Note that, the last four terms consider the correlation among pixels. Existing KD methods as shown in Table 1 are all spatial distillation methods. All these methods consider the N channel activation values of a spatial location as the feature vectors to operate on.

3.2. Channel-wise Distillation

To better exploit the knowledge in each channel, we propose to *softly* align activations of corresponding channels between the teacher and student networks. To do so, we first convert activations of a channel into a probability distribution such that we can measure the discrepancy using a probability distance metric such as the KL divergence. As demonstrated in Figure 2(c), the activations of different channels tend to encode the saliency of scene categories of an input image. Besides, a well-trained teacher network for semantic segmentation shows activation maps of clear category-specific masks for each channel—which is expected—as displayed on the right part of Figure 1. Here, we propose a novel channel-wise distillation paradigm to guide the student to learn the knowledge from a well-trained teacher.

Let us denote the teacher and student networks as T and S , and the activation maps from T and S are y^T and y^S , respectively. The channel-wise distillation loss can be formulated as in a general form:

$$\varphi(\phi(y^T), \phi(y^S)) = \varphi(\phi(y_c^T), \phi(y_c^S)). \quad (2)$$

Loss	$\varphi(u, v)$	$\phi(x)$	
		Formulation	Dimensionality
Point-wise alignment			
Attention transfer [40] Pixelwise [24, 6, 23, 33]	L_1 or L_2 KL	$\sum_{c=1}^C \ x_{ic}\ ^p$ softmax(x_i/τ)	$1 \times W \times H$ $C \times W \times H$
Pairwise or higher order alignment			
Local similarity [34]	L_1 or L_2	$\sum_{j \in N(i)} \ x_j - x_i\ $	$1 \times W \times H$
Pairwise affinity [24, 13, 23]	L_2	$\frac{x_i^T x_j}{\ x_i\ _2 \cdot \ x_j\ _2}$	$1 \times W \times H$
IFVD [33]	L_2	$\cos(x_i, \sum_{j \in S_i} x_j / S_i)$	$1 \times W \times H$
Holistic [24, 23, 33]	Wasserstein Distance	$D(x_i)$	1

Table 1 – Current spatial distillation methods. i and j indicate the pixel index. $D(\cdot)$ is a discriminator, and $N(i)$ indicates 8-neighborhood of pixel i . S_i is the pixel set having the same label as pixel i and $|S_i|$ stands for the size of the set S_i .

In our case, $\phi(\cdot)$ is used to convert the activation values into a probability distribution as below:

$$\phi(y_c) = \frac{\exp(\frac{y_{c,i}}{\mathcal{T}})}{\sum_{i=1}^{W \cdot H} \exp(\frac{y_{c,i}}{\mathcal{T}})}, \quad (3)$$

where $c = 1, 2, \dots, C$ indexes the channel; and i indexes the spatial location of a channel. \mathcal{T} is a hyper-parameter (the temperature). The probability becomes softer if we use a larger \mathcal{T} , meaning that we focus on a wider spatial region for each channel. By applying the softmax normalization, we remove the influences of magnitude scales between the large networks and the compact networks. This normalization is helpful in KD as observed in [30]. A 1×1 convolution layer is employed to upsample the number of channels for the student network if the number of channels mismatches between the teacher and the student. $\varphi(\cdot)$ evaluates the discrepancy between the channel distribution from the teacher network and the student network. We use the KL divergence:

$$\varphi(y^T, y^S) = \frac{\mathcal{T}^2}{C} \sum_{c=1}^C \sum_{i=1}^{W \cdot H} \phi(y_{c,i}^T) \cdot \log \left[\frac{\phi(y_{c,i}^T)}{\phi(y_{c,i}^S)} \right]. \quad (4)$$

The KL divergence is an asymmetric metric. From Equation (4), we can see that, if $\phi(y_{c,i}^T)$ is large, $\phi(y_{c,i}^S)$ should be as large as $\phi(y_{c,i}^T)$ to minimize the KL divergence. Otherwise, if $\phi(y_{c,i}^T)$ is very small, the KL divergence pays less attention to minimize the $\phi(y_{c,i}^S)$. Thus, the student network tends to produce similar activation distribution in the foreground saliency, while the activations corresponding to the background region of the teacher network would have less impact on the learning. We hypothesize that this asymmetry property of KL benefits the KD learning for dense prediction tasks.

4. Experiments

In this section, we first describe the implementation details and the experiment settings. Then, we compare our

channel-wise distillation method with other state-of-the-art distillation methods and conduct ablation studies on semantic segmentation. Finally, we show consistent improvements in semantic segmentation and object detection with various benchmarks and student network structures.

4.1. Experimental Settings

Datasets. Three public semantic segmentation benchmarks, namely, Cityscapes [7], ADE20K [46] and Pascal VOC [9] are used here. We also apply the proposed distillation method to object detection on MS-COCO 2017 [22], which is a large-scale dataset that contains over 120k images of 80 categories.

The Cityscapes dataset is used for semantic urban scene understanding. It contains 5,000 finely annotated images with 2,975/500/1,525 images for training/validation/testing respectively, where 30 common classes are provided and 19 classes are used for evaluation and testing. The size of each image is 2048×1024 pixels. They are gathered from 50 different cities. The coarsely labeled data is not used in our experiments.

The Pascal VOC dataset contains 1,464/1,449/1,456 images for training/validation/testing. It contains 20 foreground object classes and an extra background class. In addition, the dataset is augmented by extra coarse labeling, which has 10,582 images for training. The training split is used for training, and the final performance is measured on the validation set across 21 classes.

The ADE20K dataset covers 150 classes of diverse scenes. It contains 20K/2K/3K images for training, validation, and testing. In our experiments, we report the segmentation accuracy on the validation set.

Evaluation metrics. To evaluate the performance and efficiency of our proposed channel distribution distillation method on semantic segmentation, following the previous work [17, 23], we test each strategy via the mean Intersection-over-Union (mIoU) in all experiments under a single-scale setting. The floating-point operations per second (FLOPs) are calculated with a fixed input size

Network		Structural	Complexity	Val mIoU (%)	
				Feature map	Logits map
Teacher	—	—	—	78.56	78.56
Student	—	—	—	69.10	69.10
Spatial Distillation	AT [40]	✗	$h_x \cdot w_x \cdot (c_x)^p$	72.37(+3.27) [®]	72.32(+3.22)
	PI [6, 33, 24, 23]	✗	$h_x \cdot w_x \cdot c_x$	70.02(+0.92) [®]	71.74(+2.64)
	LOCAL [34]	✓	$8h_x \cdot w_x \cdot c_x$	69.81(+0.71)	69.75(+0.65)
	PA [24, 13, 23]	✓	$(h_x \cdot w_x)^2 \cdot c_x$	71.23(+2.13)	71.41(+2.31)
	IFVD [33]	✓	$h_x \cdot w_x \cdot c_x \cdot n$	71.35(+2.25)	70.66(+1.56)
HO [24, 23, 33]	✓	$\mathcal{O}(D)$	— [®]	72.13(+3.03)	
	CD (Ours)	✓	$h_x \cdot w_x \cdot c_x$	74.27(+5.17) [®]	74.87(+5.77)

Table 2 – Comparison between computation complexity and performance on the validation set among various distillation methods. The mIoU is calculated on the Cityscapes validation set with PSPNet-R101 as the teacher network and PSPNet-R18 as the student network. The complexity depends on the shape ($h_x \times w_x \times c_x$) of the input. $\mathcal{O}(D)$ denotes the discriminator complexity. The superscript [®] means that additional channel alignment convolution is needed. All the results are the mean of three runs.

of 512×1024 pixels. Besides, the mean class Accuracy (mAcc) is listed for Pascal VOC and ADE20K. To evaluate the performance on object detection, we report the mean Average Precision (mAP), the inference speed (FPS), and the model size (parameters) following the work in [41].

Implementation details. For semantic segmentation, the teacher network is PSPNet with ResNet101 (PSPNet-R101) as the backbone for all experiments. We employ several different architectures, including PSPNet [45], Deeplab [42] with the backbones of ResNet18, and MobileNetV2 as student networks to verify the effectiveness of our method.

In the ablation study, we analyze the effectiveness of our method based on PSPNet with the ResNet18 backbone (PSPNet-R18). Unless otherwise indicated, each training image for the student network is randomly cropped into 512×512 pixels. The batch size is set to 8, and the number of the training step is 40K. We set the temperature parameter $T = 4$, the loss weight $\alpha = 3$ for the logits map, and $\alpha = 50$ for the feature map for all experiments. For object detection, we employ the same teacher and student networks and the training settings as in [41].

4.2. Comparison with Recent Knowledge Distillation Methods

To verify the effectiveness of our proposed channel-wise distillation, we compare our method with current distillation methods listed below:

- Attention Transfer (AT) [40]: Sergey *et al.* calculate the summation of all channels at each spatial location to obtain a single channel attention map. L_2 is employed to minimize the difference between the attention map.
- Local affinity (LOCAL) [34]: For each pixel, a local similarity map is constructed, which considers the correlations between itself and its 8 neighborhood pixels. L_2 is employed to minimize the difference between the local affinity map.
- Pixel-wise distillation (PI) [24, 23, 33, 6]: KL divergence is used to align the distribution of each spatial location from two networks.
- Pair-wise distillation (PA) [24, 13, 23]: The correlations between all pixel pairs are considered.
- Intra-class feature variation distillation (IFVD) [33]: The set of similarity between the feature of each pixel and its corresponding class-wise prototype is regarded as the intra-class feature variation to transfer the structural knowledge.
- Holistic distillation (HO) [24, 23, 33]: The holistic embeddings of feature maps are computed by a discriminator, which is used to minimize the discrepancy between high-order relations.

We apply all these popular distillation methods to both the inner feature map (logits map) and the final logits map. The conventional cross-entropy loss is applied in all experiments. The computational complexity and performance of spatial distillation methods are reported in Table 2.

Given the input feature map (logits map) of the size of $h_f \times w_f \times c$ ($h_s \times w_s \times n$), where $h_f(h_s) \times w_f(w_s)$ is the shape of the feature map (logits map). c is the number of channels and n is the number of classes.

As reported in Table 2, all distillation methods can improve the performance of the student network. Our channel distillation method outperforms all spatial distillation methods. Ours outperforms the best spatial distillation method (AT) by 2.5%. Moreover, our method is more efficient as it requires less computational cost than other methods during the training phase.

Furthermore, we list the detailed class IoU of our method and two recent state-of-the-art methods, PA [24] and IFVD [17] in Table 3. These methods propose to transfer structure information in semantic segmentation. Our methods significantly improve the class accuracy of several objects, such as traffic light, terrain, wall, truck, bus, and

Method	mIoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation
PA	71.41	97.30	80.48	90.76	37.89	52.78	60.33	63.48	74.06	91.69
IFVD	71.66	97.56	81.44	91.49	44.45	55.95	62.40	66.38	76.44	91.85
Ours	75.13	97.64	81.97	91.89	49.44	56.84	62.53	68.73	77.60	92.20
Class	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
PA	58.60	93.48	78.96	55.45	93.42	63.79	78.48	60.12	51.62	74.01
IFVD	61.29	93.97	78.64	52.33	93.50	60.25	74.70	58.81	44.85	75.41
Ours	63.37	94.32	80.06	58.49	94.18	70.31	85.61	72.85	52.92	76.58

Table 3 – The class IoU of our proposed channel-wise distillation method compared with other two typical structural knowledge transfer methods on the validation set of Cityscape, where PSPNet-R18 (1.0) is used as the student network.

train, indicating that the channel distribution can well transfer the structural knowledge.

4.3. Ablation Study

We show the effectiveness of the channel-wise distillation and discuss the choice of the hyper-parameters in semantic segmentation in this section. The baseline student model is PSPNet-R18, and the teacher model is the PSPNet-R101. All the results are evaluated on the validation set of Cityscapes.

Effectiveness of channel-wise distillation. The normalized channel-wise probability map and the asymmetric KL divergence play an important role in our distillation method. We conduct experiments with four different variants to show the effectiveness of proposed methods in Table 4.

All the distillation methods are applied to the same activation maps as input; and we use the same training scheme as described in Section 4.1.

‘PI’ represents the pixel-level knowledge distillation, which normalizes the activation of each spatial location. ‘ L_2 w/o NORM’ represents that we directly minimize the difference between the feature maps from two networks, which considers the difference at all locations in all channels equally. ‘Bhat’ is the Bhattacharyya distance [3], which is a symmetrical distribution measurement. It aligns the discrepancy in each channel.

From Table 4, we can see that the asymmetric KL divergence measuring the normalized channel discrepancy achieves the best performance. Note that as the KL divergence is asymmetric, the input of the student and teacher can not be swapped. We experiment by changing the order of the input in the KL divergence, and the training does not converge.

Impact of the temperature parameter and loss weights. We conduct experiments to vary the channel-wise probability maps by adjusting the temperature parameter \mathcal{T} under different loss weights α . The experiments are conducted on the logits map. Results are illustrated in Figure 3.

All the results are the mean of three runs. The loss weight is set to 1, 2, 3, and $\mathcal{T} \in [1, 5]$. The distribution tends to be softer if we increase \mathcal{T} .

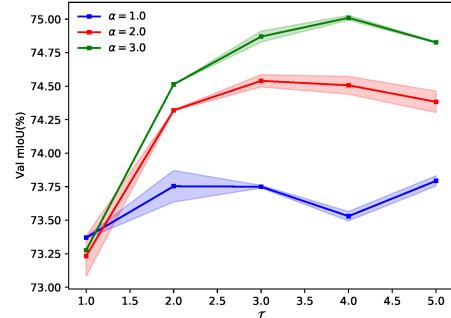


Figure 3 – Impact of the temperature parameter \mathcal{T} and the loss weight α .

From the figure, we can see that a softer probability map may help the knowledge distillation. Besides, in a certain range, the performance is stable. The performance appears to drop if \mathcal{T} is set to be small. In such cases, the method only focuses on limited salient pixels. We attain the best performance when $\mathcal{T} = 4$ and $\alpha = 3$ with the PSPNet18 on the Cityscapes validation set.

4.4. Semantic Segmentation

We demonstrate that our proposed channel wise distillation method can be combined with previous semantic seg-

Method	Norm.	Divergence	Logits map	Feature map
Teacher	-	-	78.56	78.56
Student	-	-	69.10	69.10
PI	Spatial	KL	71.74	70.02
L_2 w/o norm.	None	MSE	70.83	71.37
L_2	Channel	MSE	71.60	71.57
Bhat	Channel	Bhat	72.21	71.96
Ours	Channel	KL	74.87	74.27

Table 4 – Mean IoU on the Cityscapes validation set. We can see that with the channel normalization and the asymmetry KL divergence, the proposed channel-wise distillation achieves the best performance among other variants. All the results are the average of three runs.

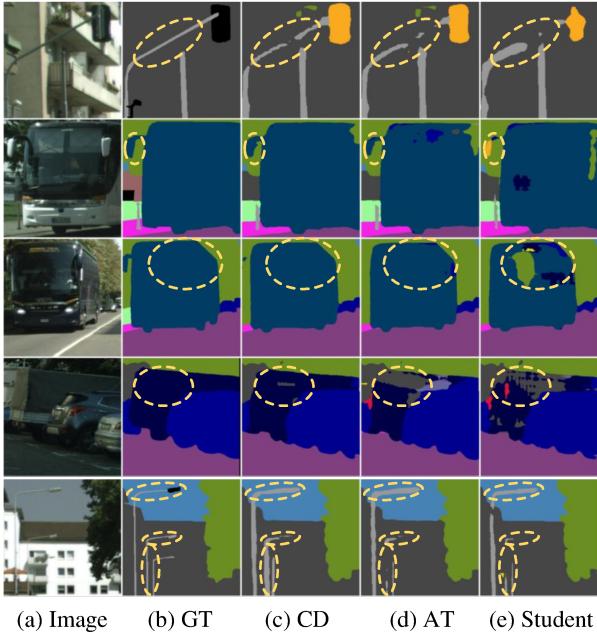


Figure 4 – Qualitative segmentation results on Cityscapes of the PSPNet-R18 model: (a) raw images, (b) ground truth (GT), (c) channel-wise distillation (CD), (d) the best spatial distillation schemes: attention transfer (AT); and (e) the output of the original student model without KD.

mentation distillation methods, *i.e.*, structural knowledge distillation for segmentation/dense prediction (SKDS [23] and SKDD [24]) and intra-class feature variation distillation (IFVD [33]), under various student networks.

We use the proposed channel-wise distillation both on the logits map (Ours-logits) and the feature map (Ours-feature). The pixel-wise distillation (PI) and the holistic distillation (HO) on the logits map are also included following previous methods [24, 17].

We first evaluate the performance of our method on the Cityscapes dataset. Various student networks with different encoders and decoders are used to verify the effectiveness of our method. Encoders include ResNet18 (initialized with or without the weights pre-trained on ImageNet, and a channel-halved variant of ResNet18 [12]), and decoders include PSPhead [45] and ASPPhead [5]. Table 5 shows the results on Cityscapes. Experiment results on Pascal VOC [9] and ADE20K [46] are shown in the supplementary materials.

Our method outperforms SKD and IFVD on five student networks and three benchmarks, which further indicates that the channel-wise distillation is effective for semantic segmentation.

For the student with the same architectural type as the teacher, *i.e.*, PSPNet-R18[°](0.5), PSPNet-R18[°] and

Method	Params (M)	FLOPs (G)	mIoU (%)	
			Val	Test
ENet [1]	0.358	3.612	—	58.3
ESPNet [28]	0.363	4.422	—	60.3
ERFNet [8]	2.067	25.60	—	68.0
ICNet [44]	26.50	28.30	—	69.5
FCN [18]	134.5	333.9	—	62.7
RefineNet [21]	118.1	525.7	—	73.6
OCNet [38]	62.58	548.5	—	80.1
Results w/ and w/o distillation schemes				
T:PSPNet [45]	70.43	574.9	78.5	78.4
S:PSPNet-R18 [°] (0.5)	3.271	31.53	61.17	—
+SKDS [23]	3.271	31.53	61.60	60.50
+SKDD [24]	3.271	31.53	62.35	—
+IFVD [33]	3.271	31.53	63.35	63.68
+Ours-feature	3.271	31.53	63.06	63.12
+Ours-logits	3.271	31.53	68.57	66.75
S:PSPNet-R18 [°]	13.07	125.8	63.63	—
+SKDS [23]	13.07	125.8	63.20	62.10
+SKDD [24]	13.07	125.8	64.68	—
+IFVD [33]	13.07	125.8	66.63	65.72
Ours-feature	13.07	125.8	66.85	66.03
Ours-logits	13.07	125.8	71.03	70.43
S:PSPNet-R18*	13.07	125.8	70.09	67.60
+SKDS [23]	13.07	125.8	72.70	71.40
+SKDD [24]	13.07	125.8	74.08	—
+IFVD [33]	13.07	125.8	74.54	72.74
+Ours-feature	13.07	125.8	74.63	73.22
+Ours-logits	13.07	125.8	75.90	74.58
S:Deeplab-R18 [°] (0.5)	3.15	31.06	61.83	60.51
+SKDS [23]	3.15	31.06	62.71	61.69
+IFVD [33]	3.15	31.06	63.12	62.37
+Ours-feature	3.15	31.06	64.61	63.18
+Ours-logits	3.15	31.06	67.44	67.12
S:Deeplab-R18 [°]	12.62	123.9	73.37	72.39
+SKDS [23]	12.62	123.9	73.87	72.63
+IFVD [33]	12.62	123.9	74.09	72.97
+Ours-feature	12.62	123.9	74.24	72.56
+Ours-logits	12.62	123.9	75.91	74.32

Table 5 – Comparison of student variants with the state-of-the-art distillation methods on the Cityscapes dataset. [°] means that the models are trained from scratch, and * indicates that models are initialized by the weights pre-trained on ImageNet. R18 stands for Resnet18.

PSPNet-R18*, the improvements are more significant. As for the student with different architectural types with the teacher, *i.e.*, Deeplab-R18[°](0.5) and Deeplab-R18*, our method achieves consistent improvement compared with SKDS and IFVD. Thus, our method works well with different teacher and student networks.

The student network of a compact model capacity (PSPNet-R18[°](0.5)) shows inferior distillation performance (68.57%) compared to the student with a larger capacity (PSPNet-R18*) (75.90%). This may be attributed to the fact that the capability of small networks is limited com-

Model		Backbone	AP (%)	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	FPS	Params.
Two-stage detector	Faster RCNN	R50	38.4	59.0	42.0	21.5	42.1	50.3	18.1	43.57
	+Chen et al. [4]		38.7	59.0	42.1	22.0	41.9	51.0	18.1	43.57
	+Wang et al. [31]		39.1	59.8	42.8	22.2	42.9	51.1	18.1	43.57
	+Heo et al. [14]		38.9	60.1	42.6	21.8	42.7	50.7	18.1	43.57
	+Zhang et al. [41]		41.5	62.2	45.1	23.5	45.0	55.3	18.1	43.57
	+Our Method		41.7	62.0	45.5	23.3	45.5	55.5	18.1	43.57
One-stage detector	RetinaNet	R50	37.4	56.7	39.6	20.0	40.7	49.7	20.0	36.19
	+Heo et al. [14]		37.8	58.3	41.1	21.6	41.2	48.3	20.0	36.19
	+Zhang et al. [41]		39.6	58.8	42.1	22.7	43.3	52.5	20.0	36.19
	+Our Method		40.8	60.4	43.4	22.7	44.5	55.3	20.0	36.19
Anchor-free detector	RepPoints	R50	38.6	59.6	41.6	22.5	42.2	50.4	18.2	36.62
	+Zhang et al. [41]		40.6	61.7	43.8	23.4	44.6	53.0	18.2	36.62
	+Our Method		42.0	63.0	45.3	24.1	46.1	55.0	18.2	36.62

Table 6 – Comparison between our methods and other distillation methods on object detection.

pared with the teacher network and can not sufficiently absorb the knowledge of the current task. For PSPNet-R18, the student initialized by the weights trained on ImageNet obtains the best distillation performance (improved from 70.09% to 75.90%), further demonstrating that the well-initialized parameters help the distillation. Thus, the better student lead to better distillation performance, but the improvement is less significant as the gap between the teacher and student network is smaller.

4.5. Object Detection

We also apply our channel-wise distillation method on the object detection task. The experiments are conducted on MS COCO2017 [22].

Various student networks under different paradigms, *i.e.*, a two-stage anchor-based method (Faster RCNN [27]), a one-stage anchor-based method (RetinaNet [20]) and anchor-free method (RepPoints [36]), are used to validate the effectiveness of our method. To make a fair comparison, we experiment on the same teacher with the same hyperparameters as in [41].

The only modification is that the feature alignment is changed to our channel-wise distillation. The results are reported in Table 6. From the table, we can see that our methods achieve consistent improvements (about 3.4% mAP) on strong baseline student networks. Compared with previous state-of-the-art distillation methods [41], our simple channel-wise distillation performs better, especially with anchor-free methods. We improve the RepPoint by 3.4% while Zhang *et al.* improve the RepPoint by 2%. Besides, we can see that the proposed distillation method can improve AP₇₅ more significantly.

5. Conclusion

In this paper, we have proposed a novel channel-wise distillation for dense prediction tasks. Different from previous spatial distillation methods, we normalize the activations of each channel to a probability map. Then, the asymmetry KL divergence is applied to minimize the discrepancy between the teacher and the student network. Experimental results show that the proposed distillation method consistently outperforms state-of-the-art distillation methods on four public benchmark datasets with various network backbones, for both semantic segmentation and object detection.

Additionally, our ablation experiments demonstrate the efficiency and effectiveness of our channel-wise distillation, and it can further complement the spatial distillation methods. We hope that the proposed simple and effective distillation method can serve as a strong baseline for effectively training compact networks for many other dense prediction tasks, including instance segmentation, depth estimation and panoptic segmentation.

References

- [1] Paszke Adam, Chaurasia Abhishek, Kim Sangpil, and Culurciello Eugenio. Enet: A deep neural network architecture for real-time semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.
- [2] Romero Adriana, Ballas Nicolas, Ebrahimi Kahou Samira, Chassang Antoine, Gatta Carlo, and Bengio Yoshua. Fitnets: Hints for thin deep nets. *Int. Conf. Learn. Represent.*, 2015.
- [3] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.
- [4] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Adv. Neural Inform. Process. Syst.*, pages 742–751, 2017.

- [5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.
- [6] Wuyang Chen, Xinyu Gong, Xianming Liu, Qian Zhang, Yuan Li, and Zhangyang Wang. Fasterseg: Searching for faster real-time semantic segmentation. *Int. Conf. Learn. Represent.*, 2020.
- [7] 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. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.
- [8] Romera Eduardo, Álvarez José M, Bergasa Luis M, and Arroyo Roberto. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Trans. Intell. Transportation Syst.*, 2017.
- [9] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 2010.
- [10] Jie Fu, Xue Geng, Zhijian Duan, Bohan Zhuang, Xingdi Yuan, Adam Trischler, Jie Lin, Chris Pal, and Hao Dong. Role-wise data augmentation for knowledge distillation. *Int. Conf. Learn. Represent.*, 2020.
- [11] Yushuo Guan, Pengyu Zhao, Bingxuan Wang, Yuanxing Zhang, Cong Yao, Kaigui Bian, and Jian Tang. Differentiable feature aggregation search for knowledge distillation. In *Eur. Conf. Comput. Vis.*, 2020.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.
- [13] Tong He, Chunhua Shen, Tian Zhi, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
- [14] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and JinYoung Choi. A comprehensive overhaul of feature distillation. In *Int. Conf. Comput. Vis.*, pages 1921–19302, 2019.
- [15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Adv. Neural Inform. Process. Syst.*, 2014.
- [16] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv:*, abs/1503.02531, 2015.
- [17] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-region affinity distillation for road marking segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12486–12495, 2020.
- [18] Long Jonathan, Shelhamer Evan, and Darrell Trevor. Fully convolutional networks for semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.
- [19] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.
- [20] Lin, Goyal Tsung-Yi, Girshick Priya, He Ross, Dollár Kaiming, and Piotr. Focal loss for dense object detection. In *Int. Conf. Comput. Vis.*, 2017.
- [21] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.
- [22] 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 *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014.
- [23] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
- [24] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.
- [25] Sangyong Park and Yong Seok Heo. Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy. *Sensors*, 20(16):4616, 2020.
- [26] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3967–3976, 2019.
- [27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2016.
- [28] Mehta Sachin, Rastegari Mohammad, Caspi Anat, Shapiro Linda, and Hajishirzi Hannaneh. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. *Eur. Conf. Comput. Vis.*, 2018.
- [29] Tian, Shen Zhi, Chen Chunhua, He Hao, and Tong. Fcos: Fully convolutional one-stage object detection. In *Int. Conf. Comput. Vis.*, 2019.
- [30] Guo-Hua Wang, Yifan Ge, and Jianxin Wu. In defense of feature mimicking for knowledge distillation. *arXiv preprint arXiv:2011.01424*, 2020.
- [31] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4933–4942, 2019.
- [32] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Adv. Neural Inform. Process. Syst.*, 33, 2020.
- [33] Yukang Wang, Zhou Wei, Jiang Tao, Bai Xiang, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. *Eur. Conf. Comput. Vis.*, 2020.
- [34] Jiafeng Xie, Bing Shuai, JianFang Hu, Jingyang Lin, and WeiShi Zheng. Improving fast segmentation with teacher-student learning. *Brit. Mach. Vis. Conf.*, 2018.
- [35] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3967–3976, 2019.

- [36] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Int. Conf. Comput. Vis.*, pages 9657–9666, 2019.
- [37] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3903–3911, 2020.
- [38] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.
- [39] Kaiyu Yue, Jiangfan Deng, and Feng Zhou. Matching guided distillation. *Eur. Conf. Comput. Vis.*, 2020.
- [40] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *Int. Conf. Learn. Represent.*, 2017.
- [41] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *Int. Conf. Learn. Represent.*, 2021.
- [42] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.
- [43] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.
- [44] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. *Eur. Conf. Comput. Vis.*, 2018.
- [45] Hengshuang Zhao1, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.
- [46] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.
- [47] Zaida Zhou, Zhuge Chaoran, Xinwei Guan, and Wen Liu. Channel distillation: Channel-wise attention for knowledge distillation. *arXiv:*, abs/2006.01683, 2020.