# Distilling Inter-Class Distance for Semantic Segmentation

**Zhengbo Zhang**[1] , **Chunluan Zhou**[2] , **Zhigang Tu**[1*]
[1]Wuhan University
[2]Wormpex AI Research
{zhangzb, tuzhigang}@whu.edu.cn, czhou002@e.ntu.edu.sg

## Abstract

Knowledge distillation is widely adopted in semantic segmentation to reduce the computation cost. The previous knowledge distillation methods for semantic segmentation focus on pixel-wise feature alignment and intra-class feature variation distillation, neglecting to transfer the knowledge of the inter-class distance in the feature space, which is important for semantic segmentation. To address this issue, we propose an Inter-class Distance Distillation (IDD) method to transfer the inter-class distance in the feature space from the teacher network to the student network. Furthermore, semantic segmentation is a position-dependent task, thus we exploit a position information distillation module to help the student network encode more position information. Extensive experiments on three popular datasets: Cityscapes, Pascal VOC and ADE20K show that our method is helpful to improve the accuracy of semantic segmentation models and achieves the state-of-the-art performance. E.g. it boosts the benchmark model ("PSPNet+ResNet18") by 7.50% in accuracy on the Cityscapes dataset.

## 1 Introduction

Semantic segmentation aims at allocating a label for each pixel of the input image. It is a basic and challenging task in computer vision, which has widely applied in many fields, e.g. autonomous driving [Dong *et al.*, 2020], ground feature changing detection [Kemker *et al.*, 2018], etc. Recently, due to the success of deep learning [Tu *et al.*, 2019] in computer vision, Convolutional Neural Networks (CNNs) based methods have greatly improved the accuracy of semantic segmentation. However, CNN based semantic segmentation algorithms usually have an expensive computational cost, which limits their application in practice, especially for the real-life tasks that demand high efficiency.

To address this issue, many lightweight models have been explored, e.g. ENet [Paszke *et al.*, 2016], ESPNet [Mehta *et al.*, 2018], ICNet [Zhao *et al.*, 2018], and STDC [Fan *et al.*, 2021]. Although researchers have designed excellent networks to reduce the cost of computation, it is difficult to reach a satisfactory compromise between accuracy and model size. Instead of redesigning the backbone, we adopt the knowledge distillation (KD) strategy to train a student network by the guidance of a teacher network, and obtain comparable result.

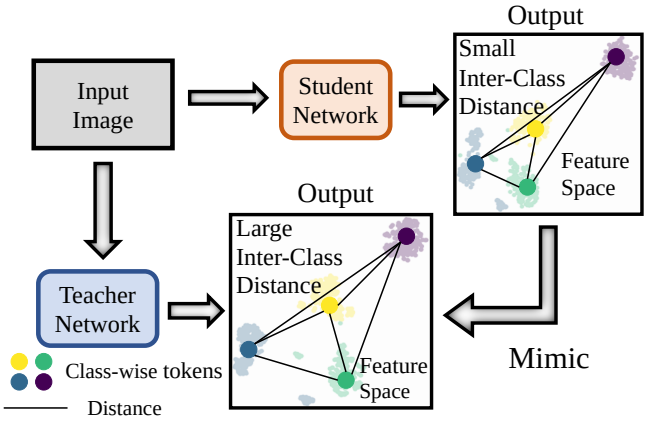


Figure 1: Limited by simple network structure and few parameters, the student network cannot have large inter-class distance like the teacher network in the feature space. Our motivation is to transfer the inter-class distance of the teacher network to help student network improve the segmentation accuracy.

KD [Hinton *et al.*, 2015], as a model compression method, is originally used in the image classification task, which is able to simplify the cumbersome model significantly. Due to the advantage of KD, some semantic segmentation approaches use KD to reduce the model size [Liu *et al.*, 2019; Wang *et al.*, 2020; Shu *et al.*, 2021]. They force the student model to learn the pixel-wise feature and intra-class feature variation from the teacher network. Representatively, Intra-Class Feature Variance Distillation (IFVD) [Wang *et al.*, 2020] focuses on transferring the variation of the intra-class feature from the teacher network to the student network. Channel-wise Knowledge Distillation (CD) [Shu *et al.*, 2021] emphasizes on distilling the most significant areas in each channel. It is worth noting that semantic segmentation is a pixel-wise category prediction task with various categories, thus the inter-class distance in the feature space is ubiquitous in semantic segmentation. Due to the help of numerous

*Corresponding author: Zhigang Tu

parameters and complex network structure, the teacher network has stronger classification ability and large inter-class distance in the feature space. However, *Issue 1*: the past KD schemes for semantic segmentation neglect to transfer the inter-class distance in the feature space of the teacher network to the student network.

Moreover, CNNs are able to encode the position information implicitly [Islam *et al.*, 2020]. Semantic segmentation is a position-dependent task. Generally, with the simple network structure and a few parameters, *Issue 2*: the student network is unable to encode as rich position information as the teacher network.

To address the above mentioned issues, we consider to distill the inter-class distance in the feature space and position information from the teacher network to the student network. Accordingly, we propose a novel method (see Figure 1) called *Inter-class Distance Distillation* (IDD). It consists of two main components. One is the *inter-class distance distillation module* (IDDM), we design a graph to encode the inter-class distance, and make the student network mimic the large inter-class distance of the teacher network. The other is the *position information distillation module* (PIDM). We design a position information network to extract the position information implicitly encoded in the feature map. Both the teacher network and the student network will predict the absolute coordinate mask via this network. By minimizing the divergence of them, the student network can encode more position information. With our IDD method, the student network learns more knowledge about inter-class distance and position information, improving the segmentation accuracy of the student network significantly.

The contributions are summarized in three-fold:

- We propose a novel approach named *Inter-class Distance Distillation* (IDD) for semantic segmentation. It is the first method to distill the inter-class distance among all KD schemes for semantic segmentation to the best of our knowledge.
- We design a *position information distillation module* (PIDM) to enhance the capability of the student network encoding position information.
- We demonstrate the effectiveness of the IDD method on three famous benchmark datasets, which not only obtains the state-of-the-art accuracy among KD schemes for semantic segmentation, but also is useful for other semantic segmentation models.

## 2 Related Work

**Semantic segmentation.** CNN based models have greatly promoted the progress of semantic segmentation. Many researchers have tried different methods to make the model to learn rich contextual information. [Zhao *et al.*, 2017] proposed a pyramid pooling strategy to collect context information from multiple scales. DeepLabv2 [Chen *et al.*, 2017] adopted the atrous spatial pyramid pooling approach to get abundant context information. An encoder-decoder module was designed to capture multilevel features and contextual information. OCNet [Yuan *et al.*, 2018] exploited a self-attention mechanism to capture relationships between all pixels. To meet the real-time semantic segmentation requirement of the mobile platform, some lightweight networks were proposed. ENet [Paszke *et al.*, 2016] used an asymmetric encoder-decoder structure and a convolution kernel decomposition operation, which greatly reduce the number of parameters and the floating point operations. Point-wise convolutions and spatial pyramid of dilated convolutions were applied in ESPNet [Mehta *et al.*, 2018] to decrease the cost of computation. ICNet [Zhao *et al.*, 2018] achieved fast semantic segmentation by designing an efficient network structure to process images with different resolutions. [Fan *et al.*, 2021] designed a new real-time segmentation architecture by reducing network redundancy. Different from [Mehta *et al.*, 2018; Zhao *et al.*, 2018], we get the lightweight semantic segmentation network with the usage of KD, which avoids to redesign the network structure, and gains high efficiency.

**KD for semantic segmentation.** [Hinton *et al.*, 2015] proposed the concept of KD, it is a process of transferring the soft-labels from the teacher network to the student network to improve the performance of the student network. Because of the remarkable performance of KD, some researchers applied KD to semantic segmentation. [Liu *et al.*, 2019] used a structured KD approach to transfer pixel-wise, pair-wise, and holistic knowledge from the teacher network. [He *et al.*, 2019] designed an autoencoder to transform knowledge into a compact form which is easier for the student network to learn. [Wang *et al.*, 2020] presented an intra-class feature variation distillation scheme to make the student network simulate the intra-class feature distribution of the teacher network. [Shu *et al.*, 2021] exploited a simple yet effective approach to minimize the channel-wise discrepancy between the teacher network and the student network. Unlike these mentioned approaches, our method pays attention to distilling the inter-class distance in the feature space, which is complementary to the previous distillation of pixel-wise feature alignment and intra-class feature variation.

## 3 Proposed Method

In this section, we first give an overview the general framework of past KD methods for semantic segmentation and our IDD model, then we describe the IDDM and PIDM in detail.

### 3.1 Overview

Semantic segmentation is a dense prediction task, aiming to assign a label to each pixel. Though the previous KD based semantic segmentation methods have achieved good progress, they mainly focus on aligning the pixel-wise feature and intra-class feature variance. Their loss function can be generally formulated as:

$$\begin{aligned} &Loss = L_{tar}(D(\mathbf{GT}), D(\mathbf{F}^S)) + \lambda \cdot L_{dis}(\varphi(\mathbf{F}^T), \varphi(\mathbf{F}^S)), \\ &L_{tar}(D(\mathbf{GT}), D(\mathbf{F}^S)) = -\sum_{k=1}^{N} D(\mathbf{GT}_k) \cdot \log(D(\mathbf{F}_k^S)). \end{aligned} \tag{1}$$

$L_{tar}$ is the cross-entropy loss, $\mathbf{GT}$ is the ground-truth, $\mathbf{F}^S$ and $\mathbf{F}^T$ denote the feature map of the student network and the teacher network, respectively. $\varphi(\cdot)$ represents a mapping

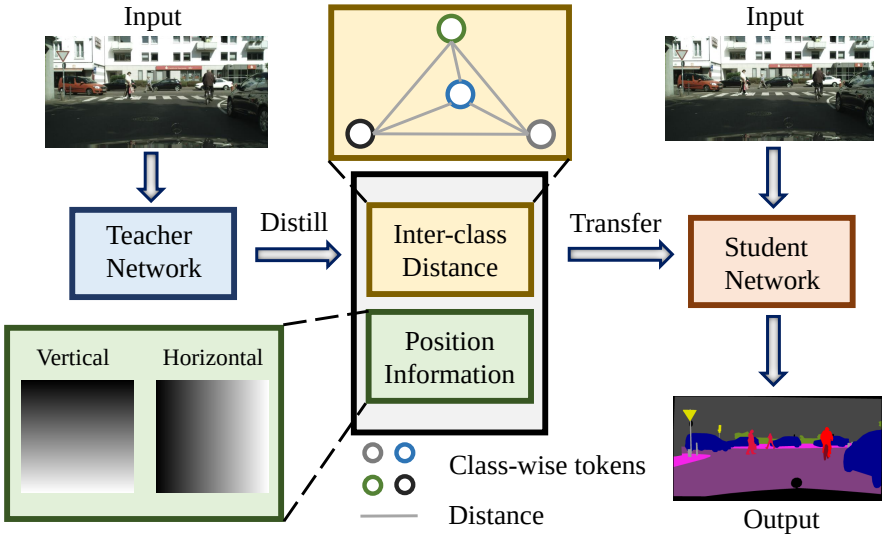


Figure 2: The network of our IDD method for semantic segmentation. We design a graph to encode the inter-class distance in the teacher network and transfer the inter-class distance to the student network. Besides, we transfer rich position information which is implicitly encoded in the teacher network to the student network.

function. $D(\mathbf{GT})$ and $D(\mathbf{F}^S)$ separately denote the ground-truth and the student network's category probability distributions of all pixels. $N$ is the number of pixels, $D(\mathbf{GT}_k)$ denotes the $k^{th}$ pixel's ground-truth category probability distribution, $D(\mathbf{F}_k^S)$ is the $k^{th}$ pixel's category probability distribution produced by the student network. $\lambda$ is a hyper-parameter to control the weight of loss. $L_{dis}(\cdot)$ is a loss function, such as the mean-squared error loss. Obviously, the prior methods ignore to transfer inter-class distance in the teacher network to the student network. Therefore, as illustrated in Figure 2, we propose the IDD method to transfer the inter-class distance and position information from the teacher to the student. We detail each module in the following subsections.

### 3.2 Inter-Class Distance Distillation Module

Semantic segmentation is a pixel-wise classification task. Limited by simple network structure and few parameters, the student network has relatively poor discriminating ability and small inter-class distance. We propose the inter-class distance distillation module to deal with this challenge.

As illustrated in Figure 2, we construct a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ to encode the inter-class category distance, where $\mathcal{V} = \{v_i \mid i = 1, ..., N\}$ is a group of nodes, $N$ denotes the total number of segmentation categories of the processed image and $\mathcal{E} = \{e_{i,j} \mid i = 1, ...N; j = 1, ...N; i \neq j\}$ represents a group of edges. $v_i$ denotes the token of the $i^{th}$ class, $v_i$ is obtained by averaging the feature of all pixels with the same category label $i$. $e_{i,j}$ is the Euclidean distance between the class-wise tokens of the $i^{th}$ and the $j^{th}$ category, which is defined as:

$$e_{i,j} = Dis(v_i, v_j). \tag{2}$$

It represents the feature distance between the $i^{th}$ class and the $j^{th}$ class, and $Dis$ is the Euclidean distance. Due to the deep network and numerous parameters, the teacher network have large inter-class distance. Inspired by this characteristic, to enable the student network to better simulate the teacher network in terms of the inter-class distance, we design an inter-class distance loss function $L_{id}$, which is defined as:

$$L_{id} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(e_{i,j}^T - e_{i,j}^S\right)^2, i \neq j, \tag{3}$$

where $e_{i,j}^T$ and $e_{i,j}^S$ stand for the $e_{i,j}$ in the teacher network and the student network, respectively.

### 3.3 Position Information Distillation Module

Semantic segmentation is a position-dependent task. It is reported in [Islam *et al.*, 2020] that CNNs have the ability to encode position information. Inspired by [Islam *et al.*, 2020], we further introduce a position information distillation module to enhance the capability of the student network predicting position information. As a result, the student network can encode more position information in its output features which could be utilized to improve segmentation accuracy.

Specifically, we use $\mathbf{A} \in \mathbb{R}^{C \times H \times W}$ to represent the input feature map. First, we input $\mathbf{A}$ into a pretrained position information network to get the position information masks $\mathbf{P}^{HOR} \in \mathbb{R}^{H \times W}$ and $\mathbf{P}^{VER} \in \mathbb{R}^{H \times W}$, which represent the abscissa and ordinate respectively. In $\mathbf{P}^{HOR}$, each column has the same value, and we use $\mathbf{V}_j^{HOR} (j \in [1, H])$ to represent the value of column $j$, where $\mathbf{V}_j^{HOR} = j$. In $\mathbf{P}^{VER}$, each row has the same value, and we use $\mathbf{V}_i^{VER} (i \in [1, W])$ to denote the value of row $i$, where $\mathbf{V}_i^{VER} = i$.

We construct a loss function $L_{pi}$ to transfer the position information of the teacher network to the student network, it is expressed as:

$$L_{pi} = \frac{1}{2} \cdot L_{pi}^{HOR} + \frac{1}{2} \cdot L_{pi}^{VER}, \tag{4}$$

| Backbone | $L_{skd}$ | $L_{cw}$ | $L_{id}$ | $L_{pi}$ | mIoU (%) |
|---|---|---|---|---|---|
| T: ResNet101 | | | | | 78.56 |
| S: ResNet18 | | | | | 70.09 |
| S: ResNet18 | ✓ | | | | 73.03 |
| S: ResNet18 | ✓ | ✓ | | | 75.78 |
| S: ResNet18 | ✓ | ✓ | ✓ | | 76.81 |
| S: ResNet18 | ✓ | ✓ | | ✓ | 76.43 |
| S: ResNet18 | ✓ | ✓ | ✓ | ✓ | 77.59 |

Table 1: Ablative studies of our loss items: $L_{skd}$, $L_{cw}$, $L_{id}$, and $L_{pi}$ on the Cityscapes validation dataset. "T: ResNet101" and "S: ResNet18" in the column of "Backbone" mean that we select ResNet101 and ResNet18 with PSPNet as the backbone for the teacher network and the student network, respectively.

| Method | mIoU (%) | Params (M) | FLOPs (G) |
|---|---|---|---|
| ENet | 58.3 | 0.358 | 3.612 |
| ESPNet | 60.3 | 0.364 | 4.422 |
| ERFNet | 68.0 | 2.067 | 25.60 |
| ICNet | 69.5 | 26.50 | 28.30 |
| FCN | 62.7 | 134.5 | 333.9 |
| RefineNet | 73.6 | 118.1 | 525.7 |
| OCNet | 80.1 | 62.58 | 548.5 |
| T: PSPNet-R101 | 78.4 | 70.43 | 574.9 |
| S: PSPNet-R18 | 67.60 | 13.07 | 125.8 |
| **S: +Ours (IDD)** | **76.33** | **13.07** | **125.8** |

Table 2: Comparison the performance of different lightweight semantic segmentation models on the Cityscapes testing set.

where

$$L_{pi}^{HOR} = \sum_{j=1}^{H} \left\| \frac{Q_j^{HOR\_T}}{\left\|Q_j^{HOR\_T}\right\|_2} - \frac{Q_j^{HOR\_S}}{\left\|Q_j^{HOR\_S}\right\|_2} \right\|_2,$$
$$L_{pi}^{VER} = \sum_{i=1}^{W} \left\| \frac{Q_i^{VER\_T}}{\left\|Q_i^{VER\_T}\right\|_2} - \frac{Q_i^{HOR\_S}}{\left\|Q_i^{HOR\_S}\right\|_2} \right\|_2 \quad (5)$$

represent $L_{pi}$ in the horizontal and the vertical directions, respectively. $Q_j^{VER\_T}$ and $Q_j^{VER\_S}$ denote the column $j$ of $\mathbf{P}^{VER}$ produced by the teacher network and the student network in the vectorized form. Analogically, $Q_i^{HOR\_T}$ and $Q_i^{HOR\_S}$ denote the row $i$ of $\mathbf{P}^{HOR}$ produced by the teacher network and the student network in the vectorized form.

### 3.4 Loss Function

Following [Shu *et al.*, 2021], we also apply the channel-wise supervision $L_{cw}$ to minimize the Kullback–Leibler (KL) divergence of the channel-wise probability map between the teacher network and the student network. The final loss function of our IDD method is formulated as:

$$L = L_{skd} + \lambda_1 \cdot L_{cw} + \lambda_2 \cdot L_{id} + \lambda_3 \cdot L_{pi}, \quad (6)$$

where $L_{skd}$ is a structured KD loss for semantic segmentation [Liu *et al.*, 2019], $\lambda_1$, $\lambda_2$ and $\lambda_3$ are the hyper parameters to balance the weight between different items.

## 4 Experiments

To verify the effectiveness of our proposed IDD based semantic segmentation method, we conduct comprehensive experiments on three popular benchmarks: Cityscapes [Cordts *et al.*, 2016], Pascal VOC [Everingham *et al.*, 2015], and ADE20K [Zhou *et al.*, 2017]. In the next subsections, we first introduce the datasets, evaluation metrics and implementation details. Next, we perform ablation experiments on the Cityscapes dataset. Finally, we compare our model with the state-of-the-art lightweight models on Cityscapes, Pascal VOC, and ADE20K.

### 4.1 Datasets and Evaluation Metrics

**Datasets.** *Cityscapes* includes 5000 finely annotated images of driving scenes in cities. It consists of 2975, 500 and 1525 images for training, validation and testing, respectively. It is labeled with 19 semantic categories. The resolution of each image is $2048 \times 1024$. In our experiments, we do not use the coarsely labeled images. *Pascal VOC* composes of 1464 images for training, 1449 images for validation and 1456 images for testing. It covers 20 foreground object classes and 1 background class. *ADE20k* is a challenging scene parsing dataset released by MIT, which contains 20K, 2K, 3K images with 150 classes for training, validation, and testing.

**Evaluation metrics.** We use the Intersection-over-Union (IoU) of each class and the mean IoU (mIoU) of all classes to measure the segmentation accuracy.The total number of model parameters (Params) is utilized to measure the model size. We adopt an input image with resolution $512 \times 1024$ to calculate the floating-point operations per second (FLOPs), which is a general metric to measure the model complexity.

### 4.2 Implementation Details

**Networks.** To make a fair and comparable evaluation, we carry out experiments on the same teacher and student network as [Liu *et al.*, 2019]. Specifically, in all of our experiments, PSPNet with ResNet101 [He *et al.*, 2016], which are pretrained on ImageNet, is used as the teacher network. For the student network, we perform experiments on different segmentation architectures, such as the representative models PSPNet and Deeplab with the backbones of ResNet18 as well as ESPNet to verify the effectiveness of our IDD method.

**Training Details.** We use the Pytorch platform to implement our method. Following [Liu *et al.*, 2019] , we train our student networks by mini-batch stochastic gradient descent (SGD) for $40000$ iterations. We set the momentum and the weight decay as 0.9 and 0.0005, respectively. We apply the polynomial learning rate policy, and the learning rate is calculated as $base\_lr \cdot \left(1 - \frac{iter}{total\_iter}\right)^{power}$. The base learning rate and power are respectively set to 0.01 and 0.9. For the input images, we crop them to $512 \times 512$. The random scaling and random flipping are applied to augment the data.

### 4.3 Ablative Study

Our loss function consists of four parts, $L_{skd}$, $L_{cw}$, $L_{id}$, and $L_{pi}$. To explore the effectiveness of each loss item, we conduct ablation experiments on the Cityscapes validation dataset with the evaluation metric mIoU (%). The teacher

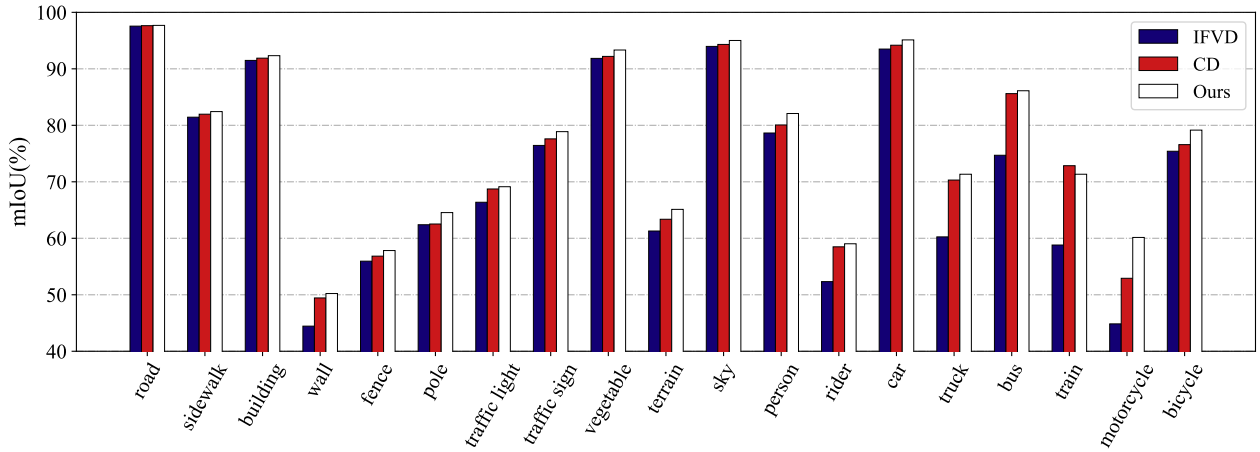


Figure 3: The class IoU scores of KD based semantic segmentation approaches on the Cityscapes validation dataset. We use PSPNet-R18(1.0) as the backbone of the student network.

| Method | mIoU (%) | | Params (M) | FLOPs (G) |
|---|---|---|---|---|
| | Val | Test | | |
| T: PSPNet-R101 | 78.50 | 78.40 | 70.43 | 574.9 |
| S: ESPNet | 61.40 | 60.30 | 0.3635 | 4.422 |
| + SKD | 63.80 | 62.00 | 0.3635 | 4.422 |
| + IFVD | 65.13 | 63.07 | 0.3635 | 4.422 |
| + CD | 67.27 | 65.32 | 0.3635 | 4.422 |
| **+ Ours** | **68.87** | **67.35** | **0.3635** | **4.422** |
| S: PSPNet-R18 (0.5) | 61.17 | - | 3.271 | 31.53 |
| + SKD | 61.60 | 60.05 | 3.271 | 31.53 |
| + IFVD | 63.35 | 63.68 | 3.271 | 31.53 |
| + CD | 68.57 | 66.75 | 3.271 | 31.53 |
| **+ Ours** | **69.76** | **68.54** | **3.271** | **31.53** |
| S: PSPNet-R18 | 70.09 | 67.60 | 13.07 | 125.8 |
| + SKD | 72.70 | 71.40 | 13.07 | 125.8 |
| + IFVD | 74.54 | 72.74 | 13.07 | 125.8 |
| + CD | 75.90 | 74.58 | 13.07 | 125.8 |
| **+ Ours** | **77.59** | **76.33** | **13.07** | **125.8** |

Table 3: Comparison of different KD based semantic segmentation methods on the Cityscapes dataset. "PSPNet-R18(0.5)" is trained from scratch.

network is PSPNet [Zhao *et al.*, 2017] with ResNet101 backbone ("T: PSPNet-R101"), and the student model is PSPNet with ResNet18 ("S: PSPNet-R18") also pretrained in the ImageNet. As can be seen in Table 1, the structured KD loss $L_{skd}$ boosts the performance of the student network "S: PSPNet-R18" from 70.09% to 73.03%. The channel-wise KD loss $L_{cw}$ further improves the student model to 75.78%. By adopting our inter-class distance distillation approach, the gain increases to 5.34% (76.43% vs 70.09%). Furthermore, after applying our position information loss $L_{pi}$, the accuracy of the lightweight student network "S: PSPNet-R18" reaches 77.59%, approximately to the accuracy of the teacher network "T: PSPNet-R101", the mIoU value of which is 78.56%. The experimental results prove that our proposed IDDM and PIDM are effective.

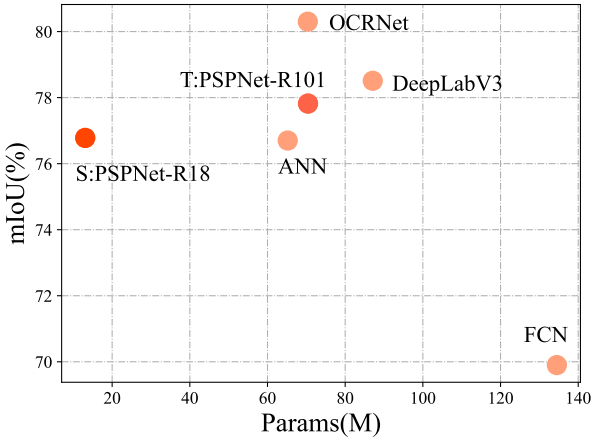


Figure 4: Comparison of Params and mIoU of different models on the Pascal VOC validation set. We use PSPNet-R18(1.0) as the backbone of the student network.

## 4.4 Results

### Cityscapes

Table 2 shows the quantitative results on the Cityscapes dataset. By using our IDD, the Params and the FLOPs of the our student network ("**Ours**") reduce by 81.44% (13.07 vs 70.43) and 78.12% (125.8 vs 574.9) compared to the teacher network, while the mIoU accuracy only decreases 2.07% (from 78.4% to 76.33%). Compared with other lightweight models, our method also has remarkable performance. For example, our IDD outperforms ENet [Paszke *et al.*, 2016] and ESPNet [Mehta *et al.*, 2018] by 18.03% and 16.03% in accuracy (mIoU), respectively. Notably, the Params of ours are only half of ICNet [Zhao *et al.*, 2018], but the accuracy of our student network is still 5.0% higher. Although the accuary of OCNet [Yuan *et al.*, 2018] is 3.77% higher than ours, the Params of ours are less than one fifth of OCNet. The results demonstrate that IDD achieves a satisfactory compromise be-

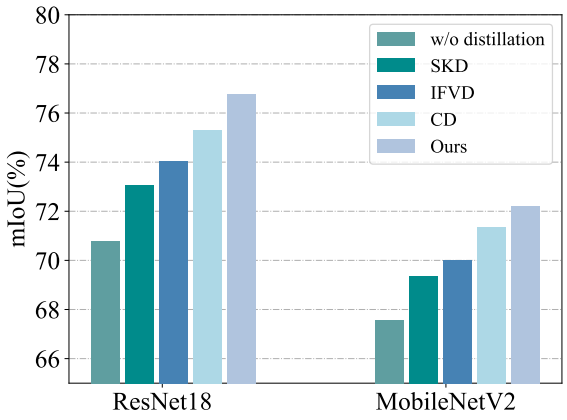


Figure 5: Comparison of different KD strategies for semantic segmentation on the Pascal VOC validation dataset.

| Method | mIoU (%) | Params (M) |
|---|---|---|
| T:PSPNet-R101 | 44.94 | 70.43 |
| S:PSPNet-R18 | 24.65 | 13.07 |
| +SKD | 25.02 | 13.07 |
| +IFVD | 25.82 | 13.07 |
| +CD | 26.80 | 13.07 |
| **+Ours** | **27.69** | **13.07** |
| S:PSPNet-MNV2 | 23.21 | 2.15 |
| +SKD | 24.89 | 2.15 |
| +IFVD | 25.43 | 2.15 |
| +CD | 27.74 | 2.15 |
| **+Ours** | **28.93** | **2.15** |

Table 4: Comparison of different KD approaches for semantic segmentation methods on the ADE20K validation dataset.

tween accuracy and model size.

We also evaluate the performance of our method and other KD based methods on the Cityscapes, e.g. SKD [He *et al.*, 2019], IFVD [Wang *et al.*, 2020] and CD [Shu *et al.*, 2021]. The student models are ESPNet, PSPNet-R18(0.5) and PSPNet-R18. Experimental results are listed in Table 3. When we adopt ESPNet as the student network, our method leads to a significant improvement of 7.47% and 7.05% on the validation set and the testing set, respectively. Compared with SKD which transfers the intra-class feature variance and CD which transfers the channel-wise feature, our method outperforms them by 3.74% and 1.60%, separately. After using our IDD, the performance of PSPNet-R18(0.5) increases from 61.17% to 69.76%, and surpasses IFVD and CD by 6.41% and 1.19% in the validation set. When PSPNet-R18 is adopted as the student model, with our IDD, the gains reach to 7.50% (70.09% to 77.59%), and outperform IFVD and CD by 3.05% and 1.69% respectively. The experimental results show that our IDD is better than the previous KD strategies for semantic segmentation.

In addition, as shown in Figure 3, we use the PSPNet-R18(1.0) as the student network to calculate the mIoU for each class compared with two state-of-the-art methods. Due to our method enables the student network to have large inter-class distance and rich position information, it performs well on some categories. For example, rider, car and bus. Table 3 shows the qualitative results, which again demonstrate the effectiveness of our IDD method.

**Pascal VOC**

As depicted in Figure 4, we adopt a dot graph to describe the parameters and accuracy of different networks, i.e. OCRNet [Yuan *et al.*, 2020], DeepLabV3, FCN [Long *et al.*, 2015], ANN [Zhu *et al.*, 2019] and PSPNet. By using our spatial knowledge distillation, the PSPNet-R18(1.0) outperforms FCN and ANN by 6.79% and 0.08%, respectively.

We adopt ResNet18 and MobileNetV2 as the student network to evaluate our approach on the validation set. The results are shown in Figure 5. With ResNet18 as the backbone of the student network, our approach improves the accuracy of the model that without distillation by 6.01%, and is better than the SKD, IFVD and CD respectively by 3.74%, 2.74% and 1.47%. For MobileNetV2, our method exceeds the benchmark model by 4.66%, and improves the SKD, IFVD and CD respectively by 2.88%, 2.21% and 0.89%.

**ADE20K**

To further verify the effectiveness of our proposed method, we carry out experiments on the challenging dataset ADE20K. The quantitative results are reported in Table 4. When the student model is built on ResNet18, our proposed approach improves the student model from 24.65% to 27.65%, and outperforms SKD, IFVD and CD by 2.67%, 1.87% and 0.89%. With MobileNetV2 as the student backbone, we achieve an improvement to 6.72% compared with the benchmark model, and improves the SKD, IFVD and CD by 4.04%, 3.50% and 1.19%, respectively.

## 5 Conclusion

In this paper, we present a novel knowledge distillation method for semantic segmentation, helping the student model have large inter-class distance in the feature space and rich position information. Specifically, we propose the inter-class distance distillation module and the position information distillation module to transfer the inter-class distance and position cue from the teacher network to the student network. Ablative experiments show that our explored two modules enable the student network to mimic the teacher network better. We demonstrate the effectiveness of our approach by conducting extensive experiments on three public datasets, i.e. Cityscapes, Pascal VOC and ADE20K.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62106177. The numerical calculation was supported by the super-computing system in the Super-computing Center of Wuhan University.

## References

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille.

Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[Cordts *et al.*, 2016] 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.

[Dong *et al.*, 2020] Genshun Dong, Yan Yan, Chunhua Shen, and Hanzi Wang. Real-time high-performance semantic image segmentation of urban street scenes. *IEEE Transactions on Intelligent Transportation Systems*, 22(6):3258–3274, 2020.

[Everingham *et al.*, 2015] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[Fan *et al.*, 2021] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9725, 2021.

[He *et al.*, 2016] 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.

[He *et al.*, 2019] 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.

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

[Islam *et al.*, 2020] Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*, 2020.

[Kemker *et al.*, 2018] Ronald Kemker, Carl Salvaggio, and Christopher Kanan. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS journal of photogrammetry and remote sensing*, 145:60–77, 2018.

[Liu *et al.*, 2019] 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.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[Mehta *et al.*, 2018] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018.

[Paszke *et al.*, 2016] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

[Shu *et al.*, 2021] 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.

[Tu *et al.*, 2019] Zhigang Tu, Hongyan Li, Dejun Zhang, Justin Dauwels, Baoxin Li, and Junsong Yuan. Action-stage emphasized spatiotemporal vlad for video action recognition. *IEEE Transactions on Image Processing*, 28(6):2799–2812, 2019.

[Wang *et al.*, 2020] 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.

[Yuan *et al.*, 2018] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[Yuan *et al.*, 2020] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.

[Zhao *et al.*, 2017] 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.

[Zhao *et al.*, 2018] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 405–420, 2018.

[Zhou *et al.*, 2017] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[Zhu *et al.*, 2019] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 593–602, 2019.