# Make a Strong Teacher with Label Assistance: A Novel Knowledge Distillation Approach for Semantic Segmentation

Shoumeng Qiu[1], Jie Chen[1], Xinrun Li[2], Ru Wan[3], Xiangyang Xue[1], and Jian Pu[1]*

[1] Fudan University, Shanghai, China
smqiu21@m.fudan.edu.cn; {chenji19,xyxue,jianpu}@fudan.edu.cn
[2] Bosch Corporate Research, China
[3] Mogo.ai Information and Technology Co., Ltd, Beijing, China
wanru@zhidaoauto.com

**Abstract.** In this paper, we introduce a novel knowledge distillation approach for the semantic segmentation task. Unlike previous methods that rely on power-trained teachers or other modalities to provide additional knowledge, our approach does not require complex teacher models or information from extra sensors. Specifically, for the teacher model training, we propose to noise the label and then incorporate it into input to effectively boost the lightweight teacher performance. To ensure the robustness of the teacher model against the introduced noise, we propose a dual-path consistency training strategy featuring a distance loss between the outputs of two paths. For the student model training, we keep it consistent with the standard distillation for simplicity. Our approach not only boosts the efficacy of knowledge distillation but also increases the flexibility in selecting teacher and student models. To demonstrate the advantages of our Label Assisted Distillation (LAD) method, we conduct extensive experiments on five challenging datasets including Cityscapes, ADE20K, PASCAL-VOC, COCO-Stuff 10K, and COCO-Stuff 164K, five popular models: FCN, PSPNet, DeepLabV3, STDC, and OCRNet, and results show the effectiveness and generalization of our approach. We posit that incorporating labels into the input, as demonstrated in our work, will provide valuable insights into related fields. Code is available at https://github.com/skyshoumeng/Label_Assisted_Distillation.

**Keywords:** Semantic Segmentation · Knowledge Distillation · Privileged Information

## 1 Introduction

Semantic segmentation is one of the most fundamental tasks that aims to classify every pixel of a given image into a specific class. It is widely applied to many applications such as autonomous driving [36, 41, 45], video surveillance [13, 50], and biomedical image diagnosis [1, 33]. Nevertheless, most of the existing models

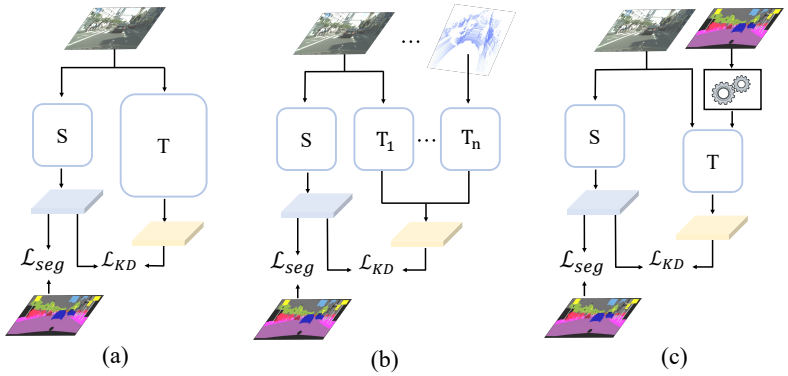



**Fig. 1:** Comparison with two main categories of distillation approaches. S and T indicate the student and teacher model, $\mathcal{L}_{seg}$ denotes the segmentation task loss and $\mathcal{L}_{kd}$ denotes the distillation loss. Additional knowledge mainly from: (a) power-pretrained teacher model; (b) teacher that takes additional modalities as input; (c) lightweight teacher model that takes processed label as input. Our approach (c) shows a clear advantage as it has no requirements for complex teacher model or other modalities.

entail high complexity and heavy computational cost for achieving superior performance [11, 44, 45, 54], which is inappropriate in real-world applications. To alleviate these issues, knowledge distillation [14] has been introduced into segmentation [3, 12, 27, 40], which aims at transferring knowledge from more complex powerful teacher models to efficient lightweight student models.

Based on the source of additional knowledge primarily deriving from the teacher model, current knowledge distillation methods for segmentation broadly fall into two categories: knowledge from the more powerful model capacity, such as [3, 12, 35], and knowledge from extra data information, such as [27, 40, 46]. However, the approaches in both categories have some apparent issues. For the first category, where additional knowledge is derived from the teacher model, this approach typically requires using a more complex and computationally expensive model as the teacher to extract more useful information from the inputs. [26,35,43,47]. For the second category, where additional knowledge comes from extra data information, this always involves other modalities [17, 46] such as infrared or LiDAR data which are often costly or unattainable in practice. For the above issues, we aim to conduct knowledge distillation learning using a lightweight teacher model and eliminate the need for additional modalities.

In this paper, we innovatively consider improving the performance of lightweight teacher model by incorporating label information into the input, as the label is always attainable in supervised learning tasks. The key difference between the existing method and our proposed approach is shown in Fig. 1. Specifically, for the teacher model, unlike previous methods that adopt complex models or extra information to achieve higher performance, we improve the performance by taking the label as part of the input to the model. However, there is a problem with directly feeding labels into the model, as the model may take shortcuts in

mapping inputs to outputs based on the labels rather than learning the intended solution as mentioned in [10, 34]. To address this, we propose to noise the label randomly before feeding it into the model. In experiments, we found that there are fluctuations in the output of the teacher model due to the random sampling in the label noising operation. To counteract the effects of the introduced noise, we further propose a dual-path consistency training strategy with a consistency loss between the outputs of two paths. When the teacher model is trained, the distillation learning for the student model is the same as the standard distillation approach. It should be noted that the noised label in this paper is different from [6, 22, 25, 30], where label noise originates from annotations, our approach intentionally introduces noise to the clean label.

In addition, our approach allows for a more flexibility choice of teacher and student models. For example, the teacher can be more complex than the student, the same as the student, or even simpler than the student. Finally, to evaluate the effectiveness and generalization of our approach, we conduct extensive experiments on five baseline segmentation models: FCN [28], PSPNet [53], DeeplabV3 [4], STDC [8], and OCRNet [49], and across five challenging datasets: Cityscapes [5], ADE20K [55], PASCAL-VOC 2012 [7], COCO-Stuff 10K [2], and COCO-Stuff 164K [2]. The experimental results show consistent performance improvement across all cases. We also perform detailed analyses of crucial components in the proposed approach and hope other researchers can gain inspiration from our study. Our contributions are summarized as follows:

- We propose a novel knowledge distillation approach with noised labels as privileged information for the semantic segmentation task. Our approach alleviates the dependency on complex teacher models or other modalities, and can effectively improve performance of knowledge distillation.
- To enhance the robustness of teacher against the noise introduced in privileged information, we propose a dual-path consistency training strategy with a distance loss to minimize discrepancies between the outputs of two paths.
- We conduct extensive experiments on five popular semantic segmentation baseline models across five challenging public datasets, and experimental results show substantial and consistent improvements on performance. We also perform analyses on crucial components of the proposed approach.

## 2 Related works

### 2.1 Semantic Segmentation

Semantic segmentation is one of the most fundamental computer vision tasks, which aims at classing every pixel on an input image into a certain class. It has a wide range of applications, such as autonomous driving and video surveillance. Deep neural network-based approaches are dominant in this task. FCN [28] is a fully convolutional model that can take input of arbitrary size and produce correspondingly-sized output with efficient inference and training. PSPNet [53]

proposed to exploit the capability of global context information by different-region-based context aggregation operation through the pyramid pooling module and the pyramid scene parsing network. DeeplabV3 [4] proposed to combine both the spatial pyramid pooling module and encode-decoder structure. In this way, the approach can have the advantage of multi-scale contextual information encoding and sharper object boundary capturing at the same time. In recent years, many computational and memory-intensive models [38, 45, 54] have been proposed to further improve the performance on this task, However, these models are not friendly to applications in real-world scenarios. In response to this problem, many real-time methods have been proposed [8, 23, 48]. For example, STDC [8] proposed an efficient Short-Term Dense Concatenate network structure by gradually reducing the dimension of feature maps and using the aggregation of them for better image representation.

### 2.2 Knowledge Distillation

Knowledge distillation is a technique for model compression and acceleration in which a smaller model, referred to as the student model, is trained to mimic the behavior of another model, usually a larger pre-trained model, known as the teacher model. It is popularized by [14] and has attracted a surge of attention in recent years [31, 37, 42]. Surveys [42] and [15] summarized methods from various perspectives, and our method can be categorized into Cross-Modal distillation (multi-modal to single modal). The key distinction between our approach and previous methods is our innovative use of noised labels as an input modality.

Based on the source of additional knowledge primarily comes from in the teacher model. We roughly categorize current works on knowledge distillation into two main categories: knowledge from a more complex teacher model and knowledge from extra information. Here we mainly focus on the distillation approaches applied to the semantic segmentation task. For the knowledge from a more complex teacher model, there is currently a lot of work devoted to further improving the efficiency and performance of the knowledge distillation approach, such as SKD [26], IFDV [43], CWD [35], CIRKD [47]. SKD [26] proposed to distill structured knowledge from the teacher model to the student model as dense prediction is a structured prediction problem. IFDV [43] proposed an Inter-class Distance Distillation method to transfer the inter-class distance in the feature space from the teacher network to the student network. CWD [35] introduced to minimize the Kullback–Leibler (KL) divergence between the channel-wise probability map of the two networks. CIRKD [26] proposed to transfer structured pixel-to-pixel and pixel-to-region relationships across entire images. For the knowledge from extra information, LGD [52] and LG3D [19] introduced distillation with label methods involving manually designed label encoder and label mapper modules, which specifically designed for object detection task, different from the above methods, our proposed approach is label encoder and label mapper free for the semantic segmentation task. KD-Net [17] proposed a framework to transfer knowledge from multi-modal to a mono-modal for medical image segmentation.

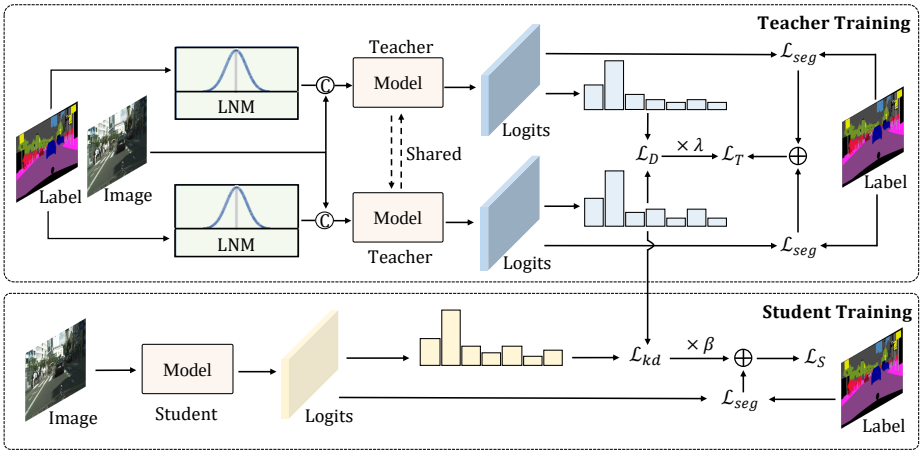



**Fig. 2:** The overall pipeline of our distillation framework. Ⓒ denotes the concatenation operation on channel, ⊕ denotes the addition operation. For the teacher model training within the dashed box above, the Label Noising Module (LNM) and the teacher model are duplicated into two copies. Then, the image and noised label are fed in respectively, The outputs of models are undergo supervision of the label. Additionally, we integrate a consistency constraint loss between the two predictions. We only need to retain one branch of teacher training for student model in distillation learning.

2DPASS [46] proposed to distill richer semantic and structural information from the multi-modal data to boost the feature learning from single-modal data.

### 2.3 Learning using Privileged Information

Learning using privileged information [39] is also a crucial technique that enables machines to learn from other machines. The framework of learning using privileged information aims to leverage the additional information at training time to help boost the performance of the model, which is not accessible at the test time. In [29], the authors proposed to unify the knowledge distillation and privileged information two into generalized distillation for machines learning from other machines. In [9], the authors proposed to translate the notion of privileged information to the unsupervised setting in order to improve clustering performance. PISR [21] proposed to leverage the high-resolution (HR) images as privileged information and transfer the important knowledge of the HR images to a student model.

## 3 Proposed Method

The overall framework of our proposed distillation approach is shown in Fig. 2. The main distinction between our framework and the conventional distillation framework lies in the input information and training strategy of the teacher

model. We do not rely on power-pretrained teacher models [26, 43] or extra information coming from other modalities [17, 46], which gives our approach obvious advantages over other distillation methods.

### 3.1 Problem Formulation

We consider improving the performance of lightweight teacher model by incorporating more task-related information into the input. However, as additional information is often unattainable or costly in practice, so we wonder if we could use the readily available information to achieve this purpose. Specifically, in this paper, we aim to introduce the label to the input of teacher model, which is always attainable in the supervised learning tasks.

The simplest approach is to directly incorporate the label into the model's input, then the objective of model learning changes from

$$\{(X_1, Y_1), ..., (X_n, Y_n)\} \sim P^n(X, Y), \tag{1}$$

to

$$\{(X_1, Y_1), ..., (X_n, Y_n)\} \sim P^n((X, Y), Y), \tag{2}$$

where each $(X_i, Y_i)$ is a image-label pair, $X_i \in \mathbb{R}^{H\times W\times 3}$, $Y_i \in \mathbb{R}^{H\times W}$ and $P(\cdot, \cdot)$ denotes the joint probability distribution. However, there is a problem with the above simple solution. Since $Y_i$ in the input and the expected output $Y_i$ are the same, the model may take a shortcut instead of learning the intended solution [10,34], which means that the deep neural network may learn a mapping from the input label to the output rather than extract some useful features in the image, as indicated in Eq. 3. Please refer to *Shortcut Learning with Clean Label Input* in subsection 4.2 for more detailed discussions.

$$\{X_i,\ Y_i\} \overset{shortcut}{\longrightarrow} Y_i \quad , \tag{3}$$

Since directly incorporating the label into the input is inappropriate, some transformation operations should be applied on the label before feeding it to the model. Here we denote the transformation operation as function $\phi(\cdot)$, then the objective of model learning become:

$$\{(X_1, \phi(Y_1), Y_1), ..., (X_n, \phi(Y_n), Y_n)\} \sim P^n((X, \phi(Y)), Y), \tag{4}$$

For the function $\phi(\cdot)$, two conditions should be satisfied. First, based on the above analysis, it should make the mapping learning from $\phi(Y_i)$ to $Y_i$ difficult. Second, $\phi(Y_i)$ should maintain some useful information for the task, otherwise, $\phi(Y_i)$ will have no benefit for the predictions. In this paper, we propose a simple but effective Label Noising Module (LNM) as function $\phi(\cdot)$ which satisfies the above two conditions well. We will give a detailed description to the implementation of LNM in the next subsection.

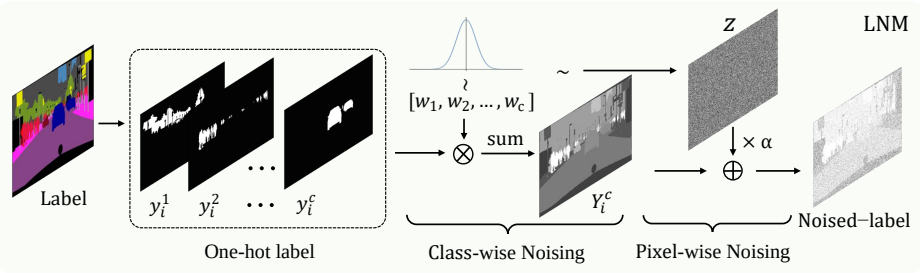



**Fig. 3:** Details of label noising module. $\bigotimes$ denotes the multiplication operation, $\bigoplus$ denotes the addition operation. The label is first represented in one-hot encoding, then the channel of each class is multiplied by a weight obtained from random sampling. After that, the results are added along the channel dimension, resulting in the distortion of the class index information. Subsequently, random noise is added for each pixel, yielding the final noised label, which is referred as *privilege information* in this paper.

### 3.2 Strong Teacher with Noised label

**Design of LNM.** We note that the information contained in the label $Y_i$ can be decomposed into two parts: unique semantic class index and index consistency among pixels within the same class. Based on the above analysis, our proposed label noising module is also two-stage: class-wise noising and pixel-wise noising, respectively. For class-wise noising operation, we aim to distort the class index. We first represent the label $Y_i$ in the form of one-hot $Y_i^{oh} \in \{y_1^1, y_i^2, ..., y_i^c\}$, where $c$ is the total number of classes. For each $y_i^j \in \mathbb{R}^{H \times W}$, we perform a multiplication operation with $w_i$ which is sampled from a predefined distribution $\Omega_1$. Then we perform sum operation across channels and get the class-wise noising label $Y_i^c$. For pixel-wise noising operation, we aim to disrupt the index consistency in each class. We first generate a random matrix $Z \in \mathbb{R}^{H \times W}$ where each element is sampled from a predefined distribution $\Omega_2$, then we scaled $Z$ with $\alpha$ and add to the $Y_i^c$. Finally, the label noising operation can be expressed as follows:

$$\phi(Y_i) = sum(W \cdot Y_i^{oh},\ \ dim{=}0) + \alpha Z\ , \tag{5}$$

Here $W = \{w_1, w_2, \ldots, w_c\}$. Up to now, the noised label is obtained. Then, we consider incorporating the noised label information into the input of model. We choose the simple concatenation operation as it is widely used in many multi-modal fusion works [24, 51]. Therefore, the teacher model takes a four-channel tensor as input, instead of the standard three RGB channels. We adopt the commonly used standard Gaussian distribution for both $\Omega_1$ and $\Omega_2$ in this paper. The pipeline of our label noising module is shown in Fig. 3.

Next, we give explanations about why the operation described in Eq. 5 meets the two conditions mentioned in subsection 3.1. First, since $\phi(\cdot)$ employs random parameters for the noising operation, so there is no direct inverse function for $\phi(Y_i)$, which makes the mapping learning from $\phi(Y_i)$ to $Y_i$ difficult. Second, for the class-wise noising operation, as each $Y_i^{oh}$ shares the same parameter $w_i$,

therefore, we can guarantee that the value within each class is still consistent after the class-wise nosing operation. For the pixel-wise noising operation, we can control the impact of this noise by adjusting parameter $\alpha$. Based on the above analysis, we can see that the generated noised label $\phi(Y_i)$ satisfies the conditions well and can be used as additional information to assist the model for better segmentation performance.

**Dual-path Consistency Training.** In experiments, we found that there is a problem when adopting the noised label from Fig. 3 for teacher training. Since parameters $W$ and $Z$ are sampled randomly in each training iteration, this will result in the fluctuation of the teacher output $O_i^1$ and $O_i^2$ with the same input image $X_i$ under two different sampled parameters $\theta_1$ and $\theta_2$ of LNM:

$$\begin{gathered}\{X_i,\ \ \phi(Y_i|\theta_1)\} \rightarrow O_i^1,\ \ \{X_i,\ \ \phi(Y_i|\theta_2)\} \rightarrow O_i^2, \\ \mathcal{D}(O_i^1,\ \ O_i^2) > 0,\end{gathered} \tag{6}$$

here we adopt a more specific form for $\phi(\cdot)$ as $\phi(\cdot|\theta)$ for the sake of clarity, where $\theta_1$, $\theta_2$ containing both $W$ and $Z$. $\mathcal{D}(\cdot,\cdot)$ is a function to measure the distribution distance between two outputs. In experiments, we found that the $\mathcal{D}(O_i^1, O_i^2)$ is a non-negligible value. This inconsistency makes it challenging for the student model to learn from the teacher, as the outputs of the teacher model vary for the same image in different iterations due to the random sampling in the LNM.

To address the above issues, we need to make the output of the teacher model become robust to the randomly sampled parameters $W$ and $Z$. Specifically, we propose an effective dual-path consistency training strategy. In teacher training, we duplicate the LNM module and the teacher model, with each path taking a noised label generated from independently sampled parameters. We then introduce a consistency loss between the outputs from these two paths. As a result, the final loss for the teacher training can be expressed as:

$$\mathcal{L}_T = \mathcal{L}_{seg}(O^1, Y) + \mathcal{L}_{seg}(O^2, Y) + \lambda \mathcal{D}(O^1, O^2), \tag{7}$$

where the above $\mathcal{L}_{seg}(\cdot,\cdot)$ is the loss function for the segmentation task, $\lambda$ is loss coefficients to balance the contribution of the consistency loss and the task loss.

### 3.3 Objective Function

The training of the student model is the same as the standard knowledge distillation methods, where the model is supervised by both the label supervision and also supervision from the teacher model, which can be expressed as:

$$\mathcal{L}_S = \mathcal{L}_{seg}(O^S, Y) + \beta \mathcal{L}_{kd}(O^S, O^T)\ , \tag{8}$$

where $O^S$ is the output of the student model, and $O^T$ is the output of the teacher model, $\mathcal{L}_{kd}(\cdot,\cdot)$ can be an arbitrary distillation loss function. $\beta$ is the loss coefficient to balance the contribution of the distillation loss and the task loss. We only adopt the knowledge distillation for the logits map for simplicity.

**Table 1:** Quantitative comparison of other distillation methods and our proposed approach, The results are reported in terms of mIoU on Cityscapes val dataset. All models are pretrained on ImageNet. $D_{\text{time}}$ denotes the distillation time per iteration. $*$ denotes the model takes the noised label as privileged information.

| Method | Teacher (P (M) / F (G) / val mIoU) | Student (P (M) / F (G) / (val/test mIoU) | $D_{\text{time}} \downarrow$ (ms) | mIoU ↑ val | mIoU ↑ test |
|---|---|---|---|---|---|
| SKD (TPAMI 2020) | DeepLabV3-R101 (61.1/2371.7/78.1) | DeepLabV3-R18 (13.6/572.0/74.2/73.5) | 780 | 75.4(+1.2) | 74.1(+0.6) |
| IFVD (ECCV 2020) | | | 1018 | 75.6(+1.4) | 74.3(+0.8) |
| CWD (ICCV 2021) | | | 415 | 75.6(+1.4) | 74.1(+0.6) |
| CIRKD (CVPR 2022) | | | 1295 | 76.4(+2.2) | 75.1(+1.6) |
| MasKD (ICLR 2023) | | | 846 | 77.0(+2.8) | 75.6(+2.1) |
| Ours | DeepLabV3-R18* (13.6/572.1/79.8) | DeepLabV3-R18 (13.6/572.0/74.2/73.5) | **270** | **77.7(+3.5)** | **76.9(+3.4)** |
| SKD (TPAMI 2020) | DeepLabV3-R101 (61.1/2371.7/78.1) | PSPNet-R18 (12.9/507.4/72.6/72.3) | 758 | 73.3(+0.7) | 73.0(+0.7) |
| IFVD (ECCV 2020) | | | 944 | 73.7(+1.1) | 72.8(+0.5) |
| CWD (ICCV 2021) | | | 402 | 74.4(+1.8) | 73.6(+1.3) |
| CIRKD (CVPR 2022) | | | 1210 | 74.7(+2.1) | 74.1(+1.8) |
| MasKD (ICLR 2023) | | | 757 | 75.3(+2.7) | 74.6(+2.3) |
| Ours | PSPNet-R18* (12.9/507.5/79.7) | PSPNet-R18 (12.9/507.4/72.6/72.3) | **259** | **76.2(+3.6)** | **75.6(+3.3)** |

# 4 Experiments

## 4.1 Experimental Setup

For thorough evaluation, we conduct experiments on five commonly used datasets: Cityscape [5], ADE20K [55], and PASCAL-VOC [7], COCO-Stuff 10K [2] and COCO-Stuff 164K [2] datasets, and five different popular segmentation baseline models, including FCN [28], PSPNet [53], DeeplabV3 [4], STDC [8], and OCRNet [49]. We use the commonly used mean Intersection over Union (mIoU) metric for evaluation. We adopt the CWD distillation loss [35] as $\mathcal{L}_{kd}(\cdot,\cdot)$ in our approach as it is a convenient plug-in loss and also very efficient and effective. For simplify, we also use the CWD loss as the consistent loss in the teacher model training. We set the hyper-parameters $\alpha, \lambda, \beta$ as 0.01, 1, and 3, respectively. More details about training setting can be found in the supplementary material.

## 4.2 Main Results

**Comparison with SOTA Methods.** We compare our proposed distillation approach with several state-of-the-art methods, including SKD [26], IFVD [43], CWD [35], CIRKD [47], MasKD [18]. They all rely on more complex models to provide additional knowledge in distillation training. The experimental results are shown in Tab. 1. It can be seen that for the PSPNet-R18 model, we achieve very competitive results among all other distillation approaches, we suppress the CWD baseline approach by 1.8% mIoU on the val set and 2.0% mIoU on the test set, For the DeepLabV3-R18 model, it can be seen that we achieve the best performance among all the approaches and suppress the CWD baseline approach by 2.1% mIoU and 2.8% mIoU, respectively. Additionally, we conducted

**Table 2:** Quantitative comparison on Cityscapes val set. Student models are trained from scratch. ∗ denotes model takes noised label as privileged information.

| Method | Teacher mIoU (%) | Student mIoU (%) | mIoU |
|---|---|---|---|
| SKD | DeepLabV3-R101 (78.1) | DeepLabV3-R18 (61.8) | 62.7(+0.9) |
| IFVD | | | 63.1(+1.3) |
| CWD | | | 64.6(+2.8) |
| CIRKD | | | 67.4(+5.6) |
| Ours | DeepLabV3-R18* (79.8) | DeepLabV3-R18 (61.8) | **68.1(+6.3)** |
| SKD | DeepLabV3-R101 (78.1) | PSPNet-R18 (65.2) | 67.1(+1.9) |
| IFVD | | | 66.0(+0.8) |
| CWD | | | 67.7(+2.5) |
| CIRKD | | | 68.2(+3.0) |
| Ours | PSPNet-R18* (79.7) | PSPNet-R18 (65.2) | **69.2(+4.0)** |

**Table 3:** Quantitative comparison on the PASCAL VOC dataset. ∗ denotes the teacher model takes the noised label as privileged information.

| Method | Teacher mIoU (%) | Student mIoU (%) | mIoU |
|---|---|---|---|
| SKD | DeepLabV3-R101 (77.7) | DeepLabV3-R18 (73.2) | 73.5(+0.2) |
| IFVD | | | 73.9(+0.7) |
| CWD | | | 74.0(+0.8) |
| CIRKD | | | 74.5(+1.3) |
| Ours | DeepLabV3-R18* (79.1) | DeepLabV3-R18 (73.2) | **75.0(+1.8)** |
| SKD | DeepLabV3-R101 (77.7) | PSPNet-R18 (73.3) | 74.1(+0.8) |
| IFVD | | | 73.5(+0.3) |
| CWD | | | 74.0(+0.7) |
| CIRKD | | | 74.8(+1.5) |
| Ours | PSPNet-R18* (79.1) | PSPNet-R18 (73.3) | **75.4(+2.1)** |

distillation time comparison on NVIDIA V100 GPU with image size of $512 \times 1024$, it can be seen that our method shows a significant advantage over others.

Tab. 2 shows the results of the performance comparison when the student model is trained from scratch. It can be seen that our approach achieves the best results among all other methods. For the PSPNet-R18 model, we suppress the baseline CWD by 1.5% mIoU, and suppress the CIRKD method by 1.0% mIoU. For the DeepLabV3-R18 model, we suppress the baseline CWD by 3.5% mIoU, and suppress the CIRKD method by 0.7% mIoU. Tab. 3 shows the experimental results on the PASCAL-VOC datasets, we can see that our approach also shows better performance compared with other methods.

We also conduct experiments on other six models on the Cityscapes dataset, results are shown in Tab. 4. For FCN-18, we suppress the baseline by 1.6%, for FCN-50, we suppress the baseline by 0.7%. For the STDC baseline, we suppress the STDC1 and STDC2 by 1.3% and 1.6% mIoU. For the OCRNet baseline, the experiments are conducted on models with OCRNet-W18s and OCRNet-W18 as the backbone, we suppress the baseline by 1.2% and 1.8% mIoU, respectively.

To further prove the generalization of our approach, we conduct experiments on ADE20K [55], PASCAL-VOC [7], COCO-Stuff 10K [2] and COCO-Stuff 164K [2] datasets using PSPNet and DeepLabV3 methods with ResNet50 backbone. The results are shown in Tab. 5. It is evident that our method significantly enhances the performance of baseline models across various datasets.

**Comparison with Stronger Teacher.** As can be seen from Tab. 1, after introducing noised label to the input, we obtain lightweight teacher models with very strong performance, even surpassing DeepLabV3-R101. However, this raises a question about the underlying reasons for our method's effectiveness: Is the performance boost due to a more powerful teacher model? To answer this question, we chose other two powerful models as a teacher: PSPNet-R101 (79.8 mIoU) and OCRNet-W48 (80.7 mIoU). The experimental results are shown in Tab. 6.

**Table 4:** Quantitative comparison between CWD baseline and our approach on more segmentation models. Our teacher model is the same model as the student model but trained with the noised label as input.

| Teacher | DeepLabV3-R101 | | | | OCRNet-W48 | |
|---|---|---|---|---|---|---|
| Student | FCN-18 | FCN-50 | STDC1 | STDC2 | OCR-W18 | OCR-W18 |
| None | 69.7 | 73.8 | 71.8 | 73.2 | 74.3 | 77.7 |
| CWD (base) | 70.8 | 74.4 | 72.3 | 74.1 | 76.9 | 78.6 |
| Ours | **71.3** | **74.5** | **73.1** | **74.8** | **78.1** | **80.4** |

**Table 5:** Quantitative results on more datasets. Our teacher model is the same model as the student model but trained with the noised label as input. PSP50 is the abbreviation for PSPNet-R50, DeepV3-50 is the abbreviation for DeepLabV3-R50.

| Dataset | ADE20K | | PASCAL-VOC 2012 | | COCO-Stuff 10K | | COCO-Stuff 164K | |
|---|---|---|---|---|---|---|---|---|
| Student | PSP50 | DeepV3-50 | PSP50 | DeepV3-50 | PSP50 | DeepV3-50 | PSP50 | DeepV3-50 |
| None | 41.1 | 42.4 | 76.8 | 76.3 | 35.7 | 34.7 | 38.8 | 39.4 |
| CWD (base) | 41.3 | 42.2 | 77.2 | 76.8 | 35.4 | 35.5 | 39.2 | 39.4 |
| Ours | **42.0** | **43.9** | **77.6** | **77.5** | **36.0** | **35.6** | **39.9** | **40.1** |

It can be seen that distillation with a more powerful teacher model does bring some performance improvements, but our method still retains a clear advantage over them. Here we give an intuitive explanation about why our performance is better than distilling with stronger teachers: the knowledge that comes from the same model structure is more transferable between each other. For more detailed experiments and discussion, please refer to the supplementary material.

**Simple to Complex Model Distillation.** As the performance of a model can be significantly improved with the privileged information, the performance of a simple model with privileged information can be even higher than a complex model. As shown in Tab. 1, the performance of the PSPNet-R18 with label information as input is 79.7%, surpassing that of the DeepLabV3 model with a ResNet101 backbone. So we wonder if we can distill knowledge from a label-assisted simple teacher to a complex student. To explore this, we conducted experiments using the STDC models, with the results presented in Tab. 7. Here we focus on the result in the third row, we can see that with the simple STDC1 model as a teacher and the complex STDC2 as a student, the performance of STDC2 model still can be improved by 1.4 % mIoU. The experimental results confirm that complex models can indeed benefit from learning from label-assisted simpler teacher models.

As large foundational models in vision gain popularity, we also extended our experiments to assess the applicability of our method to these models. We employ a relatively lightweight yet high-performance model (OCRNet-W48) with noised label input as teacher and used LoRA [16] method to finetune DINOv2 [32] on VOC dataset. The results also show performance improvements: 81.1 $\rightarrow$ **82.3** (with backbone ViT-S/14) and 82.5 $\rightarrow$ **83.2** (with backbone ViT-B/14), indicating that the teacher model with noised labels as input contains knowledge that is not present in large models.

**Table 6:** Experimental analysis about the reason of performance gains of the student models. The experiments are conducted on the Cityscapes val set. $*$ denotes the model takes the noised label as privileged information.

| Teacher (mIoU) | None | DeepV3-R101 (78.1) | PSPNet-R101 (79.8) | OCRNet-W48 (80.7) | PSPNet-R18$*$ (79.7) | DeepV3-R18$*$ (79.8) |
|---|---|---|---|---|---|---|
| PSPNet-R18 | 72.6 | 74.4(+1.8) | 75.9(+3.3) | 75.0(+2.4) | **76.2(+3.6)** | **76.2(+3.6)** |
| DeepLabV3-R18 | 74.2 | 75.6(+1.4) | 77.1(+2.9) | 77.2(+3.0) | 77.4(+3.2) | **77.7(+3.5)** |

**Table 7:** Experimental of distillation between models of different complexities. M is short for STDC model. The results are reported on Cityscapes val dataset.

| Mode | Teacher | Student | mIoU |
|---|---|---|---|
| simple->simple | M1 (79.1) | M1 (71.8) | 73.1 |
| complex->complex | M2 (80.0) | M2 (73.2) | 74.8 |
| simple->complex | M1 (79.1) | M2 (73.2) | 74.6 |
| complex->simple | M2 (80.0) | M1 (71.8) | 73.5 |

**Table 8:** Ablation study on consistency loss for teacher. $\times 100$ for better presentation. mIoU is student performance when distillation with the corresponding teacher.

| Model | $\mathcal{D}(\cdot,\cdot)$ | mKL ($\times$ 100) | mIoU |
|---|---|---|---|
| DeepLabV3-R18 | | 6.8 | 76.7 |
| DeepLabV3-R18 | ✓ | 3.4 | 77.7 |
| STDC2 | | 3.4 | 73.6 |
| STDC2 | ✓ | 2.1 | 74.8 |

In addition, as the teacher can be the same as the student, so our approach can also be seen as a self-enhanced technique, which provides greater potential in the future. Please refer to the supplementary material for more information.

**Shortcut Learning with Clean Label Input.** To demonstrate that the model will learn a shortcut instead of extracting useful information when we take clean label as input in training, we conduct experiments to assess the contribution of two different inputs to the final prediction. The results are shown in Fig. 4. It can be seen that when we input with noised label, the contribution of the final prediction obviously comes from two parts. However, when we input with clean label, the prediction predominantly relies on the input label alone, with the RGB image contributing negligibly to the output. Therefore, in this case, the model has learned a mapping from input label to output label, rather than extracting useful information from the RGB image as we have mentioned in subsection 3.1.

### 4.3 Ablation Studies

**Dual-path Consistency Loss.** We perform ablation studies to verify the effectiveness of the dual-path consistency training strategy. We conduct the experiments on models DeeplabV3-18 and STDC2. To assess the robustness of final the teacher model against the label noising operation, we perform random sampling $m$ times for the transformation parameters and obtain the corresponding output logits of the teacher with the same image $X_i$, referred as $O_i^1, O_i^2, ..., O_i^m$, where $O_i^k \in \mathbb{R}^{C\times H\times W}$. We then calculate the $KL$ divergence of distributions between each pair. We utilize the mean value of the distance as a criterion of stability: $KL_{mean} = \frac{1}{m\cdot m}\sum_{i=k}^{m}\sum_{j=1}^{m} KL(softmax(O_i^k, dim{=}0), softmax(O_i^j, dim{=}0))$. In our experiment, we set $n = 3$. The results are shown in Tab. 8. show that our consistent loss significantly reduces the distribution distance between outputs, enhancing the stability of the distillation process and notably improving the performance of the student model.

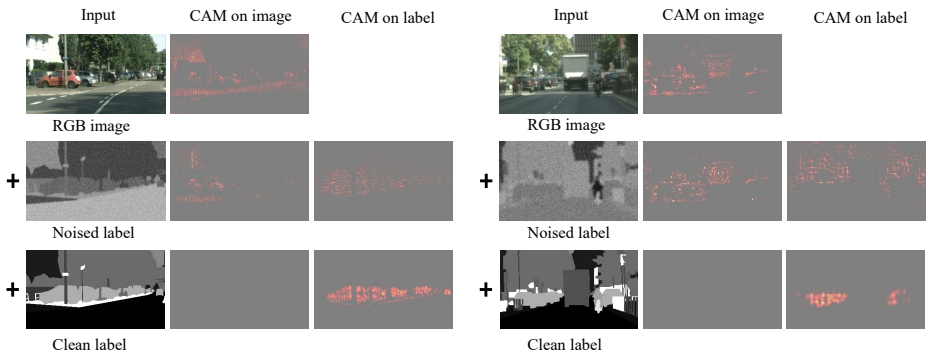



**Fig. 4:** Experiment on the contribution of different inputs to the predictions. Here we choose the contribution to the class *car* for visualization, where red regions correspond to high contribution for the class. We use LayerCAM [20] for its better performance in lower layers. Best viewed with zoomed in.

**Table 9:** Sensitive analysis about the class-wise and pixel-wise label nosing operation. The experiments are conducted with the STDC2 model on the cityscapes val set.

| | with class-wise noising | | | | | w/o class-wise noising | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Pixel-wise noising ($\alpha$) | 0 | 1e-3 | 1e-2 | 1e-1 | 1.0 | 0 | 1e-3 | 1e-2 | 1e-1 | 1.0 |
| Teacher (mIoU) | 80.5 | 80.5 | 80.0 | 74.1 | 71.7 | 90.0 | 84.3 | 86.5 | 83.2 | 76.3 |
| Student (mIoU) | 73.8 | 74.5 | 74.8 | 72.3 | 70.5 | 69.4 | 70.1 | 72.3 | 73.0 | 68.8 |

**Sensitivity Analysis.** We conducted experiments to evaluate the effects of class-wise and pixel-wise noising on the performance of both the teacher and student models. The results are shown in Tab. 9. It should be noted that if without class-wise noising operation, the labels are first normalized to achieve a similar input interval with the normalized images. With class-wise noising, if the pixel-wise noise is lower (0.001), better performance can be achieved for the teacher model as the information is cleaner. However, this knowledge cannot be effectively transferred as the predictions rely more on the leaked label information. When the pixel-wise noise is higher (0.1), the useful information in the noised label will be contaminated by the noise, degrading the teacher model's performance and subsequently affecting the performance of the student model. Without class-wise noising operation, the teacher model can achieve higher performance as the class index information is preserved. However, as predictions are more dependent on label information than on extracted features, the knowledge in the outputs cannot be effectively transferred to the student model, resulting in unsatisfactory distillation performance.

### 4.4 Discussion

Here, we offer an alternative denoising perspective to explain our approach of utilizing noised label as privileged information. For the teacher model, it receives both the image $X_i$ and the noised label $\phi(Y_i|\theta)$ as input, aiming to produce the

clean label $Y_i$ as its output. Hence, the teacher model can be considered as a denoising model whose objective is to eliminate the noise present in the input noised label. In this case, the objective of model learning is changed from

$$Y_i = mapping(X_i), \tag{9}$$

to

$$Y_i = denoising(X_i, \phi(Y_i|\theta)), \tag{10}$$

where the behavior of noising function $\phi(\cdot)$ is controlled by the parameter $\theta$, which is randomly sampled every iteration in training. The above equation shows that we transform the standard segmentation task of learning an image-to-label mapping to a label-denoising task with the RGB images serve as observations.

From the above perspective, the performance of the denoising task should heavily depend on the level of noise in the input, which aligns with our experiment results shown in Tab. 9. The results indicated that performance varied significantly depending on the noise levels. With relatively minimal noise (0.001 for pixel-wise noising), the teacher model could achieve high performance effortlessly. However, when noise levels were substantially increased (0.1 for pixel-wise noising), the performance of the teacher model declined dramatically. This decline is primarily due to the model's difficulty in extracting useful information from the excessively noised labels.

## 5 Conclusion

In conclusion, our proposed knowledge distillation approach for semantic segmentation offers a novel and effective way to leverage readily available labels as a source of knowledge. By strategically introducing noise to labels and employing a dual-path consistency training strategy, we significantly enhance the performance of lightweight teacher models. Extensive experiments demonstrate the effectiveness of our method, achieving substantial improvements in mIoU scores across various datasets and baseline models. We believe this work provides valuable insights and a promising direction for future research in knowledge distillation for semantic segmentation.

## 6 Limitations

We believe that our proposed knowledge distillation approach can significantly enhance the performance of baseline models across various computer vision tasks. Our approach has potential value in the deployment of deep learning models in practice, we do not see any potential negative effects.

## Acknowledgements

This work is supported by the Shanghai Platform for Neuromorphic and AI Chip under Grant 17DZ2260900 (NeuHelium). The computations in this research were performed using the CFFF platform of Fudan University.

## References


1. Alzahrani, Y., Boufama, B.: Biomedical image segmentation: a survey. SN Computer Science **2**, 1–22 (2021)
2. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR. pp. 1209–1218 (2018)
3. Chen, D., Mei, J.P., Zhang, Y., Wang, C., Wang, Z., Feng, Y., Chen, C.: Cross-layer distillation with semantic calibration. In: AAAI. vol. 35, pp. 7028–7036 (2021)
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
6. Das, R., Sanghavi, S.: Understanding self-distillation in the presence of label noise. arXiv preprint arXiv:2301.13304 (2023)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)
8. Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X.: Rethinking bisenet for real-time semantic segmentation. In: CVPR. pp. 9716–9725 (2021)
9. Feyereisl, J., Aickelin, U.: Privileged information for data clustering. Information Sciences **194**, 4–23 (2012)
10. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
11. He, J., Deng, Z., Zhou, L., Wang, Y., Qiao, Y.: Adaptive pyramid context network for semantic segmentation. In: CVPR. pp. 7519–7528 (2019)
12. He, T., Shen, C., Tian, Z., Gong, D., Sun, C., Yan, Y.: Knowledge adaptation for efficient semantic segmentation. In: CVPR. pp. 578–587 (2019)
13. Heo, M., Hwang, S., Oh, S.W., Lee, J.Y., Kim, S.J.: Vita: Video instance segmentation via object token association. NIPS **35**, 23109–23120 (2022)
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
15. Hu, C., et al.: Teacher-student architecture for knowledge distillation: A survey. arXiv:2308.04268 (2023)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
17. Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., Gori, P.: Knowledge distillation from multi-modal to mono-modal segmentation networks. In: MICCAI. pp. 772–781. Springer (2020)
18. Huang, T., Zhang, Y., You, S., Wang, F., Qian, C., Cao, J., Xu, C.: Masked distillation with receptive tokens. In: ICLR (2023)
19. Huang, Y., Liu, X., Zhu, Y., Xu, Z., Shen, C., Che, Z., Zhang, G., Peng, Y., Feng, F., Tang, J.: Label-guided auxiliary training improves 3d object detector. In: ECCV. pp. 684–700. Springer (2022)
20. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: Exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing **30**, 5875–5888 (2021)

21. Lee, W., Lee, J., Kim, D., Ham, B.: Learning with privileged information for efficient image super-resolution. In: ECCV. pp. 465–482. Springer (2020)
22. Li, A., Lu, Z., Wang, L., Han, P., Wen, J.R.: Large-scale sparse learning from noisy tags for semantic segmentation. IEEE transactions on cybernetics **48**(1), 253–263 (2016)
23. Li, H., Xiong, P., Fan, H., Sun, J.: Dfanet: Deep feature aggregation for real-time semantic segmentation. In: CVPR. pp. 9522–9531 (2019)
24. Li, X., Li, W., Ren, D., Zhang, H., Wang, M., Zuo, W.: Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In: CVPR. pp. 2706–2715 (2020)
25. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.J.: Learning from noisy labels with distillation. In: ICCV. pp. 1910–1918 (2017)
26. Liu, Y., Shu, C., Wang, J., Shen, C.: Structured knowledge distillation for dense prediction. TPAMI (2020)
27. Liu, Z., Qi, X., Fu, C.W.: 3d-to-2d distillation for indoor scene parsing. In: CVPR. pp. 4464–4474 (2021)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
29. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643 (2015)
30. Lu, Z., Fu, Z., Xiang, T., Han, P., Wang, L., Gao, X.: Learning from weak and noisy labels for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(3), 486–500 (2016)
31. Nguyen, T., Novak, R., Xiao, L., Lee, J.: Dataset distillation with infinitely wide convolutional networks. NIPS **34**, 5186–5198 (2021)
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
34. Scimeca, L., Oh, S.J., Chun, S., Poli, M., Yun, S.: Which shortcut cues will dnns choose? a study from the parameter-space perspective. In: ICLR (2021)
35. Shu, C., Liu, Y., Gao, J., Yan, Z., Shen, C.: Channel-wise knowledge distillation for dense prediction. In: ICCV. pp. 5311–5320 (2021)
36. Siam, M., Gamal, M., Abdel-Razek, M., Yogamani, S., Jagersand, M., Zhang, H.: A comparative study of real-time semantic segmentation for autonomous driving. In: CVPRW. pp. 587–597 (2018)
37. Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A.A., Wilson, A.G.: Does knowledge distillation really work? NIPS **34**, 6906–6919 (2021)
38. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: ICCV. pp. 7262–7272 (2021)
39. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks **22**(5-6), 544–557 (2009)
40. Vobecky, A., Hurych, D., Siméoni, O., Gidaris, S., Bursuc, A., Pérez, P., Sivic, J.: Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation. In: ECCV. pp. 478–495. Springer (2022)
41. Wang, J., Gou, C., Wu, Q., Feng, H., Han, J., Ding, E., Wang, J.: Rtformer: Efficient design for real-time semantic segmentation with transformer. NIPS **35**, 7423–7436 (2022)

42. Wang, L., Yoon, K.J.: Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE transactions on pattern analysis and machine intelligence (2021)
43. Wang, Y., Zhou, W., Jiang, T., Bai, X., Xu, Y.: Intra-class feature variation distillation for semantic segmentation. In: ECCV. pp. 346–362. Springer (2020)
44. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV. pp. 418–434 (2018)
45. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. NIPS **34**, 12077–12090 (2021)
46. Yan, X., Gao, J., Zheng, C., Zheng, C., Zhang, R., Cui, S., Li, Z.: 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In: ECCV. pp. 677–695. Springer (2022)
47. Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., Zhang, Q.: Cross-image relational knowledge distillation for semantic segmentation. In: CVPR. pp. 12319–12328 (2022)
48. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: ECCV. pp. 325–341 (2018)
49. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: ECCV. pp. 173–190. Springer (2020)
50. Zhang, J., Xu, X., Shen, F., Yao, Y., Shao, J., Zhu, X.: Video representation learning with graph contrastive augmentation. In: ACM MM. pp. 3043–3051 (2021)
51. Zhang, L., Danelljan, M., Gonzalez-Garcia, A., Van De Weijer, J., Shahbaz Khan, F.: Multi-modal fusion for end-to-end rgb-t tracking. In: ICCVW. pp. 0–0 (2019)
52. Zhang, P., Kang, Z., Yang, T., Zhang, X., Zheng, N., Sun, J.: Lgd: label-guided self-distillation for object detection. In: AAAI. vol. 36, pp. 3309–3317 (2022)
53. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017)
54. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR. pp. 6881–6890 (2021)
55. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR. pp. 633–641 (2017)