Cross-Image Relational Knowledge Distillation for Semantic Segmentation

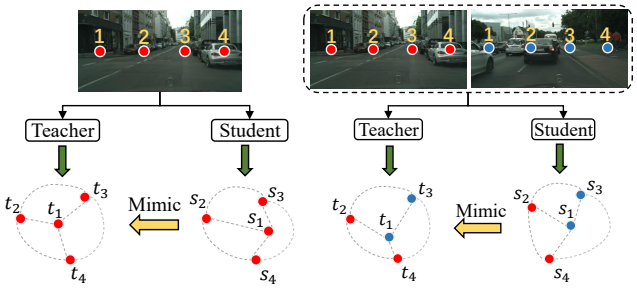
Chuanguang Yang[1,2] Helong Zhou[3] Zhulin An[1*] Xue Jiang[4]
Yongjun Xu[1] Qian Zhang[3]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Horizon Robotics [4]School of Computer Science, Wuhan University
{yangchuanguang, anzhulin, xyj}@ict.ac.cn
{helong.zhou, qian01.zhang}@horizon.ai jxt@whu.edu.cn

Abstract

Current Knowledge Distillation (KD) methods for semantic segmentation often guide the student to mimic the teacher's structured information generated from individual data samples. However, they ignore the global semantic relations among pixels across various images that are valuable for KD. This paper proposes a novel Cross-Image Relational KD (CIRKD), which focuses on transferring structured pixel-to-pixel and pixel-to-region relations among the whole images. The motivation is that a good teacher network could construct a well-structured feature space in terms of global pixel dependencies. CIRKD makes the student mimic better structured semantic relations from the teacher, thus improving the segmentation performance. Experimental results over Cityscapes, CamVid and Pascal VOC datasets demonstrate the effectiveness of our proposed approach against state-of-the-art distillation methods. The code is available at https://github.com/winycg/CIRKD.

1. Introduction

Semantic segmentation is a crucial and challenging task in computer vision. It aims to classify each pixel in the input image with an individual category label. The applications of segmentation often focus on autonomous driving, virtual reality and robots. Although popular state-of-the-art segmentation networks, such as DeepLab [3, 5], PSPNet [51] and OCRNet [47], achieve remarkable performance, they often need high computational costs. This weakness makes them difficult to be deployed for real-world scenarios over resource-limited mobile devices. Therefore, a series of lightweight segmentation networks are proposed, such as ESPet [24], ICNet [50] and BiSeNet [46]. Moreover, model compression is also an alternative field to pursue compact networks, mainly divided into quantization [37], pruning [2, 43] and knowledge distillation (KD) [16, 30, 41].



(a) Intra-image relational KD. (b) Cross-image relational KD.

Figure 1. Overview of intra-image (*left*) and our proposed cross-image relational distillation (*right*). The circles (● or ●) with the same color denote pixel embeddings from the identical image. t_i and s_i represent the pixel embeddings of the i-th pixel location tagged in an image from the teacher and student, respectively. The dotted line (– –) shows the similarity relationship between two pixels. The circles and lines construct a relational graph.

This paper investigates KD to improve the performance of a compact student network under the guidance of a high-capacity teacher network for semantic segmentation. A broad range of KD approaches [16, 18, 41, 48] have been well studied but mostly for image classification tasks. Unlike image-level recognition, the segmentation task aims at dense pixel predictions, which is more challenging. Previous researches [18, 22] have found that directly utilizing classification-based KD methods to deal with dense prediction tasks may not achieve desirable performance. This is because strictly aligning the coarse feature maps between the teacher and student networks may lead to negative constraints and ignore the structured context among pixels.

Recent works attempt to propose specialized KD methods [14, 20, 21, 30, 35, 40] for semantic segmentation. Most focus on mining correlations or dependencies among spa-

*Corresponding author.

tial pixel locations because segmentation needs a structured output. Typical knowledge can be local pixel affinity [40], global pairwise relations [14, 20] and intra-class pixel variation [35]. Such methods often perform better than the traditional point-wise alignment in capturing structured spatial knowledge. More recently, Shu *et al.* [30] revealed that each channel represents a category-specific mask and thus proposed Channel-Wise KD (CWD) [30]. CWD achieves state-of-the-art distillation performance and demonstrates the importance of channel-level information for dense prediction tasks. However, previous segmentation KD methods often guide a student to mimic the teacher's structured information generated from *individual data samples*. They ignore cross-image semantic relations among pixels for knowledge transfer, as shown in Fig. 1.

Based on this motivation, we propose Cross-Image Relational Knowledge Distillation (CIRKD) for semantic segmentation. The core idea is to construct global pixel relations across the whole training images as meaningful knowledge. A good pre-trained teacher network could often generate a well-structured pixel embedding space and capture better pixel correlations than a student network. Based on this property, we transfer such pixel relations from teacher to student. Specifically, we propose pixel-to-pixel distillation and pixel-to-region distillation to fully exploit structured relations across various images. The former aims to transfer similarity distributions among pixel embeddings. The latter focuses on transferring pixel-to-region similarity distributions complementary to the former. The region embedding is generated by averagely pooling pixel embeddings from the same class and represents that class's feature center. The pixel-to-region relations indicate the relative similarities between pixels and class-wise prototypes.

A naive way for constructing cross-image relations is to derive embeddings from the current mini-batch. However, the batch size of the segmentation task is often small, limiting the network to capture broader pixel dependencies. Motivated by previous self-supervised learning [31, 38], we introduce a pixel queue and a region queue in the memory bank to store abundant embeddings for modelling long-range pixel relations. The embeddings in queues are consistent during the distillation process, since they are generated from the pre-trained and frozen teacher network. We regard the teacher and student pixel embeddings from the current mini-batch as anchors. We randomly sample contrastive embeddings from the queues to model pixel-to-pixel as well as pixel-to-region similarity distributions. Then we align such soft relations via KL-divergence from the student to teacher.

CIRKD guides the student network to learn the global property of relative pixel structures across training images from the teacher, further improving the segmentation performance. We evaluate our method over popular DeepLabV3 [5] and PSPNet [51] architectures on three segmentation benchmark datasets: Cityscapes [7], CamVid [1] and Pascal VOC [9]. Experimental results indicate that CIRKD outperforms other state-of-the-art distillation approaches, demonstrating the value of transferring global pixel relationships in semantic segmentation.

The main contributions are summarized as follows:

- We propose cross-image relational KD to transfer global pixel relationships. We may be the first to build pixel dependencies across global images for segmentation KD.
- We propose pixel-to-pixel and pixel-to-region distillation with the memory bank mechanism to fully explore structured relations for transfer.
- Our CIRKD achieves the best distillation performance among state-of-the-art methods on the public segmentation datasets.

2. Related Work

Semantic Segmentation. Fully Convolutional Networks (FCN) [23] creates a seminal paradigm for end-to-end dense feature learning for semantic segmentation. Since contextual pixel dependencies are essential for segmentation performance [36], capturing long-range relationships becomes a critical topic. DeepLab [3] applies atrous convolution to enlarge the receptive field for learning broader context. DeepLabV3 [4] assembles convolution blocks with various atrous rates in parallel to capture multi-scale contexts. PSPNet [51] proposes a pyramid pooling module to exploit different-region-based context aggregation. RefineNet [19] preserves high-resolution predictions by long-range residual connections for the down-sampling process. More recently, SegFormer [39] utilizes a structured Transformer encoder to model global context information. However, such high-performance segmentation networks with expensive computational costs are difficult to be deployed over resource-limited mobile devices.

Efficient segmentation networks attract wide attention due to the need for real-time inference. Most works attempt to design lightweight networks with cheap operations. ENet [26] is equipped with early downsampling, small decoder size and filter factorization. ESPNet [24] factorizes the standard convolution into the spatial pyramid of dilated convolution. ICNet [50] builds a cascade structure to balance the efficiency between low-resolution and high-resolution features. BiSeNet [46] combines a spatial path and a context path to process features efficiently. Beyond designing a segmentation framework, lightweight backbone networks [29, 45, 49], *e.g.* MobileNet [29] and ShuffleNet [49], can also implement acceleration.

Knowledge Distillation. The core idea of KD is to transfer meaningful knowledge from a cumbersome teacher into a smaller and faster student. Most current KD meth-

ods deal with image classification networks, mainly divided into probability-based, feature-based and relation-based approaches. Probability-based KD [16, 52] transfers class probabilities produced from the teacher as soft labels to supervise the student. Feature-based KD focuses on intermediate feature maps [28] or their refined information [15, 48] as knowledge. Relation-based KD [10, 25, 27, 32, 42, 44] aligns correlations or dependencies among multiple instances between the student and teacher networks. Our CIRKD is related to SEED [10] that both of them are contrastive distillation manners with a shared memory bank. However, these image-level KD methods are often unsuitable for pixel-wise semantic segmentation [18, 22].

Recent KD methods for semantic segmentation often encode contextual pixel affinity as knowledge. Xie *et al.* [40] align local similarity maps constructed from 8 neighbourhood pixels between the student and teacher networks. He *et al.* [14] transfer non-local pairwise affinity maps with an autoencoder to minimize the discrepancy of features. Liu *et al.* [20, 22] perform a pairwise similarity distillation among pixels and an adversarial distillation of score maps. Wang *et al.* [35] distill intra-class feature variation to learn more robust relations with class-wise prototypes. Beyond spatial distillation, Shu *et al.* [30] propose channel-wise distillation to guide the student to mimic the teacher's semantic masks along the channel dimension. Though achieving desirable performance, these approaches only consider pixel dependencies within an individual image, ignoring global pixel relations across various images.

3. Methodology

3.1. Preliminary

Notations of Semantic Segmentation Framework. Unlike traditional image classification, semantic segmentation is a pixel-wise dense classification task. A segmentation network needs to classify each pixel in the image to an individual category label from C classes. The network can be decomposed of a feature extractor and a classifier. The former generates a dense feature map $\mathbf{F} \in \mathbb{R}^{H\times W\times d}$, where H, W and d denote the height, width and number of channels, respectively. We can derive $H \times W$ pixel embeddings along the spatial dimension. The latter further transforms the $\mathbf{F}$ into a categorical logit map $\mathbf{Z} \in \mathbb{R}^{H\times W\times C}$. The conventional segmentation task loss is to train each pixel with its ground-truth label using cross-entropy:

$$L_{task} = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W} CE(\sigma(\mathbf{Z}_{h,w}), y_{h,w}). \quad (1)$$

Here, CE denotes the cross-entropy loss, σ denotes the softmax function and $y_{h,w}$ denotes the ground-truth label of the (h, w)-th pixel.

Pixel-wise Class Probability Distillation. Motivated by Hinton's KD [16], a direct method is to align the class probability distribution of each pixel from the student to the teacher. The formulation is expressed as:

$$L_{kd} = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W} KL(\sigma(\frac{\mathbf{Z}^s_{h,w}}{T})||\sigma(\frac{\mathbf{Z}^t_{h,w}}{T})). \quad (2)$$

Here, $\sigma(\mathbf{Z}^s_{h,w}/T)$ and $\sigma(\mathbf{Z}^t_{h,w}/T)$ represent the soft class probabilities of the (h, w)-th pixel produced from the student and teacher, respectively. KL denotes the Kullback-Leibler divergence, and T is a temperature. Following previous works [20, 35], $T = 1$ is good enough.

3.2. Cross-Image Relational Knowledge Distillaton

Motivation. Although the training objectives of L_{task} and L_{kd} are widely used in semantic segmentation, they only deal with pixel-wise predictions independently but neglect semantic relations between pixels. Some segmentation KD methods [14, 20, 35] attempt to capture spatial relational knowledge by modelling pixel affinity. Nevertheless, these KD methods only construct the relationships among pixels *within a single image*, regardless of the semantic dependencies among pixels across global images. This paper demonstrates that cross-image relational knowledge is also valuable for conducting teacher-student-based KD.

Our CIRKD makes use of pixel embeddings beyond a single image. Inspired by the recent memory-based contrastive learning [6, 31, 34, 38], we may retrieve pixel embeddings of other images from *the current mini-batch* or *an online memory bank*. This paper considers both of two manners to model relationships among pixels, the details of which are shown as follows.

3.2.1 Mini-batch-based Pixel-to-Pixel Distillation

Given a mini-batch $\{\boldsymbol{x}_n\}_{n=1}^N$, the segmentation network extracts N structured feature maps $\{\mathbf{F}_n \in \mathbb{R}^{H\times W\times d}\}_{n=1}^N$ from N input images. We preprocess each pixel embedding in $\mathbf{F}_n$ by l_2-normalization. For easy notation, we reshape the spatial dimension of $\{\mathbf{F}_n \in \mathbb{R}^{H\times W\times d}\}_{n=1}^N$ to $\{\mathbf{F}_n \in \mathbb{R}^{A\times d}\}_{n=1}^N$, where $A = H \times W$. For the i-th image $\boldsymbol{x}_i$ and the j-th image $\boldsymbol{x}_j$, $i, j \in \{1, 2, \cdots, N\}$, we can calculate the cross-image pair-wise similarity matrix $\mathbf{S}_{ij} = \mathbf{F}_i\mathbf{F}_j^\top \in \mathbb{R}^{A\times A}$. The relational matrix $\mathbf{S}_{ij}$ captures the cross-image pair-wise correlations among pixels.

We guide the pair-wise similarity matrix of $\mathbf{S}^s_{ij}$ produced from the student to align that of $\mathbf{S}^t_{ij}$ produced from the teacher. The distillation process is formulated as:

$$L_{p2p}(\mathbf{S}^s_{ij}, \mathbf{S}^t_{ij}) = \frac{1}{A}\sum_{a=1}^{A} KL(\sigma(\frac{\mathbf{S}^s_{ij|a,:}}{\tau})||\sigma(\frac{\mathbf{S}^t_{ij|a,:}}{\tau})). \quad (3)$$

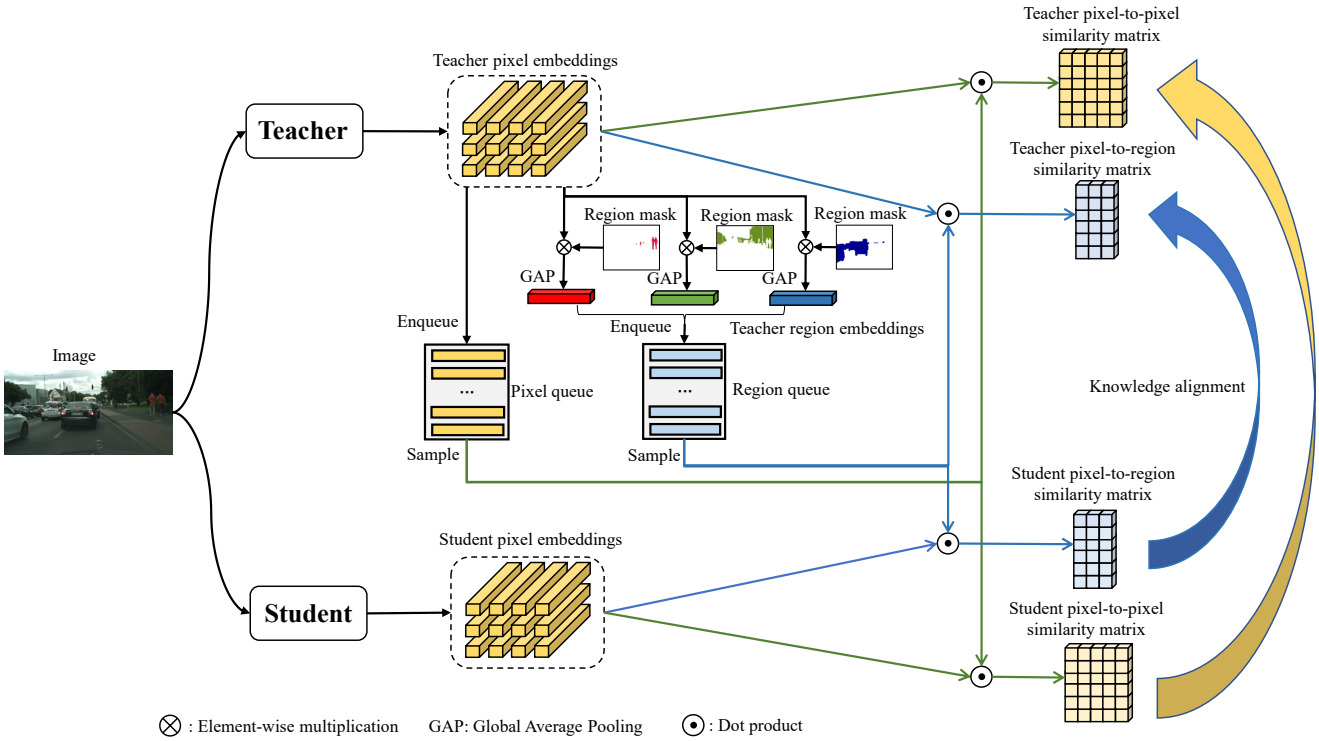


Figure 2. Overview of our proposed memory-based pixel-to-pixel distillation and pixel-to-region distillation.

Here, $\mathbf{S}_{ij|a,:}$ denotes the a-th row vector of $\mathbf{S}_{ij}$. We normalize each row similarity distribution of $\mathbf{S}_{ij}$ to a probability distribution with a temperature τ by softmax function σ. The magnitude gaps would be removed between the student and teacher networks due to the softmax normalization. KL is used to align each row-wise probability distribution. We perform pixel-to-pixel distillation every two of N images:

$$L_{batch\_p2p} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} L_{p2p}(\mathbf{S}_{ij}^{s}, \mathbf{S}_{ij}^{t}). \tag{4}$$

We show the illustration of mini-batch-based pixel-to-pixel distillation in the supplement.

3.2.2 Memory-based Pixel-to-Pixel Distillation

Although mini-batch-based distillation could capture cross-image relations to some extent, it is difficult to model dependencies among pixels from global images, since the batch size per GPU of segmentation task is often small, *e.g.* 1 or 2. To address this problem, we introduce an online pixel queue that can store massive pixel embeddings in the memory bank generated from the past mini-batches. It allows us to retrieve abundant embeddings efficiently. The usage of memory bank dates back to self-supervised learning [31, 38]. This is because a large number of negative samples are pivotal for unsupervised contrastive learning, and the mini-batch size limits available contrastive samples.

In the context of the dense segmentation task, each image would contain a vast number of pixel samples, and most pixels in the same object region are often homogeneous. Therefore storing all pixel embeddings may learn redundant relational knowledge and slow down the distillation process. Moreover, saving several last batches to the queue may also damage the diversity of pixel embeddings. Thus we maintain a class-aware pixel queue $\mathcal{Q}_p \in \mathbb{R}^{C\times N_p\times d}$, where N_p is the number of pixel embeddings per class and d is the embedding size. For each image in the mini-batch, we only randomly sample a small number, *i.e.* V ($V \ll N_p$), of pixel embeddings from the same class and push them into the pixel queue $\mathcal{Q}_p$. The queue is progressively updated under the "first-in-first-out" strategy as distillation proceeds.

Inspired by [10], we adopt a shared pixel queue between the student and teacher networks and store pixel embeddings generated from the teacher during the distillation process. Given an input image $\boldsymbol{x}_n$, the generated pixel embeddings of the student and teacher networks are $\mathbf{F}_n^s \in \mathbb{R}^{A\times d}$ and $\mathbf{F}_n^t \in \mathbb{R}^{A\times d}$, respectively. Each pixel embedding of $\mathbf{F}_n^s$ and $\mathbf{F}_n^t$ is preprocessed by l_2-normalization. We regard $\mathbf{F}_n^s$ and $\mathbf{F}_n^t$ as anchors and sample K_p contrastive embeddings $\{\boldsymbol{v}_k \in \mathbb{R}^d\}_{k=1}^{K_p}$ randomly from the pixel queue $\mathcal{Q}_p$. Here, we adopt a class-balanced sampling since the numbers of pixels from various classes often conform to a long-tailed distribution. For easy notation, $\mathbf{V}_p = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{K_p}] \in \mathbb{R}^{K_p\times d}$ is a concatenation of $\{\boldsymbol{v}_k\}_{k=1}^{K_p}$ along the row dimen-

sion. Then we model the pixel similarity matrix between the anchors and contrastive embeddings for the student and teacher as $\mathbf{P}^s$ and $\mathbf{P}^t$:

$$\mathbf{P}^s = \mathbf{F}_n^s \mathbf{V}_p^\top \in \mathbb{R}^{A \times K_p},\ \mathbf{P}^t = \mathbf{F}_n^t \mathbf{V}_p^\top \in \mathbb{R}^{A \times K_p}. \quad (5)$$

The teacher network often shows a better pixel similarity matrix than the student. We force the student's $\mathbf{P}^s$ to mimic the teacher's $\mathbf{P}^t$ for penalizing the difference. Similar to Section 3.2.1, we apply softmax normalization on each row distribution of $\mathbf{P}^s$ and $\mathbf{P}^t$ and perform pixel-to-pixel distillation via KL-divergence loss. It is formulated as follows:

$$L_{memory\_p2p} = \frac{1}{A}\sum_{a=1}^{A} KL(\sigma(\frac{\mathbf{P}_{a,:}^s}{\tau})||\sigma(\frac{\mathbf{P}_{a,:}^t}{\tau})). \quad (6)$$

After each iteration, we push V teacher pixel embeddings per class into the pixel queue $\mathcal{Q}_p$. Because the teacher is pre-trained and frozen, it can provide consistent feature embeddings during the distillation process. Therefore, we can naturally avoid the inconsistent problem between the anchor and dequeued features appeared in previous contrastive learning [12, 17, 33].

3.2.3 Memory-based Pixel-to-Region Distillation

Discrete pixel embeddings may not fully capture image content. Thus we introduce an online region queue that can store massive more representative region embeddings in the memory bank. Beyond pixel-to-pixel distillation, we further construct pixel-to-region distillation to model the relations between pixels and class-wise region embeddings across global images. Each region embedding represents the feature center of one semantic class in an image. We formulate the region embedding of class c by averagely pooling all the pixel embeddings belonging to class c in a single image.

We maintain a region queue $\mathcal{Q}_r \in \mathbb{R}^{C \times N_r \times d}$ during the distillation process, where N_r is the number of region embeddings per class and d is the embedding size. For each iteration, we sample K_r contrastive region embeddings $\{\boldsymbol{r}_k \in \mathbb{R}^d\}_{k=1}^{K_r}$ from $\mathcal{Q}_r$ in a class-balanced manner. For easy notation, $\mathbf{V}_r = [\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_{K_r}] \in \mathbb{R}^{K_r \times d}$ is a concatenation of $\{\boldsymbol{r}_k\}_{k=1}^{K_r}$ along the row dimension. Given an input image $\boldsymbol{x}_n$, we model the pixel-to-region similarity matrix from $\mathbf{F}_n^s \in \mathbb{R}^{A \times d}$ and $\mathbf{F}_n^t \in \mathbb{R}^{A \times d}$ to region embeddings $\mathbf{V}_r$ as $\mathbf{R}^s$ and $\mathbf{R}^t$:

$$\mathbf{R}^s = \mathbf{F}_n^s \mathbf{V}_r^\top \in \mathbb{R}^{A \times K_r},\ \mathbf{R}^t = \mathbf{F}_n^t \mathbf{V}_r^\top \in \mathbb{R}^{A \times K_r}. \quad (7)$$

Similar to the Equ. (6), we distill normalized pixel-to-region similarity matrix between the student and teacher networks via KL-divergence loss:

$$L_{memory\_p2r} = \frac{1}{A}\sum_{a=1}^{A} KL(\sigma(\frac{\mathbf{R}_{a,:}^s}{\tau})||\sigma(\frac{\mathbf{R}_{a,:}^t}{\tau})). \quad (8)$$

Algorithm 1 Cross-Image Relational KD (CIRKD)

Initialize the pixel queue $\mathcal{Q}_p$ and the region queue $\mathcal{Q}_r$ with random unit vectors.
while the student network has not converged **do**
  Sample a mini-batch.
  Generate the student and teacher pixel embeddings.
  Compute the mini-batch-based pixel-to-pixel distillation loss L_{batch_p2p}.
  Sample contrastive pixel and region embeddings from the pixel queue $\mathcal{Q}_p$ and the region queue $\mathcal{Q}_r$.
  Compute the memory-based pixel-to-pixel loss L_{memory_p2p} and pixel-to-region loss L_{memory_p2r}.
  Update the student w.r.t the overall loss L_{CIRKD}.
  Enqueue the current teacher pixel and region embeddings to $\mathcal{Q}_p$ and $\mathcal{Q}_r$.
  Dequeue the earliest pixel and region embeddings from $\mathcal{Q}_p$ and $\mathcal{Q}_r$.
end while

For each mini-batch, we push all teacher region embeddings into the region queue $\mathcal{Q}_r$. The overview of our proposed memory-based distillation is shown in Fig. 2.

3.3. Overall Framework

We summarize our mini-batch-based pixel-to-pixel, memory-based pixel-to-pixel and pixel-to-region distillation together to train the student network. We also employ the conventional pixel-wise cross-entropy task loss L_{task} (Equ. (1)) and class probability KD loss L_{kd} (Equ. (2)) as the basic losses. The overall loss is formulated as:

$$\begin{aligned} L_{CIRKD} =& L_{task} + L_{kd} + \alpha L_{batch\_p2p} \\ &+ \beta L_{memory\_p2p} + \gamma L_{memory\_p2r}. \end{aligned} \quad (9)$$

Here, α, β and γ are weights coefficients. We set $\alpha = 1$, $\beta = 0.1$ and $\gamma = 0.1$. Empirically, we find our CIRKD are not sensitive to coefficients when $\alpha, \beta, \gamma \in [0.1, 1]$. When the student and teacher networks mismatch the embedding size, we attach a projection head to the student network. It can map the student's pixel embeddings to match the teacher's dimension. The projection head is composed of two 1×1 convolutional layers with ReLU and batch normalization. It would be discarded at the inference phase without introducing extra costs. In Algorithm 1, we use pseudo-code to illustrate the overall training pipeline of CIRKD.

4. Experiments

4.1. Experimental Setup

Dataset. We employ three popular semantic segmentation datasets to conduct our experiments. (1) **Cityscapes** [7] is an urban scene parsing dataset that contains 5000 finely

annotated images, where 2975/500/1525 images are used for `train/val/test`. The segmentation performance is reported on 19 classes. (2) **CamVid** [1] is an automotive dataset that contains 367/101/233 images for `train/val/test` with 11 semantic classes. (3) **Pascal VOC** [9] is a visual object segmentation dataset that includes 20 foreground object categories and one background class. We adopt the augmented data with extra annotations provided by [11]. The resulting dataset contains 10582/1449/1456 images for `train/val/test`.

Evaluation metrics. Following the standard setting, we employ mean Intersection-over-Union (mIoU) to measure the segmentation performance.

Network architectures. For all experiments, we use the segmentation framework DeepLabV3 [5] with ResNet-101 (Res101) backbone [13] as the powerful teacher network. For student networks, we use various segmentation architectures to verify the effectiveness of distillation methods. Specifically, DeepLabV3 and PSPNet [51] with different backbones of ResNet-18 (Res18) and MobileNetV2 (MBV2) [29] are adopted.

Training details. Following the standard data augmentation, we employ random flipping and scaling in the range of $[0.5, 2]$. All experiments are optimized by SGD with a momentum of 0.9, a batch size of 16 and an initial learning rate of 0.02. The number of the total training iterations is 40K. The learning rate is decayed by $(1 - \frac{iter}{total\_iter})^{0.9}$ following the polynomial annealing policy [4]. For crop size during the training phase, we use 512×1024, 360×360 and 512×512 for Cityscapes, CamVid and Pascal VOC, respectively.

Evaluation details. We evaluate the segmentation performance under a single scale setting over the original image size following the general protocol [30].

Compared distillation methods. We compare our proposed CIRKD with state-of-the-art segmentation distillation methods: SKD [20], IFVD [35] and CWD [30]. We re-run all methods using author-provided code. All methods use the same pre-trained teacher DeepLabV3-ResNet101.

Hyper-parameters setup. The hyper-parameters are mainly from the pixel and region queues. For the pixel queue, we set $N_p = 20K$ for each class and enqueue $V = 16$ pixels per class for each image. For the region queue, we set $N_r = 2K$ for each class. For each mini-batch, we sample $K_p = 4096$ pixel embeddings from the pixel queue and $K_r = 1024$ region embeddings from the region queue to compute similarity matrices.

4.2. Experimental Results

4.2.1 Results on Cityscapes

In Table 1, we compare our proposed CIRKD against state-of-the-art distillation methods on Cityscapes in terms of the validation and test mIoU performance. We can observe that all structured KD methods improve student networks under the teacher's supervision. CIRKD achieves the best segmentation performance across various student networks with similar or different architecture styles. It reveals that CIRKD does not rely on architecture-specific cues. Moreover, our method outperforms the best completing CWD with an average 0.60% validation mIoU gain and 0.78% test mIoU gain across four student networks. The results demonstrate that distilling cross-image relations guides the student to achieve better segmentation performance than intra-image pixel affinity [20, 35].

Method	Params (M)	FLOPs (G)	mIoU (%) Val	mIoU (%) Test
T: DeepLabV3-Res101	61.1M	2371.7G	78.07	77.46
S: DeepLabV3-Res18	13.6M	572.0G	74.21	73.45
+SKD [20]			75.42	74.06
+IFVD [35]			75.59	74.26
+CWD [30]			75.55	74.07
+CIRKD (ours)			**76.38**	**75.05**
S: DeepLabV3-Res18*	13.6M	572.0G	65.17	65.47
+SKD [20]			67.08	66.71
+IFVD [35]			65.96	65.78
+CWD [30]			67.74	67.35
+CIRKD (ours)			**68.18**	**68.22**
S: DeepLabV3-MBV2	3.2M	128.9G	73.12	72.36
+SKD [20]			73.82	73.02
+IFVD [35]			73.50	72.58
+CWD [30]			74.66	73.25
+CIRKD (ours)			**75.42**	**74.03**
S: PSPNet-Res18	12.9M	507.4G	72.55	72.29
+SKD [20]			73.29	72.95
+IFVD [35]			73.71	72.83
+CWD [30]			74.36	73.57
+CIRKD (ours)			**74.73**	**74.05**

Table 1. Performance comparison with state-of-the-art distillation methods over various student segmentation networks on Cityscapes. * denotes that we do not initialize the backbone with ImageNet [8] pre-trained weights. FLOPs is measured based on the fixed size of 1024×2048. The bold number denotes the best result in each block. We tag the teacher as T and the student as S.

As illustrated in Fig. 3, we also show the performance of individual class IoU scores over the student network. We can observe that our CIRKD achieves better class IoU scores than baseline (w/o distillation) and CWD consistently, especially for those categories with low IoU scores. For example, our method obtains 10.4% and 9.4% relative improvements on *Wall* than baseline and CWD, respectively. We further show the qualitative segmentation results visually in Fig. 4. We can observe that our CIRKD produces more consistent semantic labels with the ground truth than baseline and CWD, indicating more meaningful pixel dependencies are captured.

T-SNE visualization of learned feature embeddings on the student network by CWD and our proposed CIRKD is shown in Fig. 5. Compared to the CWD, the network trained by CIRKD shows a well-structured pixel-wise semantic feature space. The visual result suggests that learn-

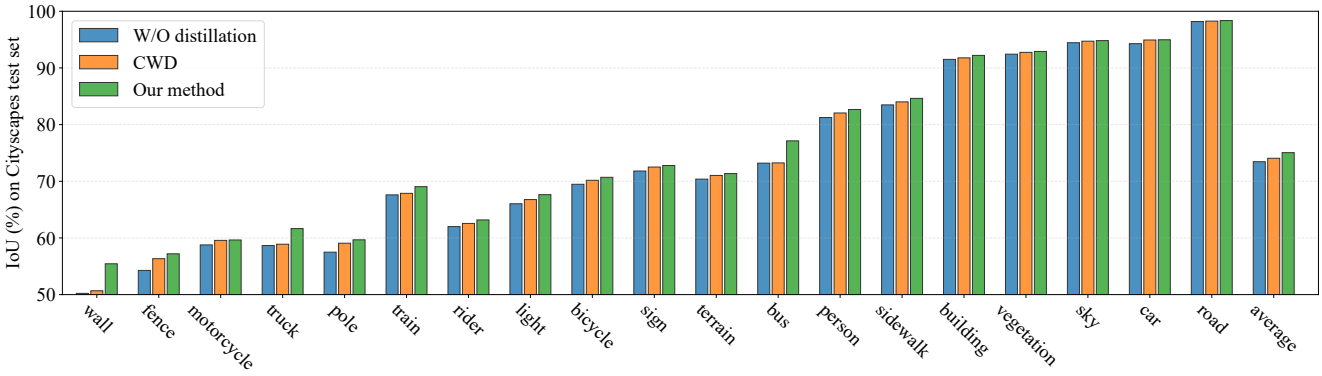
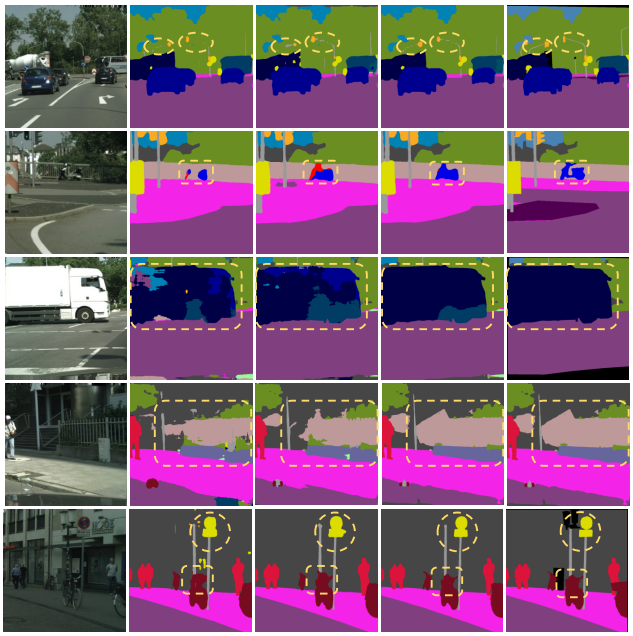


Figure 3. Illustration of individual class IoU scores over the student network DeepLabV3-ResNet18 with baseline (w/o distillation), state-of-the-art CWD and our proposed CIRKD on Cityscapes test set. Our CIRKD can consistently improve individual class IoU scores compared to the baseline and CWD, especially for those challenging classes with low IoU scores.



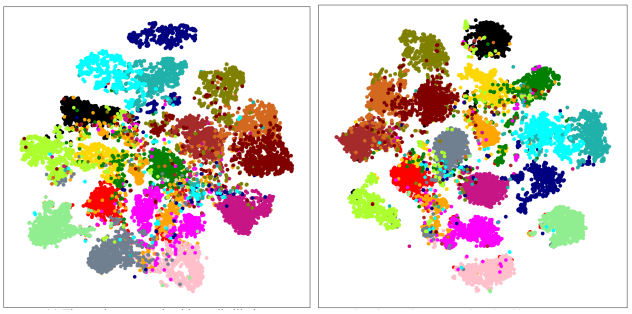
(a) Image (b) Student (c) CWD (d) Ours (e) GT

Figure 4. Qualitative segmentation results on the validation set of Cityscapes using the DeepLabV3-ResNet18 network: (a) raw images, (b) the original student network without KD, (c) channel-wise distillation, (d) our method and (e) ground truth.

ing cross-image pixel relations from the teacher network would help the student achieve better intra-class compactness and inter-class separability, thus improving segmentation performance.

4.2.2 Results on CamVid

In Table 2, we evaluate various distillation methods on CamVid. Our CIRKD achieves the best performance consistently. It outperforms the state-of-the-art CWD by 0.50% and 0.73% mIoU gains over DeepLabV3 and PSPNet, respectively.



(a) State-of-the-art CWD [30]. (b) Our CIRKD.

Figure 5. T-SNE visualization of learned features embeddings on the validation set of Cityscapes over the DeepLabV3-ResNet18 network trained with CWD (*left*) and our proposed CIRKD (*right*).

Method	Params (M)	FLOPs (G)	Test mIoU (%)
T: DeepLabV3-Res101	61.1M	280.2G	69.84
S: DeepLabV3-Res18	13.6M	61.0G	66.92
+SKD [20]			67.46
+IFVD [35]			67.28
+CWD [30]			67.71
+CIRKD (ours)			**68.21**
S: PSPNet-Res18	12.9M	45.6G	66.73
+SKD [20]			67.83
+IFVD [35]			67.61
+CWD [30]			67.92
+CIRKD (ours)			**68.65**

Table 2. Performance comparison with state-of-the-art distillation methods over various student segmentation networks on CamVid. FLOPs is measured based on the test size of 360×480.

4.2.3 Results on Pascal VOC

Beyond scene-parsing datasets, we also evaluate our CIRKD on Pascal VOC, a representative visual object segmentation dataset. As shown in Table 3, CIRKD achieves the best performance compared to other segmentation KD approaches. It surpasses the best completing CWD by 0.48% and 0.79% mIoU improvements on DeepLabV3 and

Method	Params (M)	FLOPs (G)	Val mIoU (%)
T: DeepLabV3-Res101	61.1M	1294.6G	77.67
S: DeepLabV3-Res18	13.6M	305.0G	73.21
+SKD [20]			73.51
+IFVD [35]			73.85
+CWD [30]			74.02
+CIRKD (ours)			**74.50**
S: PSPNet-Res18	12.9M	260.0G	73.33
+SKD [20]			74.07
+IFVD [35]			73.54
+CWD [30]			73.99
+CIRKD (ours)			**74.78**

Table 3. Performance comparison with state-of-the-art distillation methods over various student segmentation networks on Pascal VOC. We report the FLOPs based on the crop size of 512×512 since the validation set does not have a fixed input size.

Loss	Baseline	Distillation					
L_{kd}	-	✓	✓	✓	✓	✓	✓
L_{batch_p2p}	-	-	✓	-	-	-	✓
L_{memory_p2p}	-	-	-	✓	-	✓	✓
L_{memory_p2r}	-	-	-	-	✓	✓	✓
mIoU (%)	73.12	74.26	74.87	75.11	74.94	75.26	**75.42**

Table 4. Ablation study of distillation loss terms on Cityscapes `val`. Baseline denotes the cross-entropy loss L_{task} (Equ. (1)).

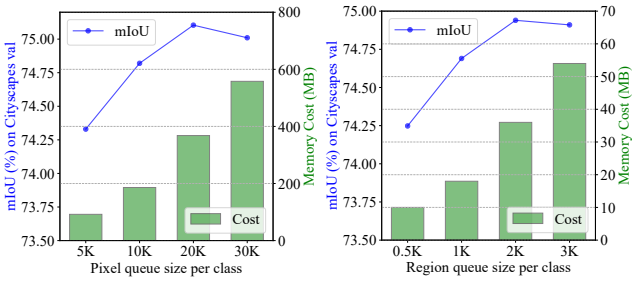
PSPNet, respectively. The results demonstrate the scalability of our CIRKD to work reasonably well on visual object segmentation.

4.3. Ablation Study and Parameter Analysis

We conduct thorough ablation experiments of our proposed CIRKD on the Cityscapes validation set, a standard benchmark for semantic segmentation. For all experiments, we choose DeepLabV3-ResNet101 as the teacher and DeepLabV3-MobileNetV2 as the student by default.

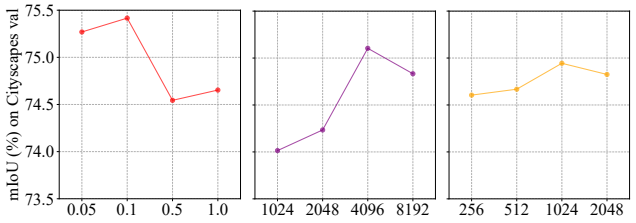
Ablation study of loss terms. As shown in Table 4, we examine the contribution of each distillation loss. The conventional KD loss L_{kd} improves the baseline by 1.14%. Applying cross-image relational KD losses of L_{batch_p2p}, L_{memory_p2p} and L_{memory_p2r} lead to 0.61%, 0.85% and 0.68% mIoU gains over L_{kd}, respectively. The results show two conclusions: (1) Pixel-to-pixel distillation is more informative than the pixel-to-region counterpart. (2) Memory-based pixel-to-pixel distillation is better than the mini-batch-based counterpart, since the former can capture broader pixel dependencies from much more images than the latter. Finally, applying all losses together maximizes the segmentation performance, reducing the gap between the student and teacher from 4.95% to 2.65%.

Impact of the queue size. We investigate the impact of memory sizes of the pixel queue and region queue. As shown in Fig. 6, distillation performance increases as the sizes of the pixel queue and region queue grow. This is because a larger queue could provide more abundant and diverse embeddings for capturing long-range dependencies.



(a) Pixel queue size N_p per class (b) Region queue size N_r per class

Figure 6. Impact of the (a) pixel queue size N_p per class and (b) region queue size N_r per class on Cityscapes `val`. 'Memory Cost' denotes the occupied GPU memory size.



(a) Temperature τ (b) Pixel-based K_p (c) Region-based K_r

Figure 7. Impact of (a) the temperature τ and (b) the number of contrastive pixel embeddings K_p and (c) the number of contrastive region embeddings K_r on Cityscapes `val`.

The results also show the distillation performance may also saturate at a certain memory capacity.

Impact of the temperature τ. Temperature τ is used to calibrate the similarity distribution for relational KD. A more significant temperature τ brings a smoother distribution. As shown in Fig. 7a, we investigate the impact of τ in our CIRKD and find $\tau = 0.1$ is the best choice.

Impact of the number of contrastive embeddings. As shown in Fig. 7b and Fig. 7c, we examine the number of contrastive embeddings to calculate pixel-to-pixel and pixel-to-region similarity matrices. The distillation performance increases as K_p and K_r grow, because the similarity distribution with a larger dimension would encode broader pixel dependencies. The upper bound of distillation performance may saturate at $K_p = 4096$ for pixel-to-pixel distillation and $K_r = 1024$ for pixel-to-region distillation.

5. Conclusion

This paper presents a novel cross-image relational KD to transfer global pixel correlations from the teacher to the student for semantic segmentation. Compared to previous KD approaches, our method helps students learn broader pixel dependencies from the teacher. Experiments on public segmentation datasets demonstrate the effectiveness of our CIRKD. We hope our work can inspire future research to explore global pixel relationships for segmentation KD.

References

[1] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*, pages 44–57. Springer, 2008.

[2] Linhang Cai, Zhulin An, Chuanguang Yang, Yangchun Yan, and Yongjun Xu. Prior gradient mask guided pruning-aware fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[3] 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.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

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

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

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

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[10] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *ICLR*, 2021.

[11] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[13] 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.

[14] 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.

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

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

[17] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16291–16301, 2021.

[18] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 6356–6364, 2017.

[19] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[20] 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.

[21] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *European Conference on Computer Vision*, pages 352–368. Springer, 2020.

[22] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[23] 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.

[24] 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.

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

[26] 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.

[27] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019.

[28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *ICLR*, 2015.

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[30] 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.

[31] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[32] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[33] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021.

[34] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2020.

[35] 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.

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[37] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4820–4828, 2016.

[38] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[39] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

[40] Jiafeng Xie, Bing Shuai, Jian-Fang Hu, Jingyang Lin, and Wei-Shi Zheng. Improving fast segmentation with teacher-student learning. *arXiv preprint arXiv:1810.08476*, 2018.

[41] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Hierarchical self-supervised augmented knowledge distillation. *International Joint Conference on Artificial Intelligence*, pages 1217–1223, 2021.

[42] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Mutual contrastive learning for visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[43] Chuanguang Yang, Zhulin An, Chao Li, Boyu Diao, and Yongjun Xu. Multi-objective pruning for cnns using genetic algorithm. In *International Conference on Artificial Neural Networks*, pages 299–305. Springer, 2019.

[44] Chuanguang Yang, Zhulin An, and Yongjun Xu. Multi-view contrastive learning for online knowledge distillation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3750–3754, 2021.

[45] Chuanguang Yang, Zhulin An, Hui Zhu, Xiaolong Hu, Kun Zhang, Kaiqiang Xu, Chao Li, and Yongjun Xu. Gated convolutional networks with hybrid connectivity for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12581–12588, 2020.

[46] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[47] 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.

[48] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

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

[50] 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.

[51] 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.

[52] Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. *ICLR*, 2021.