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

Presenter: Seonghak KIM

Introduction

- **Knowledge Distillation**
  - **Previous works**
    - A broad range of KD methods have been well studied but mostly for image classification.
    - Directly utilizing classification-based KD for **dense prediction tasks** → desirable performance X † ¶
      - Ignore of the structured context among pixels

Task of predicting a label for each pixel
(i.e., semantic and instance segmentation)

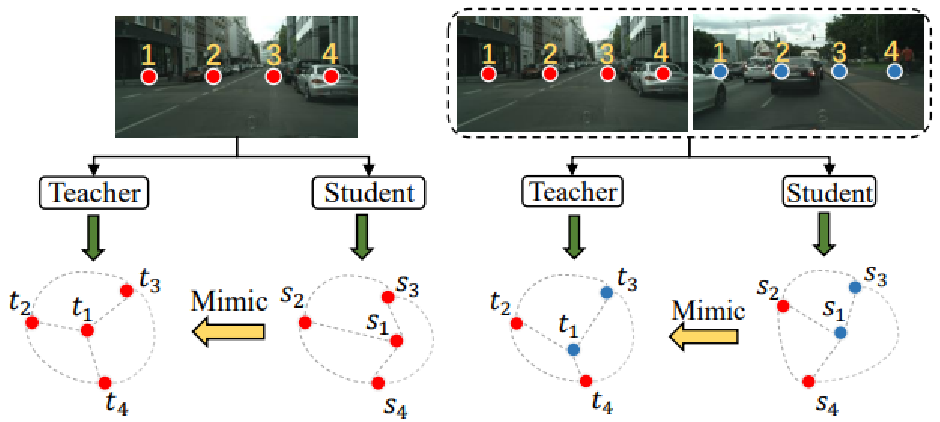
→ **Specialized KD methods for semantic segmentation!**

- Although existing segmentation-based KD employs structured spatial knowledge, this is generated from individual data samples, **ignoring cross-image semantic relations among pixels.**

† Quanquan Li *et al.*, CVPR, 2017
¶ Yifan Liu *et al.*, TPAMI, 2020

Introduction

- **Contributions**
  - **Global pixels relations across the various images**
    - Pixel-to-pixel distillation
    - Pixel-to-region distillation



(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.

† Quanquan Li *et al.*, CVPR, 2017
¶ Yifan Liu *et al.*, TPAMI, 2020

Method

- **Notations**
  - **Segmentation**
    - Feature extractor, $\mathbf{F} \in \mathbb{R}^{H \times W \times d}$
    - Classifier, $\mathbf{F} \rightarrow \mathbf{Z} \in \mathbb{R}^{H \times W \times C}$
  - **Loss functions**
    - Conventional segmentation loss, $\mathcal{L}_{\text{seg}} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathrm{CE}\left(\underbrace{\sigma(\mathbf{Z}_{h,w})}_{\text{Each pixel's logit after softmax}}, \underbrace{y_{h,w}}_{\text{Ground-truth label}}\right)$
    - Pixel-wise logit distillation, $\mathcal{L}_{\text{kd}} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathrm{KL}\left(\underbrace{\sigma\left(\frac{\mathbf{Z}_{h,w}^{\mathcal{S}}}{\boldsymbol{T}}\right) || \sigma\left(\frac{\mathbf{Z}_{h,w}^{\mathcal{T}}}{\boldsymbol{T}}\right)}_{\text{Soft class probabilities from student and teacher}}\right)$

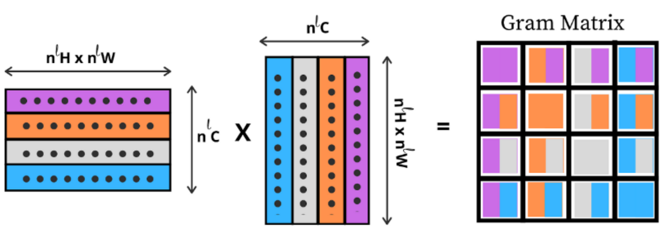
➔ (−) Only address pixel-wise predictions **independently** but **neglect semantic relations between pixels.**

Method

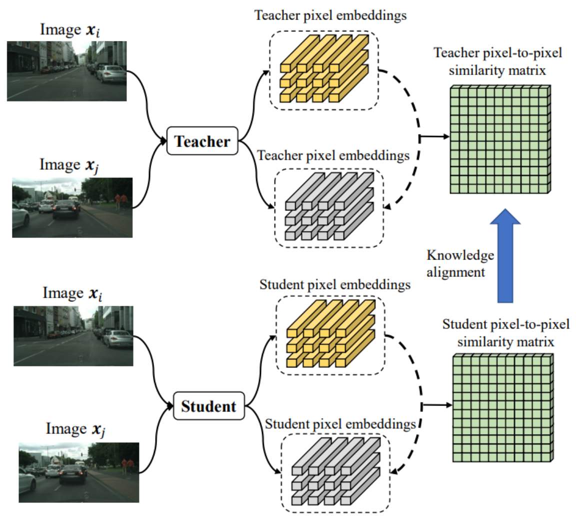
- **Cross-Image Relational KD (CIRKD)**
  - **Pixel-to-Pixel Distillation**
    - Mini-batch-based distillation
      - Mini-batch, $\{x_n\}_{n=1}^{N}$ → feature maps, $\{\mathbf{F}_n \in \mathbb{R}^{H\times W\times d}\}_{n=1}^{N} = \{\mathbf{F}_n \in \mathbb{R}^{A\times d}\}_{n=1}^{N}$
      - Cross-image pair-wise similarity matrix, $\mathbf{S}_{i,j} = \mathbf{F}_i\mathbf{F}_j^{\mathrm{T}} \in \mathbb{R}^{A\times A}$
      - $\mathcal{L}_{\text{p2p}}(\mathbf{S}_{i,j}^{\mathcal{S}}, \mathbf{S}_{i,j}^{\mathcal{T}}) = \frac{1}{A}\sum_{a=1}^{A} \text{KL}\left(\sigma\left(\boxed{\frac{\mathbf{S}_{ij|a,:}^{\mathcal{S}}}{\tau}\right) || \sigma\left(\frac{\mathbf{S}_{ij|a,:}^{\mathcal{T}}}{\tau}}\right)\right)$

        a^{th} row vector
    - Mini-batch-based Pixel-to-pixel loss, $\mathcal{L}_{\text{batch\_p2p}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{L}_{p2p}(\mathbf{S}_{i,j}^{\mathcal{S}}, \mathbf{S}_{i,j}^{\mathcal{T}})$

(−) batch size per GPU of segmentation is often small → dependencies among pixels from global images ↓



[Gram matrix]



[Overview of mini-batch based pixel-to-pixel distillation]

* Motivated by self-supervised learning

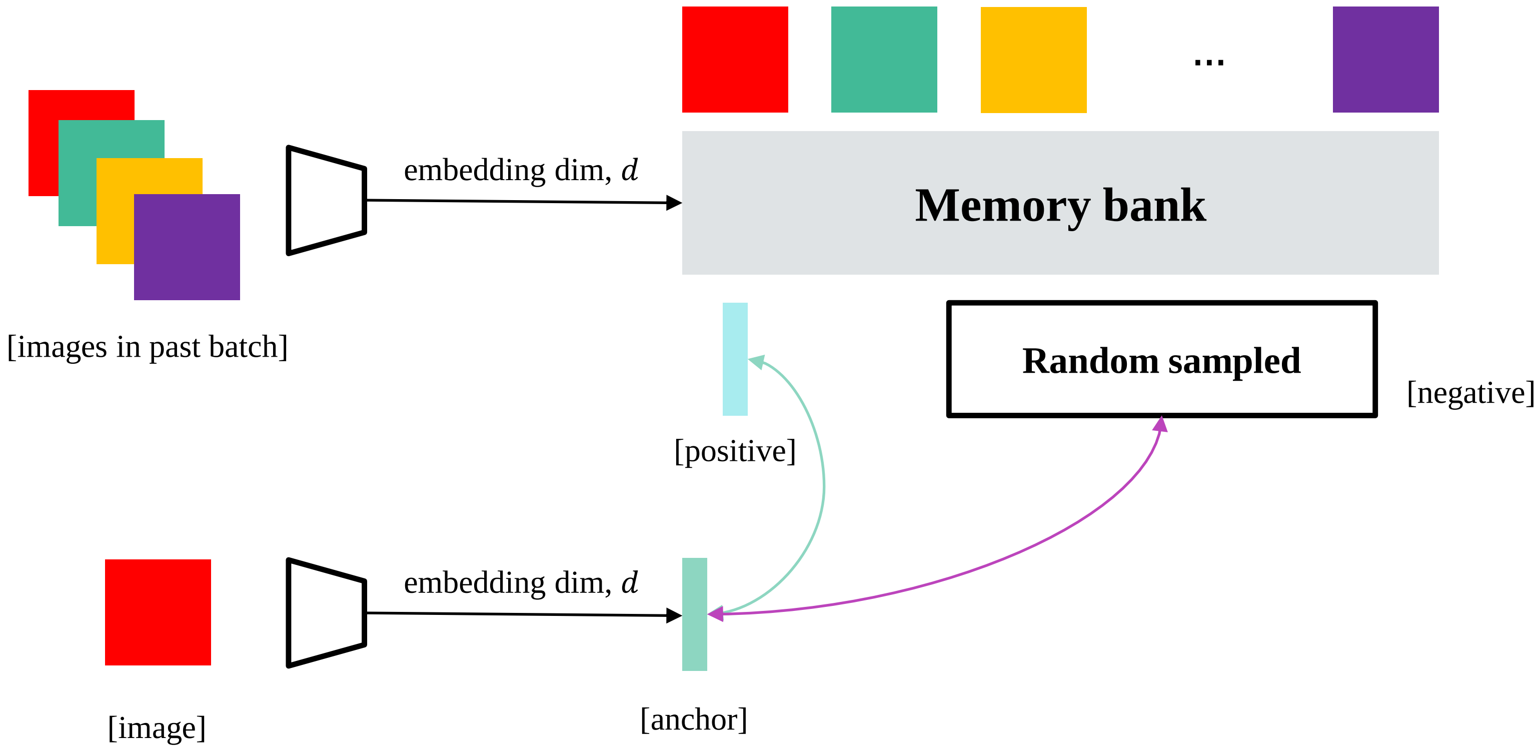
Cross-Image Relational KD (CIRKD)

Pixel-to-Pixel Distillation

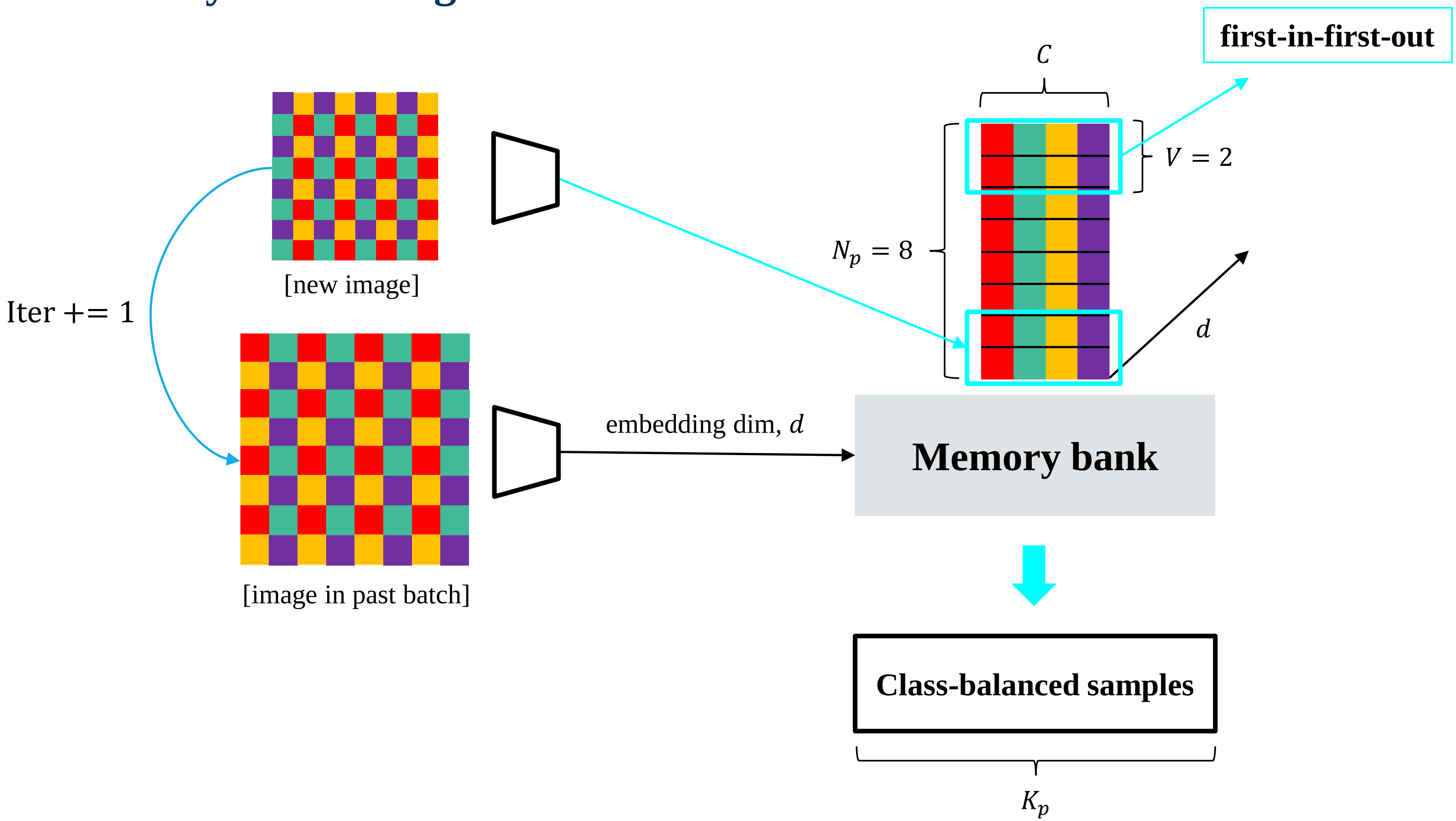
- Memory-based distillation
  - Pixel embeddings from the past mini-batches are stored in the memory bank*
  - Class-ware pixel queue, $\mathcal{Q}_p \in \mathbb{R}^{C \times N_p \times d}$
    - N_p: number of pixel embeddings per class, d: embedding size
  - Input image, $x_n \rightarrow$ feature embeddings, $\mathbf{F}_n^{\mathcal{S}}, \mathbf{F}_n^{\mathcal{T}} \in \mathbb{R}^{A \times d}$
  - Anchors $\mathbf{F}_n^{\mathcal{S}}, \mathbf{F}_n^{\mathcal{T}}$ and class-balanced sample K_p contrastive embeddings $\boxed{\{\boldsymbol{v}_k \in \mathbb{R}^d\}_{k=1}^{K_p} \text{ randomly from } \mathcal{Q}_p}$
  - $\boxed{\mathbf{V}_p = \left[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{K_p}\right] \in \mathbb{R}^{K_p \times d}}$

    Concatenation
  - Similarity matrix between the anchors and contrastive embeddings, $\mathbf{P} = \mathbf{F}_n \mathbf{V}_n^{\mathrm{T}} \in \mathbb{R}^{A \times K_p}$
- Memory-based Pixel-to-Pixel loss, $\mathcal{L}_{\text{memory\_p2p}} = \frac{1}{A}\sum_{a=1}^{A} \mathrm{KL}\left(\sigma\left(\frac{\mathbf{P}_{a,:}^{\mathcal{S}}}{\tau}\right) || \sigma\left(\frac{\mathbf{P}_{a,:}^{\mathcal{T}}}{\tau}\right)\right)$

- *Memory bank in self-supervised learning



- **Memory bank in segmentation**



* Motivated by self-supervised learning

Cross-Image Relational KD (CIRKD)

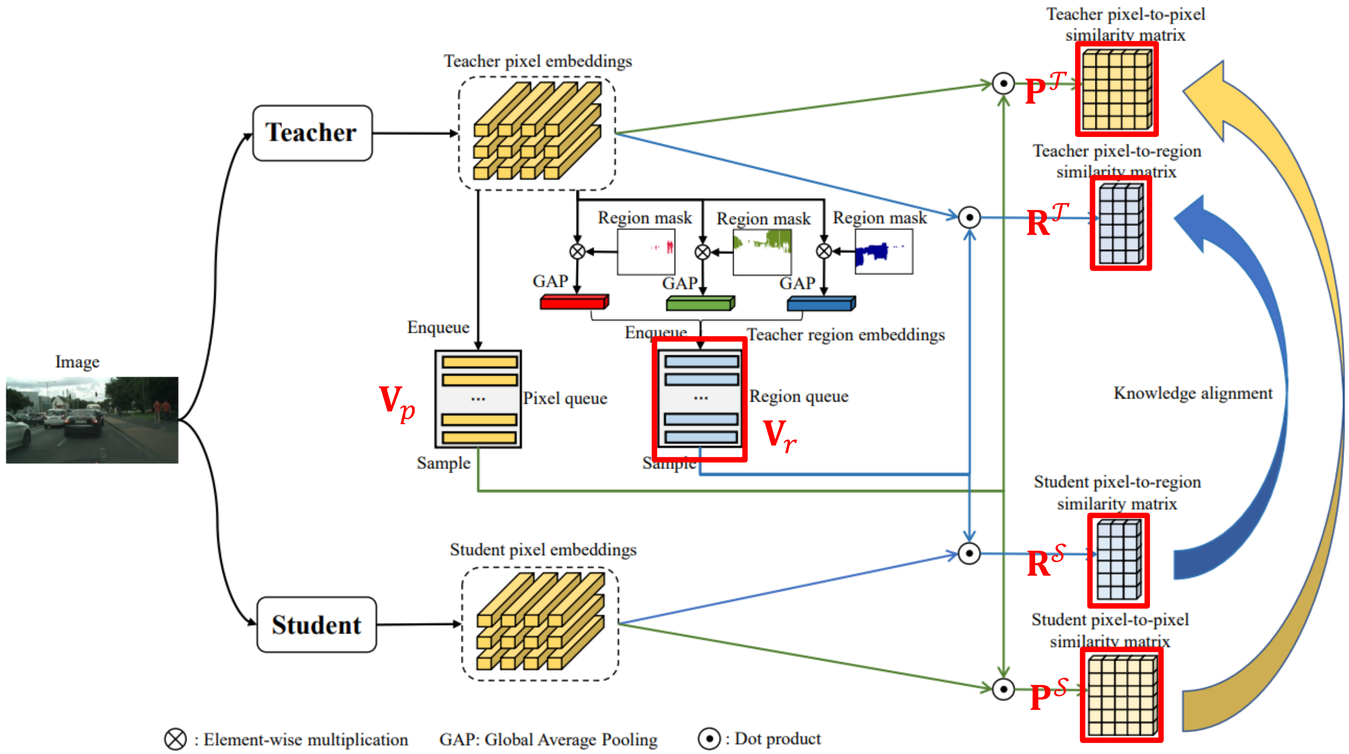
Pixel-to-Region Distillation

- Memory-based distillation
  - More representative region embeddings are stored in the memory bank*
    - by **averagely pooling** all the pixel embeddings belonging to class c in a single image
  - Region queue, $\mathcal{Q}_r \in \mathbb{R}^{C \times N_r \times d}$
    - N_r: number of region embeddings per class, d: embedding size
  - sample K_r contrastive region embeddings $\{\boldsymbol{r}_k \in \mathbb{R}^d\}_{k=1}^{K_r}$ randomly from $\mathcal{Q}_r \rightarrow \mathbf{V}_r = [\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{K_r}] \in \mathbb{R}^{K_r \times d}$
  - Pixel-to-region similarity matrix, $\mathbf{R} = \mathbf{F}_n \mathbf{V}_r^{\mathrm{T}} \in \mathbb{R}^{A \times K_r}$
- Memory-based Pixel-to-Pixel loss, $\mathcal{L}_{\text{memory\_p2r}} = \frac{1}{A} \sum_{a=1}^{A} \mathrm{KL}\left(\sigma\left(\frac{\mathbf{R}_{a,:}^{\mathcal{S}}}{\tau}\right) || \sigma\left(\frac{\mathbf{R}_{a,:}^{\mathcal{T}}}{\tau}\right)\right)$

* Motivated by self-supervised learning

- **Cross-Image Relational KD (CIRKD)**
  - **Overall framework**
    - $\mathcal{L}_{\text{CIRKD}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{kd}} + \alpha\mathcal{L}_{\text{batch\_p2p}} + \beta\mathcal{L}_{\text{memory\_p2p}} + \gamma\mathcal{L}_{\text{memory\_p2r}}$
      - If $d^{\mathcal{S}} \neq d^{\mathcal{T}}$, projection head is attached to the student model

```
self.project_head = nn.Sequential(
            nn.Conv2d(s_channels, t_channels, 1, bias=False),
            nn.SyncBatchNorm(t_channels),
            nn.ReLU(True),
            nn.Conv2d(t_channels, t_channels, 1, bias=False)
        )
```



[Overview of memory-based pixel-to-pixel and pixel-to-region distillation]

Experiments

- **Cityscapes**
  - **mIoU performance**

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.

Experiments

- **Cityscapes**
  - **Performance of individual class IoU scores**

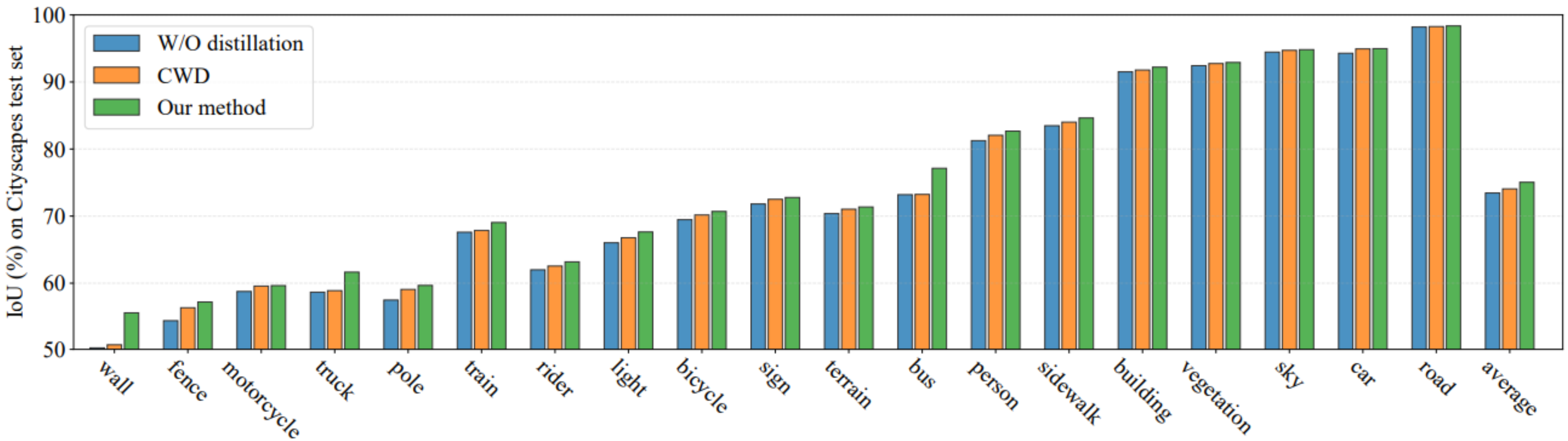


Figure 3. Illustration of individual class IoU scores over the student network DeepLabV3-ResNet18 with baseline (w/o distillation), state-of-the-art CWD and our proposed CIRKD on Cityscapes test set. Our CIRKD can consistently improve individual class IoU scores compared to the baseline and CWD, especially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

Experiments

- **Cityscapes**
  - **Qualitative results**

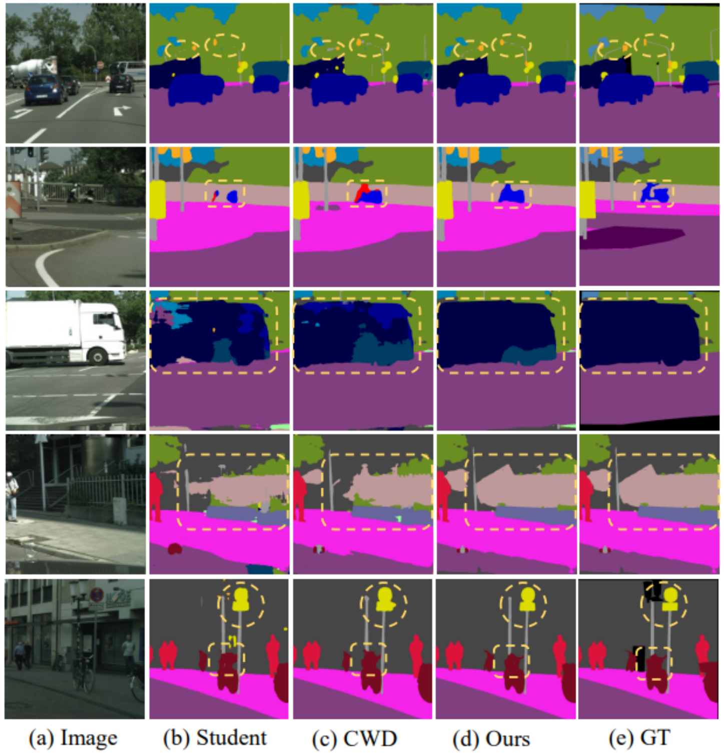


Figure 4. Qualitative segmentation results on the validation set of Cityscapes using the DeepLabV3-ResNet18 network: (a) raw images, (b) the original student network without KD, (c) channel-wise distillation, (d) our method and (e) ground truth.

Experiments

- **CamVid and Pascal VOC**
  - **mIoU performance**

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.

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.

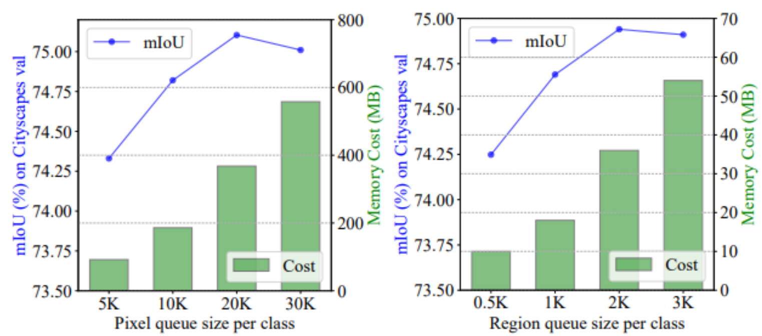
Experiments

- **Ablation study**
  - **Loss term**

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)).

  - **Queue size**
    - Larger queue provide more abundant and diverse embeddings



(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.

Experiments

- **Ablation study**
  - **Temperature τ**
  - **Number of contrastive embeddings**
    - The similarity distribution with a larger dimension encode broader pixel dependencies.

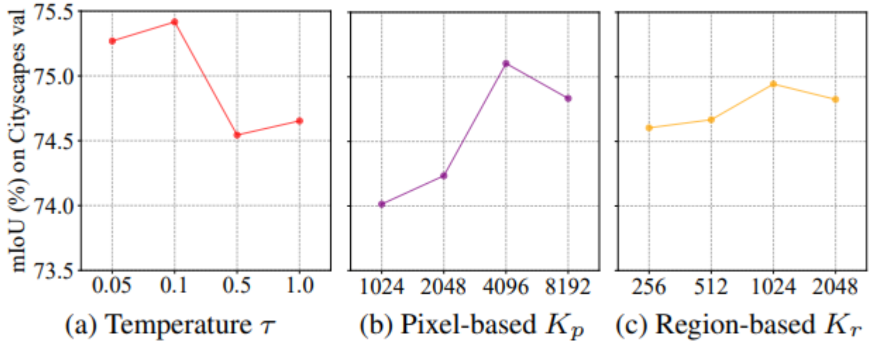


Figure 7. Impact of (a) the temperature τ and (b) the number of contrastive pixel embeddings K_p and (c) the number of contrastive region embeddings K_r on Cityscapes `val`.

Conclusion

- **Contributions**
  - **Cross-image relational KD transferring global pixel correlations**
  - **Significant improvement on various segmentation datasets**

Thank you.