南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Channel-spatial knowledge distillation for efficient semantic segmentation

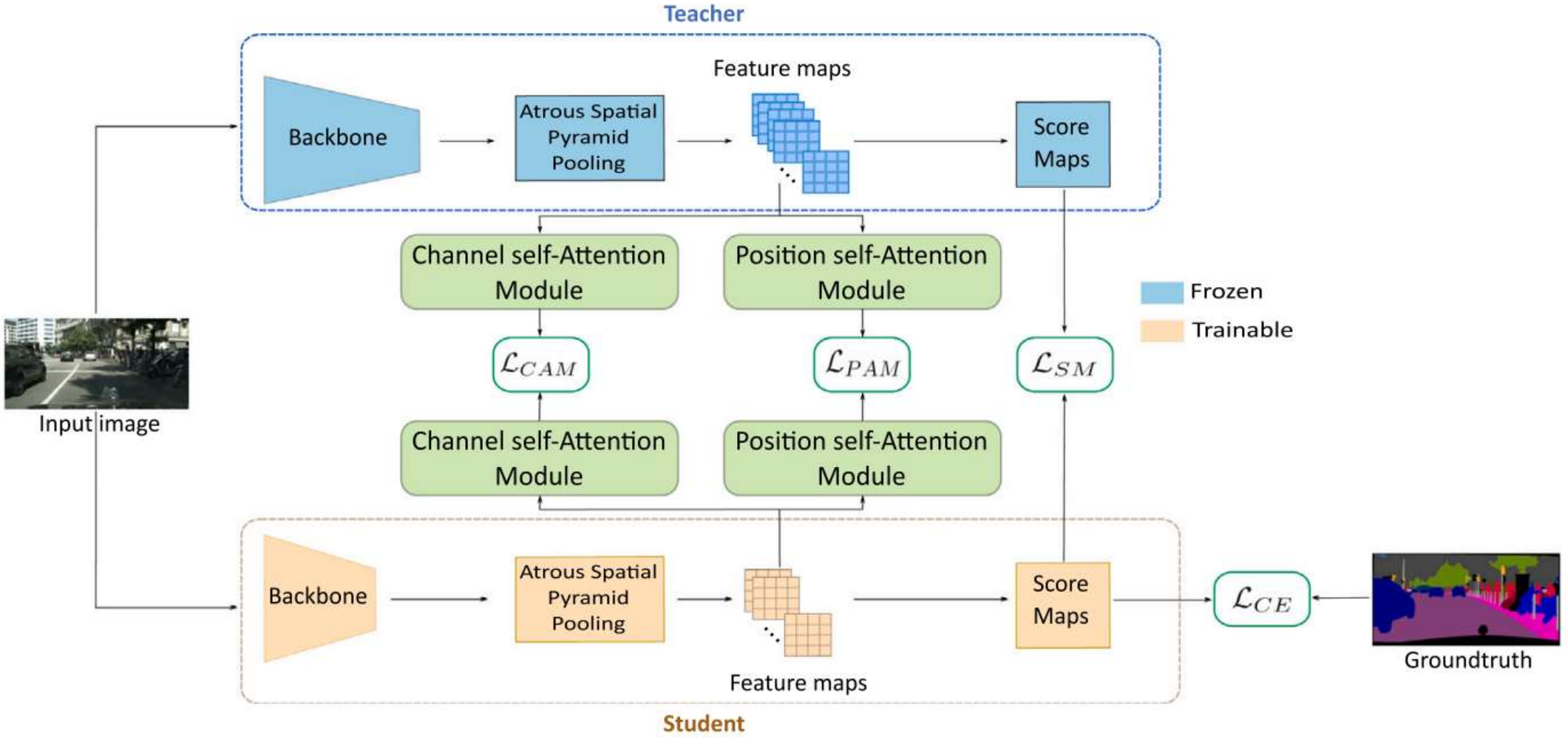
（IEEE TGRS） 2024

Contributions

1、 improve the distillation mechanism by capturing the contextual dependencies in spatial and channel dimensions through two self-attention modules.

2、adopt the Centered Kernel Alignment (CKA) metric that avoids the student to add additional leaning layers to match the teacher features size.

Flowchart of the proposed method: CSKD



Method：Channel self-attention module

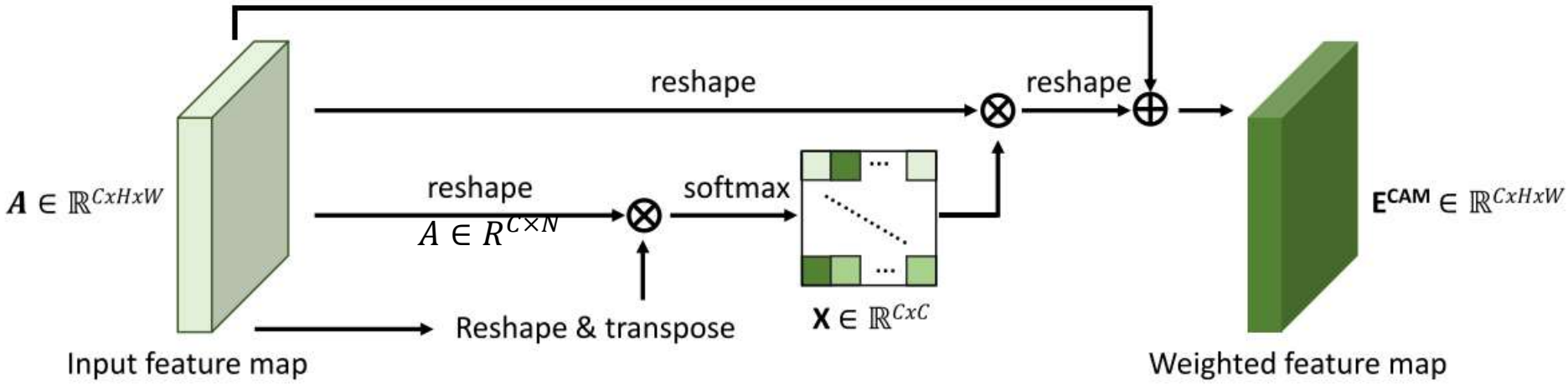


Fig. 2. The architecture of Channel self-Attention Module (CAM).

$$x_{ji} = \phi(A_j \cdot A_i) = \frac{\exp(\frac{A_j \cdot A_i}{\mathcal{T}_{FM}})}{\sum_{i=1}^{C} \exp(\frac{A_j \cdot A_i}{\mathcal{T}_{FM}})} \qquad E_j^{CAM} = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j$$

Method： Position self-attention module

$B, C \in R^{C_r \times H \times W} where C_r < C$

$B, C \in R^{C_r \times N}$

$D \in R^{C \times N}$

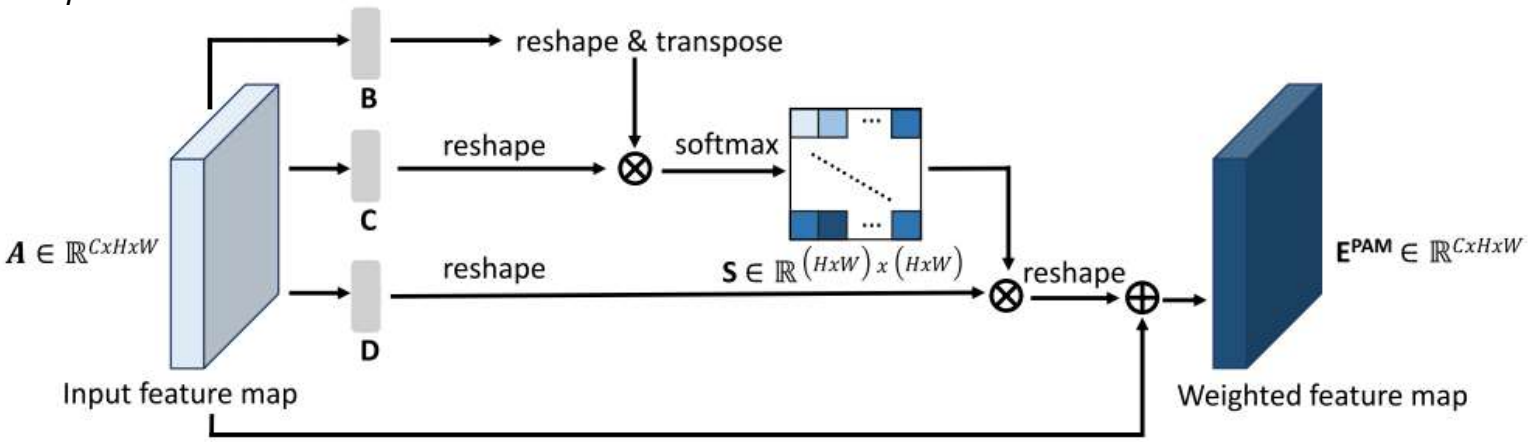


Fig. 3. The architecture of Position self-Attention Module (PAM). The gray rectangles refer to a 1 × 1 convolution layer.

$$s_{ji} = \phi(B_j \cdot C_i) = \frac{\exp(\frac{B_i \cdot C_j}{\mathcal{T}_{FM}})}{\sum_{i=1}^{N} \exp(\frac{B_i \cdot C_j}{\mathcal{T}_{FM}})} \qquad E_j^{PAM} = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + A_j$$

Optimization

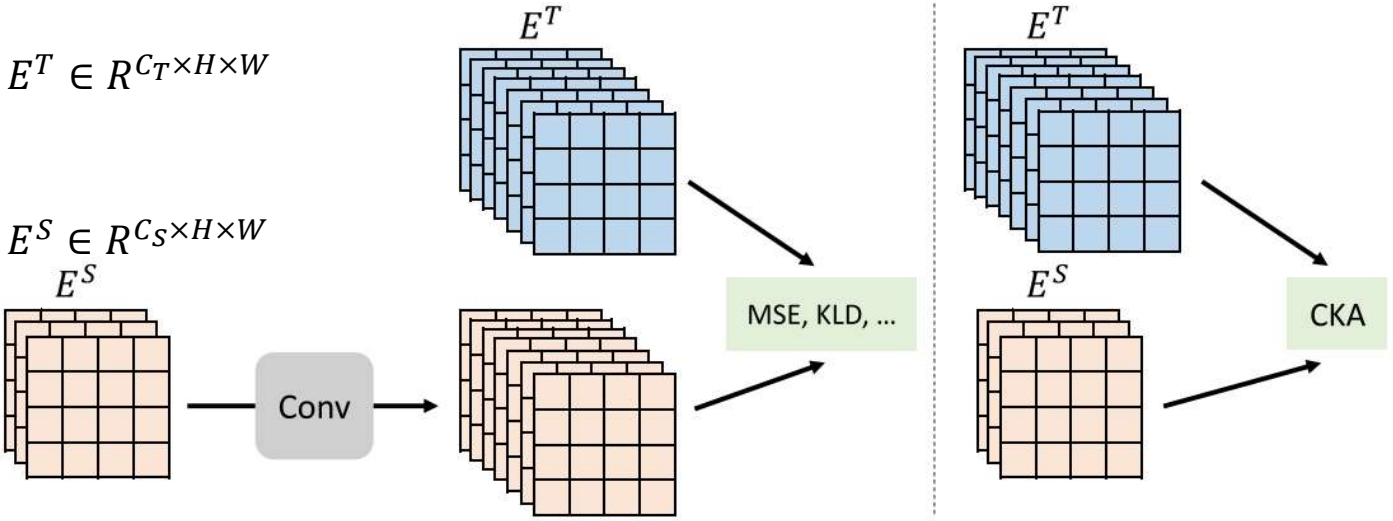
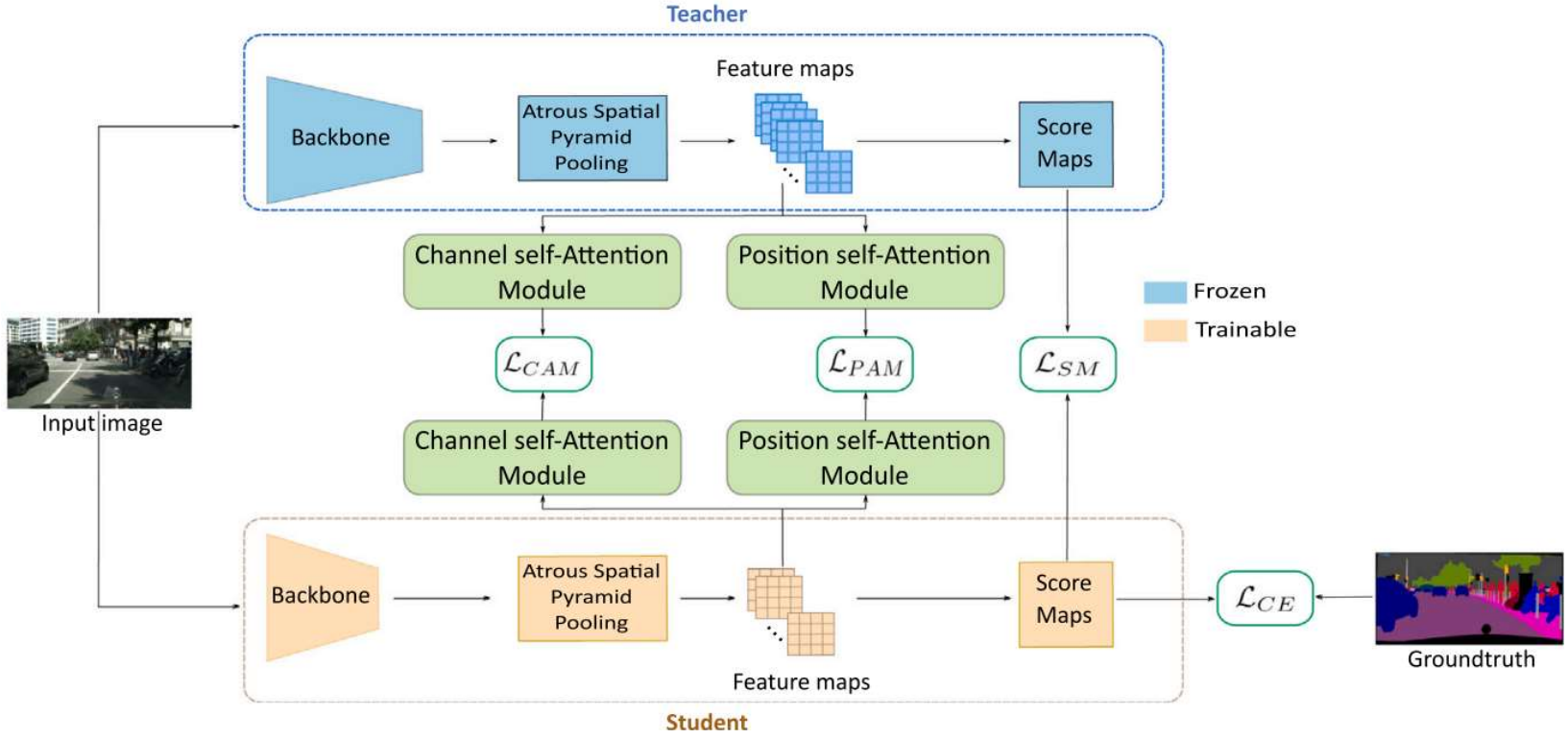


Fig.4. Left: Knowledge distillation with a loss that requires mapping the student feature maps into the same number of the teacher feature maps. Right: Knowledge distillation that computes directly the similarity between the teacher and student feature maps.

$$\mathrm{CKA}(E^S, E^T) = \frac{\|(E^T)'(E^S)\|_F^2}{\|(E^S)'(E^S)\|_F \|(E^T)'(E^T)\|_F} \qquad \mathcal{L}_{FM} = -\log(\mathrm{CKA}(E^S, E^T))$$

Kornblith, S., Norouzi, M., Lee, H. and Hinton, G., 2019, May. Similarity of neural network representations revisited. In International conference on machine learning (pp. 3519-3529). PMLR.

Optimization



$$\mathcal{L}_{CAM} = -\log(\mathrm{CKA}(E^{CAM_T}, E^{CAM_S}))$$

$$\mathcal{L}_{PAM} = -\log(\mathrm{CKA}(E^{PAM_T}, E^{PAM_S}))$$

$$\mathcal{L}_{Total} = \lambda_{CAM}\mathcal{L}_{CAM} + \lambda_{PAM}\mathcal{L}_{PAM} + \lambda_{SM}\mathcal{L}_{SM} + \mathcal{L}_{CE}$$

Experiments

1、Ablation study

Table 1
Comparison on the validation set of Cityscapes using different schemes of our method and two distinct losses. T and S refer to the Teacher and Student respectively.

Method	Val $mIoU$ (%)	
T: DeepLabV3-R101	78.07	
S: DeepLabV3-R18	74.21	
	MSE	CKA
S+CAM ($\lambda_{CAM} = 1 \mid \lambda_{PAM} = 0$)	75.00	75.49
S+PAM ($\lambda_{CAM} = 0 \mid \lambda_{PAM} = 1$)	75.26	75.43
S+CAM+PAM ($\lambda_{CAM} = 1 \mid \lambda_{PAM} = 1$)	**76.05**	**76.74**

Experiments

2、 Comparison with the state-of-the-arts

Table 2
Comparison study of our method with the state-of-the-art methods on Cityscapes dataset.

Method	Params (M)	*mIoU* (%) Val	Test
Comparison with different segmentation methods			
SegNet [32]	29.5	–	57.0
ENet [7]	0.4	–	58.3
ERFNet [33]	2.1	–	68.0
ESPNet [8]	0.4	–	60.3
ICNet [34]	26.5	–	69.5
BiseNet [9]	49.0	–	74.7
RefineNet [35]	118.4	–	73.6
DFANet [36]	7.8	–	70.3
FCN [2]	134.5	–	65.3

Comparison with different distillation methods			
T: DeepLabV3-R101	61.1	78.07	77.46
S: DeepLabV3-R18	13.6	74.21	73.45
+SKD [19]		75.42	74.06
+IFVD [20]		75.59	74.26
+CWD [26]		75.55	74.07
+CIRKD [24]		76.38	**75.05**
+CSKD (ours)		**76.74**	75.04
S: DeepLabV3-MN2	3.2	73.12	72.36
+SKD [19]		73.82	73.02
+IFVD [20]		73.50	72.58
+CWD [26]		74.66	73.25
+CIRKD [24]		75.42	74.03
+CSKD (ours)		**76.22**	**74.11**
S: PSPNet-R18	12.9	72.55	72.29
+SKD [19]		73.29	72.95
+IFVD [20]		73.71	72.83
+CWD [26]		74.36	73.57
+CIRKD [24]		**74.73**	**74.05**
+CSKD (ours)		74.47	73.71
T: PSPNet-R101	70.43	74.74	–
S: PSPNet-R18	12.9	72.55	–
+CWD [26]		73.75	–
+CIRKD [24]		72.45	–
+CSKD (ours)		**73.45**	–
S: DeepLabV3-R18	13.6	74.21	–
+CWD [26]		73.96	–
+CIRKD [24]		74.46	–
+CSKD (ours)		**74.74**	–

Experiments

Table 3
Comparison study of our method with the state-of-the-art methods on CamVid dataset.

Method	Params (M)	Test *mIoU* (%)
Comparison with different segmentation methods		
SegNet [32]	29.5	55.6
ENet [7]	0.4	51.3
ESPNet [8]	0.4	57.8
ICNet [34]	26.5	67.1
BiseNet [9]	49.0	68.7
DFANet [36]	7.8	59.3
Comparison with different distillation methods		
T: DeepLabV3-R101	61.1	69.84
S: DeepLabV3-R18	13.6	66.92
+SKD [19]		67.46
+IFVD [20]		67.28
+CWD [26]		67.71
+CIRKD [24]		68.21
+CSKD (ours)		**68.87**
S: PSPNet-R18	12.9	66.73
+SKD [19]		67.83
+IFVD [20]		67.61
+CWD [26]		67.92
+CIRKD [24]		68.65
+CSKD (ours)		**69.51**

Table 4
Comparison study of our method with the state-of-the-art methods on Pascal VOC dataset.

Method	Params (M)	Val *mIoU* (%)
Comparison with different segmentation methods		
FCN [2]	134.5	69.6
RefineNet [35]	118.1	82.4
PSANet [37]	78.13	77.9
OCRNet [38]	70.37	80.3
Comparison with different distillation methods		
T: DeepLabV3-R101	61.1	77.67
S: DeepLabV3-R18	13.6	73.21
+SKD [19]		73.51
+IFVD [20]		73.85
+CWD [26]		74.02
+CIRKD [24]		74.50
+CSKD (ours)		**76.66**
S: PSPNet-R18	12.9	73.33
+SKD [19]		74.07
+IFVD [20]		73.54
+CWD [26]		73.99
+CIRKD [24]		74.78
+CSKD (ours)		**75.77**

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

THANKS