Attention-guided Feature Distillation for Semantic Segmentation

Amir M. Mansourian, Arya Jalali*, Rozhan Ahmadi, Shohreh Kasaei*

AAAI 2025

*These authors contributed equally.

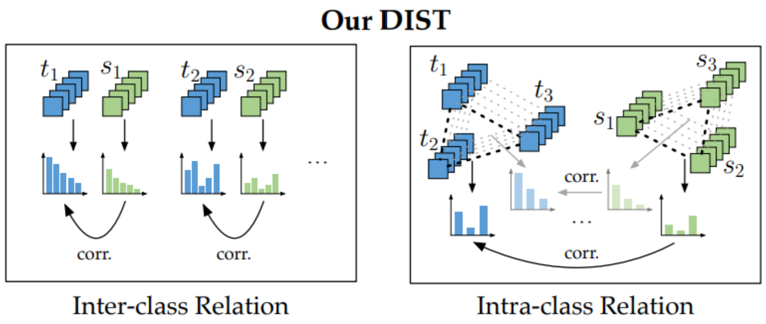
- Problem/Objective
  - Knowledge Distillation method
  - Semantic segmentation

- Contribution/Key Idea
  - Novel & Simple attention based feature distillation
  - Channel & Spatial attention
  - SOTA in 2 network

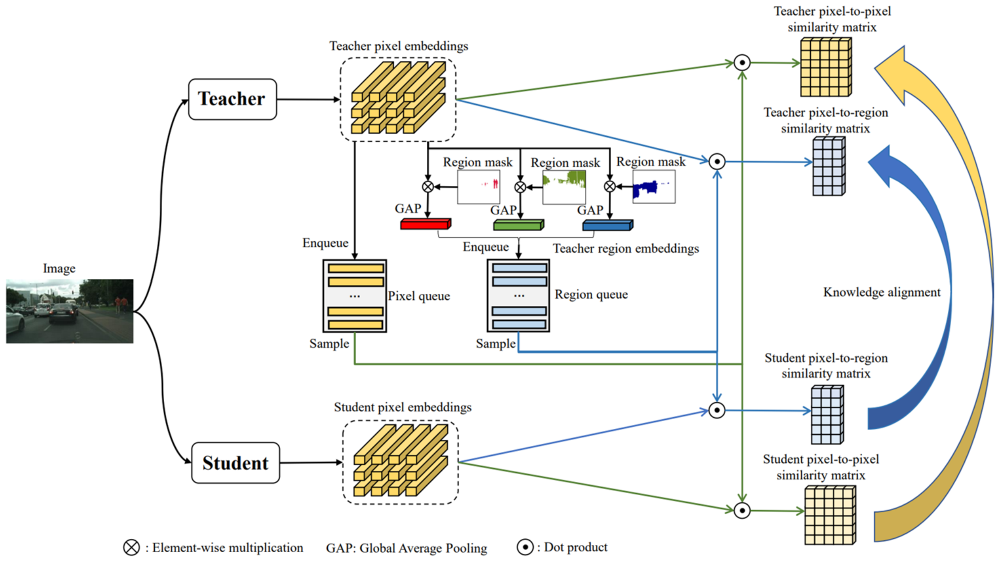
김범준

[Attention-guided Feature Distillation for Semantic Segmentation](#)
Amir M. Mansourian, Arya Jalali*, Rozhan Ahmadi, Shohreh Kasaei*

• Introduction



DistKD[1]



CIRKD[2]

- 최근 Knowledge distillation의 트렌드
  : 강력한 teacher, 멀티 teacher, 마스킹하여 distill, loss term 구성 복잡화 등

  → Computational Cost가 너무나도 늘어나서 Knowledge distillation의 <u>학습 시간 및 메모리가 거의 한계점</u>

[1] Huang, Tao, et al. "Knowledge distillation from a stronger teacher." *Advances in Neural Information Processing Systems* 35 (2022): 33716-33727.
[2] Yang, Chuanguang, et al. "Cross-image relational knowledge distillation for semantic segmentation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[Attention-guided Feature Distillation for Semantic Segmentation](#)
Amir M. Mansourian, Arya Jalali*, Rozhan Ahmadi, Shohreh Kasaei*

- **Method**

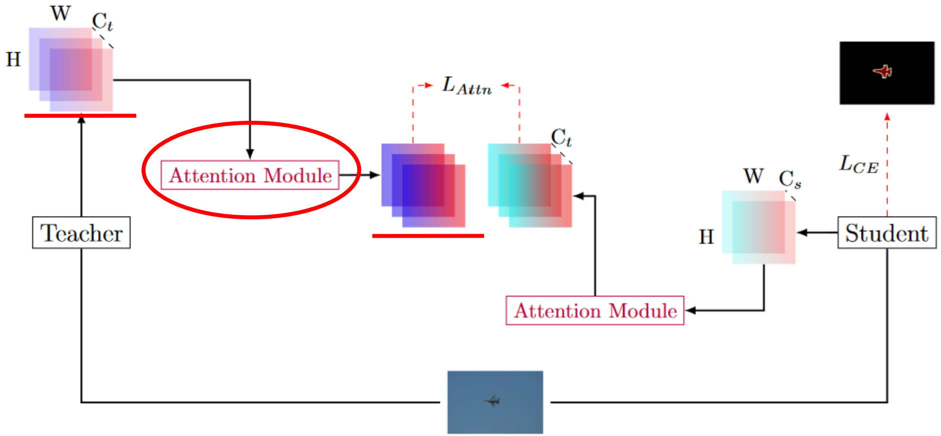


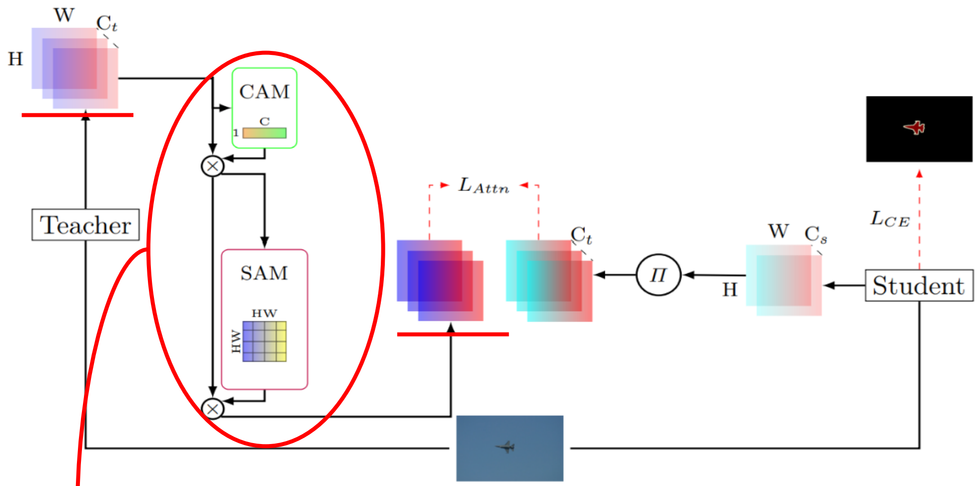
Figure 2: Proposed Attention-guided feature distillation.

- Novel + Simple Attention Module을 제안
→ Transformer 기반 attention x
→ Feature를 정제하는 module

[Attention-guided Feature Distillation for Semantic Segmentation](#)

Amir M. Mansourian, Arya Jalali*, Rozhan Ahmadi, Shohreh Kasaei*

- **Method**



$$F' = M_C(F) \otimes F$$

$$F'' = M_S(F') \otimes F'$$

$F \in \mathbb{R}^{c \times w \times h}$: feature map

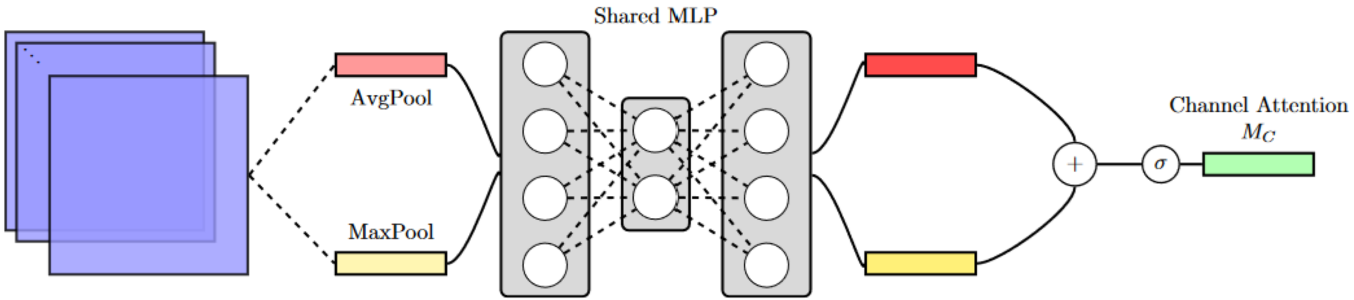
$M_C(F) \in \mathbb{R}^{c \times 1 \times 1}$: Channel Attention Module

$M_S(F) \in \mathbb{R}^{c \times h \times w}$: Spatial Attention Module

[Attention-guided Feature Distillation for Semantic Segmentation](#)

Amir M. Mansourian, Arya Jalali*, Rozhan Ahmadi, Shohreh Kasaei*

- **Channel Attention Module**

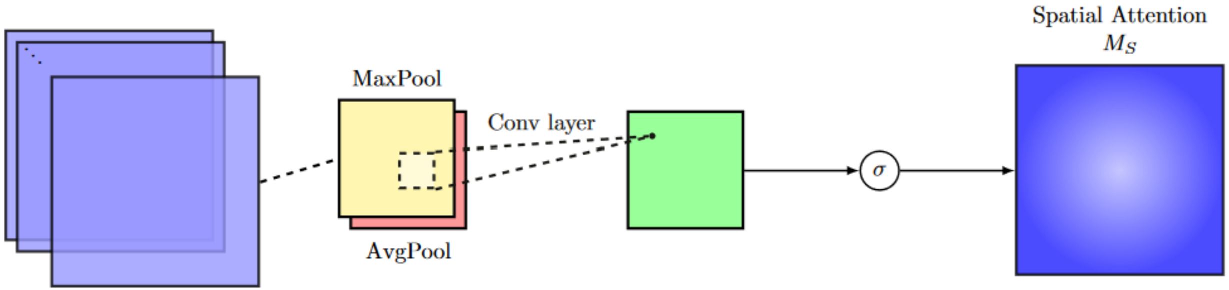


$$M_C(F) \in \mathbb{R}^{c\times1\times1}$$

$$M_C(F) = \sigma(W_1(W_0(F^C_{avg}))) + W_1(W_0(F^C_{max}))$$

- Feature map의 채널별 중요도를 계산
  → 중요한 영역(예: 객체)과 그렇지 않은 영역(예: 배경)을 구별, 배경을 나타내는 채널은 낮은 가중치

- **Spatial Attention Module**



$$M_S(F) \in \mathbb{R}^{c\times h\times w}$$

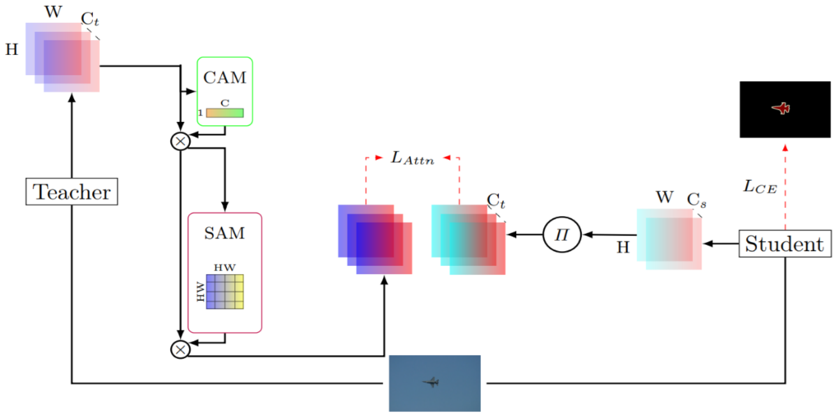
$$M_S(F) = \sigma(A^{7\times 7}([F^S_{avg}; F^S_{max}]))$$

- 공간 간 상관관계를 고려한 중요도를 계산

[Attention-guided Feature Distillation for Semantic Segmentation](#)

Amir M. Mansourian, Arya Jalali*, Rozhan Ahmadi, Shohreh Kasaei*

- **Method**



$$L_{Attn} = \frac{1}{N}\sum_{i=1}^{N}\left\| \frac{F''_{Sj}}{\| F''_{Sj} \|} - \frac{F''_{Tj}}{\| F''_{Tj} \|} \right\|$$

→ Feature map 크기 normalized

$$L_{AttnFD} = L_{CE} + \alpha L_{Attn}$$

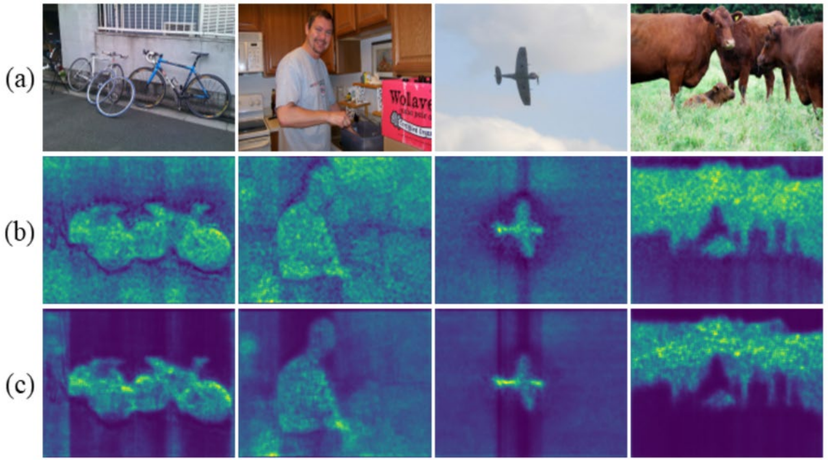


Figure 1: Visualization of images (a), raw feature maps (b), and refined feature maps (c). Channel and spatial attention is applied to raw features, emphasizing on the important regions and making them valuable distillation source.

Attention-guided Feature Distillation for Semantic Segmentation

Amir M. Mansourian, Arya Jalali*, Rozhan Ahmadi, Shohreh Kasaei*

- ## Experiment

Table 1: Quantitative results on PscalVoc Validation set.

Method	mIoU(%)	Params(M)
T: DeepLabV3-Res101	77.85	59.3
S: DeepLabV3-Res18	67.50	
S + KD	69.13 ± 0.11	
S + DistKD	69.84 ± 0.11	16.6
S + CIRKD	71.02 ± 0.11	
S+ LAD	71.42 ± 0.09	
S + AttnFD (ours)	**73.09 ± 0.06**	
S: DeepLabV3-MBV2	63.92	
S + KD	66.39 ± 0.21	
S + DistKD	67.62 ± 0.22	5.9
S + CIRKD	69.02 ± 0.16	
S + LAD	68.63 ± 0.07	
S + AttnFD (ours)	**70.38 ± 0.16**	
S: PSPNet-Res18	67.4	
S + KD	68.18 ± 0.08	
S + DistKD	68.93 ± 0.19	12.6
S + CIRKD	69.53 ± 0.11	
S + LAD	69.71 ± 0.10	
S + AttnFD (ours)	**70.95 ± 0.06**	

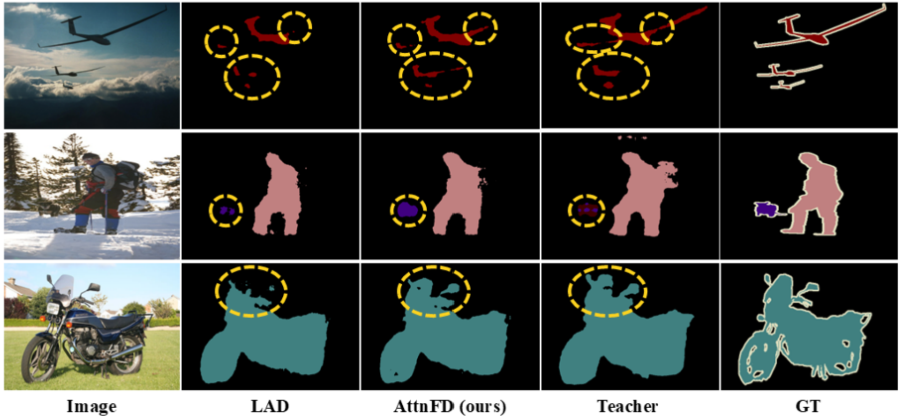


Figure 5: Some qualitative comparisons on the PascalVoc validation split.

김범준

[Attention-guided Feature Distillation for Semantic Segmentation](#)

Amir M. Mansourian, Arya Jalali*, Rozhan Ahmadi, Shohreh Kasaei*

- # Experiment

Table 2: Quantitative results on Cityscapes Validation set.

Method	mIoU(%)	Accuracy(%)
T: DeepLabV3-Res101	77.66	84.05
S: DeepLabV3-Res18	64.09	74.8
S + KD	65.21 (+1.12)	76.32 (+1.74)
S + CIRKD	70.49 (+6.40)	79.99 (+5.19)
S + DistKD	71.81 (+7.72)	80.73 (+5.93)
S + LAD	71.37 (+7.28)	80.93 (+6.13)
S + AttnFD (ours)	**73.04 (+8.95)**	**83.01 (+8.21)**
S: DeepLabV3-MBV2	63.05	73.38
S + KD	64.03 (+0.98)	75.34 (+1.96)
S + CIRKD	69.34 (+6.39)	78.66 (+5.28)
S + DistKD	69.53 (+6.48)	79.10 (+5.72)
S + LAD	69.84 (+6.79)	80.49 (+7.11)
S + AttnFD (ours)	**70.80 (+7.75)**	**81.59(+8.15)**
S: PSPNet-Res18	65.72	73.77
S + KD	66.89 (+1.17)	74.82 (+1.05)
S + CIRKD	67.51 (+1.79)	75.25 (+1.48)
S + DistKD	68.13 (+2.41)	76.25 (+2.48)
S + LAD	67.71 (+1.99)	75.63 (+1.86)
S + AttnFD (ours)	**68.86 (+3.14)**	**76.47 (+2.70)**

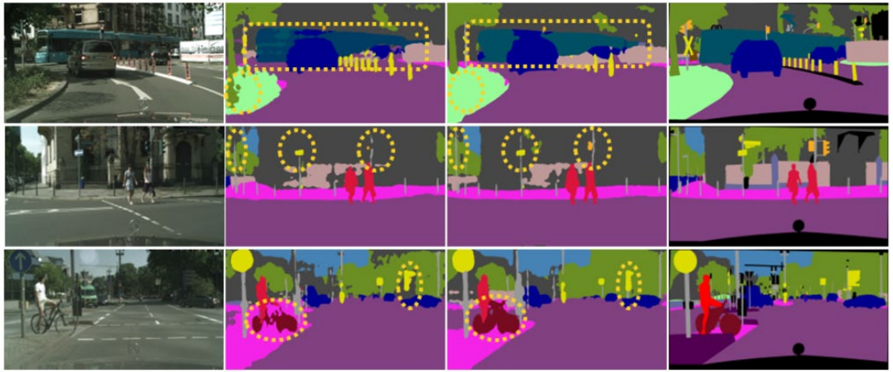


Image DistKD AttnFD (ours) GT

김범준

Amir M. Mansourian, Arya Jalali*, Rozhan Ahmadi, Shohreh Kasaei*

- **Experiment**

Table 3: Quantitative results on COCO Validation set.

Method	Params (M)	mIoU(%)
T: DeepLabV3-Res101	59.3	60.56
S: DeepLabV3-Res18	16.6	52.08
S + KD		54.6
S + CIRKD		55.60
S + DistKD		55.9
S + LAD		56.56
S + AttnFD (ours)		**57.74**
S: DeepLabV3-MBV2	5.9	47.92
S + KD		52.21
S + CIRKD		53.65
S + DistKD		53.33
S + LAD		55.29
S + AttnFD (ours)		**56.95**
S: PSPNet-Res18	12.6	52.68
S + KD		54.07
S + CIRKD		56.96
S + DistKD		55.06
S + LAD		57.50
S + AttnFD (ours)		**58.08**

Table 4: Quantitative results on CamaVid dataset.

Method	*Val* mIoU(%)	*Test* mIoU(%)
T: DeepLabV3-Res101	76.02	65.35
S: DeepLabV3-Res18	71.20	62.89
S + CIRKD	76.20	67.58
S + DistKD	75.36	68.32
S + LAD	76.13	66.57
S + AttnFD (ours)	**76.39**	**68.77**
S: PSPNet-Res18	72.64	63.02
S + CIRKD	73.89	65.03
S + DistKD	75.96	65.09
S + LAD	75.84	66.13
S + AttnFD (ours)	**76.56**	**66.74**

- # Experiment

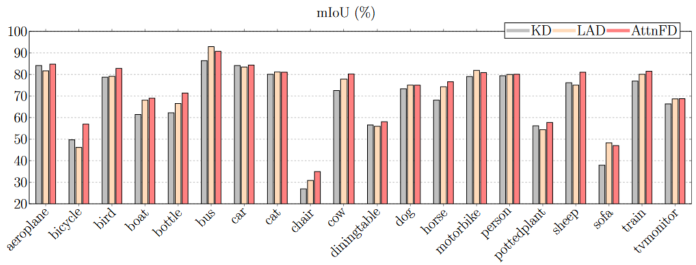


Figure 7: Visual representation of the performance of proposed method in terms of per-class mIoU using ResNet18 network on PascalVoc validation set.

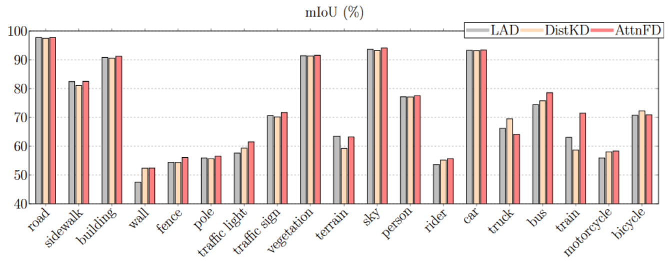


Figure 8: Comparison of mIoU per class among LAD, DistKD, and AttnFD on Cityscapes validation set, employing a ResNet18 backbone for the student network.

superior performance in classes like train (+12.91) and bus (+2.82). The top row of Figure 6 corroborates this, high-

- ## Experiment

Table 5: An ablation analysis conducted on PascalVOC validation set, examining the influence of distilling refined feature maps across various layers of the network.

Method	mIoU(%)	Accuracy(%)
T:DeepLabV3-Res101	77.85	-
S:DeepLabV3-Res18	67.50	76.49
S + B	70.25 (+2.75)	78.88 (+2.39)
S + E	72.31 (+4.81)	81.48 (+4.99)
S + D	72.47 (+4.97)	82.13 (+5.64)
s + B + E	72.58 (+5.08)	81.71 (+5.22)
S + B + D	72.82 (+5.32)	81.87 (+5.38)
S + E + D	72.92 (+5.42)	82.68 (+6.19)
S + B+ E + D	73.09 (+5.59)	82.95 (+6.46)
S:DeepLabV3-MBV2	63.92	73.98
S + B	66.68 (+2.76)	77.01 (+3.03)
S + E	68.91 (+4.99)	79.60 (+5.62)
S + D	69.55 (+5.63)	78.50 (+4.52)
s + B + E	69.17 (+5.25)	79.61 (+5.63)
S + B + D	69.46 (+5.54)	78.65 (+4.67)
S + E + D	69.96 (+6.04)	79.73 (+5.75)
S + B+ E + D	70.38 (+6.46)	81.13 (+7.21)

Table 6: Ablations for different attention modules.

Method	mIoU(%)	Params	Explanation
S: ResNet18	67.50	-	w/o attention
S + AT	68.95	-	Channel Aggregation w/o learning
S + SA	71.72	492800	Pairwise similarity of pixels
S + BAM	72.68	103235	Simultaneously channel & spatial attention
S + EMA	72.86	3986	Multi-scale attention by channel grouping
S + CBAM	73.09	50540	Channel and then spatial attention
S: MobileNet	63.92	-	w/o attention
S + AT	66.27	-	Channel Aggregation w/o learning
S + SA	68.29	292880	Pairwise similarity of pixels
S + BAM	69.96	61703	Simultaneously channel & spatial attention
S + EMA	70.05	2348	Multi-scale attention by channel grouping
S + CBAM	70.38	30368	Channel and then spatial attention