Rethinking Knowledge Distillation with Raw Features for Semantic Segmentation - Supplementary Material

Tao Liu*, Chenshu Chen*,†, Xi Yang, Wenming Tan
Hikvision Research Institute
Hangzhou, Zhejiang, China
{liutao46, chenchenshu, yangxi6, tanwenming}@hikvision.com

A. Naive feature distillation on VOC dataset

Let $\boldsymbol{F}^s \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{F}^t \in \mathbb{R}^{C\times H\times W}$ denote the feature maps of the student and teacher, respectively, where C is the number of channels, H and W are the height and width. We reformulate the loss function of the naive feature distillation [2] in Eq. (3) of the main paper. For the sake of description, we copy Eq. (3) from the main paper to Eq. (10) as follows:

$$\begin{aligned}\mathcal{L}_{naive} &= \frac{1}{N}\sum_{i=1}^{N}(m\boldsymbol{y}_i - n\boldsymbol{x}_i)^2 \\ &= \frac{1}{N}(m^2\sum_{i=1}^{N}\boldsymbol{y}_i^2 + n^2\sum_{i=1}^{N}\boldsymbol{x}_i^2 - 2mn\sum_{i=1}^{N}\boldsymbol{y}_i\boldsymbol{x}_i) \\ &= \frac{1}{N}(m^2||\boldsymbol{y}||^2 + n^2||\boldsymbol{x}||^2 - 2mn\,\boldsymbol{y}\cdot\boldsymbol{x}) \\ &= \frac{1}{N}(m^2 + n^2 - 2mn\cos\theta) \\ &= \frac{1}{N}[(m-n)^2 + 2mn(1-\cos\theta)]\end{aligned} \tag{10}$$

where $N = C\times H\times W$, n and m denote the magnitudes of $\boldsymbol{F}^s$ and $\boldsymbol{F}^t$, $\boldsymbol{x}$ and $\boldsymbol{y}$ are unit vectors, θ denotes the angle between $\boldsymbol{F}^s$ and $\boldsymbol{F}^t$. The first term in Eq. (10) minimizes the magnitude difference between $\boldsymbol{F}^s$ and $\boldsymbol{F}^t$, and the second term minimizes the angular difference between $\boldsymbol{F}^s$ and $\boldsymbol{F}^t$ but is affected by the magnitude.

We present the results of the naive feature distillation under varying loss weights on the Cityscapes dataset in Figs. 1 and 2 of the main paper. We also conducted the same experiments on the VOC dataset, as shown in Figs. 7 and 8. The same phenomenon as in the main paper can be observed:

(1) The naive feature distillation is sensitive to the loss weight. As shown in Fig. 7, when the teacher is PSPNet-R101 and the student is PSPNet-R18 or PSPNet-MV2, it requires a large loss weight (*e.g.*, 100 or 1000) to get good results. Instead, a relatively small loss weight (*e.g.*, 10) is sufficient when the teacher is UPerNet-SwinB and the student is UPerNet-SwinT.

(2) The magnitude difference term contributes little to the naive feature distillation, while it is the angular difference term that plays a crucial role. As shown in Fig. 7, Magnitude Distillation apparently fails to reach the performance of the naive feature distillation, even far worse than the baseline without distillation in some cases.

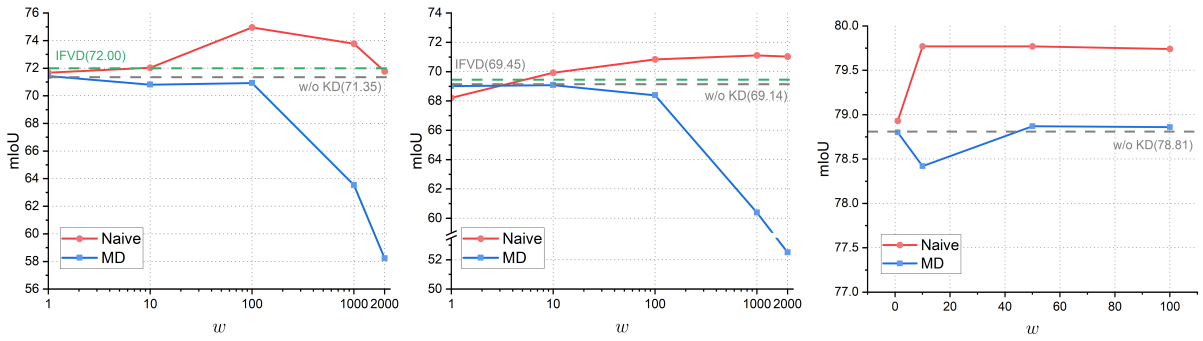
*Equal contribution
†Corresponding author

Teacher	Student	mAP	AP_S	AP_M	AP_L
RetinaNet -X101 (41.0)	RetinaNet-R50	37.4	20.6	40.7	49.7
	FKD [6]	39.6	22.7	43.3	52.5
	CWD [3]	40.8	22.7	44.5	55.3
	FGD [4]	40.7	22.9	45.0	54.7
	MGD [5]	41.0	23.4	45.3	55.7
	LAD (Ours)	41.0	23.3	45.2	55.1
	CAD (Ours)	41.1	22.9	45.2	55.2
Cascade Mask RCNN -X101 (47.3)	Faster RCNN-R50	38.4	21.5	42.1	50.3
	FKD [6]	41.5	23.5	45.0	55.3
	CWD [3]	41.7	23.3	45.5	55.5
	FGD [4]	42.0	23.8	46.4	55.5
	MGD [5]	42.1	23.7	46.4	56.1
	LAD (Ours)	41.8	23.6	45.7	55.6
	CAD (Ours)	41.8	23.6	45.6	55.9

Table 8. Comparison with state-of-the-art methods for object detection on COCO validation set. "X101" denotes ResNeXt101.

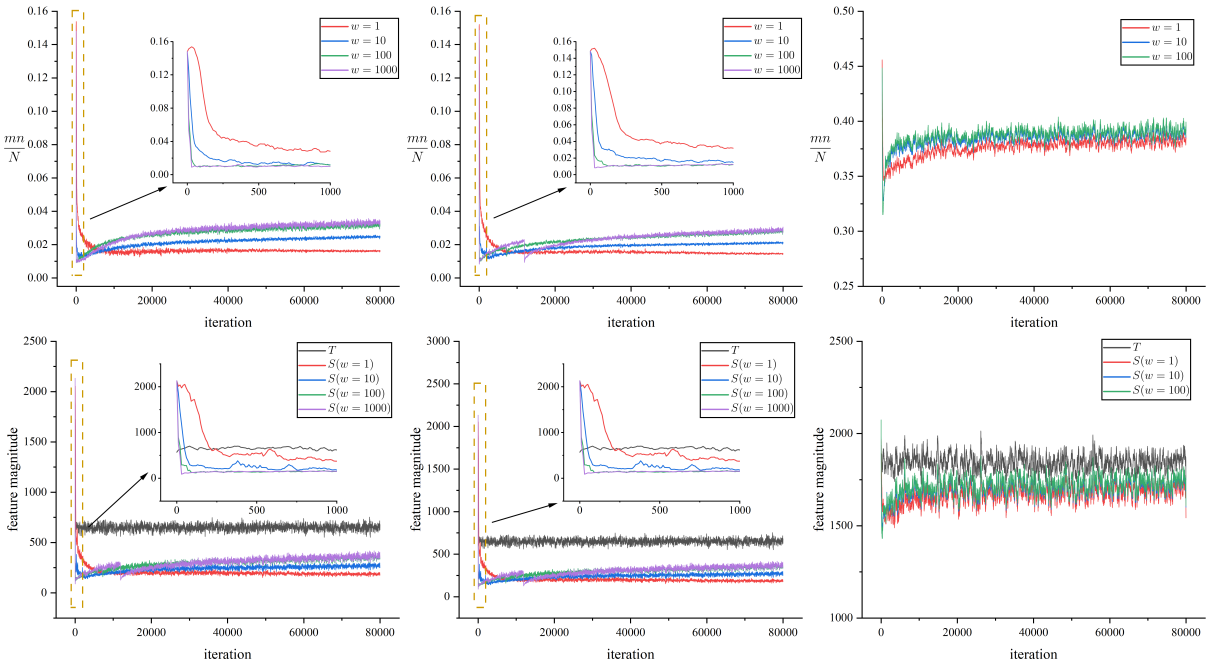
Teacher	Student	mAP	AP_S	AP_M	AP_L
Cascade Mask RCNN -X101 (41.1)	Mask RCNN-R50	35.4	16.6	38.2	52.5
	FGD [4]	37.8	17.1	40.7	56.0
	MGD [5]	38.1	17.1	41.1	56.3
	LAD (Ours)	37.6	16.9	40.4	56.1
	CAD (Ours)	37.8	17.4	40.7	55.8

Table 9. Comparison with state-of-the-art methods for instance segmentation on COCO validation set. "X101" denotes ResNeXt101. Here the AP means Mask AP.



(a) T:PSPNet-R101, S:PSPNet-R18 (b) T:PSPNet-R101, S:PSPNet-MV2 (c) T:UPerNet-SwinB, S:UPerNet-SwinT

Figure 7. Distillation results under varying loss weights on VOC validation set. The gray dashed line indicates the performance of the student model without distillation. T: Teacher. S: Student. Naive: Naive feature distillation. MD: Magnitude Distillation. w: Loss weight of $\mathcal{L}_{naive}$ (Eq. (1) in the main paper) or $\mathcal{L}_{md}$ (Eq. (4) in the main paper).



(a) T:PSPNet-R101, S:PSPNet-R18 (b) T:PSPNet-R101, S:PSPNet-MV2 (c) T:UPerNet-SwinB, S:UPerNet-SwinT

Figure 8. The values of $\frac{mn}{N}$, m, and n in Eq. (3) of the main paper during training for the naive feature distillation on VOC dataset. The first row records the values of $\frac{mn}{N}$. The second row records the magnitudes of the teacher features (m) and the student features (n). Note that the parameters of the teacher model are fixed, so the value of m for a sample is constant during training. T: Teacher. S: Student. w: Loss weight of $\mathcal{L}_{naive}$.

(3) Since the angular difference term is affected by the magnitude of the features, it is hard to determine a suitable loss weight for various models. As shown in Fig. 8, in the case where PSPNet-R101 is the teacher and PSPNet-R18 (Fig. 8a) or PSPNet-MV2 (Fig. 8b) is the student, the value of $\frac{mn}{N}$ is quite small because m and n have relatively small values. Therefore, a large loss weight is required to ensure that the weight of the angular distillation term has a reasonable value. As for the case where UPerNet-SwinB is the teacher and UPerNet-SwinT is the student (Fig. 8c), the value of $\frac{mn}{N}$ is clearly larger than that in Figs. 8a and 8b since both m and n have a relatively large value. As a result, the naive feature distillation can get good results (Fig. 7c) with a relatively small loss weight.

B. More results on other tasks

To verify the generality of our method, we conducted experiments on object detection and instance segmentation on

COCO2017 dataset [1]. Some of the results on object detection are shown in Tab. 7 of the main paper, and here we give more results on object detection and instance segmentation. Following [5], we calculate the distillation loss on all the feature maps from the neck. We train all the models for 24 epochs with SGD optimizer, where the momentum is 0.9 and the weight decay is 0.0001. It should be noted that some approaches (e.g., [5]) use carefully tuned hyper-parameters for different models. Instead, we set the weight of distillation loss (the only hyper-parameter of our method) to 3 for all models.

As shown in Tabs. 8 and 9, our method achieves competitive performance compared to state-of-the-art methods on object detection and instance segmentation. This indicates a promising generality of our method.

C. Training overhead of our method

It takes about 7 hours on 8 RTX3090 GPUs to train PSPNet-R18 for 80k iterations with PSPNet-R101 as the teacher and an input size of 512×1024 on Cityscapes. The memory footprint of each GPU is 5.9 GB when the total batch size is 16.

References

[1] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 3

[2] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[3] 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 (ICCV)*, pages 5311–5320, Oct. 2021. 1

[4] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4633–4642. IEEE, 2022. 1

[5] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 53–69, Cham, 2022. Springer Nature Switzerland. 1, 3

[6] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1