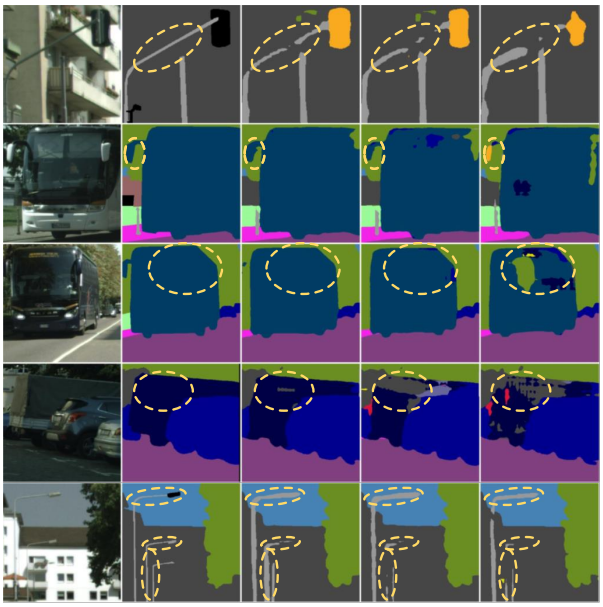
Supplementary Materials: Channel-wise Knowledge Distillation for Dense Prediction

S1. Results with feature map on Cityscapes



(a) Image (b) GT (c) CD (d) AT (e) Student

Figure 1. **Qualitative segmentation results** on Cityscapes of the PSPNet-R18 model: (a) raw images, (b) ground truth (GT), (c) channel-wise distillation (CD), (d) the best spatial distillation schemes: attention transfer (AT); and (e) the output of the original student model without KD.

S2. Results on Pascal VOC and ADE20K

To further demonstrate the effectiveness of the proposed channel distribution distillation, we only employ the proposed CD on the feature maps as our final results on Pascal VOC and ADE20K. The experiment results are reported in Table 2 and Table 3. Multi student-network variants with different encoders and decoders are used to validate the efficiency of our method. Here, encoders include ResNet18 and MobileNetV2, and decoders include PSP-head and ASPP-head.

Pascal VOC. We evaluate the performance of our method on the Pascal VOC dataset. The distillation results are listed in Table 2. Our proposed CD improves PSPNet-R18 without distillation by 3.83%, outperforms the SKDS and IFVD by 1.51% and 1.21%. Consistent improvements on other student networks with different encoders and decoders are achieved. The gains on PSPNet-MBV2 with our method is 3.55%, surpassing the SKDS and IFVD by 1.98% and 1.20%. As for Deeplab-R18, our CD improves the student from 66.81% to 69.97%, outperforming the SKDS and IFVD by 1.84% and 1.55% respectively. Besides, the performance of Deeplab-MBV2 with our distillation is increased from 50.80% to 54.62%, outperforming the SKDS and IFVD by 2.51% and 1.23% respectively.

ADE20K. We also evaluate our method on the ADE20K dataset to further demonstrate that CD works better than other structural knowledge distillation methods. The results are shown in Table 3. Our proposed CD improves PSPNet-R18 without distillation by 3.83%, and outperforms the SKDS and IFVD by 1.51% and 1.21% in several. Notable performance gains on other student with different encoders and decoders are also consistently achieved, As for PSPNet-MBV2, our method achieves a superior performance of 27.97%, surpassing the student, SKDS and IFVD by 4.82%, 3.18% and 2.64%. The gain on Deeplab-R18 with our CD is 2.48%, outperforming the SKDS and IFVD by 1.85% and 0.84%. Finally, the performance of Deeplab-MBV2 with our channel-wise distillation is increased from 24.98% to 29.18%, outperforming the SKDS and IFVD by 3.08% and 1.93% respectively.

S3. More visualization results

We list the visualization results in Figure 2 to intuitively demonstrate that, the channel distribution distillation method (CD) outperforms the spatial distillation strategy (attention transfer). Besides, to evaluate the effectiveness of the proposed channel distribution distillation, we visualize the channel distribution of the student network under three paradigms, *i.e.*, original network, distilled by the attention transfer (AT) and channel distribution distillation respectively, in Figure 3 and Figure 4. We also present the visualization results in Figure **??** to intuitively demonstrate that, the channel distillation method (CD) outperforms the spatial distillation strategy.

Method	Params (M)	FLOPs (G)	mIoU (%)	
			Val	Test
ENet [1]	0.358	3.612	–	58.3
ESPNet [28]	0.363	4.422	–	60.3
ERFNet [8]	2.067	25.60	–	68.0
ICNet [44]	26.50	28.30	–	69.5
FCN [18]	134.5	333.9	–	62.7
RefineNet [21]	118.1	525.7	–	73.6
OCNet [38]	62.58	548.5	–	80.1
Results w/ and w/o distillation schemes				
T:PSPNet [45]	70.43	574.9	78.5	78.4
S:PSPNet-R18$^{\diamond}$(0.5)	3.835	31.53	55.40	54.10
+SKDS [23]	3.835	31.53	61.60	60.50
+SKDD [24]	3.835	31.53	62.35	–
+IFVD [33]	3.835	31.53	63.35	63.68
+Ours	3.835	31.53	67.26	67.33
S:PSPNet-R18$^{\diamond}$	13.07	125.8	57.50	56.00
+SKDS [23]	13.07	125.8	63.20	62.10
+SKDD [24]	13.07	125.8	64.68	–
+IFVD [33]	13.07	125.8	66.63	65.72
Ours	13.07	125.8	70.04	70.11
S:PSPNet-R18*	13.07	125.8	69.72	67.60
+SKDS [23]	13.07	125.8	72.70	71.40
+SKDD [24]	13.07	125.8	74.08	–
+IFVD [33]	13.07	125.8	74.54	72.74
+Ours	13.07	125.8	74.87	73.86
S:PSPNet-MBV2*	1.98	16.40	58.64	57.43
+SKDS [23]	1.98	16.40	61.12	60.36
+IFVD [33]	1.98	16.40	62.74	61.92
+Ours	1.98	16.40	64.37	63.12
S:Deeplab-R18$^{\diamond}$(0.5)	3.15	31.06	61.83	60.51
+SKDS [23]	3.15	31.06	62.71	61.69
+IFVD [33]	3.15	31.06	63.12	62.37
+Ours	3.15	31.06	65.60	64.33
S:Deeplab-R18*	12.62	123.9	73.37	72.39
+SKDS [23]	12.62	123.9	73.87	72.63
+IFVD [33]	12.62	123.9	74.09	72.97
+Ours	12.62	123.9	75.25	74.12
S:Deeplab-MBV2*	2.45	20.39	65.94	65.07
+SKDS [23]	2.45	20.39	66.73	65.81
+IFVD [33]	2.45	20.39	67.04	66.12
+Ours	2.45	20.39	67.92	66.87

Table 1. Comparison of student variants with the state-of-the-art distillation methods on Cityscapes, where $^{\diamond}$ denotes to be trained from scratch and * indicates to be initialized by the weights pre-trained on ImageNet, and R18 (MBV2) is the abbreviation for Resnet18 (MobileNetV2).

Method	Params	mIoU(%)	mAcc(%)
FCN [18]	134.5	69.9	78.1
DeepLabV3 [5]	87.1	77.9	85.7
PSANet [**?**]	78.13	77.9	86.6
GCNet [**?**]	68.82	77.8	85.9
ANN [**?**]	65.2	76.7	84.5
OCRNet [**?**]	70.37	80.3	87.1
Results w/ and w/o our distillation schemes			
T:PSPNet [45]	70.43	78.52	79.57
S:PSPNet-R18	13.07	65.42	80.43
+SKDS [23]	13.07	67.73	81.73
+IFDV [33]	13.07	68.04	82.25
+Ours	13.07	69.25	83.14
S:PSPNet-MBV2	1.98	62.38	77.82
+SKDS [23]	1.98	63.95	78.93
+IFDV [33]	1.98	64.73	79.81
+Ours	1.98	65.93	81.45
S:Deeplab-R18	12.62	66.81	81.14
+SKDS [23]	12.62	68.13	82.26
+IFDV [33]	12.62	68.42	82.70
+Ours	12.62	69.97	83.47
S:Deeplab-MBV2	2.45	50.80	74.24
+SKDS [23]	2.45	52.11	75.17
+IFDV [33]	2.45	53.39	76.02
+Ours	2.45	54.62	77.13

Table 2. mIoU and mAcc on validation set of VOC 2012, R18 (MBV2) is the abbreviation for Resnet18 (MobileNetV2).

References

[1] Paszke Adam, Chaurasia Abhishek, Kim Sangpil, and Culurciello Eugenio. Enet: A deep neural network architecture for real-time semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[2] Romero Adriana, Ballas Nicolas, Ebrahimi Kahou Samira, Chassang Antoine, Gatta Carlo, and Bengio Yoshua. Fitnets: Hints for thin deep nets. *Int. Conf. Learn. Represent.*, 2015.

[3] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[4] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Adv. Neural Inform. Process. Syst.*, pages 742–751, 2017.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[6] Wuyang Chen, Xinyu Gong, Xianming Liu, Qian Zhang, Yuan Li, and Zhangyang Wang. Fasterseg: Searching for faster real-time semantic segmentation. *Int. Conf. Learn. Represent.*, 2020.

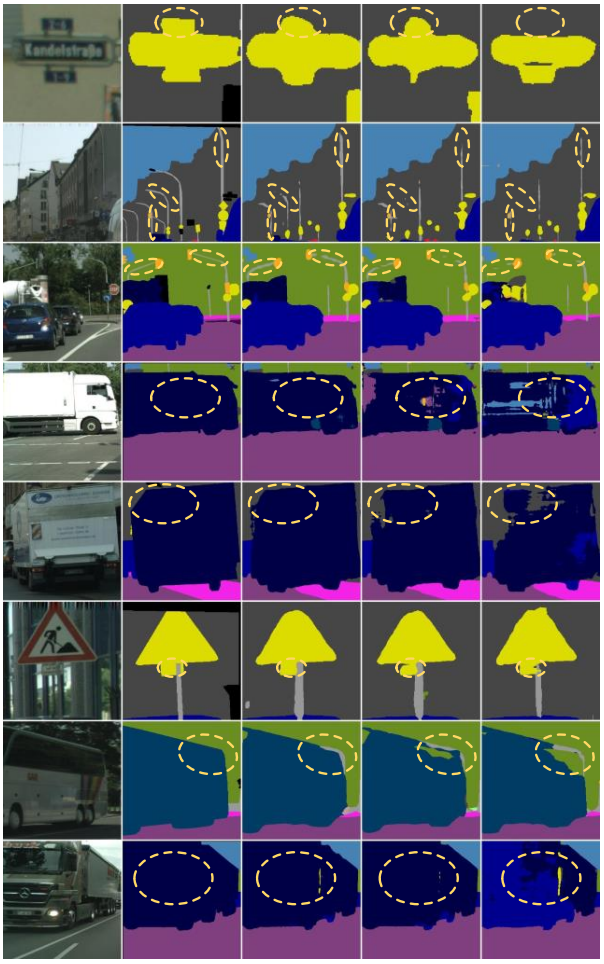
[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[8] Romera Eduardo, Álvarez José M, Bergasa Luis M, and Arroyo Roberto. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Trans. Intell. Transportation Syst.*, 2017.

[9] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 2010.

Method	Params	mIoU(%)	mAcc(%)
FCN [18]	134.5	39.91	49.62
DeepLabV3 [5]	87.1	44.99	55.81
PSANet [**?**]	78.13	43.74	54.09
GCNet [**?**]	68.82	43.68	54.28
ANN [**?**]	65.2	42.93	53.25
OCRNet [**?**]	70.37	43.70	53.74
Results w/ and w/o our distillation schemes			
T:PSPNet [45]	70.43	44.39	45.35
S:PSPNet-R18	13.07	24.65	33.66
+SKDS [23]	13.07	25.11	33.72
+IFDV [33]	13.07	25.72	33.83
+Ours	13.07	26.80	34.02
S:PSPNet-MBV2	1.98	23.15	32.93
+SKDS [23]	1.98	24.79	34.04
+IFDV [33]	1.98	25.33	35.57
+Ours	1.98	27.97	37.16
S:Deeplab-R18	12.62	24.89	33.60
+SKDS [23]	12.62	25.52	34.10
+IFDV [33]	12.62	26.53	34.79
+Ours	12.62	27.37	35.34
S:Deeplab-MBV2	2.45	24.98	35.34
+SKDS [23]	2.45	26.10	36.51
+IFDV [33]	2.45	27.25	37.23
+Ours	2.45	29.18	38.08

Table 3. mIoU and mAcc on validation set of ADE20K, R18 (MBV2) is the abbreviation for Resnet18 (MobileNetV2).



(a) Image (b) GT (c) CD (d) AT (e) Student

Figure 2. Qualitative segmentation results on Cityscapes produced from PSPNet-R18: (a) raw images, (b) ground truth (GT), (c) channel-wise distillation (CW), (d) the spatial distillation schemes: attention transfer (AT), and (e) output of the original student model.

[10] Jie Fu, Xue Geng, Zhijian Duan, Bohan Zhuang, Xingdi Yuan, Adam Trischler, Jie Lin, Chris Pal, and Hao Dong. Role-wise data augmentation for knowledge distillation. *Int. Conf. Learn. Represent.*, 2020.

[11] Yushuo Guan, Pengyu Zhao, Bingxuan Wang, Yuanxing Zhang, Cong Yao, Kaigui Bian, and Jian Tang. Differentiable feature aggregation search for knowledge distillation. In *Eur. Conf. Comput. Vis.*, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[13] Tong He, Chunhua Shen, Tian Zhi, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[14] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and JinYoung. Choi. A comprehensive overhaul of feature distillation. In *Int. Conf. Comput. Vis.*, pages 1921–19302, 2019.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Adv. Neural Inform. Process. Syst.*, 2014.

[16] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv:*, abs/1503.02531, 2015.

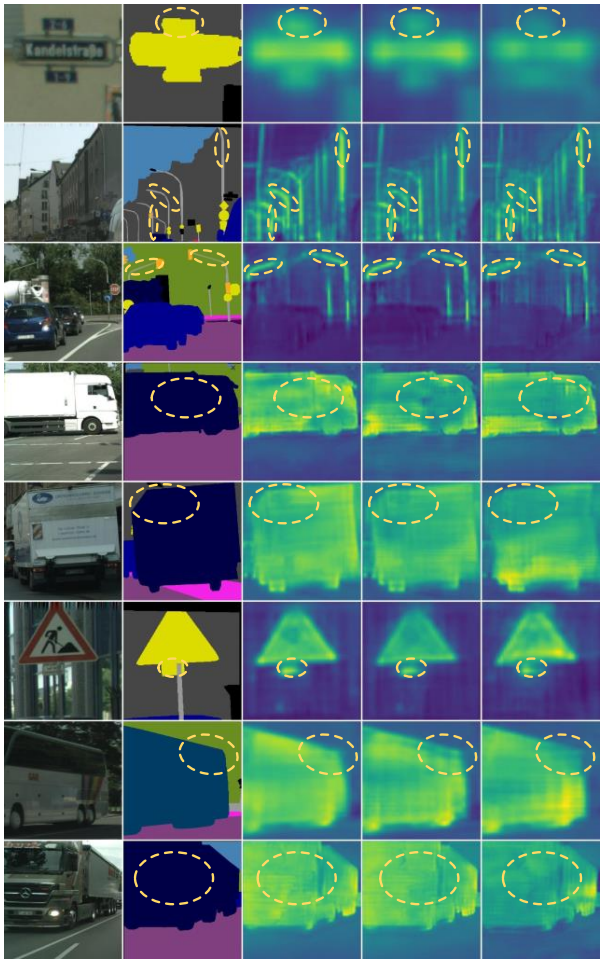
[17] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-region affinity distillation for road marking segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12486–12495, 2020.

[18] Long Jonathan, Shelhamer Evan, and Darrell Trevor. Fully convolutional networks for semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[19] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

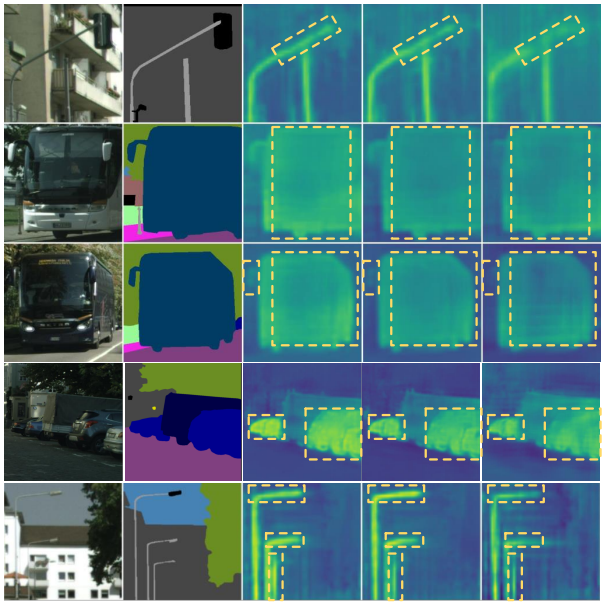
[20] Lin, Goyal Tsung-Yi, Girshick Priya, He Ross, Dollár Kaiming, and Piotr. Focal loss for dense object detection. In *Int. Conf. Comput. Vis.*, 2017.

[21] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.



(a) Image (b) GT (c) CD (d) AT (e) Student

Figure 3. The channel distribution of the student under three paradigms. (a) raw images, (b) ground truth (GT), (c) channel distillation, (d) the spatial distillation schemes: attention transfer (AT), and (e) output of the original student model.



(a) Image (b) GT (c) CD (d) AT (e) Student

Figure 4. The channel distribution of the student under three paradigms. The yellow dotted lines show the activation maps of CD are better than that in AT and the student network.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014.

[23] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[24] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[25] Sangyong Park and Yong Seok Heo. Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy. *Sensors*, 20(16):4616, 2020.

[26] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3967–3976, 2019.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2016.

[28] Mehta Sachin, Rastegari Mohammad, Caspi Anat, Shapiro Linda, and Hajishirzi Hannaneh. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. *Eur. Conf. Comput. Vis.*, 2018.

[29] Tian, Shen Zhi, Chen Chunhua, He Hao, and Tong. Fcos: Fully convolutional one-stage object detection. In *Int. Conf. Comput. Vis.*, 2019.

[30] Guo-Hua Wang, Yifan Ge, and Jianxin Wu. In defense of feature mimicking for knowledge distillation. *arXiv preprint arXiv:2011.01424*, 2020.

[31] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi. Feng. Distilling object detectors with fine-grained feature imitation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4933–4942, 2019.

[32] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Adv. Neural Inform. Process. Syst.*, 33, 2020.

[33] Yukang Wang, Zhou Wei, Jiang Tao, Bai Xiang, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. *Eur. Conf. Comput. Vis.*, 2020.

[34] Jiafeng Xie, Bing Shuai, JianFang Hu, Jingyang Lin, and WeiShi Zheng. Improving fast segmentation with teacher-student learning. *Brit. Mach. Vis. Conf.*, 2018.

[35] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3967–3976, 2019.

[36] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Int. Conf. Comput. Vis.*, pages 9657–9666, 2019.

[37] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3903–3911, 2020.

[38] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[39] Kaiyu Yue, Jiangfan Deng, and Feng Zhou. Matching guided distillation. *Eur. Conf. Comput. Vis.*, 2020.

[40] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *Int. Conf. Learn. Represent.*, 2017.

[41] Linfeng Zhang and Kaisheng. Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *Int. Conf. Learn. Represent.*, 2021.

[42] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[43] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[44] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. *Eur. Conf. Comput. Vis.*, 2018.

[45] Hengshuang Zhao1, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[46] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio. Torralba. Scene parsing through ade20k dataset. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[47] Zaida Zhou, Zhuge Chaoran, Xinwei Guan, and Wen Liu. Channel distillation: Channel-wise attention for knowledge distillation. *arXiv:*, abs/2006.01683, 2020.