Supplementary Material for Rethinking Local and Global Feature Representation for Semantic Segmentation

Mohan Chen[1]
mhchen19@fudan.edu.cn
Xinxuan Zhao[1]
zhaoxx19@fudan.edu.cn
Bingfei Fu[3]
bffu18@fudan.edu.cn
Li Zhang*[2]
lizhangfd@fudan.edu.cn
Xiangyang Xue[3]
xyxue@fudan.edu.cn

[1] Academy for Engineering & Technology, Fudan University
[2] School of Data Science, Fudan University
[3] School of Computer Science, Fudan University

1 Per-class results on Cityscapes

Table 1 shows the per-class results on Cityscapes test set of our DSCT and alternatives. Our model can achieve higher accuracy and better parsing of the images.

2 Per-class results on ADE20K

Our final model DSCT-SAPP and DSCT-SAC achieve 48.18% mIoU and 48.66% mIoU on ADE20K [10] validation set with single scale testing. Table 2 shows the per-class results on ADE20K validation set of our DSCT and alternatives. We can observe that our model achieves better performance on most of the categories.

Methods	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU
FCN [5]	97.4	78.4	89.2	34.9	44.2	47.4	60.1	65.0	91.4	69.3	93.9	77.1	51.4	92.6	35.3	48.6	46.5	51.6	66.8	65.3
BFPNet [2]	98.7	87.1	93.5	59.8	63.4	68.9	76.8	80.9	93.7	72.8	95.5	87.0	72.1	96.0	77.6	89.0	**86.9**	69.2	77.6	81.4
PSPNet [8]	98.6	86.2	92.9	50.8	58.8	64.0	75.6	79.0	93.4	72.3	95.4	86.5	71.3	95.9	68.2	79.5	73.8	69.5	77.2	78.4
DenseASPP [6]	98.7	87.1	93.4	60.7	62.7	65.6	74.6	78.5	93.6	72.5	95.4	86.2	71.9	96.0	78.0	90.3	80.7	69.7	76.8	80.6
DANet [3]	98.6	87.1	93.5	56.1	63.3	69.7	77.3	81.3	93.9	72.9	95.7	87.3	72.9	96.2	76.8	89.4	86.5	72.2	78.2	81.5
DGCNet [7]	98.7	87.4	93.9	**62.4**	63.4	70.8	**78.7**	81.3	94.0	73.3	95.8	87.8	73.7	**96.4**	76.0	**91.6**	81.6	71.5	78.2	82.0
DSCT (SAPP)	**98.8**	**87.7**	94.0	57.5	**64.4**	**71.5**	78.5	**81.9**	**94.1**	73.4	**96.1**	**88.3**	**75.0**	96.3	76.6	90.9	84.9	**74.0**	**79.1**	82.3
DSCT (SAC)	98.7	87.3	**94.1**	56.8	63.3	71.1	**78.7**	**81.9**	93.8	**74.1**	96.0	88.1	73.9	96.2	**81.3**	91.4	86.4	73.7	78.9	**82.4**

Table 1: Per-class results on Cityscapes validation set. Our methods can achieve higher accuracy and better parsing of the images.

Class	FCN	PVT	DSCT (SAPP)	DSCT (SAC)
wall	73.26	**76.44**	76.26	74.97
building	78.73	**81.78**	80.77	79.26
sky	93.49	94.10	**94.34**	93.76
floor	76.53	**80.70**	79.80	78.48
tree	72.01	73.00	**73.32**	72.30
ceiling	80.97	**83.13**	82.98	82.27
road	78.02	**84.79**	82.36	83.01
bed	86.28	**88.57**	86.81	86.97
windowpane	58.51	**61.17**	60.87	60.21
grass	64.02	68.12	**68.34**	64.40
cabinet	55.66	**62.21**	61.48	58.19
sidewalk	61.52	**67.40**	63.01	66.78
person	75.60	79.44	**80.39**	76.64
earth	29.87	**38.06**	37.24	31.53
door	43.02	46.36	**46.56**	45.82
table	57.59	59.28	**61.84**	58.74
mountain	50.93	**53.85**	51.61	48.18
plant	49.56	50.65	**51.20**	49.29
curtain	66.98	72.74	**73.29**	67.45
chair	53.13	59.33	**59.97**	56.40
car	79.93	84.62	**85.35**	82.11
water	46.56	**57.11**	53.82	50.14
painting	69.60	74.54	**75.20**	68.70
sofa	63.58	**65.71**	61.71	60.16
shelf	35.25	44.87	**47.31**	41.00
house	31.17	**54.68**	49.91	28.06
sea	53.10	**56.38**	50.90	56.22
mirror	60.39	**65.73**	64.92	62.70
rug	48.55	58.56	**59.86**	54.95
field	19.92	**33.20**	24.83	20.98
armchair	37.87	42.87	**48.63**	37.31
seat	41.82	56.54	**61.92**	56.50
fence	33.35	39.53	**44.57**	40.16
desk	40.55	47.61	**48.49**	47.57
rock	32.38	**35.81**	34.22	30.06
wardrobe	27.68	**50.13**	49.43	41.54
lamp	61.61	**64.85**	64.23	60.29
bathtub	70.13	79.05	**79.94**	74.67
railing	29.05	28.84	**34.91**	30.09
cushion	51.54	**59.28**	56.54	51.96
base	12.22	30.49	**28.58**	22.53
box	20.44	**28.01**	27.51	21.44
column	38.14	35.61	42.92	**46.51**
signboard	32.53	**38.17**	36.90	32.07
chest of drawers	33.57	41.39	**43.52**	34.20
counter	16.22	22.94	**35.31**	26.20

Class	FCN	PVT	DSCT (SAPP)	DSCT (SAC)
sand	25.09	32.75	**38.11**	33.88
sink	65.29	**72.75**	70.08	66.50
skyscraper	25.22	47.48	**65.88**	29.89
fireplace	67.18	**70.12**	67.36	68.19
refrigerator	68.40	73.56	**76.84**	71.18
grandstand	11.22	34.03	36.66	**43.75**
path	17.05	**24.56**	11.83	20.61
stairs	31.37	27.53	31.56	**32.83**
runway	10.33	65.15	**70.21**	61.52
case	17.12	40.95	**57.97**	43.79
pool table	90.91	93.12	**93.37**	90.99
pillow	51.21	54.52	**59.67**	51.23
screen door	50.91	67.41	67.26	**68.62**
stairway	19.46	31.23	**35.89**	23.56
river	9.55	13.95	**22.39**	14.99
bridge	**53.29**	32.32	36.39	43.24
bookcase	32.35	**40.99**	37.90	34.37
blind	**46.26**	43.72	38.44	34.10
coffee table	61.67	53.94	**65.00**	58.12
toilet	82.06	**86.10**	85.31	85.11
flower	37.15	40.09	**43.72**	33.42
book	41.89	**49.89**	46.92	43.24
hill	5.15	4.42	7.17	**8.08**
bench	43.07	41.48	**46.09**	43.36
countertop	52.54	**65.92**	57.96	52.60
stove	67.68	73.80	**77.15**	70.90
palm	45.47	49.56	**50.76**	45.79
kitchen island	37.01	**45.62**	41.16	29.63
computer	54.56	64.46	**70.99**	53.95
swivel chair	38.69	39.64	**46.72**	39.64
boat	32.77	**64.87**	36.62	37.05
bar	12.09	29.63	30.19	**31.67**
arcade machine	5.71	54.28	**67.63**	17.60
hovel	27.47	35.29	**38.71**	22.97
bus	58.74	82.29	**88.23**	85.57
towel	48.73	60.99	**63.55**	48.67
light	50.71	**56.36**	51.16	49.25
truck	24.11	25.13	30.08	**33.94**
tower	**19.78**	18.63	8.39	12.27
chandelier	64.31	64.41	**65.86**	64.86
awning	19.66	25.12	**29.27**	21.48
streetlight	22.27	26.08	**27.35**	21.83
booth	22.57	**57.54**	48.25	54.41
television receiver	59.61	**69.71**	67.84	66.98
airplane	50.69	**61.51**	55.68	55.00
dirt track	6.05	4.65	6.33	**19.17**
apparel	23.84	25.96	**29.62**	23.60

Class	FCN	PVT	DSCT (SAPP)	DSCT (SAC)
pole	19.09	27.31	**25.79**	22.81
land	**3.99**	1.26	3.66	3.74
bannister	12.58	10.14	**15.60**	6.82
escalator	4.51	17.40	**25.01**	14.91
ottoman	45.69	**48.76**	48.74	41.15
bottle	18.41	21.66	**28.99**	19.20
buffet	32.15	**48.48**	36.57	38.04
poster	24.89	22.80	**25.30**	24.00
stage	13.29	16.03	12.47	**19.90**
van	32.20	**38.09**	37.33	43.88
ship	6.03	**54.72**	46.89	6.81
fountain	8.63	**21.41**	21.06	1.91
conveyer belt	57.22	**75.07**	73.11	60.28
canopy	19.06	**34.19**	12.65	14.09
washer	48.63	72.78	**79.69**	63.26
plaything	27.41	**34.67**	28.31	24.46
swimming pool	28.34	**66.61**	64.12	40.90
stool	34.69	43.32	**45.24**	41.20
barrel	18.34	**59.57**	55.68	37.78
basket	28.95	**30.28**	30.07	28.71
waterfall	21.82	60.83	**67.24**	54.50
tent	88.35	86.45	**95.71**	93.21
bag	10.65	8.84	10.78	**11.10**
minibike	49.71	66.11	**74.95**	60.74
cradle	74.33	**77.39**	67.03	75.97
oven	42.32	38.52	**44.15**	23.57
ball	38.86	37.75	**50.20**	38.52
food	38.43	44.62	46.47	**53.49**
step	15.46	**18.79**	13.89	16.99
tank	28.57	**45.09**	38.00	29.84
trade name	23.21	22.91	24.02	**24.09**
microwave	39.57	39.59	**43.62**	38.34
pot	38.05	35.76	39.72	**41.63**
animal	53.56	48.19	**59.90**	50.73
bicycle	46.41	50.43	**55.27**	48.98
lake	19.03	**61.20**	59.77	49.48
dishwasher	72.25	62.49	65.75	**74.30**
screen	**65.52**	62.63	58.96	62.49
blanket	10.01	13.31	**15.33**	8.59
sculpture	22.10	**57.15**	50.94	37.53
hood	61.07	**64.14**	63.21	55.11
sconce	40.79	46.08	**47.35**	41.53
vase	36.98	**43.08**	41.58	36.49
traffic light	27.31	**35.95**	33.83	32.32
tray	6.53	10.28	**15.59**	8.01
ashcan	39.26	37.21	**43.87**	40.99
fan	58.53	63.78	**65.42**	55.82

Class	FCN	PVT	DSCT (SAPP)	DSCT (SAC)
pier	19.26	32.48	33.81	**49.47**
crt screen	2.61	**3.87**	2.73	0.55
plate	44.99	56.61	**56.68**	44.99
monitor	15.27	**25.06**	20.94	9.69
bulletin board	30.71	**37.89**	25.96	35.71
shower	0.85	**4.94**	3.00	0.00
radiator	50.43	**61.64**	58.01	46.51
glass	11.36	15.85	**16.60**	11.31
clock	27.67	**47.72**	43.58	24.85
flag	30.11	36.27	**44.00**	38.38
mIoU	39.91	43.49	48.18	**48.66**

Table 2: Per-class results on ADE20K validation set. Our method can achieve higher accuracy at most of the categories.

3 Detailed architecture of DSCT encoder

Detailed architecture of DSCT encoder is shown in Table 3. We have feature fusion to exchange information in such a flow: *conv residual –> transformer –> another conv residual*. We thus have to maintain the number of residual block twice of the Transformer block. For example, there are 8 residual blocks and 4 Transformer blocks in stage 3.

	Tiny				Small				Base			
stage	Spatial Stream	Fusion	Context Stream		Spatial Stream	Fusion	Context Stream		Spatial Stream	Fusion	Context Stream	
1	7×7, 64, stride 2				7×7, 64, stride 2				7×7, 64, stride 2			
	3×3 max pooling, stride 2				3×3 max pooling, stride 2				3×3 max pooling, stride 2			
	$\begin{bmatrix} 1\times1, 16 \\ 3\times3, 16 \\ 1\times1, 64 \end{bmatrix}$	-	4×4, 384, stride 4 $\begin{bmatrix} 384\text{-d} \\ 6\text{-h} \end{bmatrix}$	×1	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$	-	4×4, 384, stride 4 $\begin{bmatrix} 384\text{-d} \\ 6\text{-h} \end{bmatrix}$	×1	$\begin{bmatrix} 1\times1, 96 \\ 3\times3, 96 \\ 1\times1, 384 \end{bmatrix}$	-	4×4, 384, stride 4 $\begin{bmatrix} 576\text{-d} \\ 9\text{-h} \end{bmatrix}$	×1
2	$\begin{bmatrix} 1\times1, 16 \\ 3\times3, 16 \\ 1\times1, 64 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 16 \\ 3\times3, 16 \\ 1\times1, 64 \end{bmatrix}$	$[1\times1, 384] \longrightarrow$ $\longleftarrow [1\times1, 16]$	$\begin{bmatrix} 384\text{-d} \\ 6\text{-h} \end{bmatrix}$	×3	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$	$[1\times1, 384] \longrightarrow$ $\longleftarrow [1\times1, 64]$	$\begin{bmatrix} 384\text{-d} \\ 6\text{-h} \end{bmatrix}$	×3	$\begin{bmatrix} 1\times1, 96 \\ 3\times3, 96 \\ 1\times1, 384 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 96 \\ 3\times3, 96 \\ 1\times1, 384 \end{bmatrix}$	$[1\times1, 576] \longrightarrow$ $\longleftarrow [1\times1, 96]$	$\begin{bmatrix} 576\text{-d} \\ 9\text{-h} \end{bmatrix}$	×3
3	$\begin{bmatrix} 1\times1, 32 \\ 3\times3, 32 \\ 1\times1, 128 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 32 \\ 3\times3, 32 \\ 1\times1, 128 \end{bmatrix}$	$[1\times1, 384] \longrightarrow$ $\longleftarrow [1\times1, 32]$	$\begin{bmatrix} 384\text{-d} \\ 6\text{-h} \end{bmatrix}$	×4	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}$	$[1\times1, 384] \longrightarrow$ $\longleftarrow [1\times1, 128]$	$\begin{bmatrix} 384\text{-d} \\ 6\text{-h} \end{bmatrix}$	×4	$\begin{bmatrix} 1\times1, 192 \\ 3\times3, 192 \\ 1\times1, 768 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 192 \\ 3\times3, 192 \\ 1\times1, 768 \end{bmatrix}$	$[1\times1, 576] \longrightarrow$ $\longleftarrow [1\times1, 192]$	$\begin{bmatrix} 576\text{-d} \\ 9\text{-h} \end{bmatrix}$	×4
4	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$	$[1\times1, 384] \longrightarrow$ $\longleftarrow [1\times1, 64]$	$\begin{bmatrix} 384\text{-d} \\ 6\text{-h} \end{bmatrix}$	×3	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}$	$[1\times1, 384] \longrightarrow$ $\longleftarrow [1\times1, 256]$	$\begin{bmatrix} 384\text{-d} \\ 6\text{-h} \end{bmatrix}$	×3	$\begin{bmatrix} 1\times1, 384 \\ 3\times3, 384 \\ 1\times1, 1536 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 384 \\ 3\times3, 384 \\ 1\times1, 1536 \end{bmatrix}$	$[1\times1, 576] \longrightarrow$ $\longleftarrow [1\times1, 384]$	$\begin{bmatrix} 576\text{-d} \\ 9\text{-h} \end{bmatrix}$	×3
5	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$	$[1\times1, 384] \longrightarrow$ $\longleftarrow [1\times1, 64]$	$\begin{bmatrix} 384\text{-d} \\ 6\text{-h} \end{bmatrix}$	×1	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}$	$[1\times1, 384] \longrightarrow$ $\longleftarrow [1\times1, 256]$	$\begin{bmatrix} 384\text{-d} \\ 6\text{-h} \end{bmatrix}$	×1	$\begin{bmatrix} 1\times1, 384 \\ 3\times3, 384 \\ 1\times1, 1536 \end{bmatrix}$ $\begin{bmatrix} 1\times1, 384 \\ 3\times3, 384 \\ 1\times1, 1536 \end{bmatrix}$	$[1\times1, 576] \longrightarrow$ $\longleftarrow [1\times1, 384]$	$\begin{bmatrix} 576\text{-d} \\ 9\text{-h} \end{bmatrix}$	×1

Table 3: Architecture of DSCT encoder. -d: the hidden dimension. -h: the number of heads in the self-attention block.

4 More qualitative results on Cityscapes

We show more qualitative results on Cityscapes [1] validation set in Figure 1. The odd rows are predicted by SETR [9], and the even rows are predicted by our DSCT. All these results are predicted on Cityscapes validation set using the model trained for 40K iterations with batch size of 8. From the visulization, our DSCT achieves better segmentation results on small objects and boundaries, which demonstrates the important role of our spatial stream.

5 Qualitative results on ADE20K

We also show the qualitative results of ADE20K [10] in Figure 2 and 3. The odd rows are predicted by CCNet [4], and the even rows are predicted by our DSCT. All these results are predicted on ADE20K validation set using the model trained for 160K iterations with batch size of 16. Comparing with CCNet, our DSCT guarantees the integrity of segmentation because of long-range feature dependencies on context stream. Moreover, DSCT has the ability to accurately separate small objects from large ones.

Figure 1: Comparison of qualitative results on Cityscapes validation set. The odd rows are predicted by SETR, and the even rows are predicted by our DSCT.

Figure 2: Comparison of qualitative results on ADE20K validation set. The odd rows are predicted by CCNet, and the even rows are predicted by our DSCT.

Figure 3: More comparison results on ADE20K validation set. The odd rows are predicted by CCNet, and the even rows are predicted by our DSCT.

References

[1] 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. In *CVPR*, 2016.

[2] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *ICCV*, 2019.

[3] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[4] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *CVPR*, 2019.

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[6] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018.

[7] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip HS Torr. Dual graph convolutional network for semantic segmentation. In *BMVC*, 2019.

[8] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[9] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.

[10] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019.