# PointFlow: Flowing Semantics Through Points for Aerial Image Segmentation

Xiangtai Li[1]*, Hao He[2,3]*, Xia Li[4], Duo Li[5], Guangliang Cheng[6,8]
Jianping Shi[6,7] Lubin Weng[2], Yunhai Tong[1] Zhouchen Lin[1]

[1] Key Laboratory of Machine Perception (MOE), Peking University
[2] NLPR, Institute of Automation, Chinese Academy of Sciences
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences [4] ETH Zurich [5] HKUST
[6] SenseTime Research [7] Qing Yuan Research Institute, SJTU [8] Shanghai AI Laboratory

## 1. Supplementary

**Overview:** In this supplementary, we will present more experimental details on Aerial datasets in the first section. Then we will provide detailed descriptions and visualization results on general semantic segmentation datasets.

### 1.1. Experiments on Aerial Datasets

In this section, we give supplementary aerial dataset experiments for the main paper due to the space limitation.

**Application on Various Methods:** For Deeplabv3 [1], we append our PFNet decoder after the ASPP output which makes the total network as an encoder-decorder framework like U-net [13]. For CCnet [5], we follow the same pipeline of Deeplabv3 by replacing ASPP head with CC-head. The GFlops in the table are calculated with $512 \times 512$ inputs.

**Effectiveness on Segmentation Boundaries:** We present the our subtraction based boundary prediction in Fig. 2. As shown in the figure, our subtraction based prediction leads to thinner and more sharpen prediction in the first row and second row of Fig. 2 and avoids the inner part noise shown in the third row of Fig. 2.

**Effect of Context Head:** We also verify the effectiveness of context head by replacing PPM [21] into ASPP [1] where we obtain 66.0 mIoU (0.9 mIoU drop compared with PPM head). Due to the efficiency of PPM, we choose it as our final context head.

**Foreground Points Ratio of Propagation:** We count the sampled points from the three different PFMs by finding the **intersection** of all matched points from Dual Point Matcher in each PFM. We match all the indexes into the high resolution map during the calculation. Thus the propagated points with no-overlap are considered.

**More Visualization on iSAID dataset:** Fig.1 provides more visualization results on iSAID datasets. Compared with previous work, our method has better segmentation results and less false positives in the scene. Fig. 3 shows more matched points visualization on full cropped images.

**Detailed Results:** Tab. 1 reports the per-class results on iSAID datasets. From that table, our PFNet achieves best results on 14 classes (total 15 classes). Tab. 3 and Tab. 2 report per-class results on Postdam and Vaihingen dataset receptively. Our PFNet also achieves top performance.

**Dataset Split on Potsdam and Vaihingen:** We provide the detailed dataset split as follows:

The Potsdam dataset consists of 38 high resolution aerial images. To train and evaluate networks, we utilize 24 images for training and 14 images for testing. For the training set, the image IDs are 2_10, 2_11, 2_12, 3_10, 3_11, 3_12, 4_10, 4_11, 4_12, 5_10, 5_11, 5_12, 6_7, 6_8, 6_9, 6_10, 6_11, 6_12, 7_7, 7_8, 7_9, 7_10, 7_11, 7_12. For the test set, the image IDs are 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, 7_13.

The Vaihingen dataset has 33 high resolution aerial images. We select 16 images for trainand and another 17 images as the testset to evaluate the models. The image IDs for training set are 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 30, 32, 34, 37 and image IDs for test set are 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, 38.

**Speed test details:** We test our models on single V100 GPU by averaging inference time with 100 images. The Pytorch version is 1.5 with CUDA-10.1.

### 1.2. Experiments on Generation Segmentation Dataset

In this section, we will first give the implementation details on Cityscapes [3], BDD [18] and ADE20k [23].

**Implementation Details**

**Cityscapes** dataset is composed of a large set of high-resolution $(2048 \times 1024)$ images in street scenes. This dataset has 5,000 images with high quality pixel-wise annotations for 19 classes, which is further divided into 2975,

*The first two authors contribute equally. Email: lxtpku@pku.edu.cn. Corresponding to: Yunhai Tong, Guangliang cheng

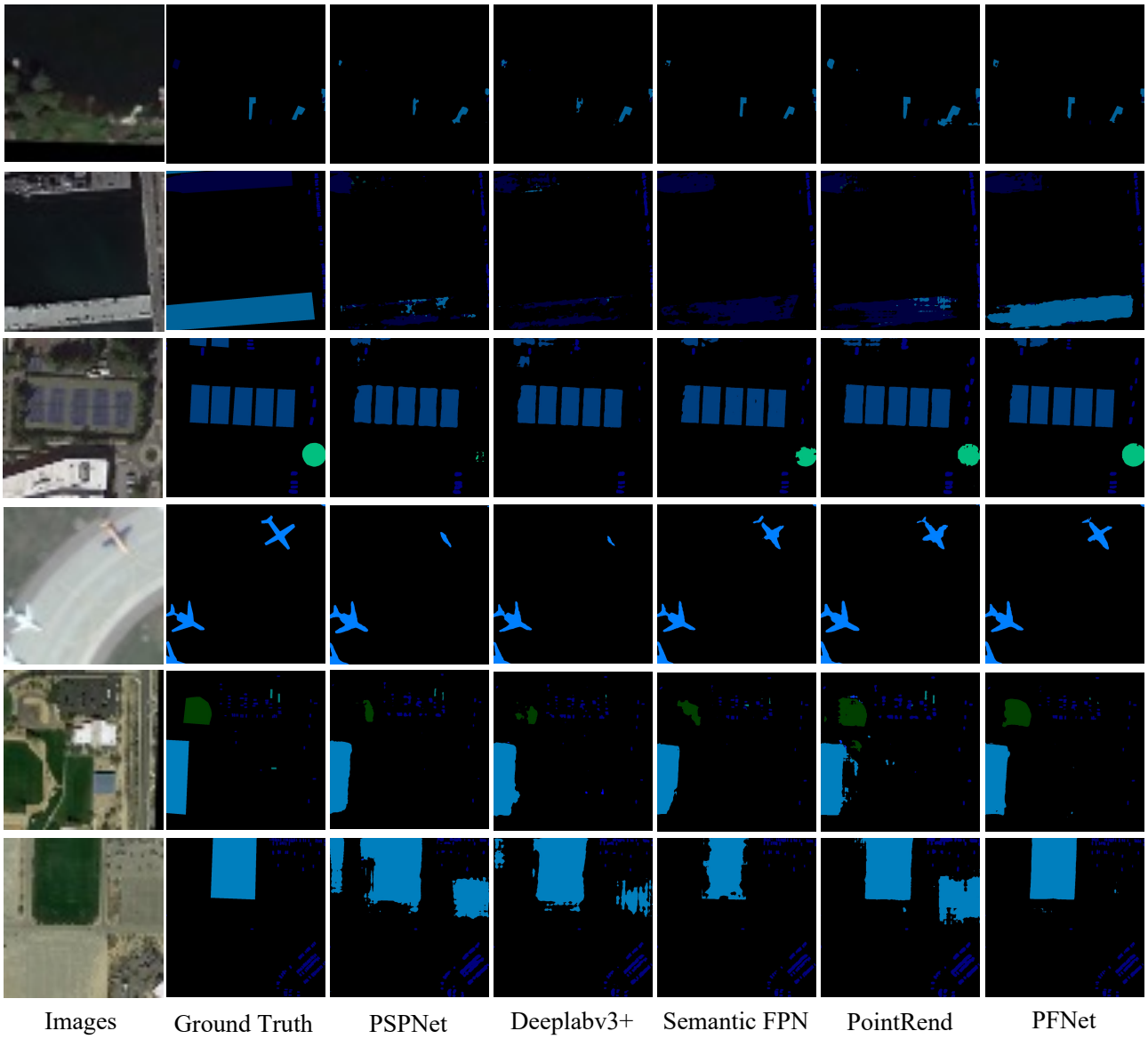



Figure 1: Visualization results on iSAID validation dataset. Best view it on screen.

500, and 1525 images for training, validation and testing. We only use the fine-data for training. Following the common practice, the "poly" learning rate policy is adopted to decay the initial learning rate by multiplying $(1-\frac{\text{iter}}{\text{total\_iter}})^{0.9}$ during training. Data augmentation contains random horizontal flip, random resizing with scale range of $[0.75,\ 2.0]$, and random cropping with crop size of $864 \times 864$ and we train totally 300 epochs for strong baseline.

**BDD** is a new road scene benchmark consisting $7,000$ images for training and $1,000$ images for validation. We follow the same setting as Cityscapes dataset.

**ADE20k** is a more challenging scene parsing dataset annotated with 150 classes, and it contains 20K/2K images for training and validation. It has the various objects in the scene. We train the network for 120 epochs with batch size 16, crop size 512 and initial learning rate 1e-2. For final testing, we perform multi-scale testing with horizontal flip operation.

**Visualization and Comparison Results** We provide visualization and comparison results on Cityscapes and ADE-20k in Fig. 4 and Fig. 5. From both figures, our PFM can better handle the segmentation tails and missing objects in the scene.

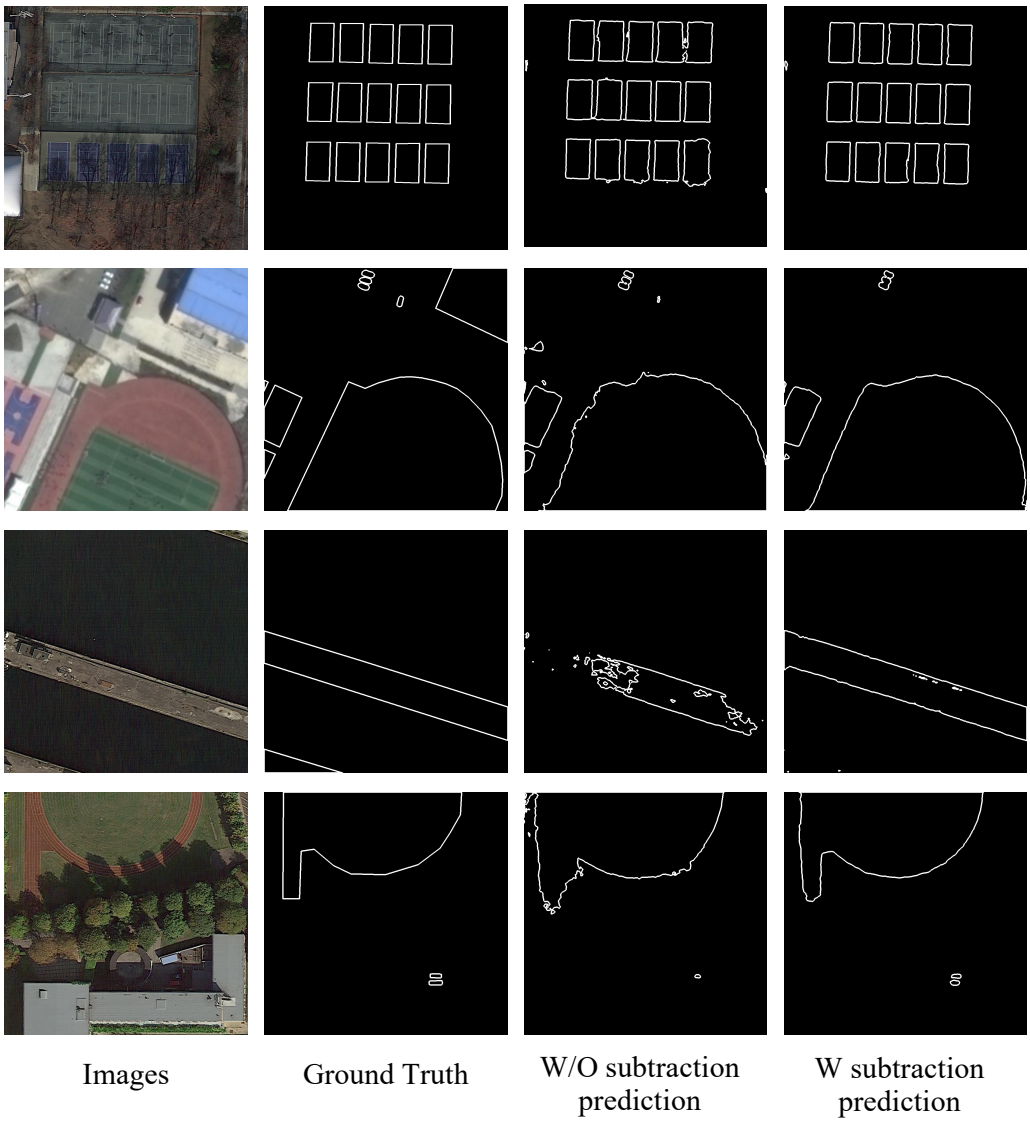



Figure 2: Visualization boundary results on iSAID validation dataset. Best view it on screen.

## References


[1] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint*, 2017.

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[4] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[5] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

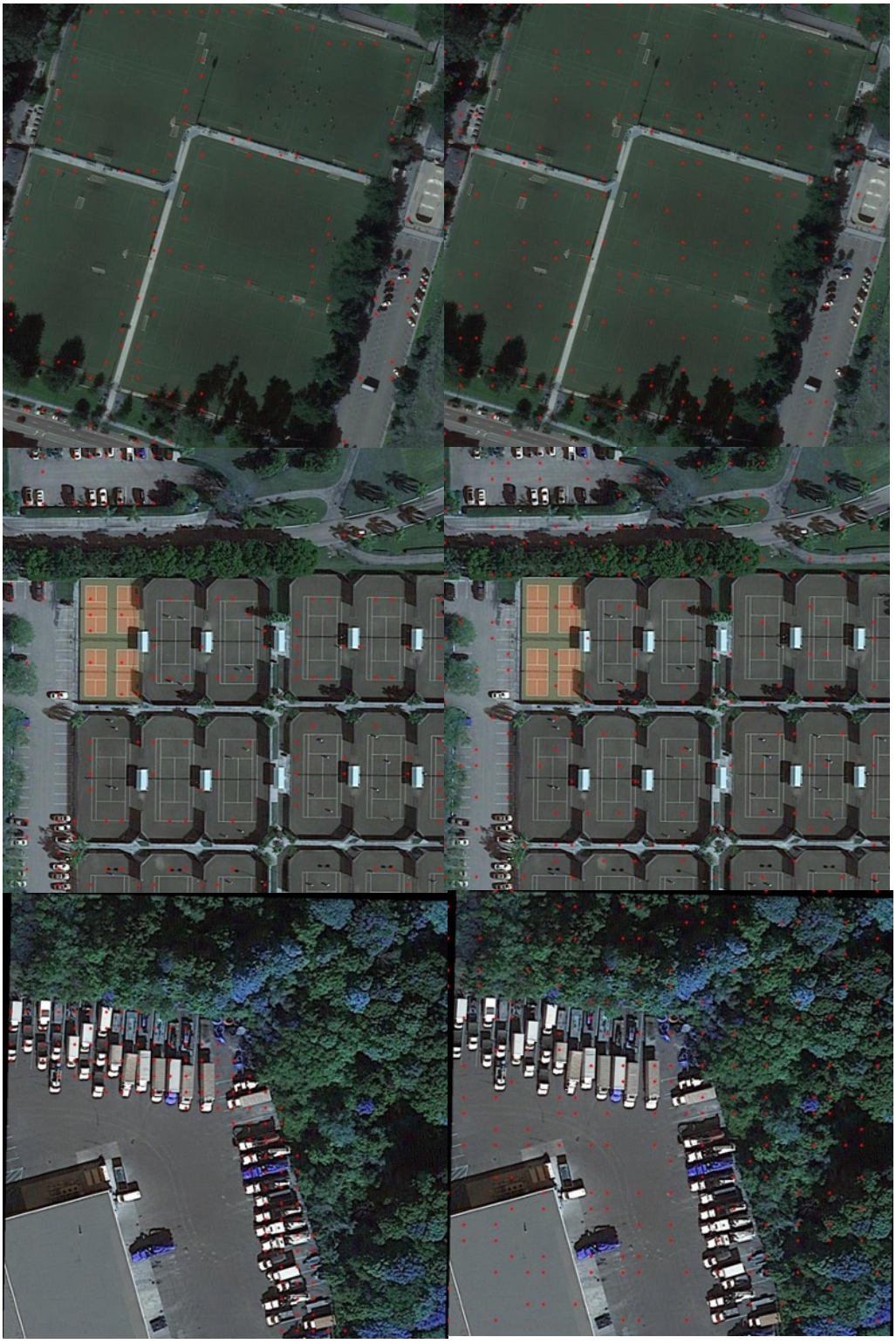
[6] A. Kirillov, R. Girshick, K. He, and P. Dollár. Panoptic feature pyramid networks. In *CVPR*, pages 6399–6408, 2019.

[7] A. Kirillov, Y. Wu, K. He, and R. Girshick. PointRend: Image segmentation as rendering. In *CVPR*, 2020.

[8] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, and Y. Tong. Semantic flow for fast and accurate scene parsing. *ECCV*, 2020.

[9] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, 2019.

[10] G. Lin, A. Milan, C. Shen, and I. D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017.



Boundary Points Salient Points

Figure 3: Visualization sampled points on iSAID validation dataset. Best view it on screen.

| Method | backbone | mIoU(%) | IoU per category(%) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Ship | ST | BD | TC | BC | GTF | Bridge | LV | SV | HC | SP | RA | SBF | Plane | Harbor |
| DenseASPP [17] | RenNet50 | 57.3 | 55.7 | 63.5 | 67.2 | 81.7 | 54.8 | 52.6 | 34.7 | 55.6 | 36.3 | 33.4 | 37.5 | 53.4 | 73.3 | 74.7 | 46.7 |
| RefineNet [10] | ResNet50 | 60.2 | 63.8 | 58.6 | 72.3 | 85.3 | 61.1 | 52.8 | 32.6 | 58.2 | 42.4 | 23.0 | 43.4 | 65.6 | 74.4 | 79.9 | 51.1 |
| PSPNet [21] | ResNet50 | 60.3 | 65.2 | 52.1 | 75.7 | 85.6 | 61.1 | 60.2 | 32.5 | 58.0 | 43.0 | 10.9 | 46.8 | 68.6 | 71.9 | 79.5 | 54.3 |
| OCNet-(ASP-OC) [19] | ResNet50 | 40.2 | 47.3 | 40.2 | 44.4 | 65.0 | 24.1 | 29.9 | 2.71 | 46.3 | 13.6 | 10.3 | 34.6 | 37.9 | 41.4 | 68.1 | 38.0 |
| EMANet [9] | ResNet50 | 55.4 | 63.1 | 68.4 | 66.2 | 82.7 | 56.0 | 18.8 | 42.1 | 58.2 | 41.0 | 33.4 | 38.9 | 46.9 | 46.4 | 78.5 | 47.5 |
| CCNet [5] | ResNet50 | 58.3 | 61.4 | 65.7 | 68.9 | 82.9 | 57.1 | 56.8 | 34.0 | 57.6 | 38.3 | 31.6 | 36.5 | 57.2 | 75.0 | 75.8 | 45.9 |
| EncodingNet [20] | ResNet50 | 58.9 | 59.7 | 64.9 | 70.0 | 84.2 | 55.2 | 46.3 | 36.8 | 57.2 | 38.7 | 34.8 | 42.4 | 59.8 | 69.8 | 76.1 | 48.0 |
| SemanticFPN [6] | ResNet50 | 62.1 | 68.9 | 62.0 | 72.1 | 85.4 | 54.1 | 48.9 | 44.9 | 61.0 | 48.6 | 37.4 | 42.8 | 70.2 | 58.6 | 84.7 | 54.9 |
| UPerNet [6] | ResNet50 | 63.8 | 68.7 | 71.0 | 73.1 | 85.5 | 55.3 | 57.3 | 43.0 | 61.3 | 45.6 | 30.3 | 45.7 | 68.7 | 75.1 | 84.3 | 56.2 |
| HRNet[16] | HRNetW18 | 61.5 | 65.9 | 68.9 | 74.0 | 86.9 | 59.4 | 61.5 | 33.8 | 62.1 | 46.9 | 14.9 | 44.2 | 52.9 | 75.6 | 81.7 | 52.2 |
| SFNet[8] | ResNet50 | 64.3 | 68.8 | 71.3 | 72.1 | 85.6 | 58.8 | 60.9 | 43.1 | 62.9 | 47.7 | 30.4 | 47.8 | 69.8 | 75.1 | 83.1 | 57.3 |
| GSCNN[14] | ResNe50 | 63.4 | 65.9 | 71.2 | 72.6 | 85.5 | 56.1 | 58.4 | 40.7 | 63.8 | 51.1 | 33.8 | 48.8 | 58.5 | 72.5 | 83.6 | 54.4 |
| RANet[12] | ResNet50 | 62.1 | 67.1 | 61.3 | 72.5 | 85.1 | 53.2 | 47.1 | 45.3 | 60.1 | 49.3 | 38.1 | 41.8 | 70.5 | 58.8 | 83.1 | 55.6 |
| FarSeg [22] | ResNet50 | 63.7 | 65.4 | 61.8 | 77.7 | 86.4 | 62.1 | 56.7 | 36.7 | 60.6 | 46.3 | 35.8 | **51.2** | 71.4 | 72.5 | 82.0 | 53.9 |
| PFNet | ResNet50 | **66.9** | **70.3** | **74.7** | **77.8** | **87.7** | **62.2** | **59.5** | **45.2** | **64.6** | **50.2** | **37.9** | 50.1 | **71.7** | **75.4** | **85.0** | **59.3** |

Table 1: Experimental results on iSAID $val$ set. The bold values in each column mean the best entries. The category are defined as: ship (Ship), storage tank (ST), baseball court (BC), ground field track (GTF), bridge (Bridge), large vehicle (LV), small vehicle (SV), helicopter (HC), swimming pool (SP), roundabout (RA), soccerball field (SBF), plane (Plane), harbor (Harbor). All the models are trained under the same setting following the FarSeg [22].

| Method | mIoU(%) | mean-$F_1$ | $F_1$ per category | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Imp.surf. | Build. | Low veg. | Tree | Car | Cluster |
| PSPNet [21] | 65.1 | 76.8 | 88.4 | 92.8 | 79.2 | 85.9 | 73.5 | 41.0 |
| FCN [11] | 64.2 | 75.9 | 87.6 | 91.6 | 77.8 | 84.6 | 73.5 | 40.3 |
| OCNet(ASP-OC) [19] | 65.7 | 77.4 | 88.8 | 92.9 | 79.2 | 85.8 | 73.9 | 43.8 |
| Deeplabv3+ [2] | 64.3 | 76.0 | 88.7 | 92.8 | 78.9 | 85.6 | 72.4 | 37.6 |
| DANet [4] | 65.3 | 77.1 | 88.5 | 92.7 | 78.8 | 85.7 | 73.7 | 43.2 |
| CCNet [5] | 64.3 | 75.9 | 88.3 | 92.5 | 78.8 | 85.7 | 73.9 | 36.3 |
| SemanticFPN [6] | 66.3 | 77.6 | 89.6 | 93.6 | 79.7 | **86.3** | 75.7 | 40.7 |
| UPerNet [16] | 66.9 | 78.7 | 89.2 | 93.0 | 79.4 | 86.0 | 74.9 | 49.7 |
| PointRend [7] | 65.9 | 78.1 | 88.2 | 92.4 | 78.9 | 84.5 | 73.5 | 51.1 |
| HRNet-W18 [15] | 66.9 | 78.2 | 89.2 | 92.6 | 78.7 | 85.7 | 77.1 | 45.9 |
| GSCNN [14] | 67.7 | 79.5 | 89.4 | 92.6 | 78.8 | 85.4 | 77.9 | 52.9 |
| SFNet [8] | 67.6 | 78.6 | 90.0 | **94.0** | **80.3** | 86.5 | 78.9 | 41.9 |
| EMANet [9] | 65.6 | 77.7 | 88.2 | 92.7 | 78.0 | 85.7 | 72.7 | 48.9 |
| RANet [12] | 66.1 | 78.2 | 88.0 | 92.3 | 79.1 | 86.0 | 78.8 | 53.1 |
| EncodingNet [20] | 65.5 | 77.4 | 88.6 | 92.5 | 78.5 | 85.7 | 73.6 | 45.5 |
| Denseaspp [17] | 64.7 | 76.4 | 87.3 | 91.1 | 76.2 | 83.4 | 77.1 | 43.3 |
| PFNet | **70.4** | **81.9** | **90.1** | 93.6 | 77.7 | 85.4 | **80.0** | **64.6** |

Table 2: Experimental results on the Vaihingen Dataset. The results are reported with single scale input. The bold values in each column mean the best entries. The category are defined as: impervious surfaces (Imp.surf.), buildings (Build), low vegetation (Low veg), trees (Tree), cars (Car), cluster/background (Cluster).


[11] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[12] L. Mou, Y. Hua, and X. X. Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *CVPR*, pages 12416–12425, 2019.

[13] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

[14] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. *ICCV*, 2019.

[15] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019.

[16] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018.

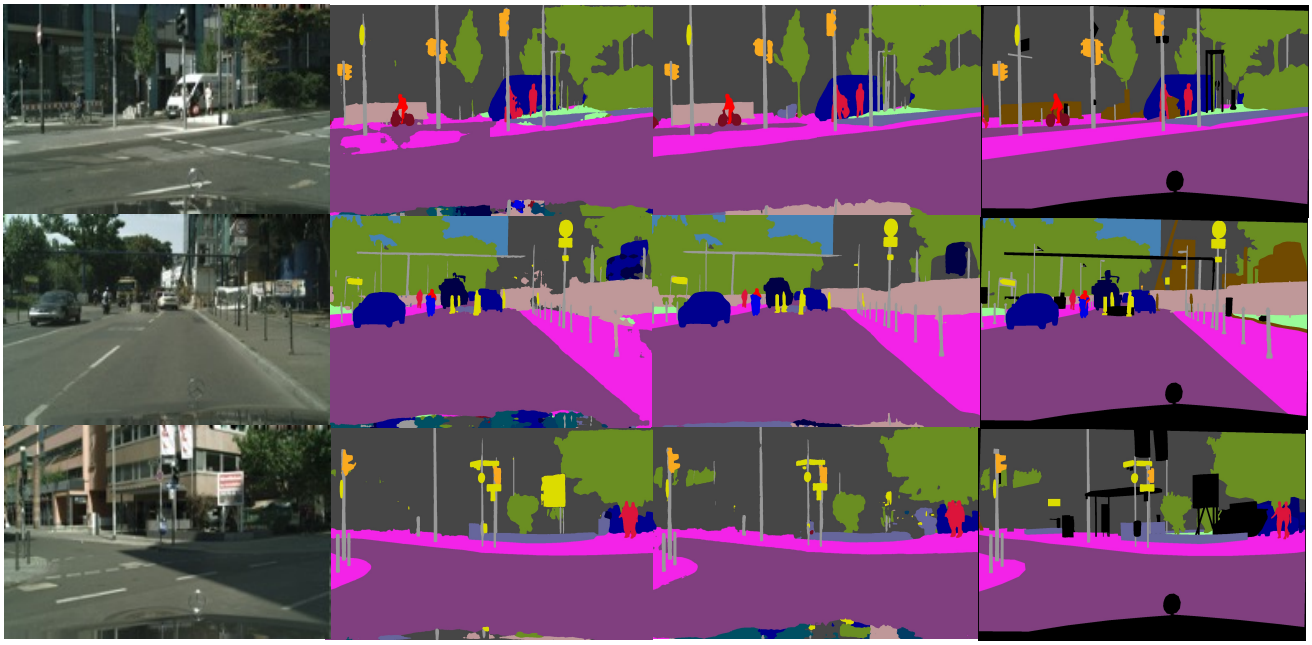
[17] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018.

[18] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020.

[19] Y. Yuan and J. Wang. Ocnet: Object context network for scene parsing. *arXiv preprint*, 2018.

| Method | mIoU(%) | mean-$F_1$ | $F_1$ per category | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Imp.surf. | Build. | Low veg. | Tree | Car | Cluster |
| PSPNet [21] | 73.9 | 83.9 | 90.8 | 95.4 | 84.5 | 86.1 | 88.6 | 58.0 |
| FCN [11] | 73.1 | 83.1 | 90.2 | 94.7 | 84.1 | 85.6 | 89.2 | 54.8 |
| OCnet(ASP-OC) [19] | 74.2 | 84.1 | 90.9 | 95.5 | 84.8 | 86.0 | 89.2 | 58.2 |
| Deeplabv3+ [2] | 74.1 | 83.9 | 91.0 | 95.6 | 84.6 | 86.0 | 90.0 | 56.2 |
| DAnet [4] | 74.0 | 83.9 | 91.0 | 95.6 | 84.9 | 86.2 | 88.7 | 57.0 |
| CCnet[5] | 73.8 | 83.8 | 90.7 | 95.5 | 84.7 | 86.0 | 88.5 | 57.3 |
| SemanticFPN [6] | 74.3 | 84.0 | 91.0 | 95.5 | 84.9 | 85.9 | 90.4 | 56.3 |
| UPerNet [16] | 74.3 | 84.0 | 90.9 | 95.7 | 85.0 | 86.0 | 90.2 | 56.2 |
| PointRend [7] | 72.0 | 82.7 | 89.8 | 94.6 | 82.8 | 85.2 | 85.2 | 58.6 |
| HRNet-W18 [15] | 73.4 | 83.4 | 90.4 | 94.9 | 84.2 | 85.4 | 90.0 | 55.5 |
| GSCNN [14] | 73.4 | 84.1 | 91.4 | 95.5 | 84.8 | 85.8 | **91.2** | 55.9 |
| SFNet [8] | 74.3 | 84.0 | 91.0 | 95.5 | 85.1 | 86.0 | 90.9 | 55.5 |
| EMANet [9] | 72.9 | 83.1 | 90.4 | 94.9 | 84.2 | 85.7 | 88.3 | 55.1 |
| RANet [12] | 73.8 | 83.9 | 90.8 | 92.1 | 84.3 | 86.8 | 88.9 | 56.0 |
| EncodingNet [20] | 73.4 | 83.5 | 90.6 | 95.1 | 84.5 | 86.0 | 88.2 | 56.6 |
| Denseaspp [17] | 73.9 | 83.9 | 90.8 | 95.4 | 84.6 | 86.0 | 88.5 | 58.1 |
| PFNet | **75.4** | **84.8** | **91.5** | **95.9** | **85.4** | **86.3** | 91.1 | **58.6** |

Table 3: Experimental results on Postdam Dataset. The results are reported with single scale input. The bold values in each column mean the best entries. The category are defined as: impervious surfaces (Imp.surf.), buildings (Build), low vegetation (Low veg), trees (Tree), cars (Car), cluster/background (Cluster).



Images w/o PFM w PFM Ground Truth

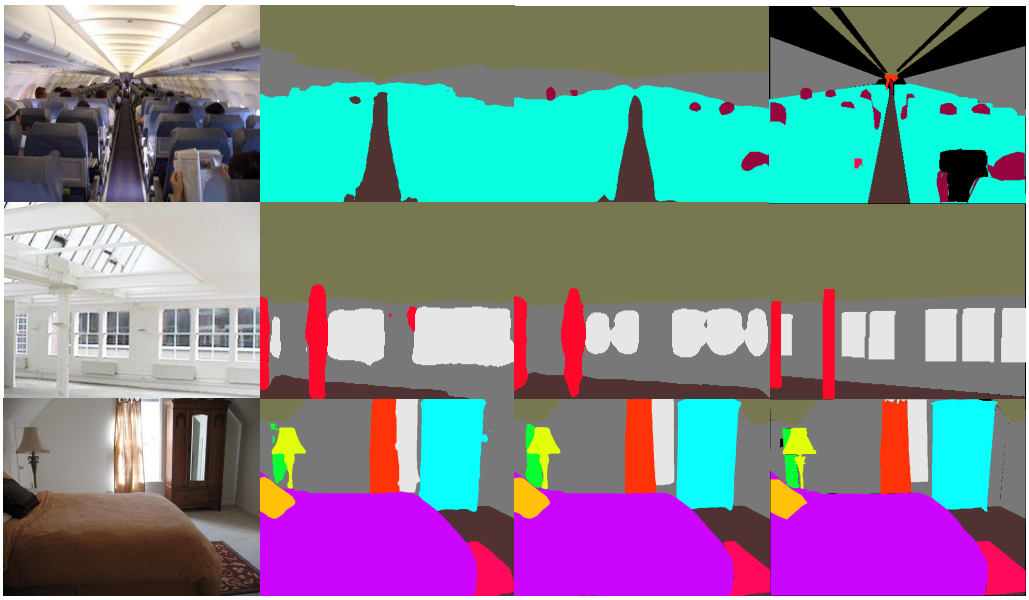
Figure 4: Visualization results on Cityscapes dataset. Best view it on screen.


[20] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018.

[21] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[22] Z. Zheng, Y. Zhong, J. Wang, and A. Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *CVPR*, pages 4096–4105, 2020.

[23] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20K dataset. *arXiv preprint*, 2016.



Images w/o PFM w PFM Ground Truth

Figure 5: Visualization results on ADE20k dataset. Best view it on screen.