Learning Implicit Feature Alignment Function for Semantic Segmentation
Supplementary Material

Hanzhe Hu[1]*, Yinbo Chen[2]*, Jiarui Xu[2], Shubhankar Borse[3], Hong Cai[3], Fatih Porikli[3], and Xiaolong Wang[2]

[1] Peking University
[2] University of California, San Diego
[3] Qualcomm AI Research

In this supplementary material, we provide additional experimental results. In section 1, we present results of additional ablation studies. In section 2, qualitative results on PASCAL Context dataset are presented. In section 3, qualitative results on ADE20K dataset are provided.

1 Additional Ablation Study

Extension to Different Backbones. We implement experiments to assess the effectiveness of the proposed IFA with different backbone networks on the `val` set of Cityscapes [1] dataset. We use the FPN decoder (involving four levels of features from the backbone) and different backbones as the encoder, including ResNet [3], HRNet [9] and ResNeSt [11]. As shown in Table 1, IFA improves FPN with ResNet-50 by 0.9% in mIoU, FPN with ResNet-101 by 0.8%, HRNet-W18 by 0.5% and FPN with ResNeSt-50 by 0.6%, indicating the extensive ability of the proposed IFA.

Method	Backbone	mIoU(%)
FPN	ResNet-50	77.19
FPN (IFA)	ResNet-50	78.02
FPN	ResNet-101	78.70
FPN (IFA)	ResNet-101	79.49
FPN	HRNet-W18	77.60
FPN (IFA)	HRNet-W18	78.10
FPN	ResNeSt-50	78.20
FPN (IFA)	ResNeSt-50	78.85

Table 1: Performance effect of IFA with different backbones on Cityscapes `val` set.

Pos. Enc	mIoU(%)
None	76.88
Coord	77.01
Sine	77.61
Cosine	77.56
Ours (fixed)	77.89
Ours (learned)	**78.02**

Table 2: Results for different formations of position encodings on Cityscapes `val` set.

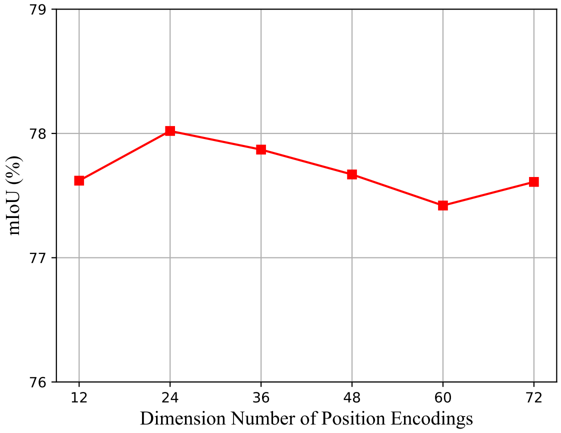

Fig. 1: Effect of dimensions of position encoding.

Impact of Position Encoding. We further perform experiments to validate the effectiveness of the position encoding inside our proposed Implicit Feature Alignment function (IFA). As illustrated in Table 2, we experiment with various formations. We first study adding relative coordinates directly, which brings 0.2% improvement ('Coord'). We also encode the relative coordinates with 'Sine' or 'Cosine' function, which further improve the results. The learnable frequencies achieves the best performance. The results also demonstrate that position encodings can effectively obtain better results than directly using the spatial coordinates. Moreover, we also investigate the relationship between the dimension number of the position encoding and model's performance. We test a total of six variations: 12, 24, 36, 48, 60 and 72. As shown in Figure 1, though the influence of dimension number is not significant, 24 yields the highest performance. Hence, we choose 24 as the dimension number by default.

Impact of Evaluation Strategies. When comparing with other state-of-the-art methods, we use multi-scale and flipping strategy like previous methods

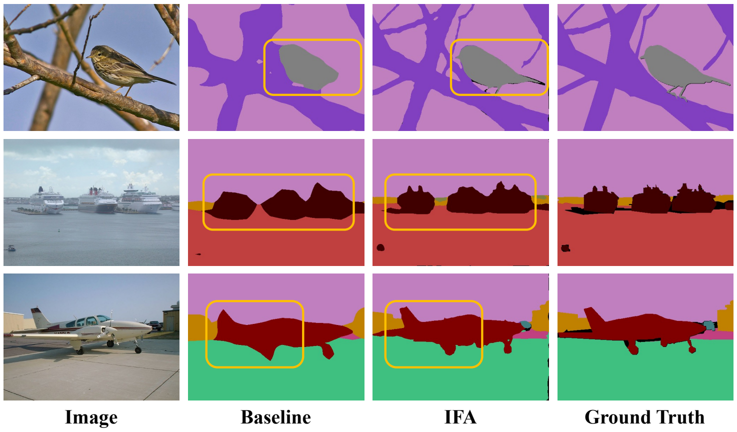

Fig. 2: Visualization results on PASCAL Context `test` set. From left to right: input image, predictions made by the FPN baseline, predictions made by the FPN with the proposed IFA and groundtruth map. Yellow squares denote the challenging regions that can be resolved by our proposed IFA.

[12,10,2,6,5,4] to further improve the performance. In Table 3, we provide results of IFA with the backbone ResNet-101 under left-right flipping and multi-scale [0.5, 0.75, 1.0, 1.25, 1.5, 2.0] evaluation strategies, where MSFlip improves the performance by 1.5% in mIoU.

Method	MS	Flip	mIoU(%)
IFA			79.49
IFA	✓		80.61
IFA		✓	80.29
IFA	✓	✓	81.02

Table 3: Performance effect of IFA with different evaluation strategies on Cityscapes `val` set. 'MS' denotes multi-scale inference and 'Flip' denotes left-right flipping strategy.

2 Qualitative Results on PASCAL Context Dataset

In this section, we provide qualitative results on PASCAL Context [8] dataset. As shown in Figure 2, we compare the segmentation results of our proposed IFA with the baseline method FPN [7]. With the baseline method producing a coarse

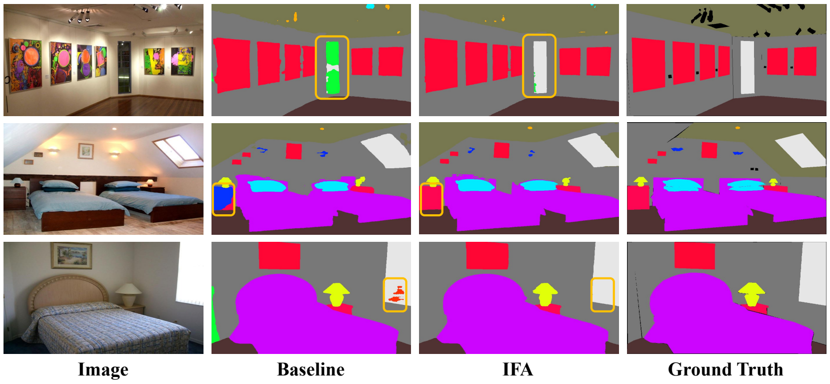

Fig. 3: Visualization results on ADE20K `val` set. From left to right: input image, predictions made by the FPN baseline, predictions made by the FPN with the proposed IFA and groundtruth map. Yellow squares denote the challenging regions that can be resolved by our proposed IFA.

prediction of the target object, our proposed IFA is able to preserve more precise appearance information of the target object.

3 Qualitative Results on ADE20K Dataset

We also present visualization results on ADE20K [13] dataset. As shown in Figure 3, we compare the segmentation results of our proposed IFA with the baseline method FPN [7]. The qualitative results indicate that our proposed IFA is capable of resolving high-level category ambiguity issues, where mis-classified regions produced by the baseline method are successfully resolved by IFA.

References

1. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016) 1
2. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2019) 3
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1
4. Hu, H., Cui, J., Wang, L.: Region-aware contrastive learning for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16291–16301 (2021) 3

5. Hu, H., Ji, D., Gan, W., Bai, S., Wu, W., Yan, J.: Class-wise dynamic graph convolution for semantic segmentation. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII. vol. 12362, pp. 1–17. Springer (2020) 3
6. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. arXiv preprint arXiv:1811.11721 (2018) 3
7. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6399–6408 (2019) 3, 4
8. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 891–898 (2014) 3
9. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence (2020) 1
10. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: Denseaspp for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3684–3692 (2018) 3
11. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. arXiv preprint arXiv:2004.08955 (2020) 1
12. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017) 3
13. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017) 4