Predictive Feature Learning for Future Segmentation Prediction Supplementary Material

Zihang Lin[1]*, Jiangxin Sun[1]*, Jian-Fang Hu[1,4,5]†, Qizhi Yu[2], Jian-Huang Lai[1,4], Wei-Shi Zheng[1,3,5]
[1]Sun Yat-sen University, China [2]Zhejiang Laboratory, China [3]Pengcheng Laboratory, China
[4]Guangdong Province Key Laboratory of Information Security Technology, Guangzhou, China
[5] Key Laboratory of Machine Intelligence and Advanced Computing, MOE
{linzh59, sunjx5}@mail2.sysu.edu.cn, hujf5@mail.sysu.edu.cn
qyu@ieee.org, stsljh@mail.sysu.edu.cn, wszheng@ieee.org

1. Experiment details of Figure 2

In order to investigate the influence of feature resolution for future segmentation prediction, in our early attempts, we conducted several experiments on Cityscapes[1] dataset using different segmentation models (i.e., Semantic FPN[5], PSPNet[9], HRNet[8] and DANet[2]). We summarize the results in Figure I. In our experiments, we first input the image frame of resolution 1024×2048 to the segmentation model, and the resolution of output feature map is 256×512, then we simply downsample these feature maps to obtain the smaller resolution ones, i.e., 128×256, 64×128 and 32×64. For each feature resolution, we train a ConvLSTM to predict the feature maps of future unobserved frames. The predicted feature maps are inputted to the segmentation head of the segmentation model to generate semantic segmentation results for future unobserved frames. As shown in Figure I, for all the segmentation models we used, as the feature resolution increases, the prediction performances first increase and then decrease. This implies that increasing feature resolution can be harmful for future segmentation prediction, although it is beneficial for image segmentation.

Considering that in the above experiments, simply downsampling the feature maps to obtain the low-resolution ones will lose some information, we further conducted an experiments using our proposed model to extract feature maps with different resolutions. The corresponding results are shown in Figure I (termed "Ours") and are used to illustrate Figure 2 in the main text.

2. Results on Cityscapes test set

We evaluate our model on the Cityscapes[1] test set by submitting the model predictions to the online evaluation server. We use the same model parameters (only trained on the train set) as the one used in the main text and the results are shown in Table I and Table II. For future instance segmentation prediction, our approach outperforms existing methods by a large margin, which demonstrates the effectiveness and robustness of the proposed ap-

*Equal contributions, † Corresponding author.

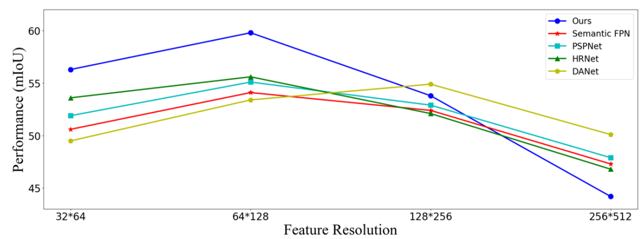

Figure I. Influence of feature resolution for future semantic segmentation prediction.

Table I. Future instance segmentation prediction performance on the Cityscapes test set.

	Short-term		Mid-term	
	AP50	AP	AP50	AP
Mask R-CNN [4] Oracle	58.1	31.9	58.1	31.9
F2F[6]	/	/	17.5	6.7
PSF[3]	31.3	14.9	19.8	8.4
Ours	**42.2**	**21.6**	**27.1**	**12.8**

Table II. Future semantic segmentation prediction performance on the Cityscapes test set using mIoU as the evaluation metric. ALL: all classes. MO: moving objects. †: Trained on both train and validation set.

	Short-term		Mid-term	
Method	ALL	MO	ALL	MO
Semantic FPN [5] Oracle	75.3	73.4	75.3	73.4
PSF[3]	67.3	58.8	57.7	48.8
F2MF[7]†	70.2	**68.7**	59.1	**56.3**
Ours	**70.3**	66.8	**59.2**	53.1

proach. For future semantic segmentation prediction, we also achieve state-of-the-art performance for all classes prediction. For moving objects, compared to the results on the validation set, we observe that the performance of our oracle (can be seen as an upper bound) decreases considerably, which leads to a drop of our future prediction performance. F2MF[7] achieves a better performance than ours but they use the validation set as an additional training source.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[2] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[3] Colin Graber, Grace Tsai, Michael Firman, Gabriel Brostow, and Alexander G. Schwing. Panoptic segmentation forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12517–12526, June 2021.

[4] K He, G Gkioxari, P Dollar, and R Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 2018.

[5] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019.

[6] Pauline Luc, Camille Couprie, Yann Lecun, and Jakob Verbeek. Predicting future instance segmentation by forecasting convolutional features. In *Proceedings of the European Conference on Computer Vision*, pages 584–599, 2018.

[7] Josip Saric, Marin Orsic, Tonci Antunovic, Sacha Vrazic, and Sinisa Segvic. Warp to the future: Joint forecasting of features and feature motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10648–10657, 2020.

[8] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6230–6239, 2017.