Supplementary Material for Submission “APANet: Auto-Path Aggregation for Future Instance Segmentation Prediction”

Jian-Fang Hu, Jiangxin Sun, Zihang Lin, Jianhuang Lai, Wenjun Zeng, and Wei-Shi Zheng

Abstract—In this supplementary document, we provide more experimental investigations and analysis on the future instance segmentation prediction system developed in our main submission, which is excluded from the main submission due to space limitation.

In our main submission, we developed our future instance segmentation prediction system based on the proposed auto-path aggregation network (APANet). Our approach can collaboratively predicting multi-level pyramid features via selectively and adaptively aggregating the task-specific hierarchical spatio-temporal contextual information obtained on the features of each individual pyramid level. We have demonstrated the effectiveness of our method on three video segmentation benchmark datasets and show that the proposed future instance segmentation prediction model outperforms existing models significantly. In the following, we provide more experimental investigations and analysis.

1 More Discussion

More evaluation on attention. Our approach intends to learn a proper operation for each auto-path connection. And the set of possible operations is defined in the form of “Operation X” and “Operation X + Attention”. Theoretically, the “Operation X + Attention” has already embedded the “Operation X” in the case of all-one attention. Indeed, the “Operation X” is needed in our approach as for the cases that the algorithm can not determine a reliable attention map due to the data limitation, so that our model can still choose the operation “Operation X” as an alternative. Therefore, the “Operation X” can be viewed as a regularizer to avoid overfitting to the rare video patterns. To experimentally verify this, we further conduct comparison experiments with three different settings, where the operations to be searched are provided by “Operation X”, “Operation X + Attention” and both (“Operation X” and “Operation X + Attention”), respectively. The results are presented in Table R.1. The result indicates that although “Operation X + Attention” can theoretically contain “Operation X”, combining them can better capture contextual information and thus obtain better segmentation prediction results in practice. Figure R.1 presents the visualization segmentation results of some samples, in which the “Operation X” is selected by the algorithm rather than “Operation X + Attention”. The results verify that using “Operation X” as a regularizer is really beneficial for the prediction of hard samples.

TABLE R.1
Evaluation on the influence of attention.

	Short-term		Mid-term	
Candidate Operation	AP50	AP	AP50	AP
Operation X	44.2	21.8	24.9	10.7
Operation X + Attention	45.7	22.9	28.6	12.5
Both	**46.1**	**23.2**	**29.2**	**12.9**

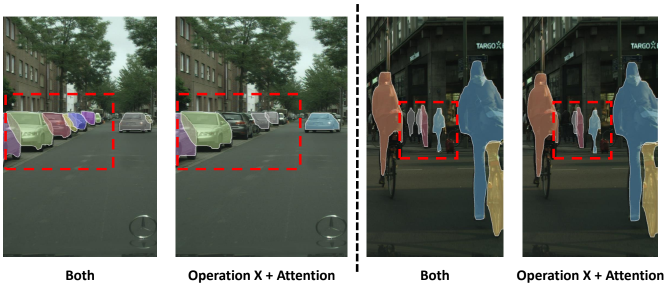

Fig. R.1. Visualized results of our approaches under the settings “Operation X + Attention” and “Both”.

Operation learning for intra-level connections. In the main manuscript, we only consider learning a proper operation for the connections among the cells from different ConvLSTMs and manually set the connections between the cells of the same ConvLSTM as identity operation, which forms a traditional ConvLSTM. Here, we provide a further study on learning the operation for intra-level connections. Similar to that for inter-level connection search, we define the set of possible operations for the intra-level connection search in the form of “Operation X” and “Operation X + Attention”, where the candidates of “Operation X” are defined in Table R.2.

The detailed results are presented in Table R.3. As shown, searching the intra-level or inter-level connections could both

- J.-F. Hu, J. Sun, Z. Lin, J. Lai, and W.-S. Zheng are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. J.-F. Hu is also with the GuangDong Province Key Laboratory of Information Security Technology, Guangzhou, China. W.-S. Zheng is also with the Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China.
E-mail:hujf5@mail.sysu.edu.cn, {sunjx5, linzh59}@mail2.sysu.edu.cn, stsljh@mail.sysu.edu.cn, and wszheng@ieee.org
- W. Zeng is with the Microsoft Research Asia.
E-mail:wezeng@microsoft.com

TABLE R.2
Candidates of "Operation X" for intra-level connections.

3×3 convolution	5×5 convolution
3×3 depthwise-separable convolution	5×5 depthwise-separable convolution
3×3 atrous convolution with rate 2	5×5 atrous convolution with rate 2
Identity	No connection (zero)

TABLE R.3
Evaluation on the influence of architecture learning.

	Short-term		Mid-term	
	AP50	AP	AP50	AP
No Search	44.9	22.4	26.3	11.7
Intra-level	45.5	22.8	27.2	12.3
Inter-level	46.1	23.2	29.2	12.9
Both	**46.4**	**23.4**	**29.8**	**13.2**

improve the system performance. We observe that the performance gain obtained by searching for inter-level connections is larger than that of searching for intra-level connections. By examining the parameters learned by searching intra-level connections, we find that more than 95% of the connections has selected the operation "Identity". We can also observe that searching for the intra-level and inter-level connections together can obtain the best performance.

More visualization results. Here, we provide more visualization results of both short-term prediction (Figure R.2) and mid-term prediction (Figure R.3). As shown, our approaches outperform F2F [1] and our preliminary work [2] for the tasks of both short-term and mid-term future instance segmentation prediction, and it demonstrates that collaboratively predicting multi-level features is beneficial for the future instance segmentation prediction. We can also observe that our extension work performs better than our preliminary work. This indicates that adaptively aggregating contextual information gained in the multi-level features also benefits future instance segmentation prediction, which can better capture the diverse appearance variations in different videos. Additionally, we also present the visual results of our approach for predicting long-term instance segmentation in Figure R.4. As shown, our method can still obtain acceptable results for predicting the segmentation of frames after 0.5s, 1.0s, 1.5s and 2.0s.

References

[1] P. Luc, C. Couprie, Y. Lecun, and J. Verbeek, "Predicting future instance segmentation by forecasting convolutional features," in *European Conference on Computer Vision*, 2018, pp. 584–599.

[2] J. Sun, J. Xie, J.-F. Hu, Z. Lin, J. Lai, W. Zeng, and W.-S. Zheng, "Predicting future instance segmentation with contextual pyramid convlstms," in *ACM International Conference on Multimedia*, 2019, pp. 2043–2051.

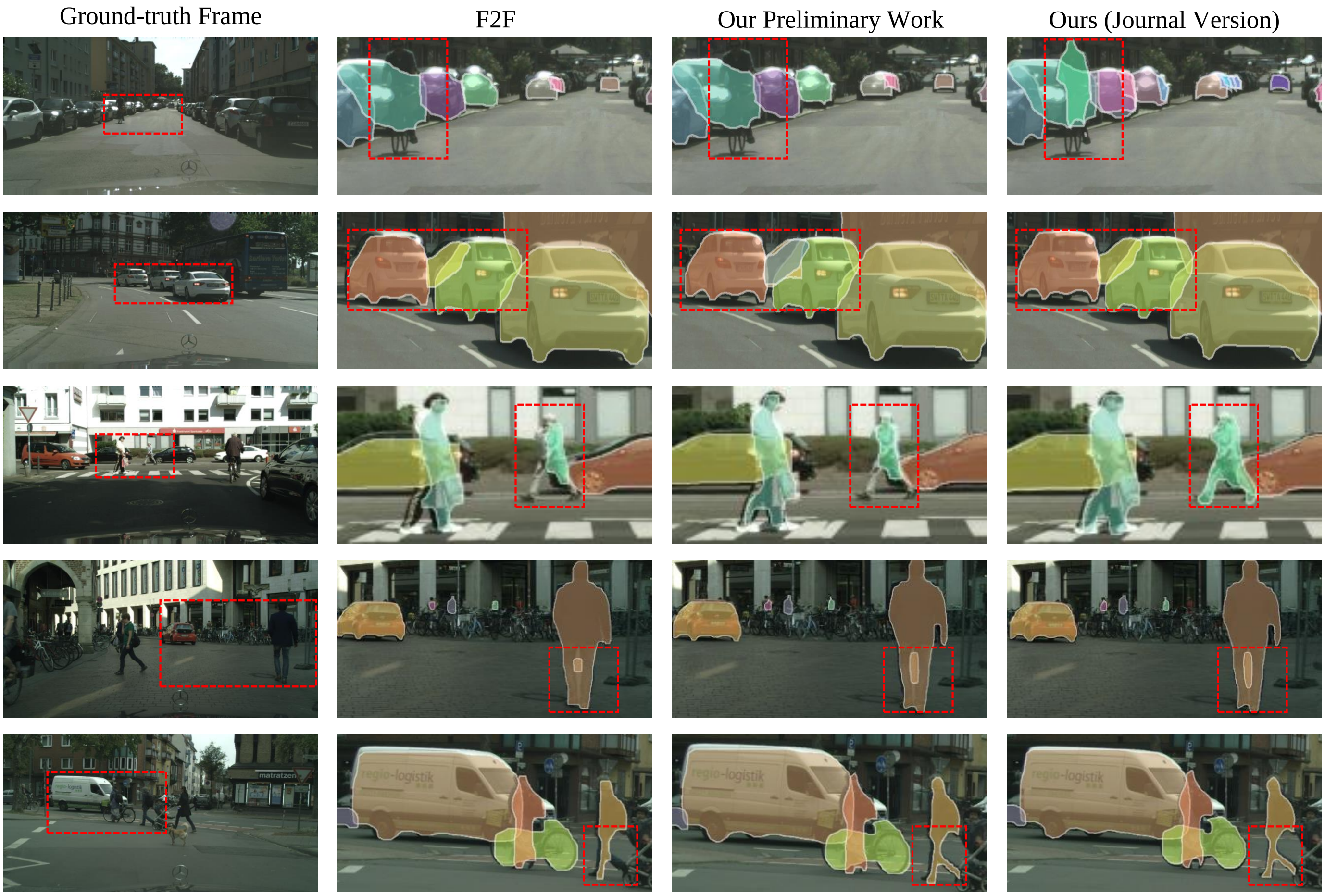

Fig. R.2. Visualized results for the short-term future instance segmentation prediction.

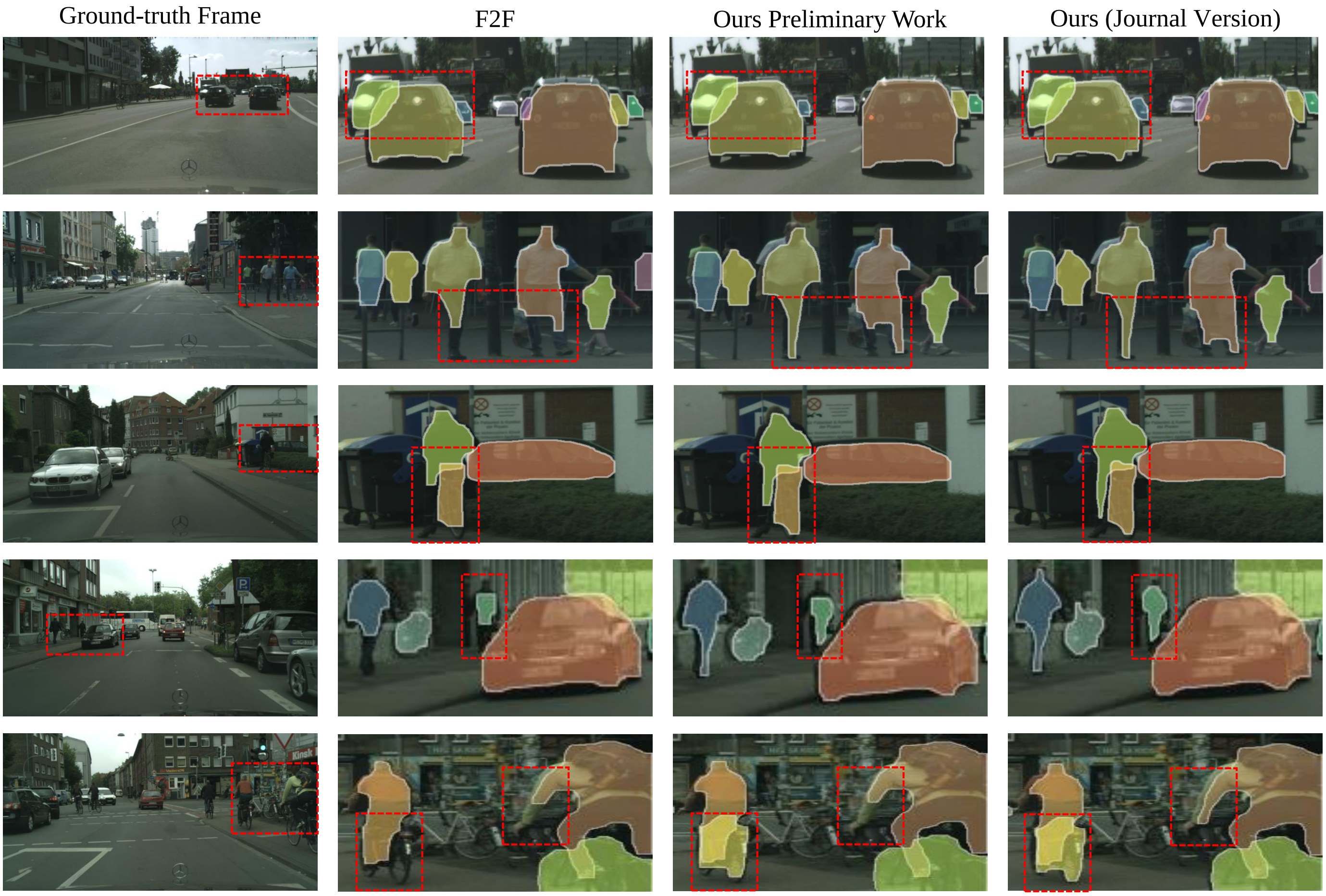

Fig. R.3. Visualized results for the mid-term future instance segmentation prediction.

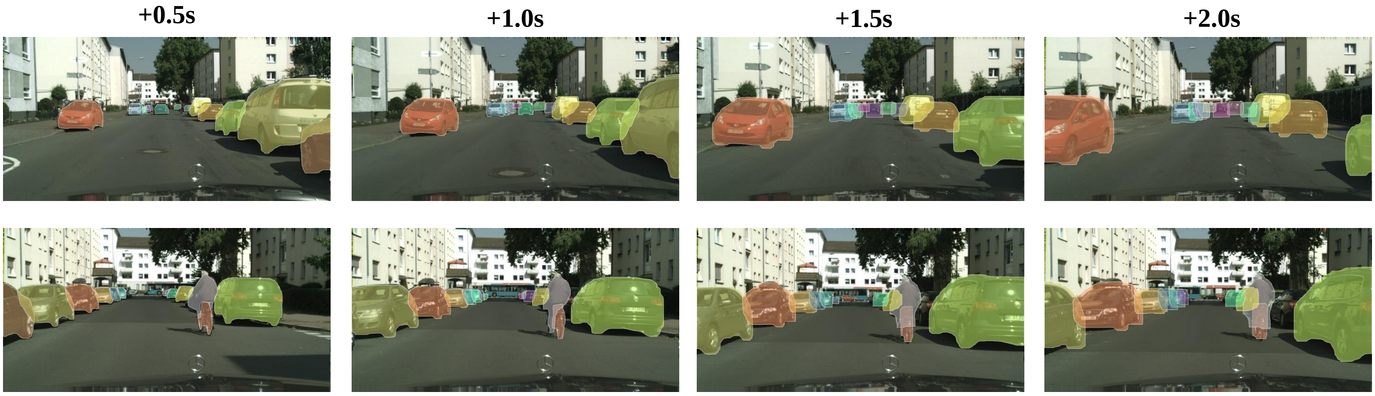

Fig. R.4. Visualized results for the long-term future instance segmentation prediction.