STD2P: RGBD Semantic Segmentation Using Spatio-Temporal Data-Driven Pooling Supplementary Material

Yang He[1], Wei-Chen Chiu[1], Margret Keuper[2] and Mario Fritz[1]
[1]Max Planck Institute for Informatics,
Saarland Informatics Campus, Saarbrücken, Germany
[2]University of Freiburg, Freiburg, Germany

In the supplementary material, we present the analysis of semantic boundary accurary in Section 1. In section 2, we evaluate the oracle performance on NYUDv2 40-class task with our spatio-temporal data-driven pooling. In section 3, we analyze the groundtruth annotations of the NYUDv2 40-class task. In section 4, we provide the qualitative results of the semantic segmentation results of the NYUDv2 4-class and 13-class tasks. In section 5, we provide more qualitative examples of the semantic segmentation results of the NYUDv2 40-class task. In section 6, we show some failure cases which do not achieve better performance than FCN.

1. Analysis of semantic segmentation boundary accuracy

In order to quantify the improvement on semantic boundary localization based on the proposed data-driven pooling scheme, we use Boundary Precision Recall (BPR), as also used in image or video segmentation benchmark [1, 2] for evaluation. Figure 1 shows the resulting semantic boundary average precision-recall curve. We conclude that our method generates more accurate boundaries than FCN, which achieve 0.477 BPR score while our method achieves 0.647. Besides, our method even improves on the superpixel [3] we build on, which means our method can successfully merge over-segmentations or non-semantic boundaries between adjacent instances of the same semantic class.

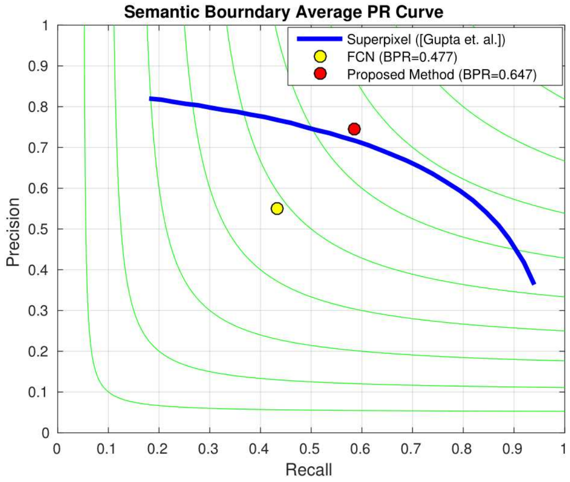

Figure 1: Precision-recall curve on semantic boundaries on the NYUDv2 dataset.

Table 1: The performance of oracle case using groundtruth to label the regions.

Groundtruth	Pixel Acc.	Mean Acc.	Mean IoU	f. w. IoU
Current Frame	96.2	94.0	90.2	92.7
Next Frame	84.7	76.2	63.4	74.4

2. Oracle performance using groundtruth labels

We perform two best-case analysis by computing an oracle performance where groundtruth labels are available for either reference or target frames. The first row of Table 1 shows the achievable performance by performing a majority vote of the groundtruth pixel labels on the employed superpixels from [3]. Thereby we achieve an upper bound of 96.2% on the pixel accuracy that is implied by the superpixel over-segmentation. In order to evaluate the effectiveness of our region correspondence, we use groundtruth labels of reference frames in the sequence. We collect 143 views to conduct this experiment in NYUDv2, which have corresponding regions in target frames. We ignore regions without correspondence in the next frame to compute the quantitative results, which are presented in Table 1. This best-case analysis for correspondence results in a pixel accuracy of 84.7%. Both oracle performances indicate a strong potential for performance improvements in our setup in all 4 reported measures.

3. Groundtruth analysis

At a closer look, it turns out that at least part of the performance loss in the best-case analysis for the correspondence is not due to bad matches between regions. In Fig. 2, we present some examples of the annotations provided in the dataset. In several cases, as the ones shown in the figure, the labeling is inconsistent and object labels are changed during the sequence. From left to right in Fig. 2, table changes to desk, table changes to dresser, floor changes to floor mat, bookshelf changes to shelves, cabinet changes to other-furniture, and window changes to blinds. Consequently, we see mistakes in the last two rows corresponding to the best case results due to inconsistent labelings.

4. Qualitative results on NYUDv2 4-class and 13-class task

We provide the qualitative results of 4-class and 13-class tasks of NYUDv2 dataset in Figure 3 and Figure 4 respectively.

5. Qualitative results on NYUDv2 40-class task

We provide more qualititative results in the following figures. We pick up some major scene categories from the test set including bedroom (Figure 5), living room (Figure 6), dining room (Figure 7), kitchen (Figure 8), bathroom (Figure 9), office (Figure 10) and classroom (Figure 11).

6. Failure cases

In this section, we present some failure cases of our methods in Figure 12. In those views, our method does not achieve better result. In the first two rows, we cannot segment the regions marked with white bounding box. This is because the superpixel in this two views cannot successful segment the regions. We use the same parameter for all views, so it fails to provide good superpixels for our system, but we believe that it is not difficult to get better superpixels for those failure views by adjusting the parameter of superpixel. In the third and fourth rows, we recognize the region as "cabinet" and "floormat" while groundtruth are "dresser" and "floor", which are also difficult for human beings to classify. In the last two rows, we show some challenges, which make our system fail to correctly recognize the region.

References

[1] Galasso, F., Nagaraja, N.S., Cardenas, T.J., Brox, T., Schiele, B.: A unified video segmentation benchmark: Annotation, metrics and analysis. In: ICCV. (2013) 1

[2] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. TPAMI (2011) 1

[3] Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: ECCV. (2014) 1

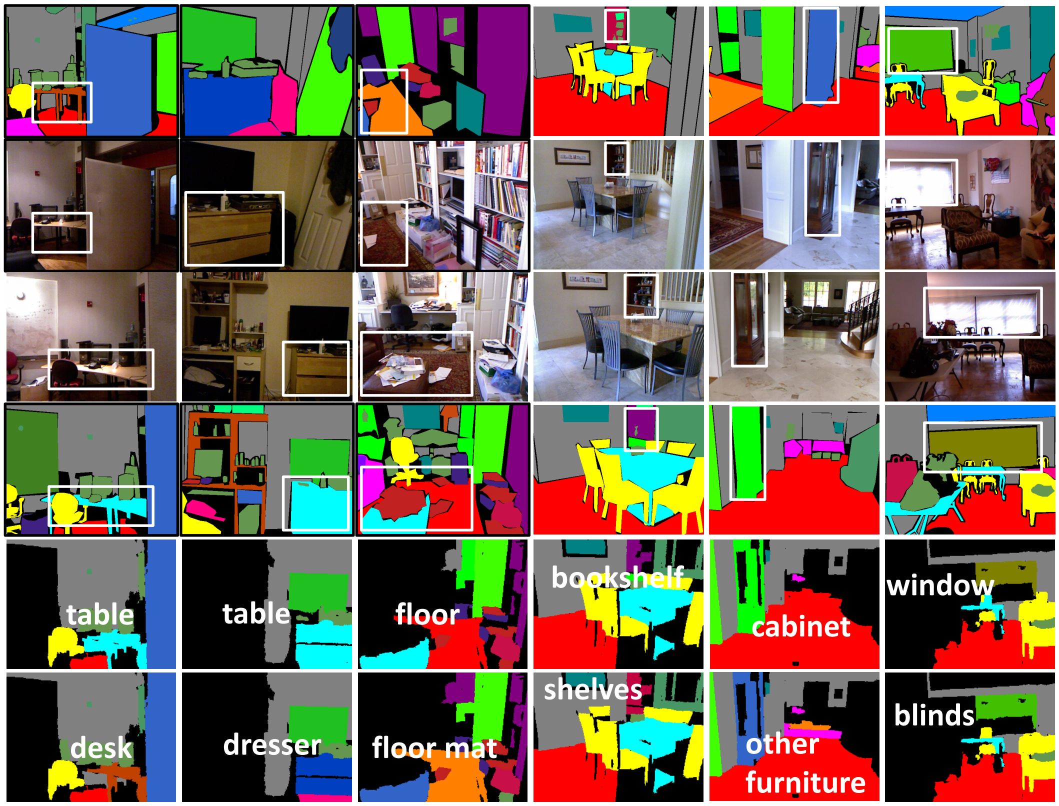

Figure 2: Example of groundtruth limitation and segmentation results of oracle case. Row 3 and 2 draw color images of target frame and next labeled frame, respectively. And row 4 and 1 draw their groundtruth. The segmentation result with groundtruth of target frame is shown in row 5, and the result with groundtruth of next frame is shown in row 6. We point out the regions in different frames with white bounding box, which are the same object of different views but labeled as different classes.

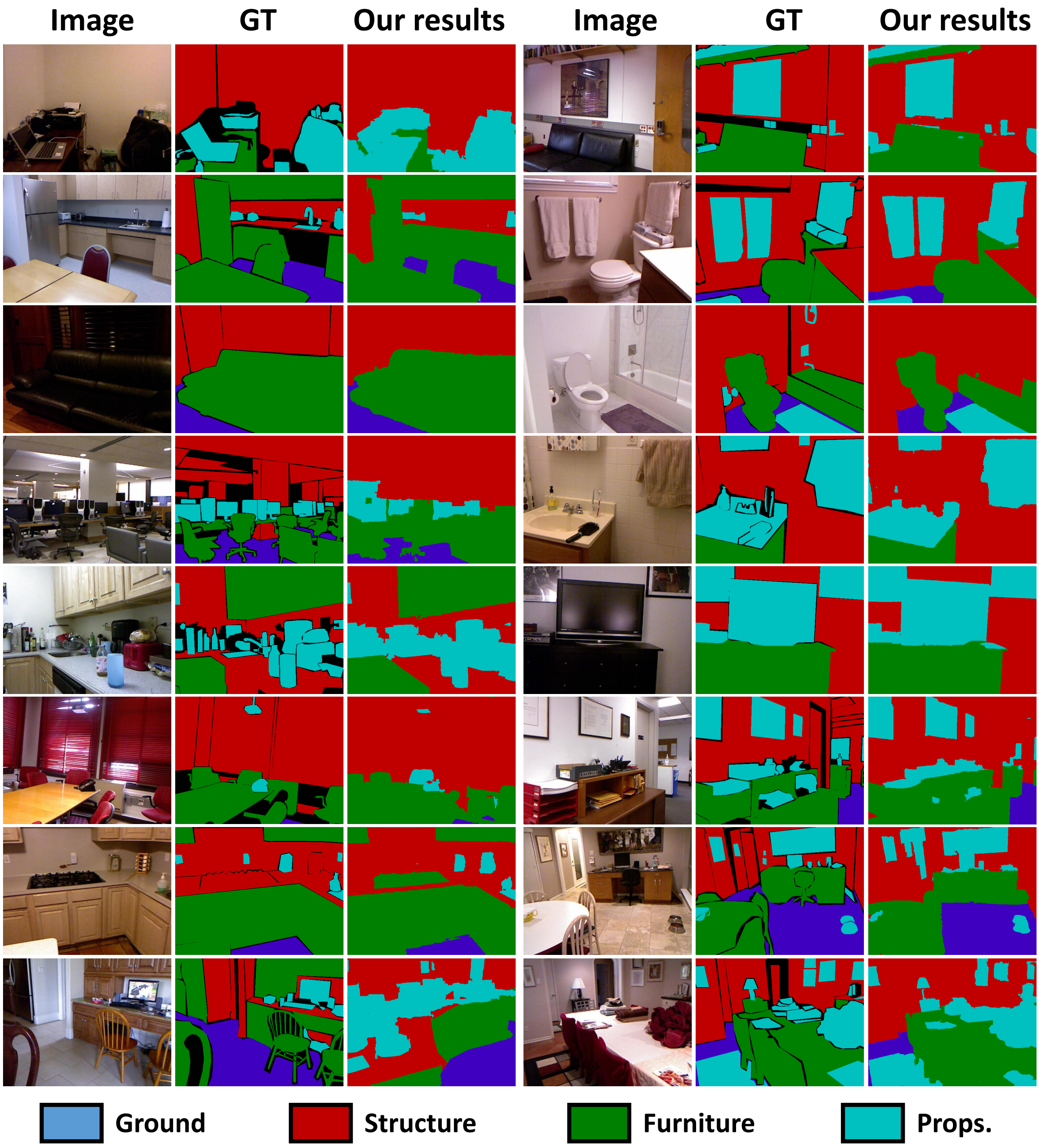

Figure 3: Semantic segmentation results of 4-class task on NYUDv2.

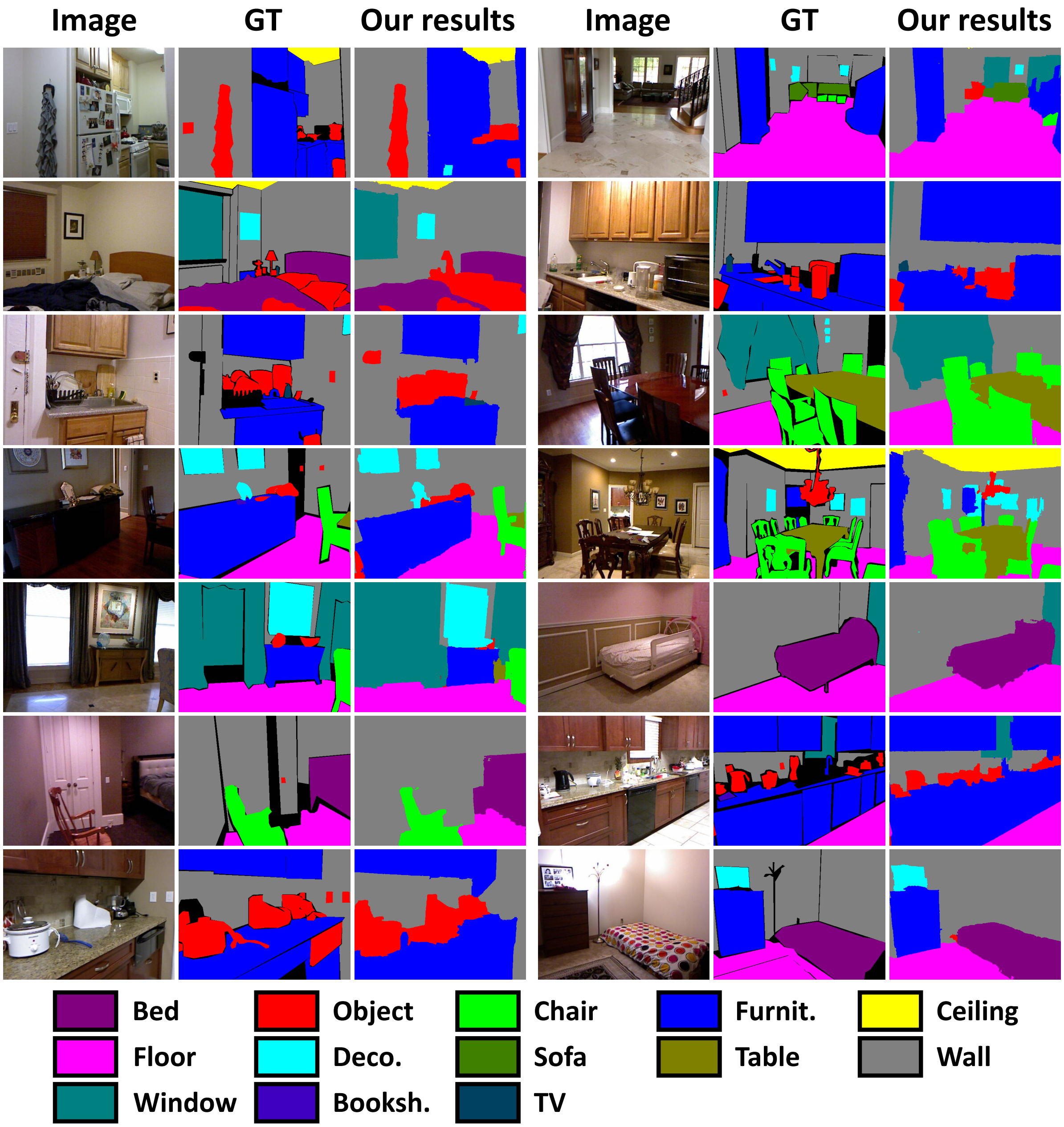

Figure 4: Semantic segmentation results of 13-class task on NYUDv2.

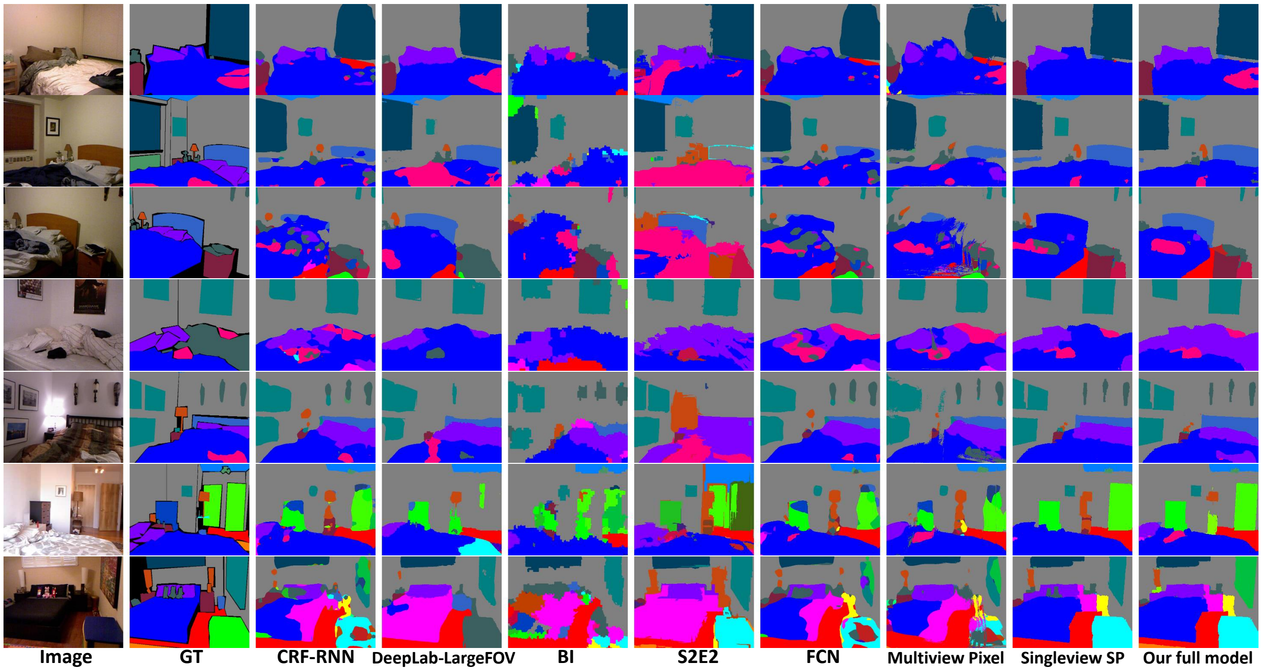

Figure 5: Semantic segmentation results of bedroom scenes on NYUDv2.

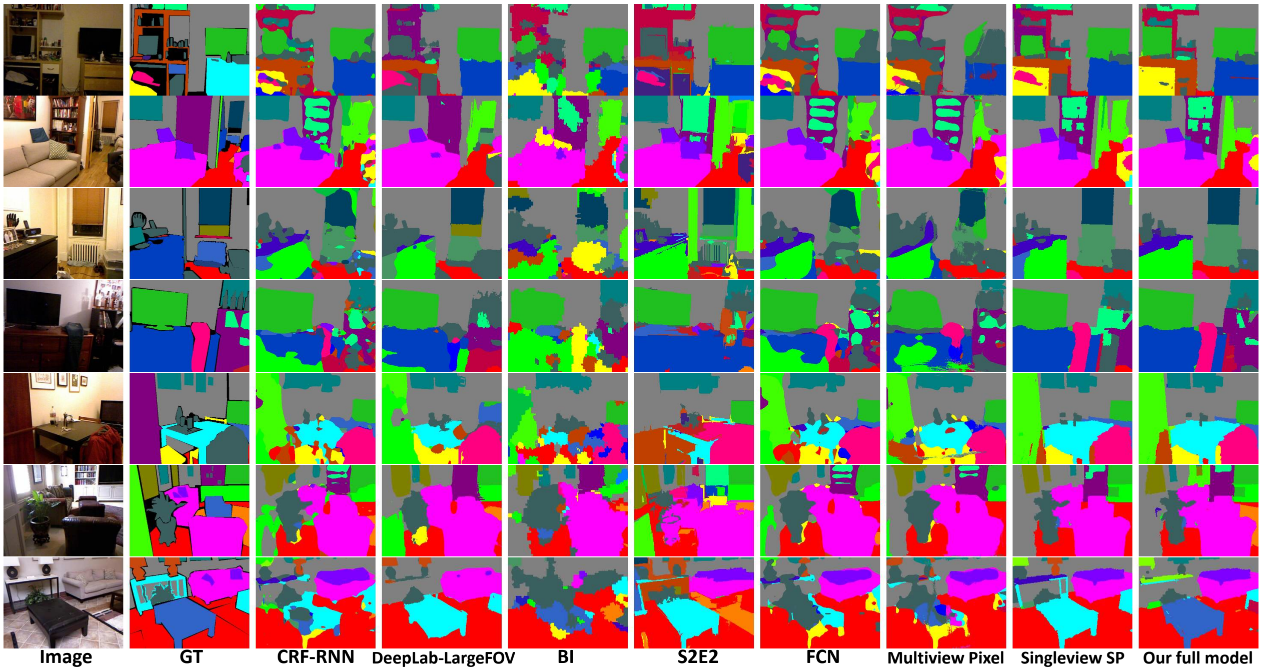

Figure 6: Semantic segmentation results of living room scenes on NYUDv2.

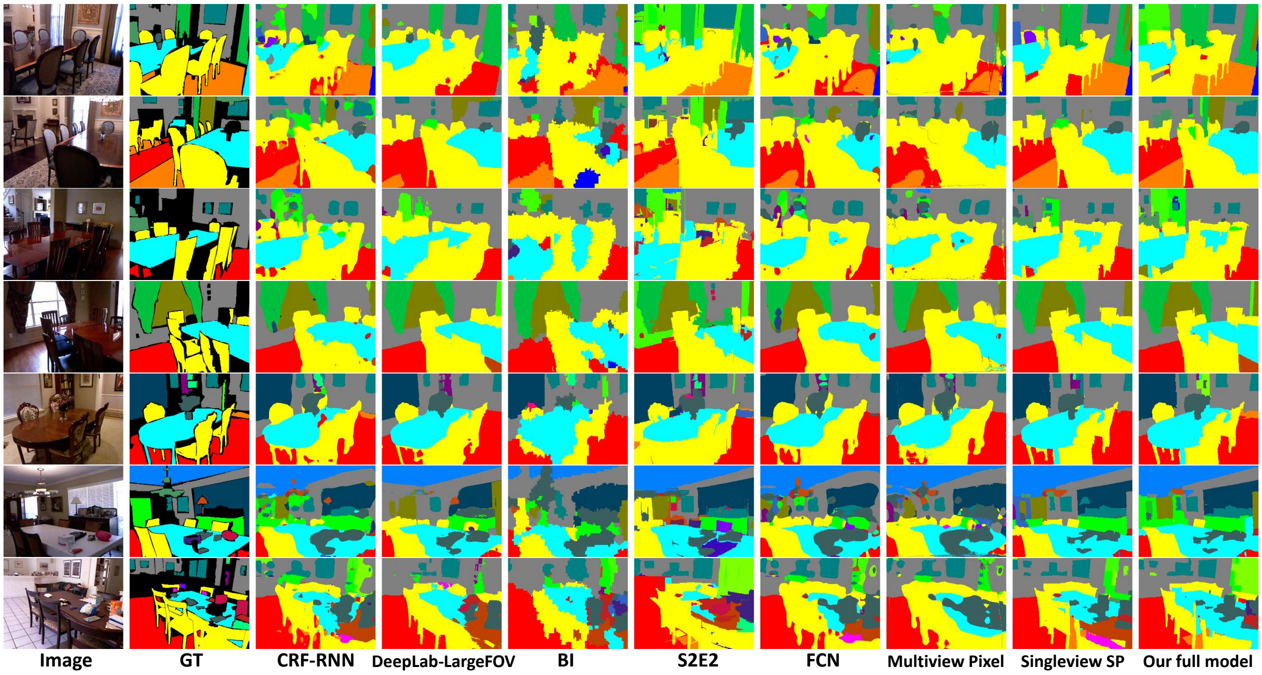

Figure 7: Semantic segmentation results of dining room scenes on NYUDv2.

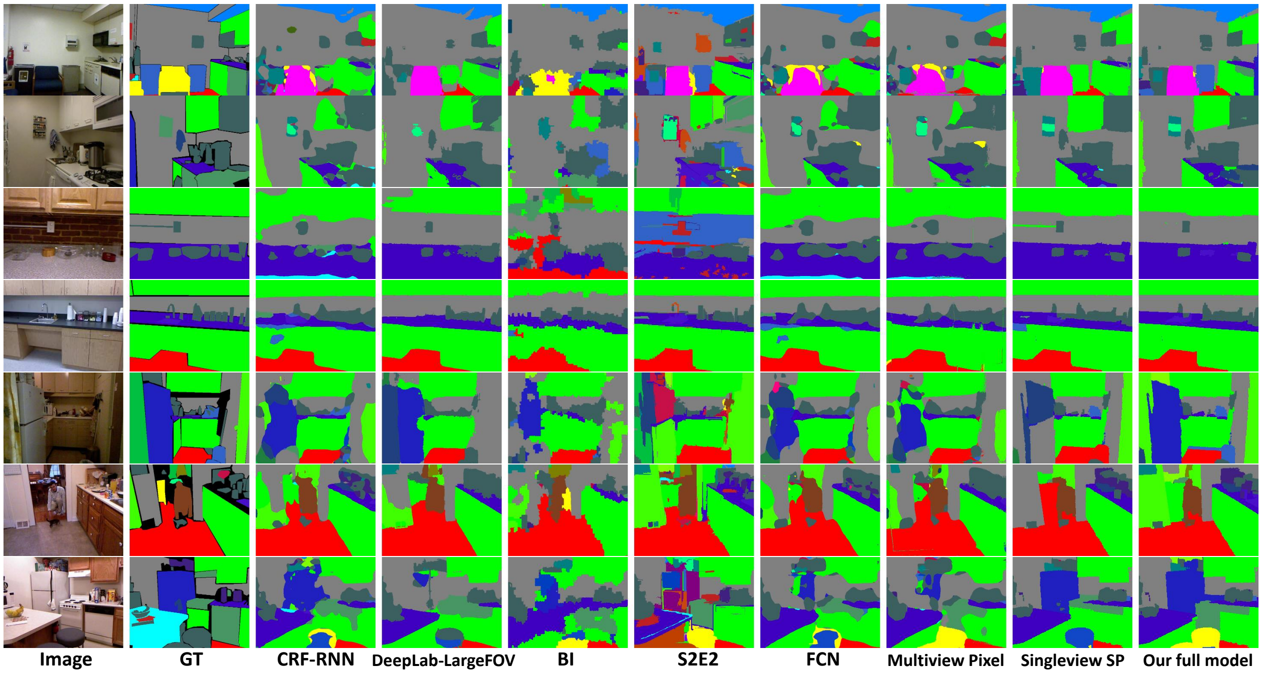

Figure 8: Semantic segmentation results of kitchen scenes on NYUDv2.

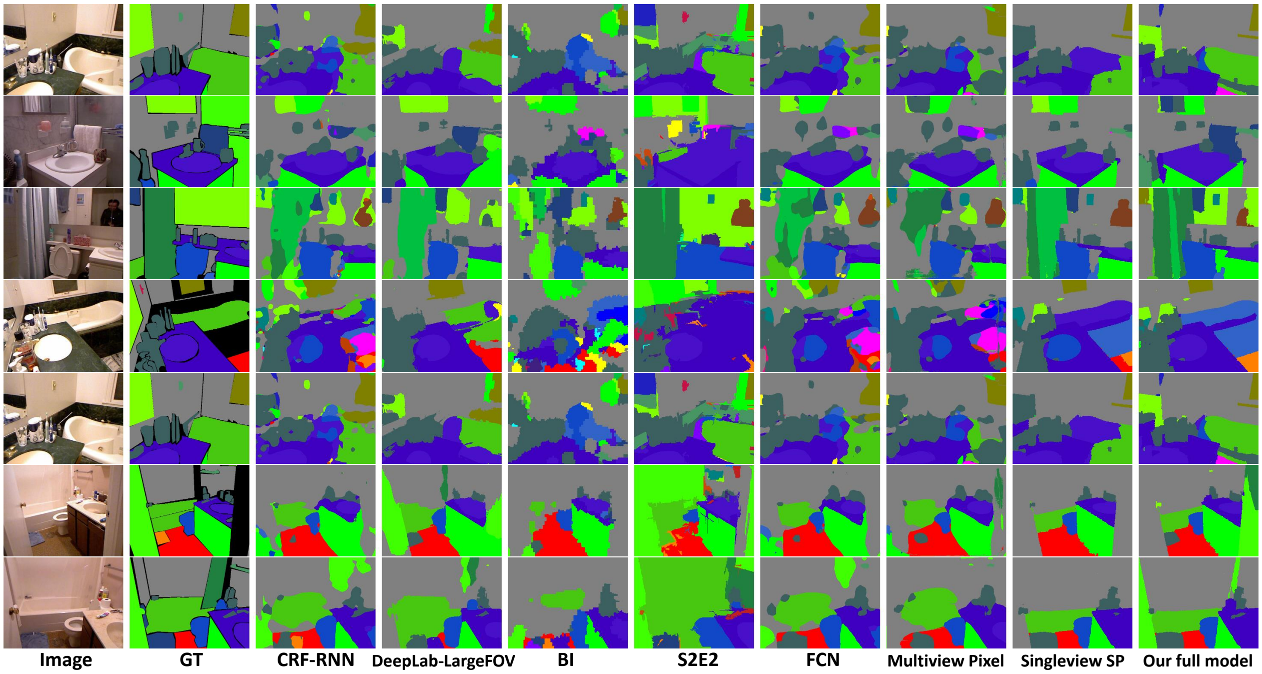

Figure 9: Semantic segmentation results of bathroom scenes on NYUDv2.

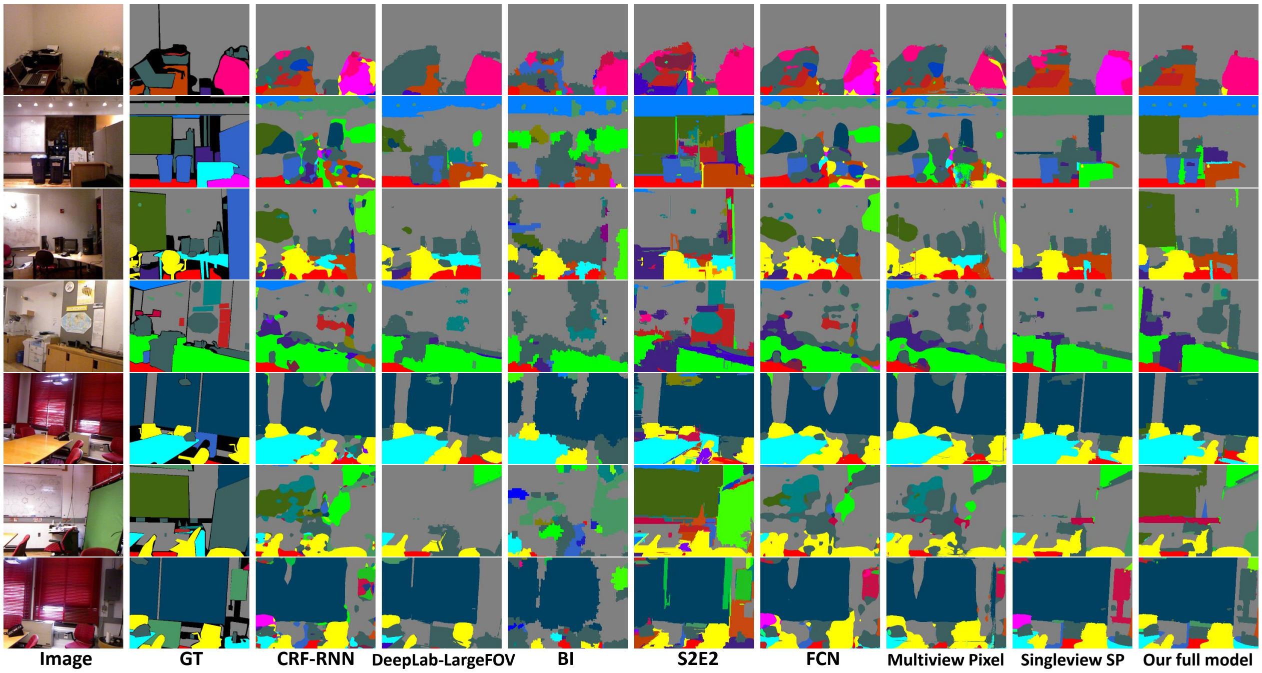

Figure 10: Semantic segmentation results of office scenes on NYUDv2.

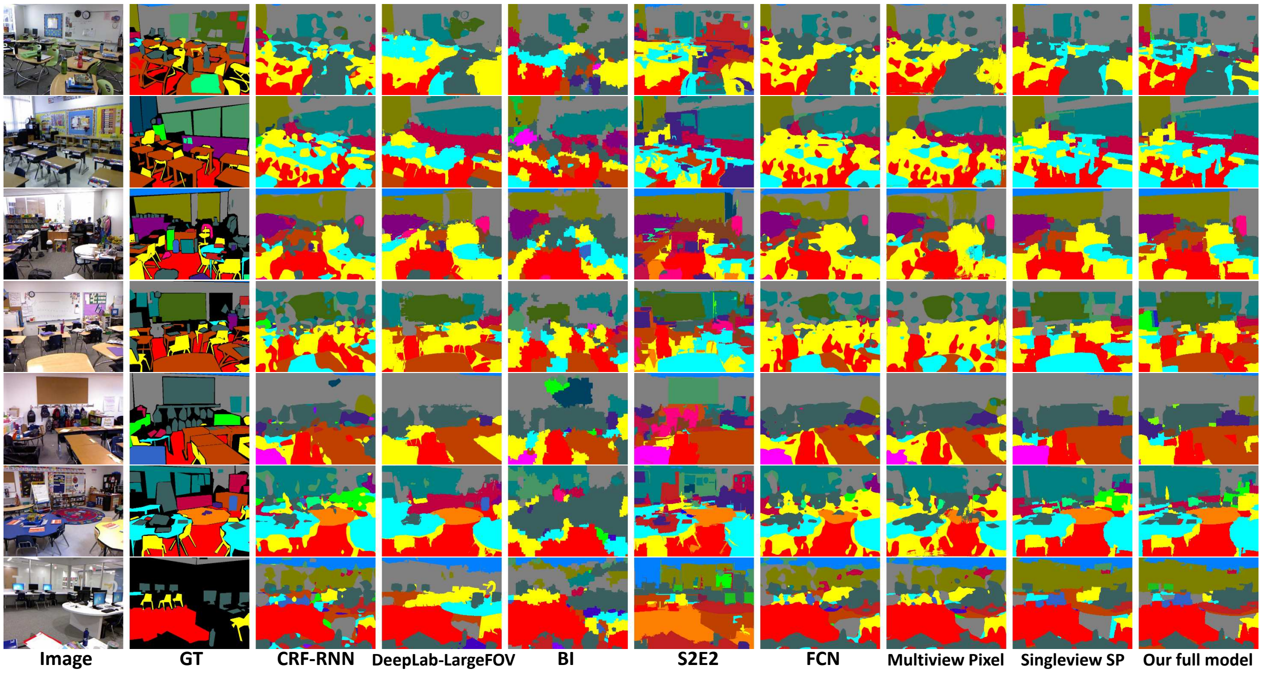

Figure 11: Semantic segmentation results of classroom scenes on NYUDv2.

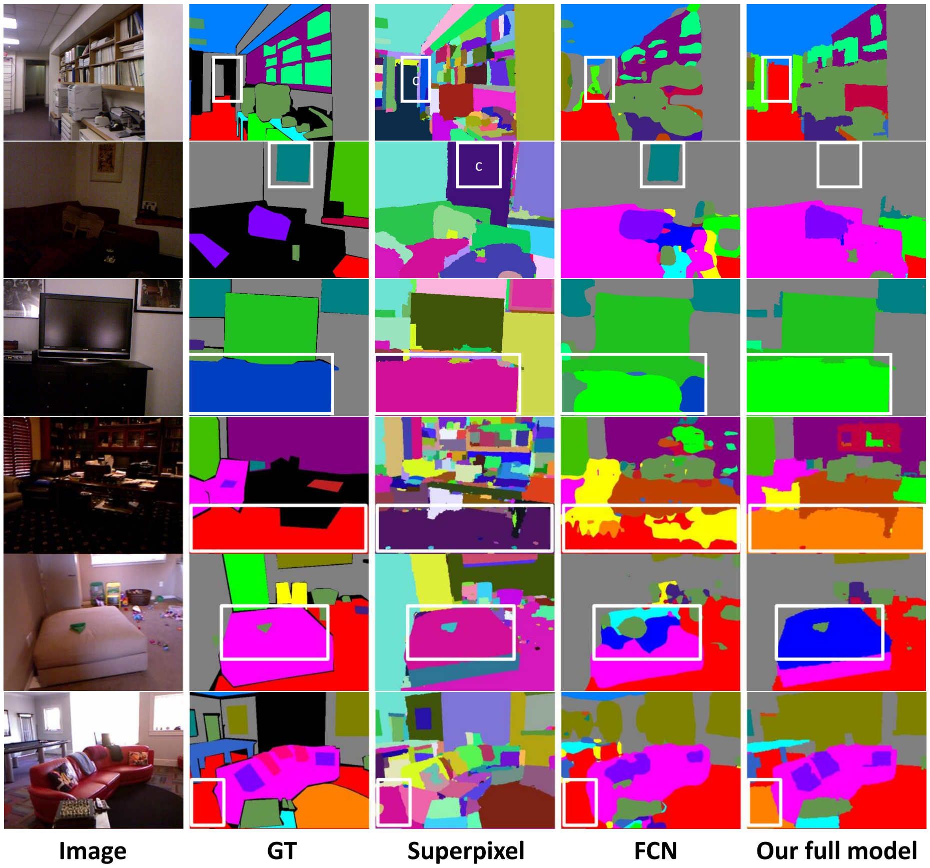

Figure 12: Some failure cases that our method is not able to improve FCN.