VoxelSplat: Dynamic Gaussian Splatting as an Effective Loss for Occupancy and Flow Prediction

Supplementary Material

1. Supplementary

In the supplementary material, we provide additional details to complement the main paper. These include:

- **Deeper Analysis of Rendering Losses:** An exploration of the impact of rendering losses on the convergence of 3D occupancy and scene flow.
- **Visualization of Rendering Results:** Examples of rendering outputs on the validation set, illustrating what the rendering branch learns after training.
- **Additional Qualitative Results:** A demonstration of the predicted 3D occupancy and scene flow through multi-view video visualizations, showcasing the quality of our method.

1.1. Deeper Analysis of Rendering Losses

In Fig. 1, we compare the occupancy and flow loss curves with and without the rendering loss $\mathcal{L}_{2D}$.

Detailed Experimental Settings. We conduct our loss curve experiments based on the model architecture of FB-Occ [3]. Following the original settings, the occupancy loss $\mathcal{L}_{occ}$ consists of *cross-entropy loss*, *Lovász-Softmax loss* [1], and *scaling loss*. As mentioned in the main paper, we employ the L1 loss as the scene flow loss function $\mathcal{L}_{flow}$. To prevent training collapse, we start computing the flow loss at 3500 iterations, after which FB-Occ begins using temporal information. We train the model with and without our rendering loss $\mathcal{L}_{2D}$ for 70,000 iterations and compare the convergence of the loss curves.

Effect of Rendering on 3D Losses. From the upper figure in Fig. 1, we observe that $\mathcal{L}_{occ}$ converges faster with the inclusion of $\mathcal{L}_{2D}$. In the middle figure, the flow loss $\mathcal{L}_{flow}$ starts converging after 40,000 iterations. This is likely due to the small proportion of dynamic objects in the scenes, which makes it challenging for the model to capture motion information. However, with our $\mathcal{L}_{2D}$, which specifically addresses dynamic objects, the $\mathcal{L}_{flow}$ converges significantly faster.

This experiment demonstrates that our strategy of explicit modeling of the occupancy field with 3D Gaussians and splat rendering supervision helps the original loss functions find a better convergence direction.

1.2. Visualization of Rendering Results

Although our rendering branch is not used during inference, we conduct a simple visualization experiment on the validation set of [2] to help understand what the rendering branch learns during training. Specifically, based on FB-Occ [3], 640,000 semantic Gaussians initialized from all voxel centers are predicted by the decoder in the rendering branch. Gaussians with opacity higher than 0.2 are splatted into the camera view. The rendering semantics and depth results in Fig. 2 demonstrate that our rendering branch successfully predicts high-quality semantics and depths, even under adverse weather conditions.

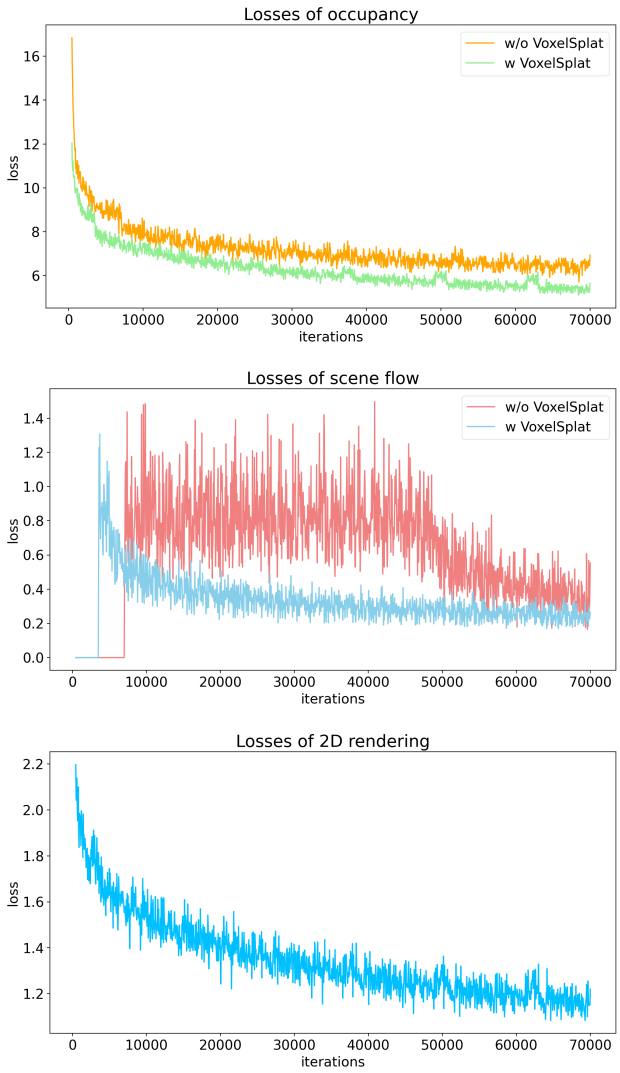

Figure 1. The comparison of loss curves with and without our VoxelSpat.

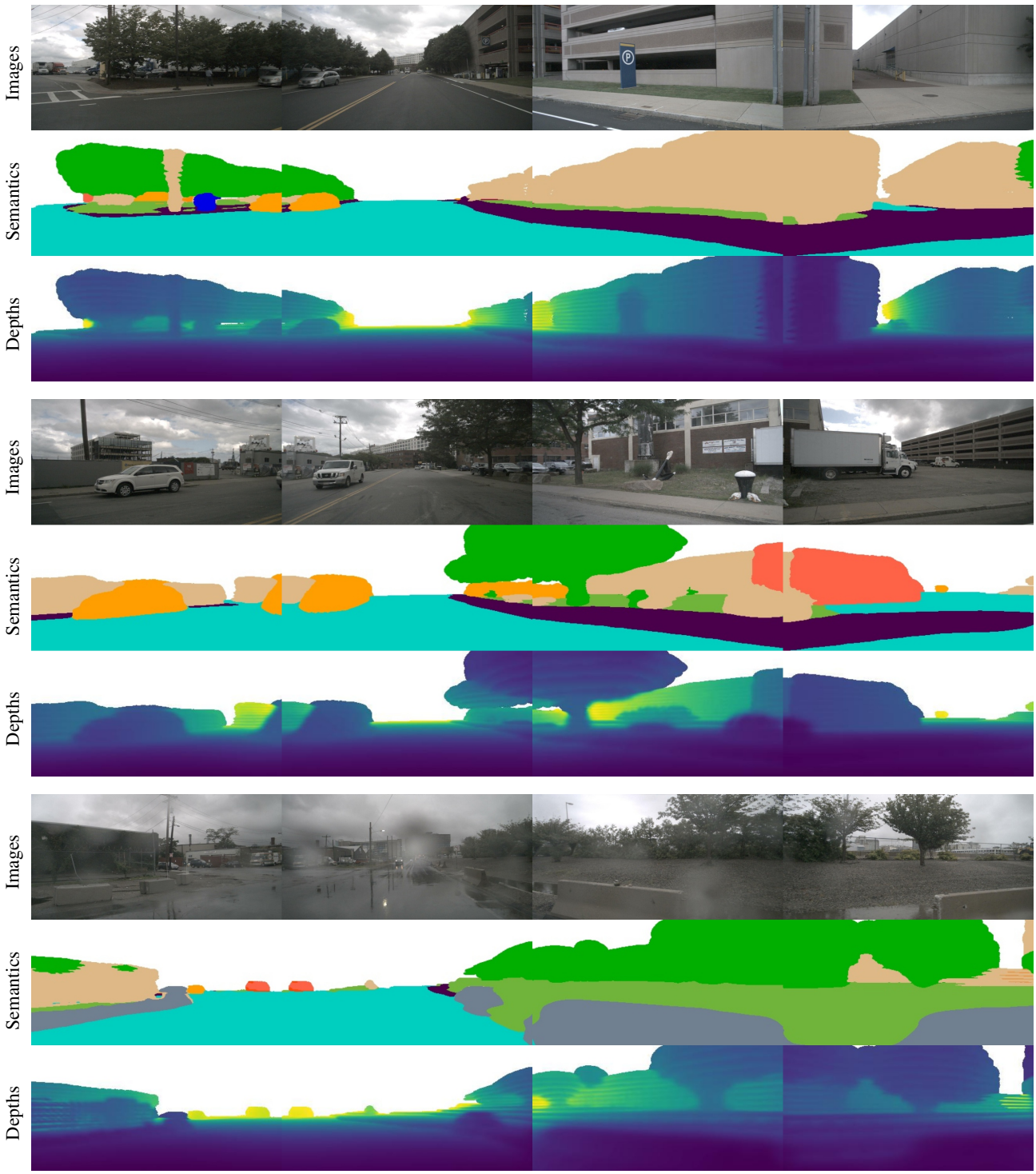

Figure 2. The visualization results of rendering semantics and depths on the validation dataset [2] are presented.

1.3. Additional Qualitative Results

In Fig. 3, we illustrate the image inputs and a more comprehensive visualizations of our predicted occupancy and flow from different viewpoints. Further, a series of videos are provided in the supplementary material to validate our accuracy and stability, which is crucial in self-driving safety.

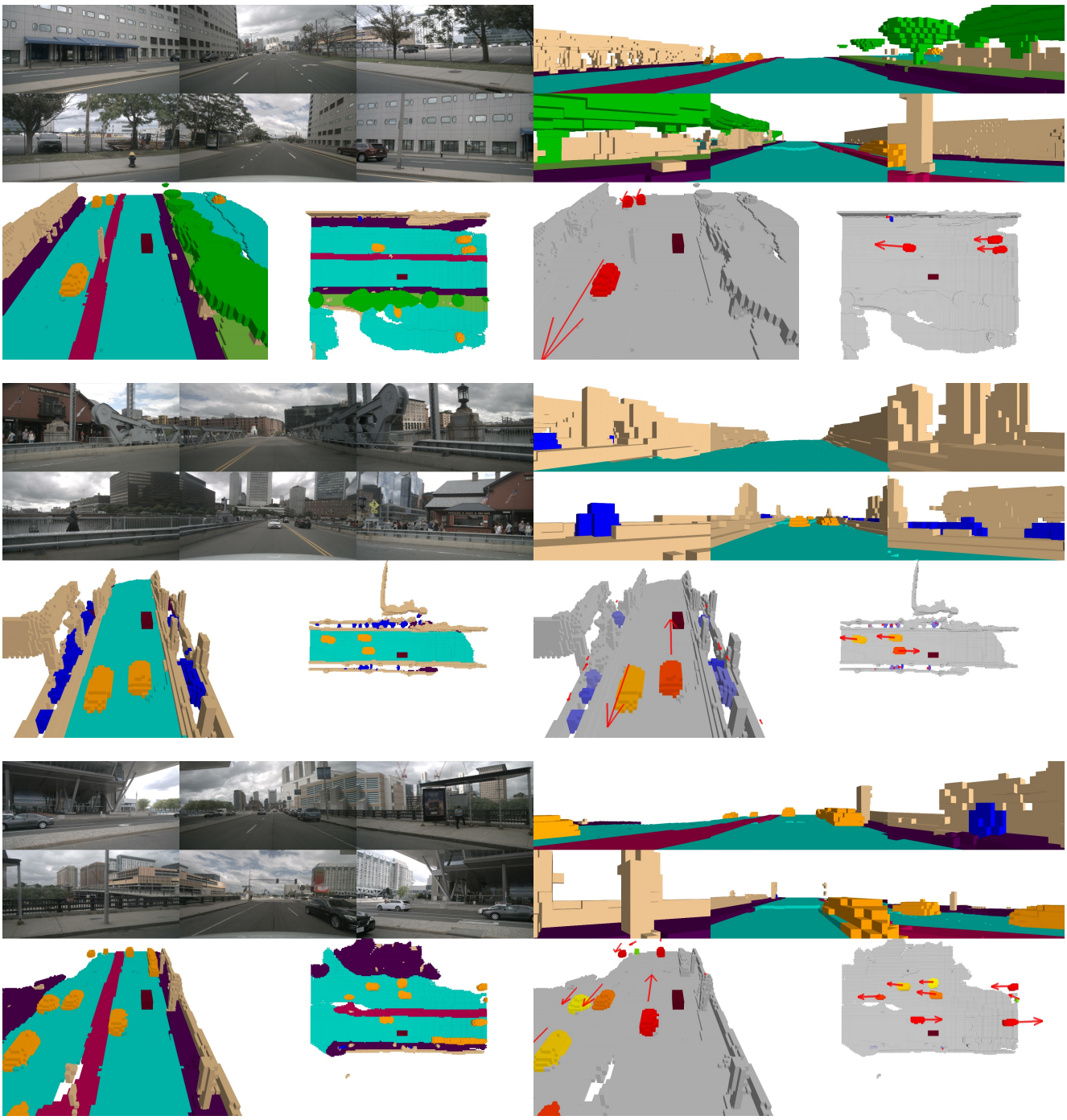

Figure 3. We provide a series of videos in the supplementary material, which demonstrate the predicted occupancy and flow from different viewpoints.

References

[1] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018. 1

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pages 11621–11631, 2020. 1, 2

[3] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 1