Towards Open-World Generation of Stereo Images and Unsupervised Matching

Supplementary Material

1. Implementation Details

We use Stable Diffusion 1.5 for a fair comparison with the previous methods. We also report results with SD v2.1. The fine-tuning of the original Stable Diffusion model requires approximately 70 hours across 3 epochs, utilizing a configuration of 2 NVIDIA A100 80GB GPUs. The training process employs a batch size of 6, with an initial learning rate set to 1×10^{-5}, following a linear warm-up schedule for the first 1000 steps. The model optimization is conducted using the Adam optimizer, characterized by parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. For Stable Diffusion v1.5 and v2.1, all datasets during both training and inference phases utilize consistently sized input images at 512×512 and 768×768, respectively. During inference, we use 50 inference steps and set the guidance scale to 1.5.

2. Cross Attention

The left view features are conditioned on the concatenated input (I_l, C_l), while the right view features are conditioned on the concatenated input (I_{warp}, C_r). As illustrated in Fig. 1, we enhance the original self-attention mechanism by incorporating an additional cross-attention, which significantly enhances feature representation. By aggregating both attention mechanisms simultaneously, we achieve not only computational efficiency but also empower the model to effectively determine where to generate and where to keep the same as the warped image.

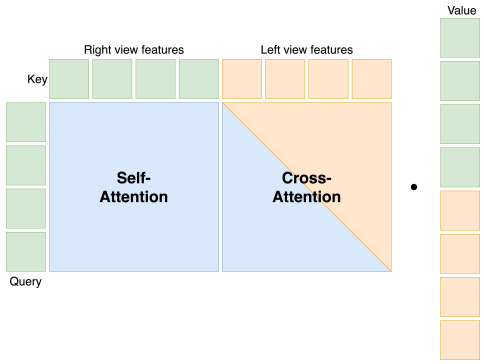

Figure 1. Cross Attention.

In our approach, we meticulously select the training datasets, deliberately excluding real-world datasets due to their potential adverse effects on model performance. The challenges associated with calibrating dual cameras and depth sensors (e.g., LIDAR) necessitate this exclusion. Furthermore, we implement a filtering process to remove images with inaccurate disparity ground truth from the synthetic datasets, ensuring the integrity and quality of the training data.

3. Visualizations

3.1. Fine-tune Performance

In this work, we fine-tune all models for 3 epochs, which is sufficient to achieve a competent model performance. As shown in Figure Fig. 3, the evaluation metrics exhibit a slight improvement as the number of training epochs increases. These two datasets are not used during the training process, demonstrating the robustness and generalization of our model.

3.2. Qualitative Results of Generation

In Figure Fig. 5, we provide a detailed qualitative comparison of our proposed method against several state-of-the-art diffusion-based generation techniques on the KITTI 2015 dataset. This comparison facilitates an in-depth analysis of performance across a range of scale factors, showcasing the robustness of our approach. Furthermore, Figure Fig. 6 demonstrates the generation performance across a diverse set of datasets, underscoring the generalization capabilities and overall effectiveness of our method in various contexts.

3.3. Qualitative Results of Stereo Matching

Figures Fig. 7, Fig. 8, and Fig. 9 present a comprehensive comparison of various unsupervised stereo matching methods. Our results highlight not only the superior performance of our approach but also its robustness across diverse datasets, underscoring its effectiveness in real-world applications.

3.4. Qualitative Comparison with NeRF-Stereo

NeR-Stereo trains NeRF on multi-view static scenes, limiting compatibility with single-view (Middlebury) or dynamic (KITTI) datasets. Besides, as illustrated in Fig. 2, it also struggles with scene boundaries, and noisier disparity maps than monocular depth estimation methods.

Figure 2. Comparison with NeRF-Stereo.

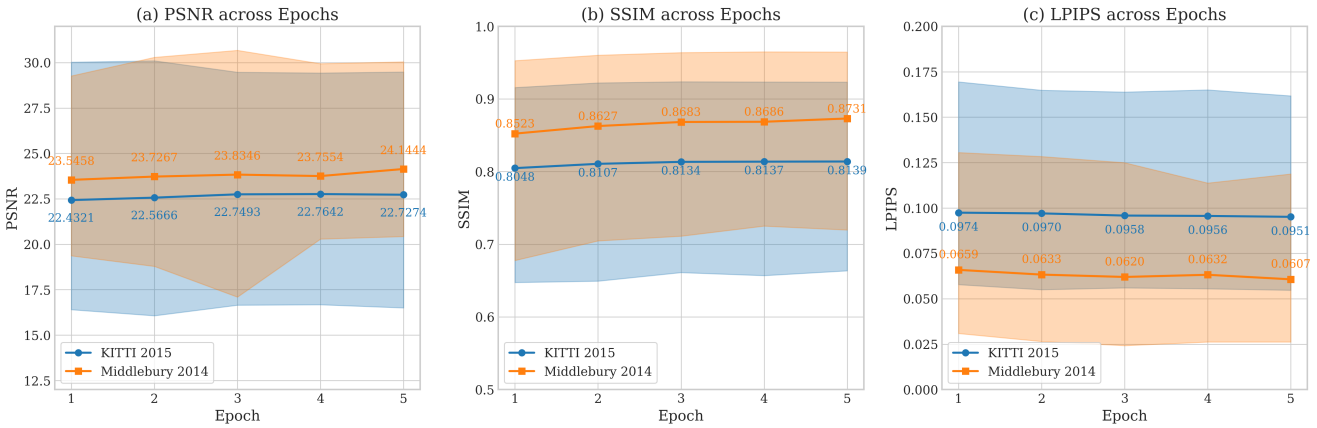

Figure 3. Evolution of image quality metrics across training epochs on the KITTI 2015 and Middlebury 2014 datasets.

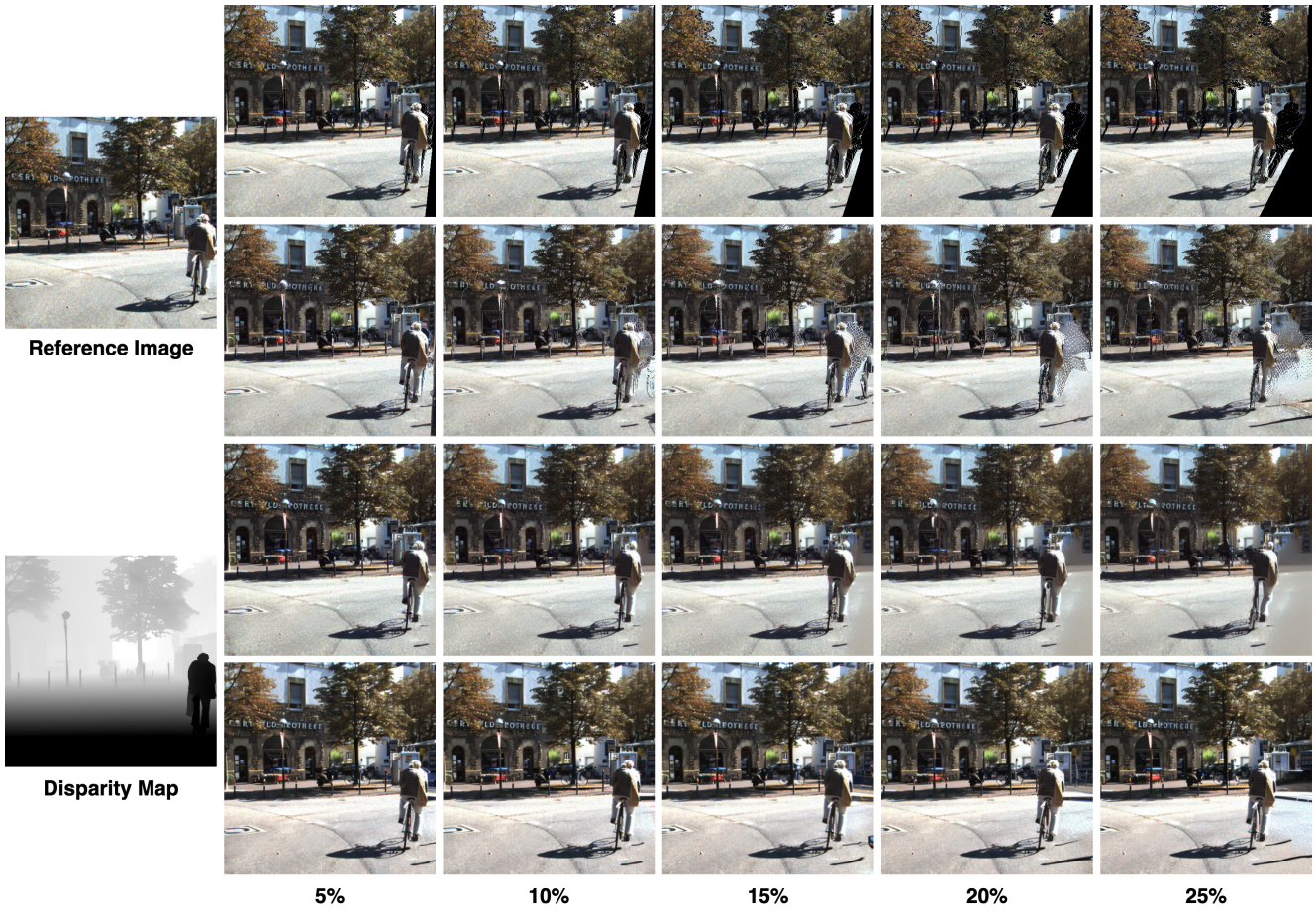

Figure 4. Performance of our generation method on the KITTI 2015 dataset. The scale factor varies from 5% to 25%. From top to the bottom are leave blank, SD-Inpainting, StereoDiffusion, and ours.

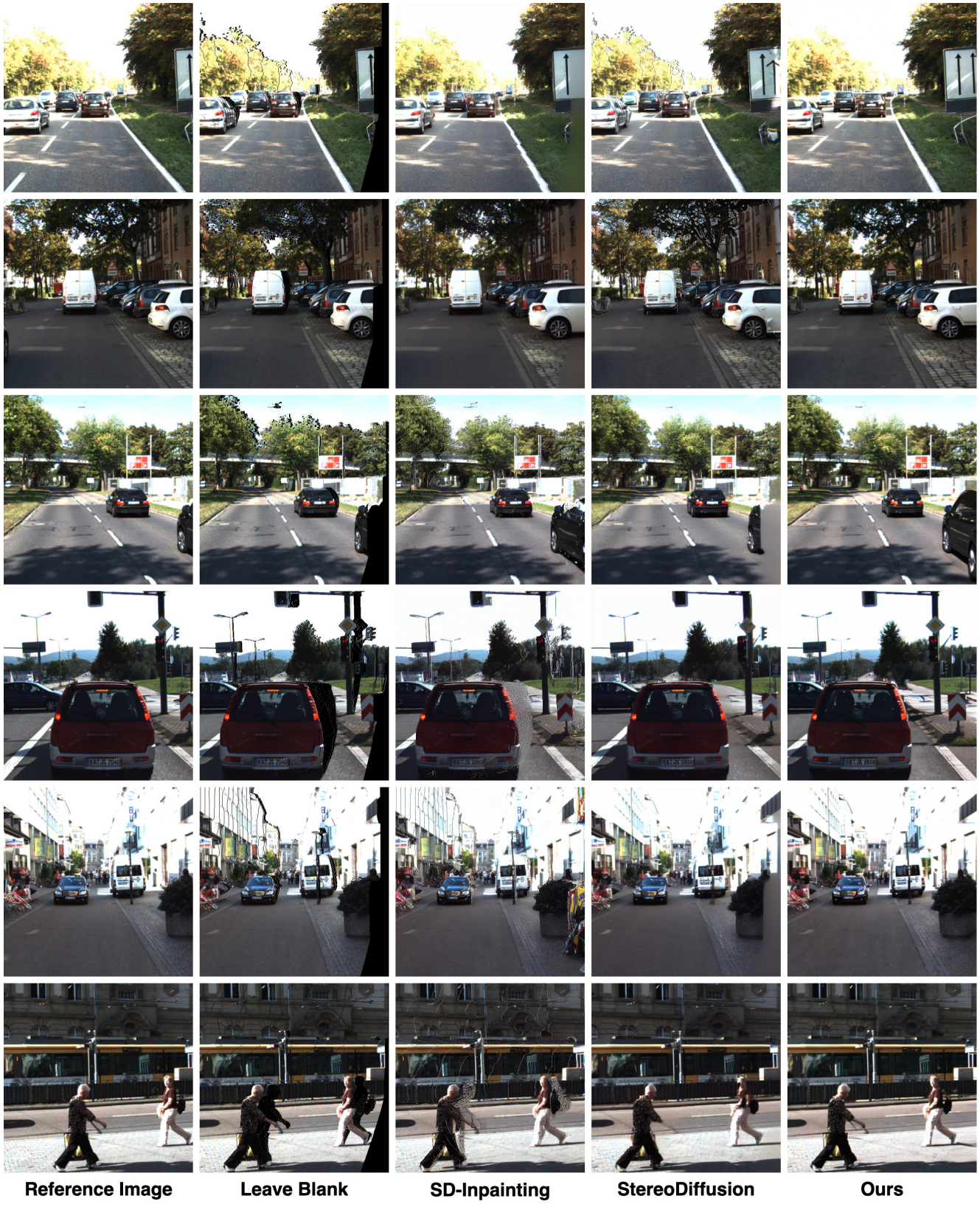

Figure 5. Qualitative comparison on KITTI 2015 with ground truth disparity maps.

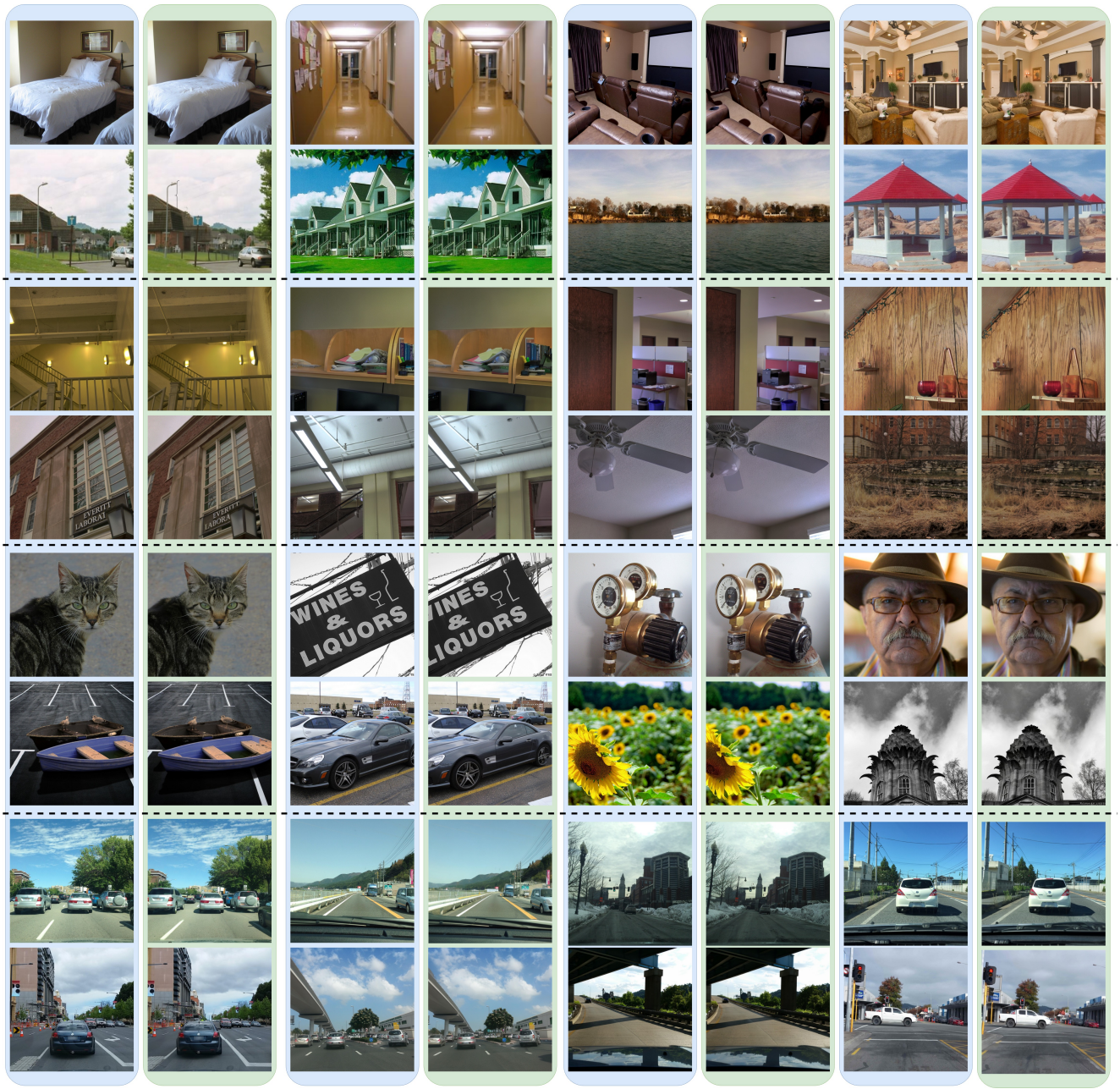

Figure 6. Performance of our generation method across various datasets, including ADE20K, DOIDE, Depth in Wild, and Mapillary Vistas. The images in the blue blocks represent the reference left images, while the images in the green blocks depict the generated right images.

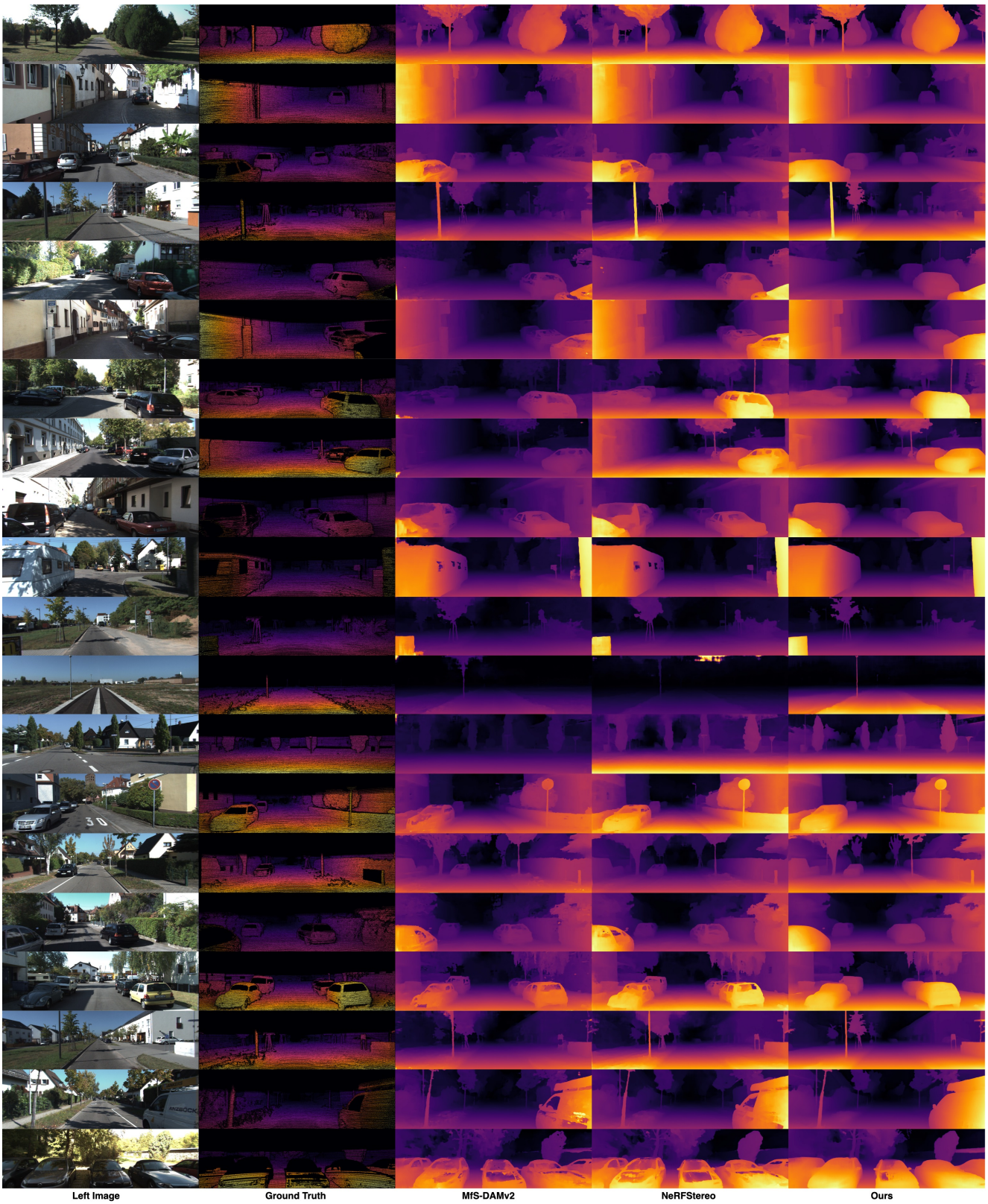

Figure 7. Qualitative comparison of unsupervised stereo matching methods on KITTI 2012 using PSMNet.

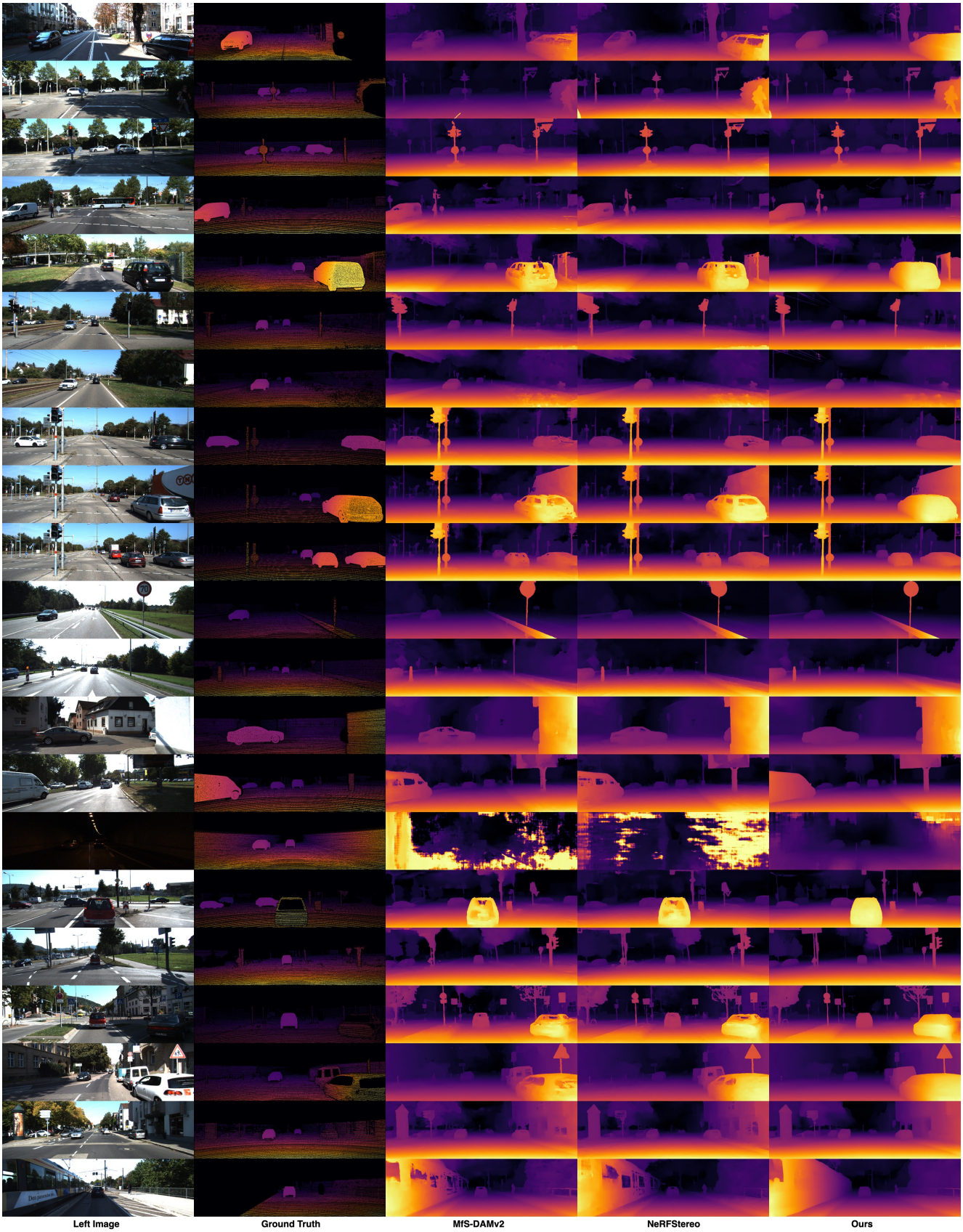

Figure 8. Qualitative comparison of unsupervised stereo matching methods on KITTI 2015 using PSMNet.

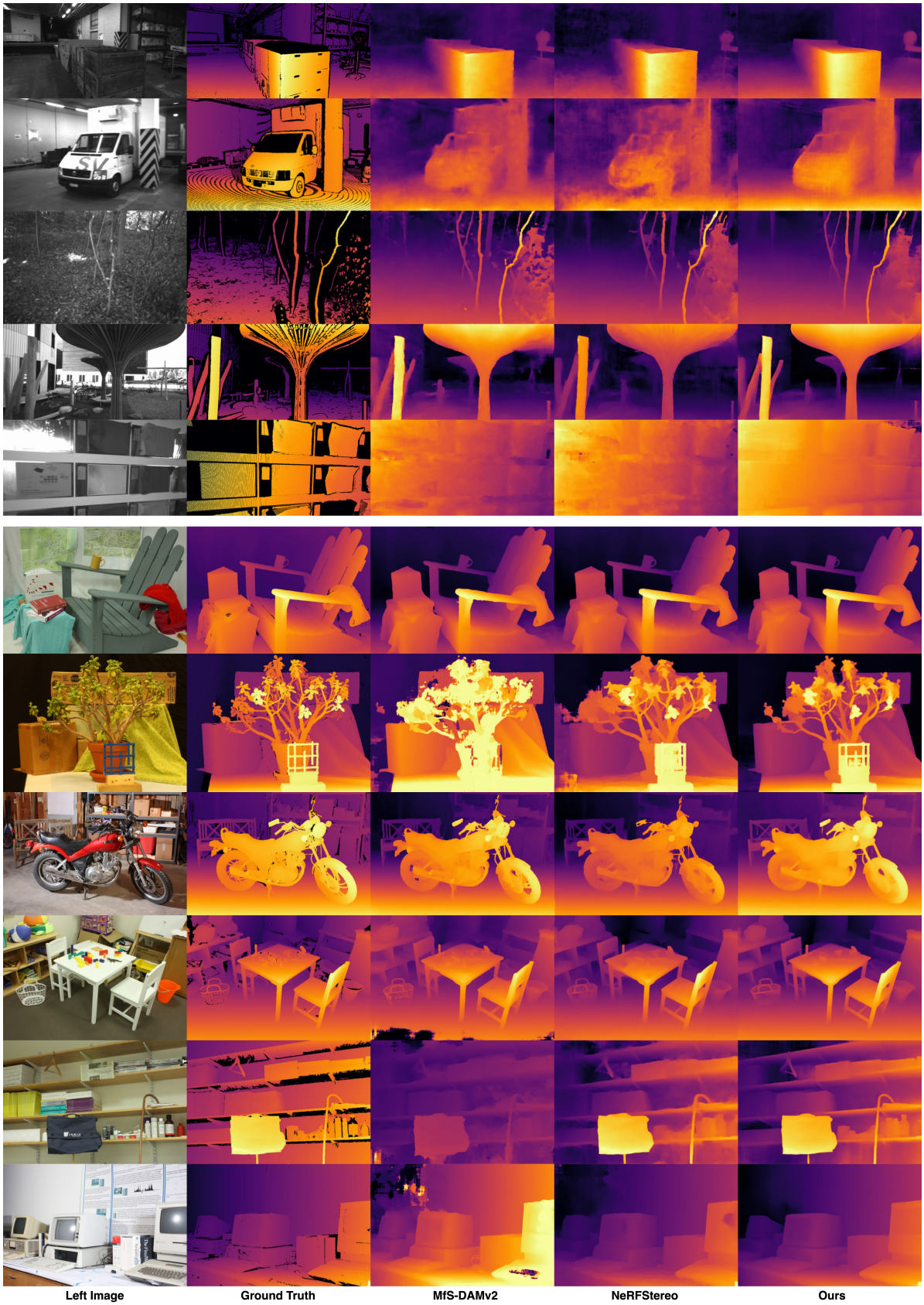

Figure 9. Qualitative comparison of unsupervised stereo matching methods on ETH3D and Middlebury using PSMNet.