IDESplat: Iterative Depth Probability Estimation for Generalizable 3D Gaussian Splatting

Supplementary Material

In this supplementary material, we provide additional details on model training, model architecture, ablation study on DPBU, experimental results on the DL3DV dataset, and more visual comparison results. Specifically, in Section A, we present the training details of the IDESplat model. In Section B, we provide the model architecture details of IDESplat and Warp-Index Epipolar Attention. In Section C, we present the experimental results of IDESplat on the DL3DV dataset, along with the ablation study on Depth Probability Boosting Units. Finally, in Section D, we include more visual comparison results for novel view synthesis and depth prediction.

A. Training Details

For a fair comparison, we trained our IDESplat using a standard setup [5, 21, 47]. The training was done on 8 RTX 4090 GPUs with a batch size of 16, using the AdamW [24] optimizer for 300,000 iterations, which took approximately 3 days. We used a cosine learning rate schedule. For the pre-trained Depth Anything V2 [50] backbone, the learning rate was 2×10^{-6}, while other layers were trained with a learning rate of 2×10^{-4}. The network was trained with a combination of MSE and LPIPS losses between the rendered and ground truth images. Following [47], for the newly added DL3DV dataset, we trained at a resolution of 256×448. First, we pre-trained on RE10k and then fine-tuned on the DL3DV dataset for 100K iterations, with a total batch size of 4, and the number of input views was randomly sampled from 2 to 6. During inference, we evaluated the model's performance on different numbers of input views.

B. Model Architecture Details

We provide a detailed description of the IDESplat network architecture, as shown in Figure 6. It consists of three main parts: a feature extraction backbone, an iterative depth estimation process, and a Gaussian focus module. The backbone has two branches: a multi-view branch using the pre-trained Unimatch [44] and a monocular branch using the ViT-small version of DepthAnything V2 [50]. The outputs from both branches are fused to provide multi-view geometry and texture information for the next modules. The depth estimation process includes three Depth Probability Boosting Units (DPBU) that sequentially generate optimized depth results. Each DPBU contains two cascaded Warp-Index Epipolar Attention layers, which use the Hadamard product to enhance depth probabilities. The process is repeated for six transformations at resolutions of 64×64, 128×128, and 256×256. The GFM has six layers, using a shifting window strategy with a window size of 16. After each attention calculation, the top half of the most relevant Gaussian positions are retained. The number of retained Gaussian weights per layer is [256, 256, 128, 128, 64, 64], and the module uses 6 attention heads with 256 channels in total.

To address the memory issues in existing warp computations for cross-view similarity, we introduce Warp-Index Epipolar Attention. Unlike the existing method, which samples target view features for each depth candidate and consumes a lot of memory, our approach only stores transformation indices for similarity matrix multiplication and uses Sparse Matrix Multiplication (SMM) for efficient computation. This design enables IDESplat to perform multiple rounds of warp and depth estimation more efficiently.

C. More Experimental Results

We conducted additional experiments on the DL3DV dataset to further evaluate the proposed IDESplat method. DL3DV is a large-scale real-world multi-view video dataset, which helps validate our method's reconstruction capability in more complex and larger scenes. The experimental results, shown in Table 7, demonstrate outstanding performance of our method compared to existing MVSplat and DepthSplat methods. IDESplat outperforms DepthSplat by **0.62dB**, **0.41dB**, and **0.42dB** when using 2, 4, and 6 input views, respectively. These results clearly show that our IDESplat provides better reconstruction performance in large, complex scenes with multiple input views compared to existing methods.

We also conducted ablation experiments on DPBU with different numbers of Warp-Index Epipolar Attention. The results in Table 8 show that depth probabilities can only undergo Multiplicative Boosting when more than one Warp-Index Epipolar Attention is used. Our IDESplat performs well when the number of attention layers exceeds one, and the performance improves as the number of layers increases. Considering both efficiency and performance, we chose the model with two Warp-Index Epipolar Attention layers.

D. More Visual Comparison Results

We also provide more visual comparison results in this supplementary material, as shown in Fig. 8 and Fig. 9. Through these qualitative visual comparisons, it can be observed that our IDESplat outperforms existing methods in novel view synthesis. Even in complex lighting and textured regions,

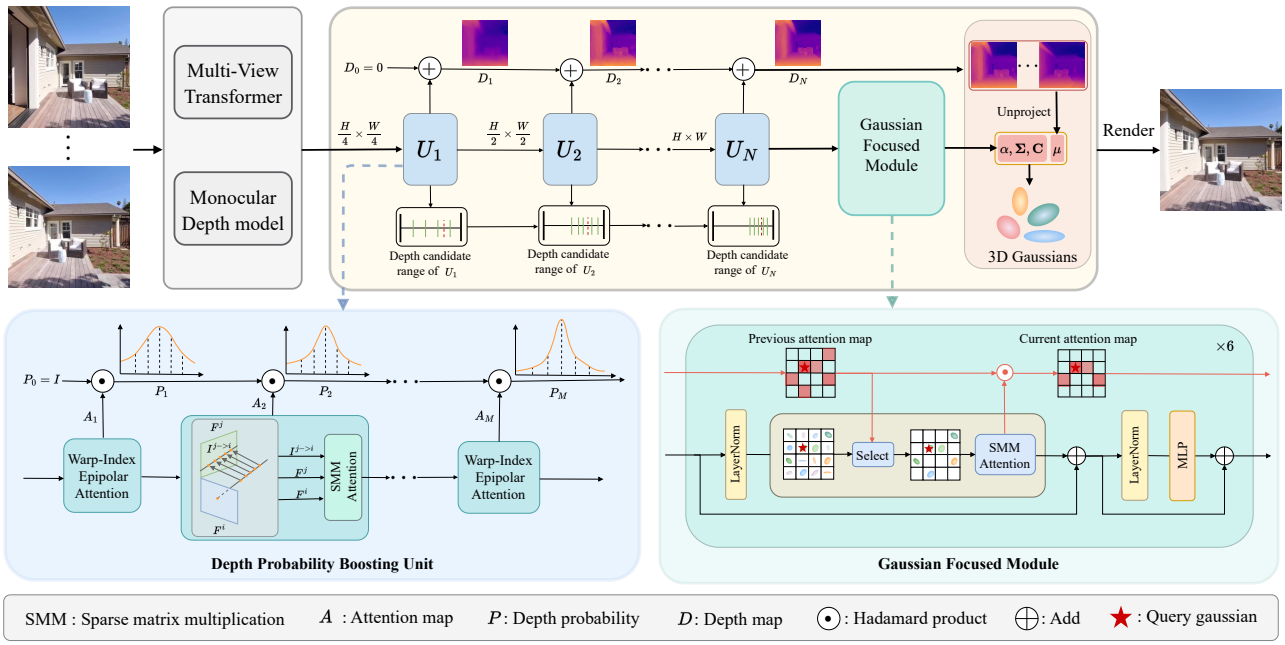
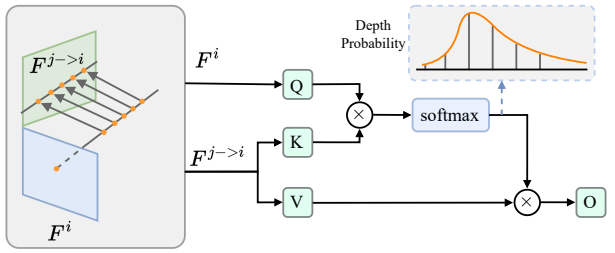
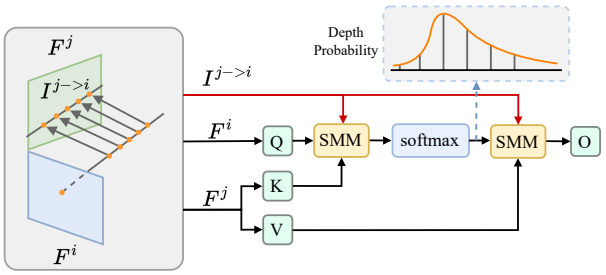


Figure 6. The architecture of IDESplat. IDESplat comprises a feature extraction backbone, an iterative depth estimation process with cascaded Depth Probability Boosting Units (DPBUs), and a Gaussian Focused Module (GFM).



(a): **Epipolar Attention**



(b): **Warp-Index Epipolar Attention**

Figure 7. The difference between Warp-Index Epipolar Attention and Epipolar Attention.

our method achieves better reconstruction results. Furthermore, in Fig. 10 and Fig. 11, we provide more qualitative depth map comparisons. The results show that our method significantly improves both the consistency and fine texture details of the depth maps, whether in indoor or outdoor environments, compared to existing methods. To show how IDESplat refines depth maps over iterations, we compare

Table 7. **Quantitative Experimental Results and Comparisons on DL3DV**. Our IDESplat consistently outperforms MVSplat and DepthSplat across different numbers of input views.

Method	#Views	PSNR ↑	SSIM ↑	LPIPS ↓
MVSplat [5]	2	17.54	0.529	0.402
DepthSplat [47]		19.31	0.615	0.310
IDESplat		**19.93**	**0.635**	**0.300**
MVSplat [5]	4	21.63	0.721	0.233
DepthSplat [47]		23.12	0.780	0.178
IDESplat		**23.53**	**0.789**	**0.176**
MVSplat [5]	6	22.93	0.775	0.193
DepthSplat [47]		24.19	0.823	0.147
IDESplat		**24.61**	**0.829**	**0.146**

Table 8. **Ablation results for DPBU with different numbers of Warp-Index Epipolar Attention.** All results are reported on the RealEstate10K dataset.

Number of WIEA	Params (M)	Mem. (M)	Time (s)	PSNR↑	SSIM↑	LPIPS↓
1	36.4	2033	0.082	27.11	0.882	0.116
2	37.6	2336	0.110	27.56	0.889	0.110
3	38.9	2642	0.139	27.65	0.890	0.109
4	40.2	2954	0.172	27.72	0.893	0.107

results visually in Fig. 12. Each iteration represents one pass through the Depth Probability Boosting Unit. We can see that with more steps, the depth map gets better and more detailed. After 3 steps, the model can work at the original image size and the depth map is clearer. With more iterations, the depth map improves. This leads to more accurate Gaussian centers, which creates a better scene reconstruction.

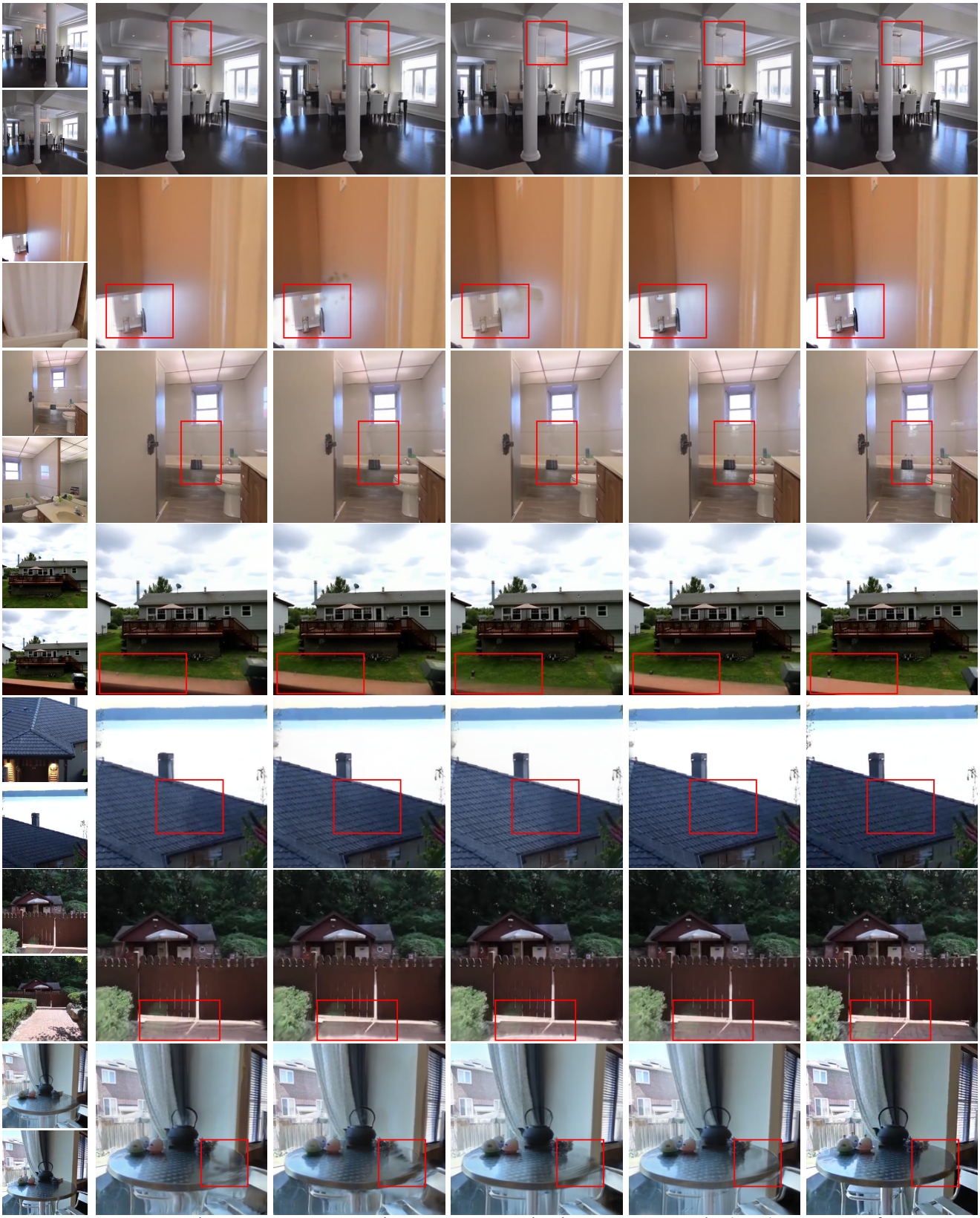


Figure 8. The comparison of visualization results for novel view synthesis on the RealEstate10K dataset.

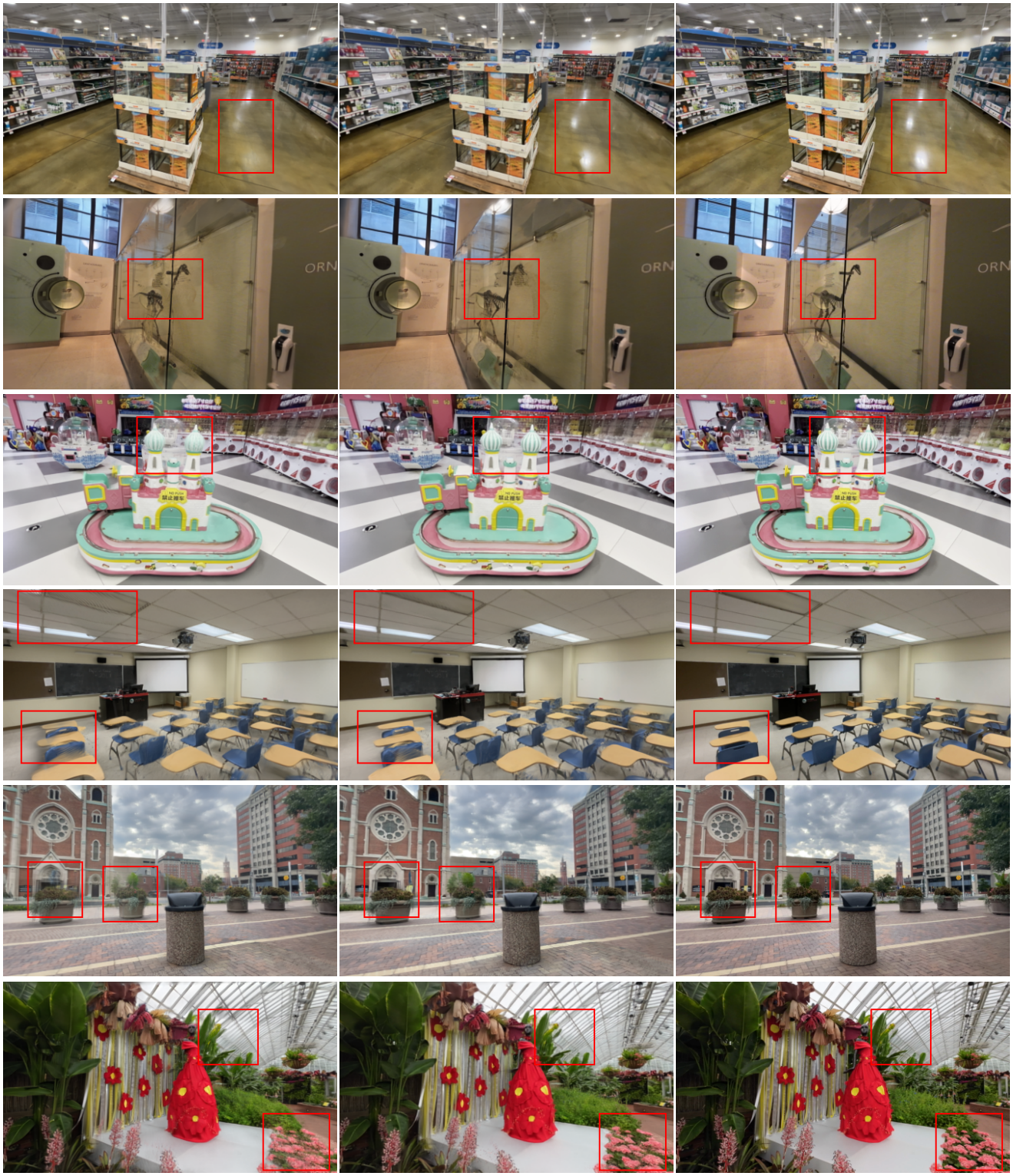


Figure 9. The comparison of visualization results for novel view synthesis on the DL3DV dataset.

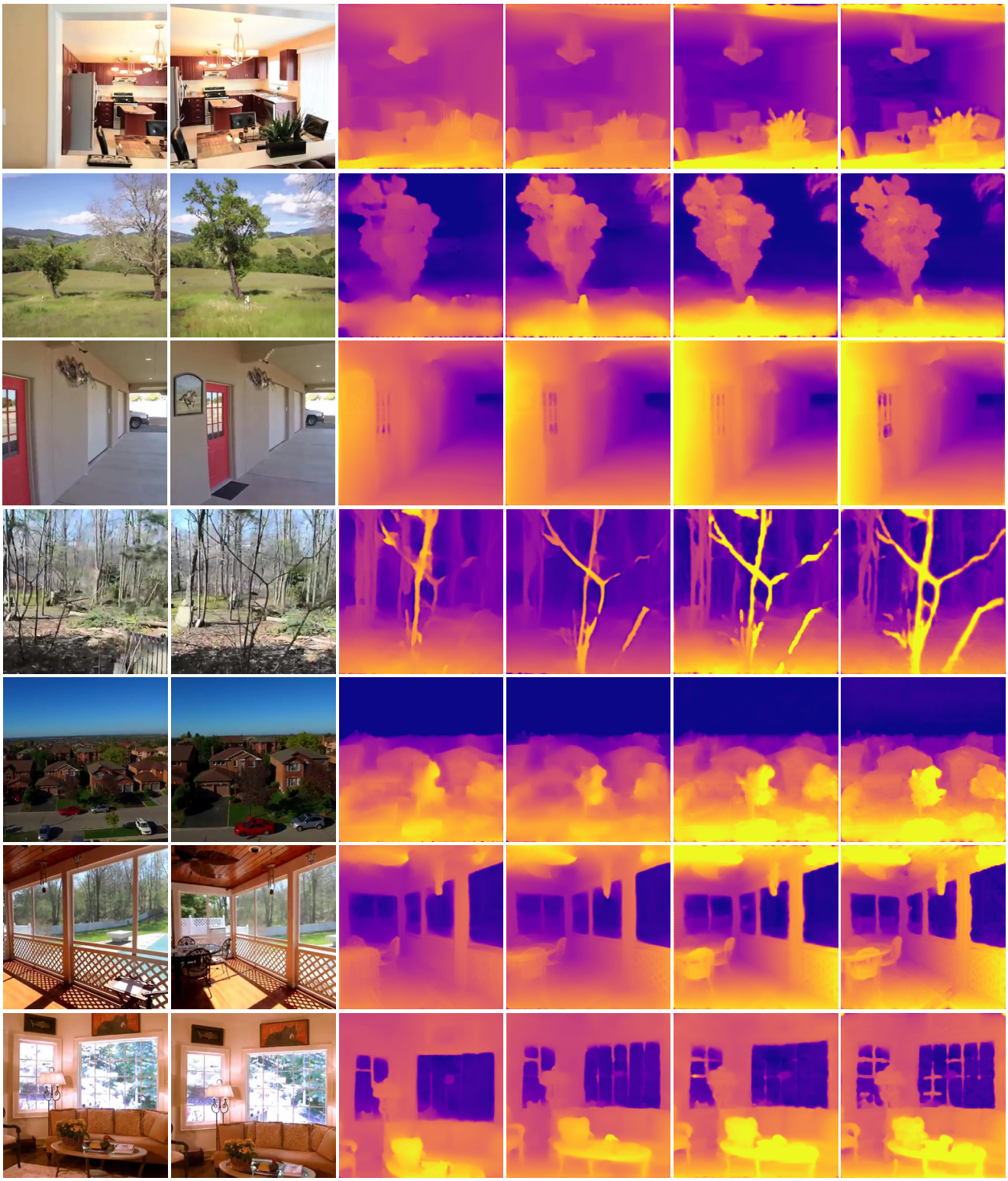


Figure 10. Comparison of depth prediction maps for different models on the RE10K dataset.

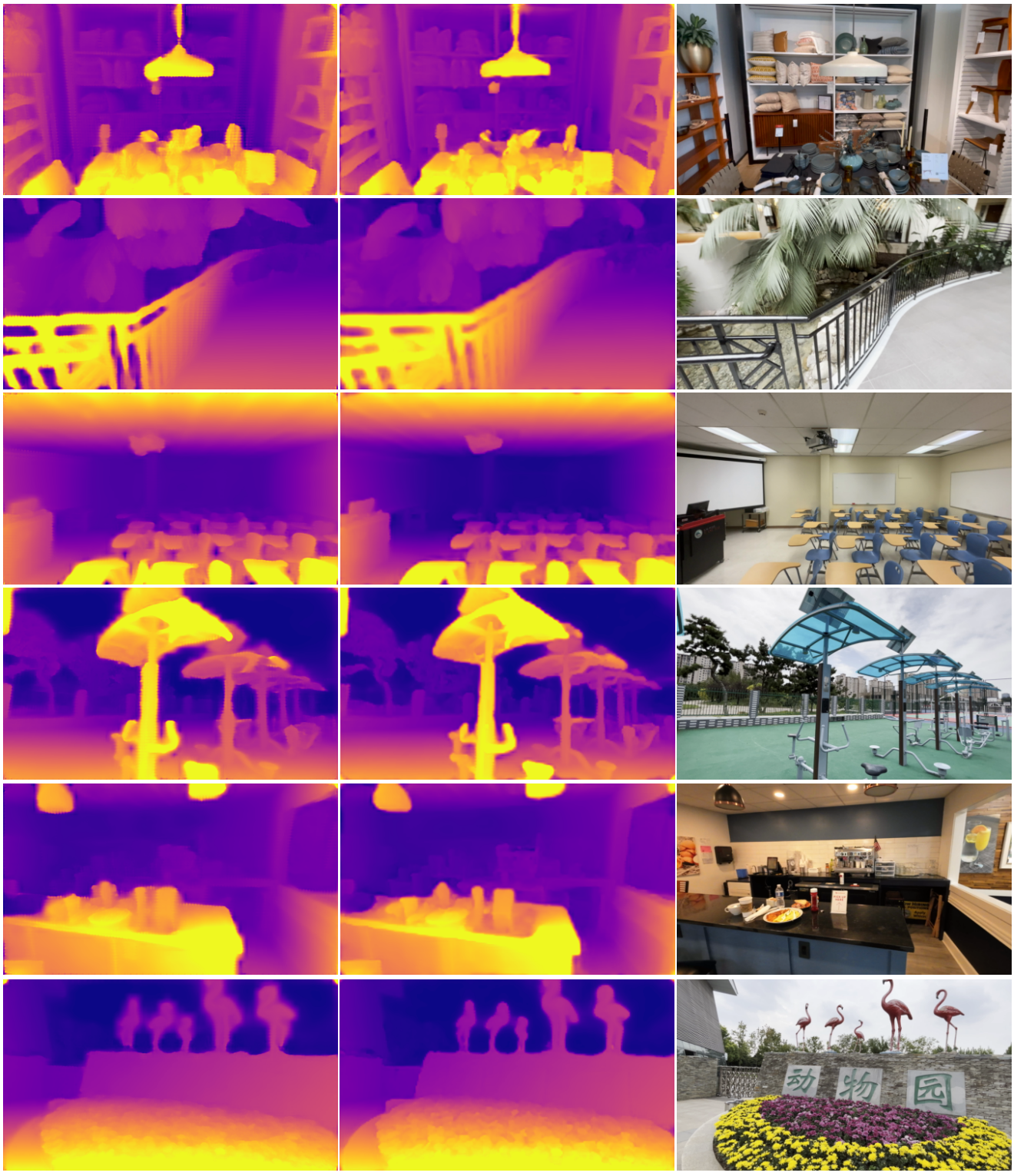


Figure 11. Comparison of depth prediction maps for different models on the DL3DV dataset.

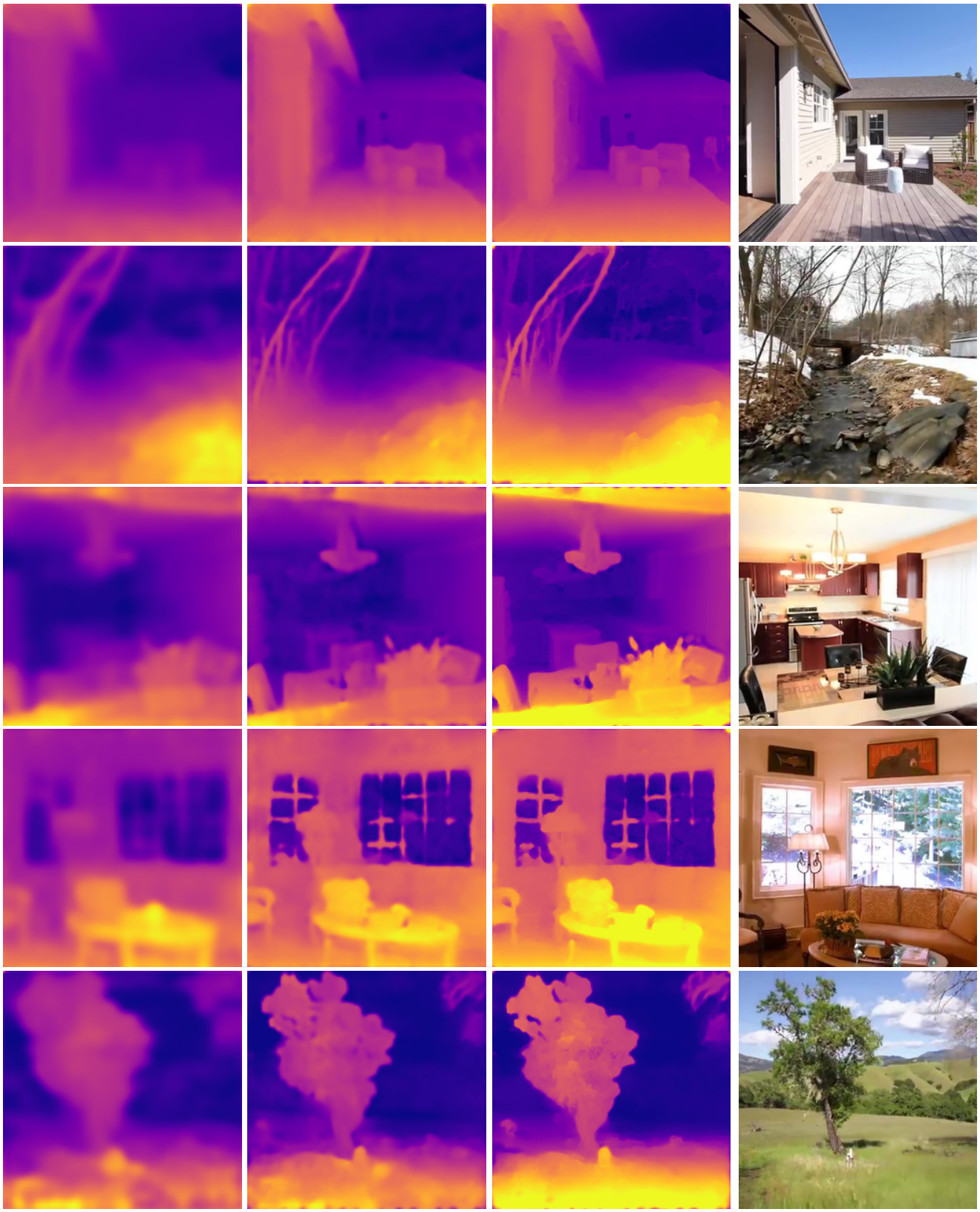


Figure 12. Visualization of intermediate depth prediction maps at different iterations in the IDESplat network.