MonoSplat: Generalizable 3D Gaussian Splatting from Monocular Depth Foundation Models

Supplementary Material

A. More Details

A.1. Data

During training, we employ our custom data loaders for all methods and progressively increase the spacing between reference views. Specifically, we implement a linear increase in view distances over the first 150,000 training steps: the minimum distance between reference views increases from 25 to 45, while the maximum distance expands from 45 to 192. To ensure a fair comparison with previous works [1, 6], input images are resized to a resolution of 256×256. Our pre-processing filters out invalid images, including those with misaligned sizes and images where the maximum field of view exceeds 100 degrees.

A.2. Model

For the frozen depth encoder, we adopt three variants of Depth Anything V2 [50], which are based on ViT-S, ViT-B, and ViT-L, respectively. For these variants, we extract features of intermediate layers of [2, 5, 8, 11], [2, 5, 8, 11], and [4, 11, 17, 23], respectively, and feed these features to the following DPT decoder and the original depth decoder. For DPT, we set the final output dimension as 64 to balance the efficiency and effectiveness. Following DPT, the multi-view transformer adds position encoding for different views and outputs features of dimension 64.

For the cost volume construction, we use 128 planes and calculate the 2D cost volume, similar to [6], followed by the UNet-style refinement. The cost volume UNet uses a base feature dimension of 128, which was empirically found to balance model capacity and efficiency well. We maintain consistent channel dimensions across three downsampling stages using multipliers [1,1,1]. Self-attention is applied at 1/4 resolution to enhance feature correlation.

After obtaining the predicted depths, we combine them with features for further Gaussian parameter prediction, achieved by a depth UNet. The depth UNet employs a base feature dimension of 32, with channel multipliers [1,1,1,1,1] across five downsampling stages. Attention mechanisms are incorporated at resolution 1/16 to capture long-range dependencies. Finally, we constrain the Gaussian scale within the range [0.5, 15.0]. The minimum scale of 0.5 ensures sufficient detail capture at fine levels. The maximum scale of 15.0 prevents overly large Gaussians while allowing coverage of broader regions. We use spherical harmonics of degree 4 to represent view-dependent appearance.

Table 4. **Ablation on the backbone.** We perform ablation studies using different backbones trained solely on Re10K [55] with 200k iterations, and test on the in-domain test set of Re10k and out-domain test set of DTU [12].

Method	Re10k		Re10k→DTU	
	PSNR ↑	SSIM↑	PSNR↑	SSIM↑
DINOv2	26.14	0.872	13.45	0.342
UniMatch	26.03	0.868	14.92	0.465
DAMv2-S	26.50	0.870	15.24	0.604
DAMv2-B	26.83	0.875	15.62	0.620
DAMv2-L	27.12	0.878	15.95	0.608

A.3. Training

Our default model training is conducted on a single A100 GPU with a batch size of 14. Each batch comprises one training scene consisting of two input views and four target views. Following pixelSplat [1] and MVSplat [6], we progressively increase the frame distance between input views throughout the training process. The near and far depth planes are empirically set to 0.5 and 100 for both RealEstate10K and ACID datasets. For the DTU dataset, we utilize the depth bounds of 2.125 and 4.525.

B. More Experimental Analysis

All experiments in this section follow the same settings as in Sec. 4.1 unless otherwise specified, which are trained and tested on RealEstate10K [55]. To investigate our hypothesis regarding the advantages of monocular depth foundation model features for generalizable Gaussian reconstruction, we conducted experiments with various frozen backbones, including DINOv2 (used in pixelSplat [1]) and UniMatch (employed in MVSplat [6]). The results, as presented in the Table 4, show substantial deterioration in both in-domain and out-of-domain generalization performance, validating the essential role of depth foundation models. Furthermore, our exploration of different Depth Anything v2 variants revealed a positive correlation between model size and performance, with larger models achieving better reconstruction quality and generalization capabilities.

C. More Visual Comparisons

In this section, we provide more visual comparisons of geometry reconstruction in Figure 7 and cross-dataset generalization results in Figure 8.

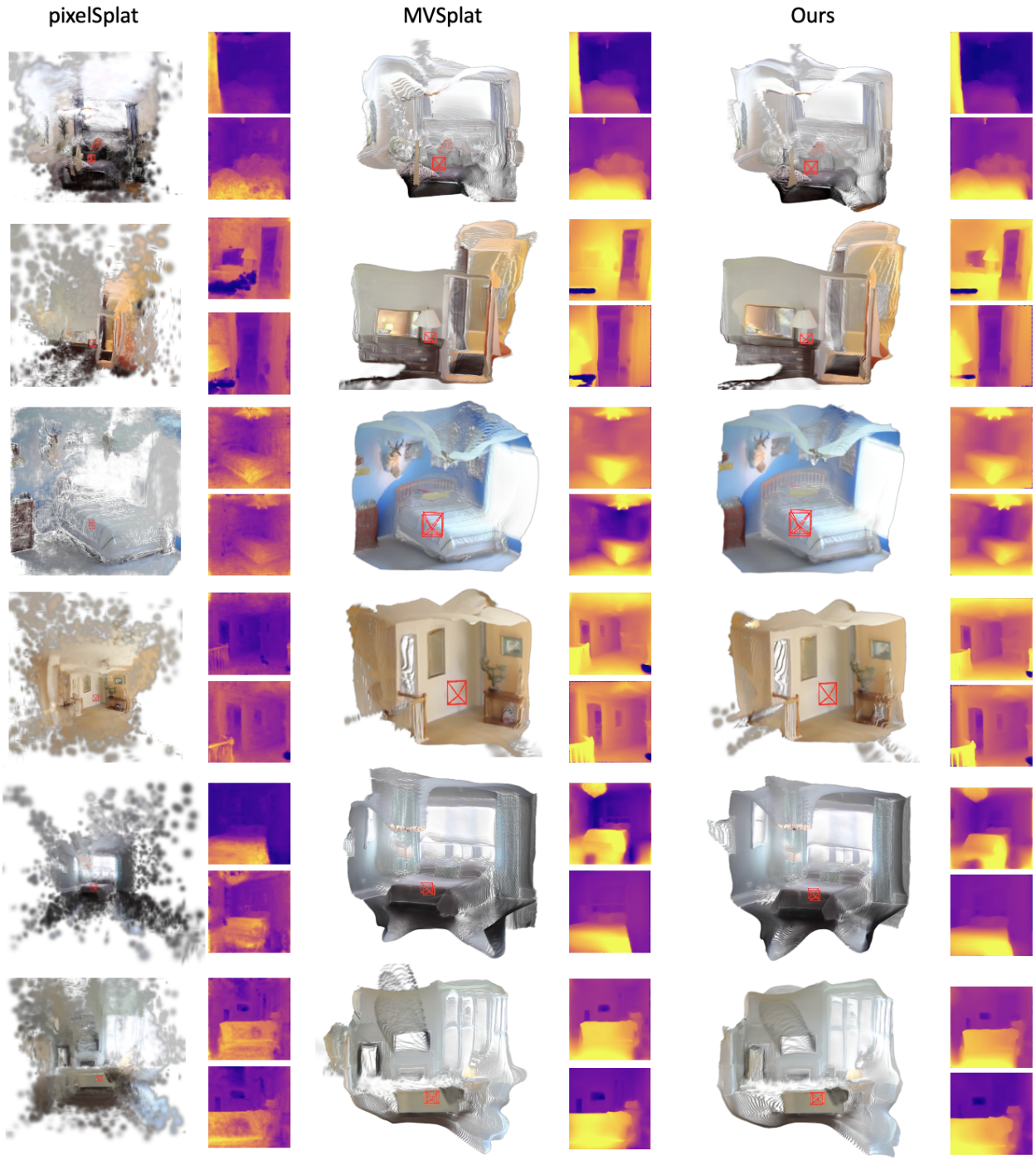

Figure 7. **Additional visual comparisons of geometry reconstruction on RealEstate10k [55].** All models were trained on a diverse collection of RealEstate10k scenes and evaluated on previously unseen scenes from the test split. For reference, we provide two rendered depth maps from the input views. Our method demonstrates superior reconstruction quality across different viewpoints and scene structures.

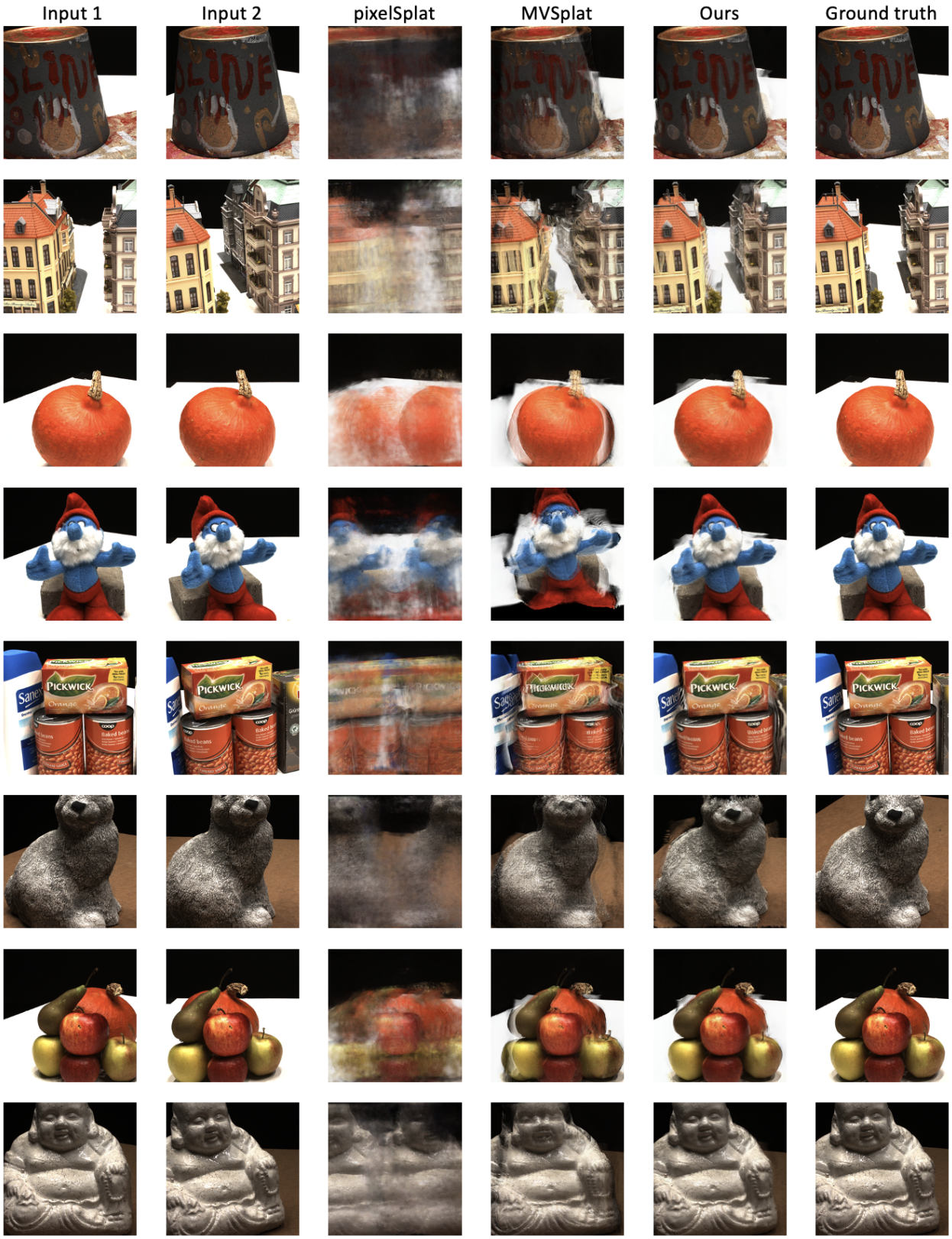

Figure 8. **More visual comparisons of cross-dataset generalization from RealEstate10k [55] to DTU [55].** Models are trained solely on RealEstate10k, and tested on novel scenes from DTU. Our method shows superior rendering quality compared to previous methods.