From None to All: Self-Supervised 3D Reconstruction via Novel View Synthesis

Supplementary Material

A. More Implementation Details

More Architecture and Training Details. In our experiments, we apply NAS3R to two representative 3D models, MASt3R and VGGT. In Sec. 4.3, NAS3R is trained from random initialization based on VGGT architecture, while in Sec. 4.4, we train NAS3R on both VGGT and MASt3R architectures and utilize their respective pretrained weights for initialization. Across all experiments, we use an initial learning rate of 1×10^{-4} and set the LPIPS loss weight to 0.05. The batch size is 10 for VGGT-based models and 16 for MASt3R-based models.

For the VGGT-based variant, we add a DPT-based Gaussian head while keeping other components as in the original VGGT: ViT-Large encoder (patch 14), alternating frame/global self-attention decoder, DPT depth head, and a camera head with self-attention layers and linear projection. For the MASt3R-based variant, we use the original MASt3R ViT-Large encoder (patch 16) and extend the pairwise ViT-Base decoder to multi-view. We add DPT-based depth and Gaussian heads, and use a 3-layer MLP camera head that predicts rotation (6D), translation (4D homogeneous coordinates), and intrinsics (FOV). For both variants, the decoder's cross-attention is modified for masked context-to-target attention, and the camera head assumes shared intrinsics for all images within the same scene to ensure stable convergence.

More Details on Baselines. All baselines use an input resolution of 256×256, except those based on the VGGT architecture (SPFSplatV2-L and the NAS3R VGGT variant), which operate at 224×224. For the prior-free comparison in Sec. 4.3, we retrain NoPoSplat and SPFSplatV2/V2-L with a 10k-step warm-up stage using the DUSt3R point-cloud distillation loss, following [22, 23, 72]. In Tab. 3, SuperPoint + SuperGlue computes feature correspondences to estimate Essential Matrices and derive relative poses. Since SelfSplat defines the target image as the reference frame, we evaluate relative poses between each pair of context images by setting the frist context image as the target image. For Fig. 4, the multi-view inference is decomposed to several two-view inference for SelfSplat. In Tab. 6, DUSt3R, MASt3R, and NoPoSplat estimate poses via PnP [16] with RANSAC [14], whereas PF3Splat, VGGT, and all SPFSplat variants directly regress camera poses.

More Details on Downstream Finetuning. For Tab. 9, following MapAnything [31], we first compute the ground-truth pointmaps from the ground-truth poses and depth maps, as well as predicted pointmaps from the predicted poses and depth maps. Using the ground-truth validity masks, we compute scaling factors for both the ground-truth and up-to-scale predicted pointmaps, which are then used to normalize the depth maps and translations. The depth loss is adopted directly from MapAnything [31].

For pose supervision, following [20], we combine a geodesic rotation loss and an L_2 translation loss. The rotation and translation terms are weighted by 0.1 and 0.01, respectively.

More Details on Multi-View Experiments. For the multi-view experiments in Tab. 8, we fix the first and last views and gradually increase the intermediate input views.

B. More Experimental Analysis

Progressive Interval Curriculum. We adopt a simple frame-interval-based strategy for view selection, a common practice in prior work [8, 72], in which the interval between context frames is progressively increased during training. We also evaluate a *no-curriculum* variant that samples from a fixed frame-interval range throughout training (Tab. 10). Results show that, although performance is slightly lower without the curriculum, NAS3R still converges stably, demonstrating the robustness of our method.

Table 10. Ablation on progressive interval curriculum on RE10K.

Method	NVS			Pose		
	PSNR↑	SSIM↑	LPIPS↓	5° ↑	10° ↑	20° ↑
No Curriculum	21.020	0.650	0.255	20.3	35.8	50.2
Progressive Interval Curriculum	**23.130**	**0.764**	**0.193**	**32.6**	**51.0**	**64.9**

Ground-truth Intrinsics. Tab. 6 show that ground-truth intrinsics improve pose estimation. Ablation results in Tab. 11 indicates intrinsics have minimal impact on NVS quality, which is mainly because photometric supervision adjusts other 3D attributes to to be consistent with the self-learned intrinsics to achieve better rendering quality.

Comparison with AnySplat. Although both of our method and AnySplat [28] adopt a local-to-global paradigm and VGGT-like architecture, there are two key differences. (1) *Depth Activation*: AnySplat follows the exponential depth activation used in VGGT, whereas our method predicts depth using a sigmoid activation followed by linear interpolation between near and far planes. This bounded formulation avoids an unbounded or explosive depth space, enabling stable training from random initialization. (2) *View Supervision*: AnySplat enforces the rendering loss only on the input context views, which leads overfitting to those views and frequent failures in depth and camera prediction. Consequently, it relies on pretrained VGGT weights

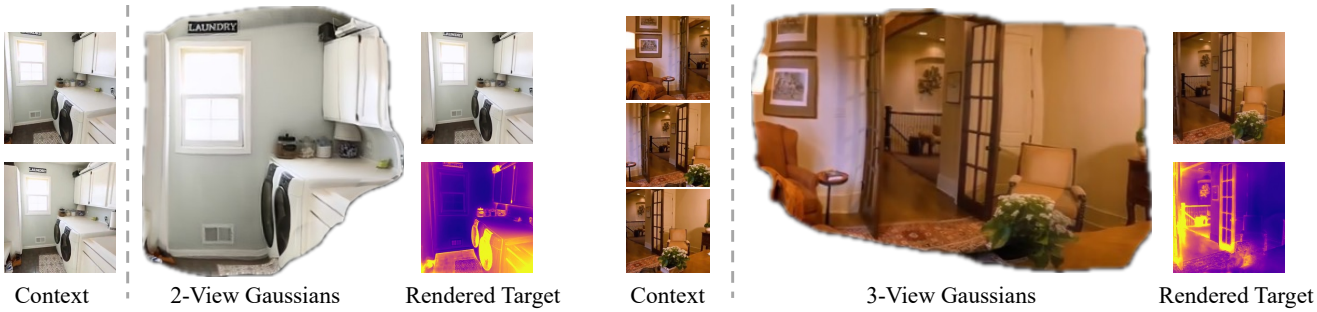

Figure 6. Examples of 3D Gaussians and rendered RGB and depth results on RE10K.

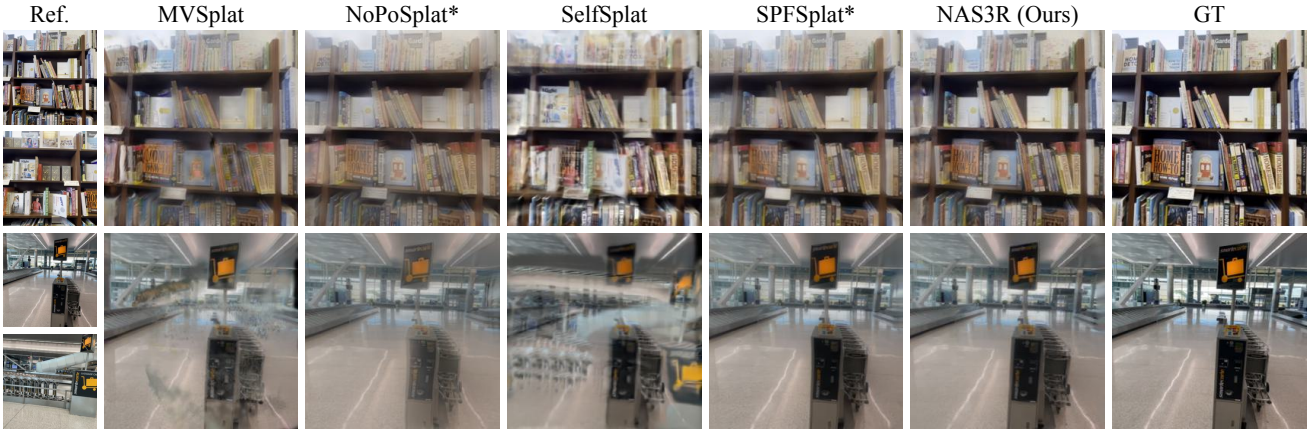

Figure 7. More novel view synthesis qualitative comparisons on DL3DV.

Table 11. Ablation on ground-truth intrinsics on RE10K.

Methods	w/o priors			w/ VGGT priors		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NAS3R	23.130	0.764	0.193	25.888	0.861	0.136
NAS3R-I	23.144	0.758	0.196	25.872	0.861	0.135

and pseudo labels for stabilization. In contrast, our method processes novel target views via masked attention and enforces photometric supervision on these unseen views, encouraging better generalization and stable training without any pretrained priors. As shown in Tab. 12, despite being trained with much less data and no pretrained priors, NAS3R outperforms AnySplat in the 2-view setting and achieves comparable performance with more views.

Table 12. Comparison with AnySplat on DL3DV dataset.

Method	Training Data	2 View			5 Views		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
(a) AnySplat	DL3DV+8 datasets	16.905	0.514	0.249	**21.680**	**0.694**	**0.218**
(b) NAS3R	DL3DV	**20.069**	**0.588**	**0.281**	21.307	0.646	0.240

Inference Efficiency. In Tab. 13, we report the inference efficiency of the VGGT-based and MASt3R-based NAS3R variants, measured on an A6000 GPU for reconstructing 3D Gaussians and predicting camera poses from two input images.

Table 13. Inference efficiency on an NVIDIA A6000 GPU.

Methods	Params (M)	FLOPs (GMac)	Time (s)
NAS3R (MASt3R)	613.47	404.52	0.042
NAS3R (VGGT)	1223.2	607.65	0.073

C. More Visualizations

Gaussian Reconstruction. Fig. 6 shows reconstructed Gaussians from two or three context views, along with the rendered RGB and depth maps, illustrating our model's ability to perform multi-view reconstruction and produce high-quality renderings.

Novel View Synthesis. We provide additional NVS qualitative comparisons with baselines on DL3DV (Fig. 7), RE10K (Fig. 8) and ACID (Fig. 9). Across all cases, our method delivers strong NVS performance, even for image pairs with very small or no viewpoint overlap.

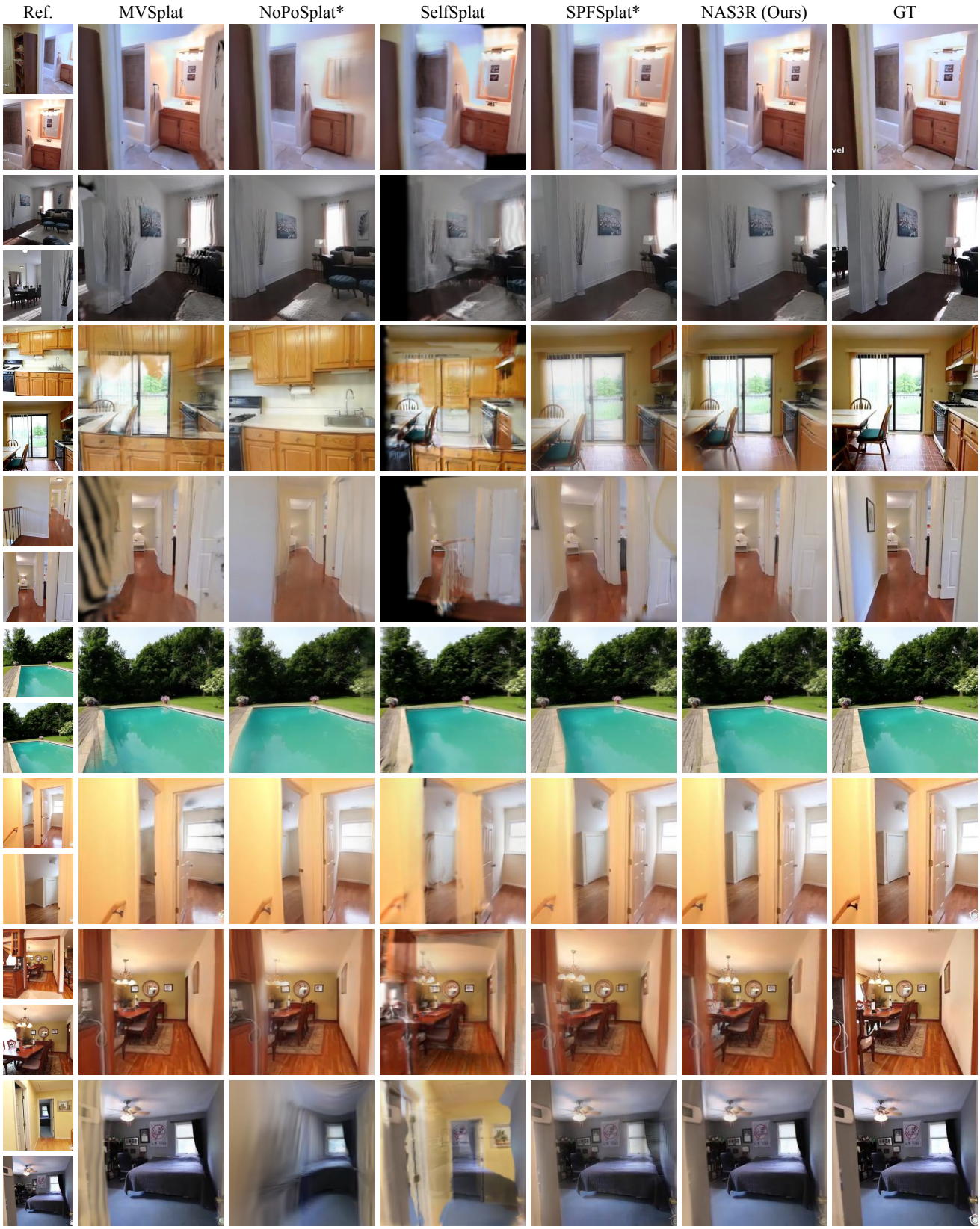

Figure 8. More novel view synthesis qualitative comparisons on RE10K.

Figure 9. More novel view synthesis qualitative comparisons on ACID.