

MuGS: Multi-Baseline Generalizable Gaussian Splatting Reconstruction

Supplementary Material

A. Implementation Details

In our MVS branch, the depth probability volume is regressed from the augmented cost volume, which is constructed by concatenating the cost volume, the MVS features, the view differences and the color samples. This augmentation is proved to be efficient by MuRF [4]. The MVS feature maps are at multiple resolution, which are $1/2$, $1/4$ and $1/8$ of the original image height and width, while the corresponding dimensions are 64, 96 and 128. An additional 128-dimension feature map is obtained by performing multi-view transformer [4], which is used to enhance the spatial cues. The cost volume is calculated by the pair-wise similarity as demonstrated in *preliminary*, where the features are warped from all above multi-view feature maps. The view differences are calculated by the angle and distance between the depth candidate point and the input view camera plane. The color samples are obtained by concatenating the color within a sample window radius w , and its dimension is $d_c = 3 * (2 * w + 1)^2$. In our experiments, we take $w = 4$ for the coarse model and $w = 3$ for the fine model.

For the MDE model, we use the metric depth version of Depth Anything V2 [7], with ViT-L feature encoder. Our feature enhancement module takes the the final layer of the Dino Vision Transformer feature maps and up-sample to the same size as the MVS feature maps through interpolate. Then, this 1024 dimension feature map is projected to 128 dimension for the subsequent steps.

The input of the 3D U-net is 3 dimension depth information as demonstrated in *methodology*, while its spatial resolution is $H * W * D$, where H , W and D are the original image height, original image width, and depth sample number. The feature dimension of the four layers are [16, 32, 64, 128], while the output is projected to 128 dimension by convolution layers.

B. Additional Comparison Experiments

Although DepthSplat[5] is a concurrent work, further experimental results demonstrate that our method still outperforms DepthSplat on multi-baseline tasks. Specifically, we reproduced their results using the code released on [github](#) under our experimental setup. The results in the table show that our method achieves over 1.5 dB PSNR improvement

on small-baseline tasks and also exhibits superior performance on large-baseline tasks. Additionally, the accuracy of depth estimation indicates that our method can predict more precise geometry.

C. Additional Ablation Experiments

Different Depth Refinement Methods. We further explore the different manner to fuse the projected depth and the sampled depth. Apart from the U-net that we used for the final full model, we try CNN and MLP to find if the performance will change. As shown in Tab. 2, the U-net performs the best since it can effectively capture the information of our proposed projected and sampled depth. Though MLP and CNN do not contribute as well as the U-net, the performance is still better than the model without the whole depth refinement module, which proves that our proposed depth fusion strategy indeed provide useful information for better rendering quality.

Different Feature Enhancement Methods. We also compare different ways to fuse MVS features and MDE features. Apart from the concatenation we used in the final full model, we also try to add them together. Specifically, the MDE features are projected to the same dimension as MVS features by a projection matrix, and the projected MDE features are add to MVS features, while the sum is used for constructing the feature volume. As shown in Tab. 2, the concatenation works better, while the adding method achieves close performance.

Different MDE Models. To explore the robustness of our method in respect of the MDE model, we select different pre-trained MDE model for our MDE branch. Note that any encoder-decoder architecture MDE model is feasible in our method. Specifically, we use the original version of Depth Anything model [6] and the DPT model [3] for further comparison. As shown in Tab. 2, the alternation of the MDE backbone does not influence the performance a lot, though the state-of-the-art Depth Anything V2 outperform DPT by large margin in monocular depth estimation metrics [7]. This indicates that our model is robust to MDE model, which highlights the effectiveness of our proposed framework.



Figure 1. Generalization on LLFF.

	DTU NVS			RE10K NVS			DTU depth (ref-view)		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	Abs err \downarrow	Acc(2) \uparrow	Acc(10) \uparrow
DepthSplat[5]	25.66	0.928	0.113	24.47	0.851	0.168	3.52	0.812	0.926
Ours	27.56	0.958	0.084	24.82	0.873	0.153	3.23	0.872	0.963

Table 1. Comparing with DepthSplat in our settings.

Module	Method	DTU (Small-Baseline)			RealEstate10K (Large-Baseline)		
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Depth Refinement	U-net	27.56	0.958	0.084	24.82	0.873	0.153
	MLP	27.42	0.955	0.086	24.69	0.870	0.158
	CNN	27.48	0.956	0.085	24.73	0.871	0.156
Feature Enhancement	Concatenate	27.56	0.958	0.084	24.82	0.873	0.153
	Add	27.51	0.957	0.085	24.77	0.872	0.156
MDE Model	Depth Anything V2 [7]	27.56	0.958	0.084	24.82	0.873	0.153
	Depth Anything [6]	27.54	0.957	0.084	24.80	0.873	0.154
	DPT [3]	27.49	0.956	0.085	24.75	0.872	0.155

Table 2. Additional Ablation.

D. Additional Visual Results

More visual comparison of zero-shot generalization performance on LLFF dataset [2] is shown in Fig. 1, and additional comparison on DTU [1] and RealEstate10K [8] in video format are available in the folder.

References

- [1] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 3
- [2] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3
- [3] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1, 3
- [4] Haofei Xu, Anpei Chen, Yuedong Chen, Christos Sakaridis, Yulun Zhang, Marc Pollefeys, Andreas Geiger, and Fisher Yu. Murf: multi-baseline radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20041–20050, 2024. 1
- [5] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depth-splat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024. 1, 2
- [6] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 1, 3
- [7] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911, 2024. 1, 3
- [8] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 3