AerialMegaDepth: Learning Aerial-Ground Reconstruction and View Synthesis

SUPPLEMENTARY MATERIAL

https://aerial-megadepth.github.io

For more 3D interactive visualizations and results, please visit our website!

Data and training details: We use Google Earth Studio (GES, https://earth.google.com/studio) to render pseudo-synthetic images. Following the method outlined in the main paper, we automatically generate viewpoints for each scene and render the images using GES. Google Earth images are rendered at 1920×1080 resolution with a randomly varied hFOV between 45°and 90°to introduce intrinsics diversity. Additionally, images are rendered at different times of day to capture appearance variations. While Google Earth doesn't provide ground-truth depth maps, we have access to ground-truth camera poses. Since we can render the scene from any viewpoint, with dense coverage, depth maps generated using classical MVS are sufficiently accurate to be used as supervision data. For all experiments, we fine-tune the publicly released DUSt3R/MASt3R checkpoint on 8 RTX A6000 GPUs for 1 day.

Generalization under varying lighting conditions: As both Google Earth and MegaDepth images contain lighting variation, our model generalizes reasonably well under these conditions as shown in Figure 1 (image pair from WxBS [2]).

Figure 1. Our model generalizes well to lighting variations.

Expanded evaluation: In addition to the ground-aerial setting evaluated in the main paper, we also include results for *ground-ground* and *aerial-aerial* configurations in Table 1. This table shows that DUSt3R and MASt3R maintain strong performance on ground-ground and aerial-aerial pairs, which typically have higher visual overlap. This demonstrates that fine-tuning with our varying-altitude data does not degrade performance on these easier, similar-viewpoint settings.

References

[1] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 2

[2] Dmytro Mishkin, Jiri Matas, Michal Perdoch, and Karel Lenc. Wxbs: Wide baseline stereo generalizations. In *BMVC*, 2015. 1

[3] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2

[4] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 2

[5] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2

Eval. Split	Method	*Camera Rotation Accuracy*			*Camera Translation Accuracy*			*3D Pointmap Accuracy*		
		RRA@5°	RRA@10°	RRA@15°	RTA@5°	RTA@10°	RTA@15°	δ@0.5m	δ@1m	δ@2m
ground-ground	LoFTR* [4]	93.60	95.97	96.92	78.20	88.63	92.65	-	-	-
	SP+SG* [3]	92.89	95.50	95.73	72.75	88.39	93.60	-	-	-
	MASt3R [1] (released)	86.49	92.89	95.02	53.32	76.07	81.99	43.42	64.23	74.82
	DUSt3R [5] (released)	90.52	95.26	96.45	**61.61**	77.96	83.18	49.11	66.70	73.93
	MASt3R + PSynth (Ours)	91.94	95.02	96.92	45.50	69.91	82.23	51.58	68.43	75.74
	DUSt3R + PSynth (Ours)	92.42	95.73	96.92	57.82	**78.20**	83.41	50.99	67.63	74.74
	MASt3R + Hybrid (Ours)	91.71	96.21	97.39	48.34	69.67	77.25	**52.32**	**69.23**	**76.31**
	DUSt3R + Hybrid (Ours)	**94.55**	**97.63**	**98.10**	55.69	77.96	**85.31**	50.18	67.98	75.63
aerial-aerial	LoFTR* [4]	96.00	96.62	96.62	92.92	95.38	96.31	-	-	-
	SP+SG* [3]	95.08	95.69	95.69	92.00	95.69	95.69	-	-	-
	MASt3R [1] (released)	**99.69**	100.00	100.00	77.23	91.69	96.31	19.10	45.37	68.09
	DUSt3R [5] (released)	98.77	100.00	100.00	55.08	90.15	95.38	11.24	37.68	64.78
	MASt3R + PSynth (Ours)	98.46	100.00	100.00	**84.92**	**96.31**	96.31	26.65	54.99	74.33
	DUSt3R + PSynth (Ours)	98.46	100.00	100.00	84.92	94.46	96.31	16.07	44.40	70.37
	MASt3R + Hybrid (Ours)	98.46	**100.00**	**100.00**	80.62	95.08	**96.31**	**27.93**	**57.76**	**75.55**
	DUSt3R + Hybrid (Ours)	98.46	100.00	100.00	80.62	92.31	95.08	14.14	41.05	66.25

Table 1. **Expanding on the results from Table 1 of the main paper, we include evaluations for *ground-ground* and *aerial-aerial* settings.** While Table 1 emphasizes the significant improvements achieved in the challenging ground-aerial setting through fine-tuning with our data, this table shows that DUSt3R and MASt3R also maintain strong performance on ground-ground and aerial-aerial pairs, which typically have higher visual overlap. This demonstrates that fine-tuning with our varying-altitude data does not degrade performance on these easier, similar-viewpoint cases.