SPIDER: Spatial Image CorresponDence Estimator for Robust Calibration

Supplementary Material

A. More Implementation Details

We adopt a VGG-19 backbone for the fine encoder. Intermediate features are extracted from layers 40, 27, 14, 7, which correspond to spatial scales 8, 4, 2, 1, respectively. These features are integrated with coarse feature from 3D VFMs at scale 16 into a five-level feature pyramid at scales 16, 8, 4, 2, 1. For the warp head, feature channels are projected to dimensions 512, 512, 256, 64, 9, respectively. For the feature head, before projecting features into the final hidden space of dimension 128, we first map the pyramid features to intermediate dimensions 256, 128, 128, 64, 64.

At inference time, for each image pair, we perform a bidirectional evaluation by running the network on both input orders: (I^A, I^B) and (I^B, I^A).

B. Linear Probing 2D and 3D VFMs

To justify our use of 3D-pretrained representations, we conduct a linear-probe experiment comparing both 2D- and 3D-pretrained backbones on an image matching task. The evaluated backbones include: VGG19, ResNet-50, Stable Diffusion, AM-RADIO, DINOv2, DINOv3, DUNE, DUSt3R, MASt3R, Aerial-DUSt3R, Aerial-MASt3R, and VGGT. Following the protocol of [2], we freeze the pretrained encoder and/or decoder and train a single linear projection layer on top of it, while correspondences are then established via kernel nearest-neighbor matching.

We evaluate performance on MegaDepth (outdoor scenes) and Aerial-MegaDepth (aerial-to-ground) test sets, using two standard metrics: (1) end-point-error (EPE), averaged at a standardized resolution of 448×448 to measure precision, and (2) PCK@32, the percentage of matches with reprojection error below 32 pixel to measure robustness.

As shown in Table 1 in the manuscript, 3D-pretrained backbones outperform conventional 2D-pretrained ones. Even though the 3D-pretrained models are exposed to only around 10 million image pairs during training—orders of magnitude fewer than the 142 million and 1.7 billion single-view images used by DINOv2 and DINOv3—their encoder representations perform comparably well. This indicates that multi-view supervision provides a far more informative learning signal than purely 2D appearance-based training. As observed by Chen et al. [1], the cross-attention mechanism in the decoder implicitly learns rigid view transformations and can extract motions from these layers. We find that a similar mechanism also benefits image matching. When equipped with a decoder, 3D-pretrained models outperform both their encoders and all 2D counterparts by a substantial margin. Among all candidates, Aerial-MASt3R decoded feature achieves the best overall performance, with the lowest EPE (8.3) and highest PCK@32 (96.6%), while VGGT decoder ranks second.

Table 1. **Comparison of MASt3R and Aerial-MASt3R under different pose-estimation protocols.** We report AUC@5° for three in-domain benchmarks. "GA" denotes the official global-alignment procedure, while "E-Matrix" denotes our unified essential-matrix protocol used for all baselines.

Method	MegaDepth	ScanNet	AerialMega
MASt3R (GA)	35.7	31.9	20.8
MASt3R (E-Matrix)	**40.0**	**33.7**	**32.8**
Aerial-MASt3R (GA)	**42.8**	29.3	38.7
Aerial-MASt3R (E-Matrix)	40.0	**34.1**	**49.3**

C. Baselines Evaluation Protocol

All baseline methods are evaluated under the same protocol to ensure comparability. For each image pair, we first obtain image correspondences from the baseline method and then estimate the relative camera pose by computing the essential matrix using the ground-truth camera intrinsics. This procedure is applied consistently across all local and dense matching methods.

MASt3R and Aerial-MASt3R, in their official implementation, estimate camera pose via global alignment (GA) by aligning 2D–3D correspondences with its predicted pointmap. For completeness, we additionally report MASt3R's GA-based pose results for in-domain benchmarks in Tab. 1. Across all three datasets, the E-Matrix protocol yields substantially higher AUC@5° scores for MASt3R, indicating that its native GA alignment underperforms when applied to large-baseline or noisy correspondences. Aerial-MASt3R also benefits on ScanNet and Aerial-MegaDepth, where the E-Matrix protocol improves robustness in scenes with strong viewpoint changes or complex geometry. Because GA is incompatible with the essential-matrix protocol used for all other baselines—and empirically produces inferior pose estimates—the main paper reports MASt3R and Aerial-MASt3R using the unified E-Matrix protocol for a fair comparison.

VGGT differs from other baselines because it includes a dedicated camera-pose head that directly regresses the relative rotation and translation. For VGGT, we use its predicted pose without applying essential-matrix estimation.

Baseline methods also differ in their training and inference resolutions. RoMa and GIM-RoMa are trained at a fixed resolution of 672×672 and can be upsampled to

Table 2. **Ablation on Multi-scale Descriptor design.** Comparison of different refinement variants on MegaDepth-1500 using AUC@5°, AUC@10°, and AUC@20°at *low resolution*. The difference between "MSF $-\alpha$" and "MSF $+\alpha$ (Ours)" is the presence of the predicted gating coefficient α.

Method	AUC@5°	AUC@10°	AUC@20°
Coarse-only	42.0	60.1	75.0
FPN-style	39.9	57.8	73.0
MSF $-\alpha$	39.5	58.2	73.8
MSF* $+\alpha$ *(Ours)	**43.1**	**61.2**	**75.9**

1344×1344 during inference. DUSt3R, MASt3R, AerialMASt3R, VGGT, and our SPIDER are trained with a maximum image dimension of 512 pixels. We treat this 512-pixel input size as the *low-resolution* setting for all baselines. MASt3R additionally employs a coarse-to-fine cropping strategy that splits the image into overlapping 512-pixel crops and fuses their predictions to recover dense correspondences at the original resolution. We use whose predictions are fused into dense, full-resolution correspondences. We use this coarse-to-fine variant, capped at a maximum input dimension of 1600 pixels, as the *high-resolution* setting. Finally, to avoid directional bias, all symmetric matching methods are run in both input orders (I^A, I^B) and (I^B, I^A), with predictions fused after geometric verification.

D. More Ablations

Multi-scale Descriptor Head Design. Coarse-only and FPN-style serve as two straightforward baselines: the former relies solely on the coarsest resolution, while the latter adopts a standard top-down refinement without learning adaptive weights [3]. Our Multi-Scale Fusion (MSF) framework introduces a structured multi-scale aggregation mechanism. Within MSF, the variant "MSF $-\alpha$" removes the predicted gating coefficient α, while "MSF $+\alpha$" (Ours) incorporates α to adaptively weight contributions from different scales.

As shown in Table 3, Coarse-only and FPN-style both underperform, indicating that neither single-scale features nor uniform top-down fusion can provide robust representations under large viewpoint or geometric variations. MSF $-\alpha$ improves over FPN-style, confirming the benefit of structured multi-scale fusion; however, its lack of scale-adaptive weighting limits the ability to suppress noisy or unreliable scales. Introducing the α-gated fusion in MSF $+\alpha$ consistently boosts performance across all angular thresholds, demonstrating that adaptive gating effectively selects informative scales while mitigating the influence of ambiguous ones. Equipping the fusion module with α gating leads to the most reliable and discriminative representations, yielding the best overall performance.

In Fig. 1, we demonstrate that high-resolution features are critical in wide-baseline scenarios, e.g. aerial–ground matching, where overlap occupies only a few aerial patches. In AerialMegaDepth, SPIDER improves AUC@5 by 7% over coarse-only (Tab. 3) and substantially increases correspondence density from *93* to *1241* matches.

Table 3. Comparison with coarse-only on AerialMegaDepth.

Method	AUC@5°	AUC@10°	AUC@20°
Coarse-only	51.6	64.1	74.4
Ours	**58.7**	**71.4**	**80.9**

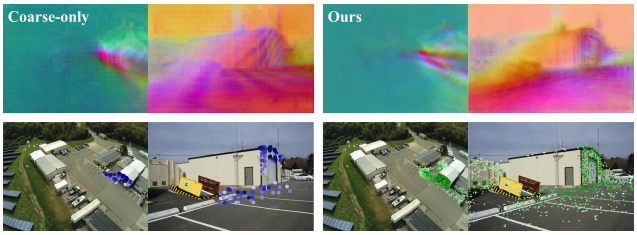

Figure 1. Visualization of features and correspondences

Ensemble Methods. We evaluate two levels of fusion: feature-level and match-level. For feature-level fusion, "$\hat{P}$ as F" replace the raw fine features F in Eq. (15) with the refined pyramid $\hat{P}$. For match-level methods, the *Warp head* and *Descriptor head* operate independently, producing dense matching flow and descriptors respectively. On top of these two heads, we test two fusion strategies. *Region guidance* assumes that the sparse correspondences from the descriptor head are highly reliable; it uses these descriptor-based matches as anchor points and samples the warp head's dense flow field in local neighborhoods around these anchors to produce guided correspondences. The *Confidence ensemble*, in contrast, performs fusion in a more principled manner. Instead of directly pooling all matches and selecting the globally highest-confidence ones, it operates *within each head* by selecting only the high-confidence correspondences from the warp head and from the feature head separately. During merging, it further enforces a balanced number of matches from the two heads, preventing either the warp or descriptor branch from dominating the final set.

Table 4 shows that replacing F with the refined $\hat{P}$ leads to weaker results on AerialMega, suggesting that refined high-resolution features are less stable under extreme viewpoint changes. Among match-level methods, using either head alone already provides competitive performance, with the warp head slightly stronger on outdoor and aerial benchmarks. Region guidance, however, fails to provide consistent improvements and degrades performance on aerial

Table 4. **Ablation on ensemble strategies.** We evaluate different feature-level fusion and match-level ensembles on outdoor (MegaDepth), indoor (ScanNet), and aerial-to-ground (AerialMega) benchmarks at *low resolution*, measured by AUC@5°. "$\hat{P}$ as F" denotes replacing our fused pyramid $\hat{P}$ with the raw feature pyramid F. Match-level methods include independent heads and their fusion variants.

Method	MegaDepth	ScanNet	AerialMega	Mean ↑
Feature-level fusion				
$\hat{P}$ as F	46.0	34.0	48.7	42.9
Match-level methods				
Warp head only	45.5	33.8	49.5	42.9
Descriptor head only	43.1	33.7	48.8	41.9
Region guidance	42.2	32.8	44.5	39.8
Confidence ensemble (Ours)	**46.4**	**34.2**	**50.0**	**43.5**

scenes, suggesting that assuming all descriptor matches are reliable leads to error propagation, especially in challenging wide-baseline scenarios. The Confidence ensemble achieves the best performance across all benchmarks, benefiting from head-wise confidence filtering and match-number balancing, which together produce a more stable and diverse correspondence set.

For a more directly comparison to existing methods, we include a baseline that directly combines GIM-RoMa and Aerial-MASt3R match results (G.R. + A.M.); our dual-head design consistently outperforms this combination on zero-shot benchmarks. The average performance of the two heads and fusion appears more similar across all datasets, while this masks large variances across datasets. As shown in Tab. 5, the warp head excels on ETH3D and Multi-FoV but underperforms on other benchmarks, particularly KITTI, whereas the descriptor head exhibits the opposite trend with notably weaker results on ETH3D and Multi-FoV. We emphasize that fusion and gated aggregation are essential for improving robustness and generalization across *different patterns and resolutions*.

Table 5. Ablation results on more zero-shot benchmarks.

Method	GL3D	BlendedMVS	ETH3D	KITTI	Multi-FoV	Mean
G.R.+A.M.	63.4	57.6	74.8	46.3	64.0	61.2
Warp Head	64.3	57.4	82.0	46.7	68.9	63.9
Descriptor Head	66.0	59.0	80.3	53.9	65.1	64.9
Ours	**66.4**	**59.8**	**82.3**	**54.7**	**69.7**	**66.6**

References

[1] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3r: Estimating disentangled motion from dust3r without training. *arXiv preprint arXiv:2503.24391*, 2025. 1

[2] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2