GEOMETRY AT THE LAST MILE: BEYOND LOCAL APPEARANCE-ONLY CORRESPONDENCE SEARCH IN FEATURE MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-dense matching has seen substantial progress, yet fine-stage localization remains predominantly driven by local appearance cues and lacks explicit pairwise geometry, leaving it vulnerable to ambiguous local evidence, especially under weak texture, repetitive patterns, and large viewpoint changes. We present GeMiL, an efficient local feature matcher that introduces shared image-pair geometry into the final stage of correspondence localization, moving beyond independent refinement driven solely by local appearance. At its core, a Structured Geometry Head, trained under direct supervision from camera pose, consolidates confident cross-view correspondences into a shared geometry prior. The resulting fundamental matrix acts as a gated and bounded geometric bias over fine candidates, suppressing inconsistent locations while reinforcing locally plausible ones. Since the quality of this geometry prior depends on the cross-view relation evidence from which it is inferred, we regularize the final cross-attention responses toward token-level correspondence targets through Correspondence-Grounded Cross-Attention. Sharp Relation Injection then propagates these correspondence-aware relation features across scales through content adaptive reassembly. Across diverse geometric tasks, GeMiL achieves consistently strong accuracy and leading efficiency among semi-dense matchers, running $1.8\times$ faster than ELoFTR with only 42% of its peak memory and 8.8% of its FLOPs.

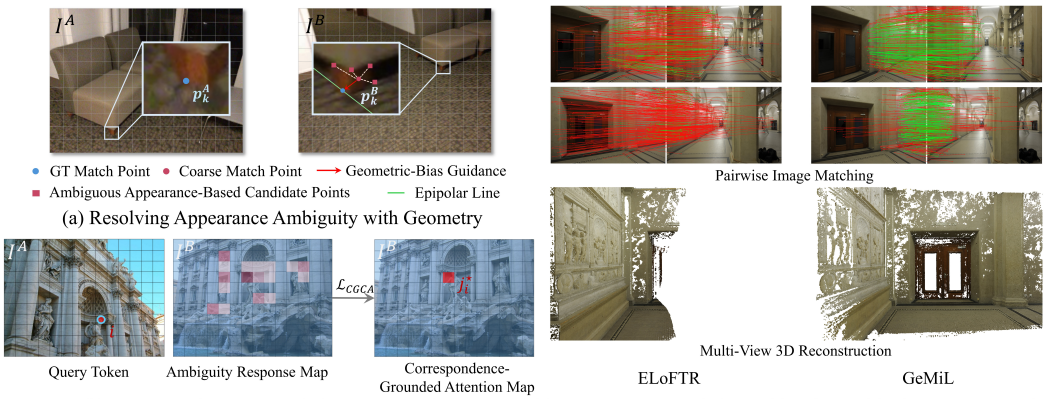
1 INTRODUCTION

Image matching is a fundamental computer vision task that underpins a wide range of geometric applications, including 3D reconstruction (Schönberger & Frahm, 2016; He et al., 2024), visual localization (Sarlin et al., 2019; 2021), and simultaneous localization and mapping (Mur-Artal et al., 2015; Campos et al., 2021). Traditional methods (Lowe, 2004; Rublee et al., 2011) typically follow a detect-describe-match pipeline based on handcrafted keypoint detectors and local descriptors. Although deep learning has substantially improved this pipeline through learned detectors, descriptors (DeTone et al., 2018; Revaud et al., 2019; Zhao et al., 2022), and matching modules (Sarlin et al., 2020; Lindenberger et al., 2023), its reliability still depends heavily on repeatable keypoint detection and discriminative local appearance, limiting its robustness under challenging appearance and geometric variations.

End-to-end detector-free matching has consequently emerged as a powerful alternative. Benefiting from the long-range modeling capability of Transformers (Vaswani et al., 2017), recent methods (Sun et al., 2021; Li et al., 2025b) combine global cross-view interaction with heuristic correspondence priors, selective feature processing, and dedicated refinement mechanisms, achieving remarkable matching performance. However, attention-based global interaction over large feature maps often incurs quadratic computational and memory costs, restricting scalability to high-resolution images. More recent methods (Wang et al., 2024; Ni et al., 2024; Li et al., 2025a) alleviate this burden by reducing redundant interactions and simplifying matching and refinement, substantially improving efficiency while maintaining impressive accuracy.

Despite rapid progress in both accuracy and efficiency, existing semi-dense matchers still determine their final fine-stage matches primarily through local appearance-based coordinate regression, as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Resolving Appearance Ambiguity with Geometry

(b) Correspondence-Grounded Cross-Attention

(c) Geometry-Grounded Matching for Multi-View Reconstruction

Figure 1: (a) Repetitive textures induce appearance ambiguity, shifting matches away from the true cross-view geometry. (b) Semantically similar regions yield diffuse attention responses. Correspondence supervision concentrates responses on true correspondences and enriches the geometric information encoded in cross-view features. (c) Under repetitive textures and large viewpoint changes, GeMiL provides reliable, geometrically consistent matches for robust multi-view 3D reconstruction.

illustrated in Fig. 1(a). Under weak textures, repetitive patterns, or large viewpoint changes, local appearance often supports multiple visually plausible locations in the target image, making a single coordinate prediction inherently ambiguous and obscuring which candidate is geometrically valid. Moreover, independently refining each coarse correspondence cannot ensure that the resulting matches remain consistent with a shared image-pair geometry.

To move beyond appearance-only fine-stage localization, we propose GeMiL, an efficient matcher that explicitly conditions final correspondence localization on shared image-pair geometry. Its central design is a Structured Geometry Head that consolidates confident cross-view correspondences into a pair-level representation and predicts the relative camera geometry under direct pose supervision. The resulting fundamental matrix is converted into a gated and bounded geometric bias over fine candidates, allowing visual evidence and pairwise geometry to jointly determine the final coordinates. This geometry-conditioned refinement requires reliable cross-view relation evidence. We therefore supervise the final cross-attention responses with token-level correspondence targets through Correspondence-Grounded Cross-Attention (CGCA). We additionally adapt content-aware neighborhood reassembly to relation propagation through Sharp Relation Injection (SRI), which transfers the interacted features from the low-resolution interaction stage to the coarse-matching scale while jointly reassembling their modulation gates. GeMiL achieves strong accuracy across multiple geometric tasks while running $1.8\times$ faster than ELoFTR with 42% of its peak memory and 8.8% of its FLOPs, as detailed in Sec. 4.1. In summary, our main contributions are:

- An efficient and geometry-grounded local feature matcher, GeMiL, featuring a Structured Geometry Head that infers camera geometry from cross-view relations under pose supervision and converts it into a geometry prior to suppress invalid fine-level candidates.
- A Correspondence-Grounded Cross-Attention mechanism that suppresses ambiguous cross-view responses and produces geometrically meaningful relation evidence through token-level correspondence supervision.
- A Sharp Relation Injection module that supports geometry prediction by preserving fine-grained relation evidence across scales through content-adaptive reassembly.

2 Related Work

Detector-Based Local Feature Matching. Sparse matching has evolved from handcrafted representations such as SIFT (Lowe, 2004) to learned local features that jointly optimize keypoint detection and description. SuperPoint (DeTone et al., 2018) pioneered this learning-based formulation, while DISK (Tyszkiewicz et al., 2020) directly optimizes feature extraction for downstream

matching. Subsequent methods further specialize this paradigm, with ALIKE (Zhao et al., 2022) emphasizing accurate keypoint localization and XFeat (Potje et al., 2024) pursuing lightweight feature extraction. Building on these features, SuperGlue (Sarlin et al., 2020) and SemaGlue (Zhang et al., 2025) formulate correspondence as context-aware assignment between sparse keypoint sets, using attentional interaction and, in the latter, additional semantic cues. Nevertheless, their performance remains fundamentally bounded by keypoint detection: correspondences cannot be recovered when the relevant locations are not detected, motivating detector-free matching over dense feature grids.

Detector-Free Local Feature Matching. Detector-free methods estimate correspondences on dense feature grids rather than isolated keypoints. LoFTR (Sun et al., 2021) established the coarse-to-fine paradigm with global cross-view interaction and local refinement. MatchFormer (Wang et al., 2022) and ASpanFormer (Chen et al., 2022) enhance cross-view modeling through hierarchical or adaptive attention, while TopicFM (Giang et al., 2023) introduces semantic topics and CoMatch (Li et al., 2025b) models dynamic covisibility with bilateral subpixel refinement. Dense matchers such as DKM (Edstedt et al., 2023) and RoMa (Edstedt et al., 2024) broaden coverage through pixel-wise warp prediction. For efficiency, QuadTree (Tang et al., 2022) hierarchically restricts attention, ELoFTR (Wang et al., 2024) and ETO (Ni et al., 2024) reduce redundant interaction and refinement, JamMa (Lu & Du, 2025) and SLiM (Choo & Li, 2026) adopt state-space modeling, and EDM (Li et al., 2025a) streamlines the overall pipeline. Nevertheless, most semi-dense matchers still refine correspondences independently within local windows using appearance evidence, leaving visually similar candidates ambiguous and potentially inconsistent with the shared image-pair geometry.

Geometry in Local Feature Matching. Geometry has been introduced into feature matching through epipolar supervision and geometry-guided coarse matching. Patch2Pix (Zhou et al., 2021) uses epipolar consistency as weak supervision for pixel-level refinement, while SCENES (Kloepfer et al., 2024) supervises both coarse and fine matching with epipolar losses. At inference time, Chang et al. (2023) estimates relative pose from high confidence anchors and iteratively restricts coarse interaction and matching to epipolar bands. Chen et al. (2025) similarly uses sparse anchors and Sampson distance constraints to optimize coarse correspondences produced by detector-free matchers. However, these methods either use geometry only as training supervision or introduce it only at the coarse stage, leaving the final coordinates primarily determined by appearance-based refinement. GeMiL instead predicts camera geometry for each image pair and converts it into an epipolar bias over fine candidates, allowing geometry to participate in the final localization decision.

3 METHOD

Given an image pair I_A and I_B, GeMiL follows a coarse-to-fine semi-dense matching pipeline. A shared backbone first extracts multi-scale features. At the coarse level, Correspondence-Grounded Cross-Attention in Sec. 3.2 produces reliable cross-view relation features for geometry estimation, which are propagated across scales by Sharp Relation Injection. The Structured Geometry Head in Sec. 3.4 then estimates the relative camera geometry and derives the fundamental matrix $\widehat{F}$. Finally, $\widehat{F}$ is converted into a geometry prior over local fine candidates to suppress geometrically inconsistent responses and refine the final correspondences in Sec. 3.5. Figure 2 presents an overview of GeMiL.

3.1 FEATURE EXTRACTION

We employ a lightweight ResNet-18 (He et al., 2016) to extract multi-scale features for I_A and I_B at resolutions from $1/8$ to $1/32$. At the $1/8$ resolution, the backbone produces coarse features F_c for coarse matching and fine features F_f that preserve the spatial details required for fine localization.

3.2 FEATURE INTERACTION

After extracting multi-scale features, we stack L Transformer blocks over the $1/32$-resolution features F_t^A and F_t^B to efficiently model intra-view context and cross-view relations. The long-range interaction of attention allows the network to capture high-level semantics and shared structures across the image pair. Following ELoFTR (Wang et al., 2024), we incorporate 2D rotary positional encoding into self-attention, making the attention sensitive to the relative 2D positions of tokens.

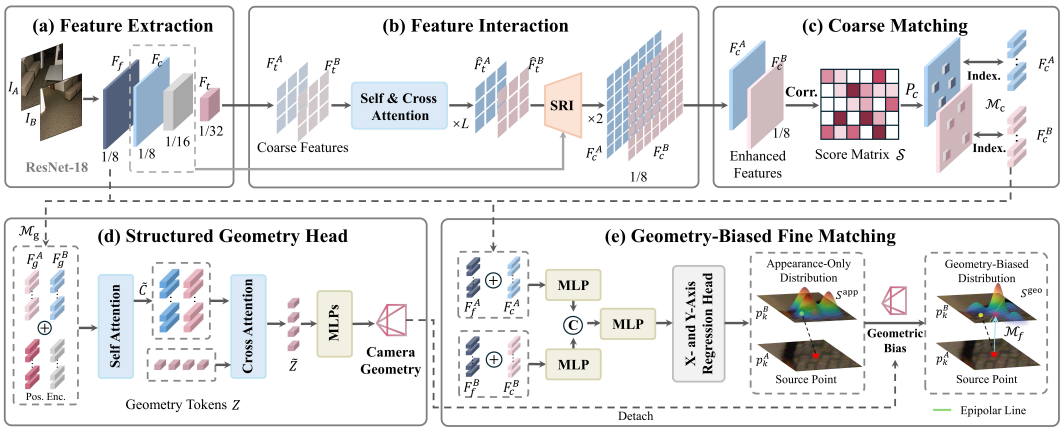


Figure 2: **Overview of GeMiL.** (a) A CNN backbone extracts multi-scale features from the input image pair. (b) The $1/32$-resolution coarse features undergo iterative self- and cross-attention, while Sharp Relation Injection propagates cross-view relations to $1/8$ resolution. (c) The enhanced features F_c^A and F_c^B are correlated and normalized by dual-softmax to obtain the confidence matrix P_c, from which coarse correspondences $\mathcal{M}_c$ are selected. (d) Fused features indexed by $\mathcal{M}_g$ are combined with their positions and aggregated through geometry tokens by the Structured Geometry Head to estimate camera geometry. (e) The coarse features F_c are added to the fine features F_f and fused to regress an appearance-based distribution, which is subsequently refined by the geometric bias derived from the estimated fundamental matrix to produce the final correspondences $\mathcal{M}_f$.

Correspondence-Grounded Cross-Attention. Although cross-attention establishes global interactions between the two views, its responses may remain diffuse or focus on visually similar but incorrect regions, particularly under repetitive patterns and weak textures. Such ambiguity weakens the cross-view relation evidence used by the Structured Geometry Head in Sec. 3.4. Building on the principle of attention response alignment (Kwon et al., 2026), we adapt it to feature matching and design Correspondence-Grounded Cross-Attention, which uses geometry-derived token supervision to sharpen the final cross-attention toward geometrically valid matches, as shown in Fig. 1(b).

The supervision targets are constructed at the $1/32$ feature resolution. For each token i in image I^A, we reproject its center into image I^B using the ground-truth depth, camera intrinsics, and relative pose, and quantize the projected location to its nearest token, denoted by $j_i^\star$. To exclude unreliable targets, we reproject $j_i^\star$ back to I^A and retain the correspondence only when the round trip returns to the original token i. The set of valid query tokens is denoted by $\mathcal{V}_A$. Targets and valid queries for the reverse direction are constructed analogously. For a query token i in I^A and a key token j in I^B, the pre-softmax responses from all heads of the final cross-attention layer are mapped by a lightweight shared projection head to a scalar correspondence logit $z_{ij}^{A\to B}$. For each valid query $i \in \mathcal{V}_A$, its corresponding token $j_i^\star$ is treated as the target class among all non-padded tokens in I^B:

$$\mathcal{L}_{A\to B} = -\frac{1}{|\mathcal{V}_A|} \sum_{i\in\mathcal{V}_A} \log \frac{\exp\left(z_{i,j_i^\star}^{A\to B}\right)}{\sum_{j\in\Omega_B} \exp\left(z_{ij}^{A\to B}\right)}, \tag{1}$$

where Ω_B denotes the set of non-padded tokens in I^B. The loss $\mathcal{L}_{B\to A}$ is defined analogously for the reverse direction, and they are averaged as $\mathcal{L}_{\mathrm{CGCA}} = \frac{1}{2}(\mathcal{L}_{A\to B} + \mathcal{L}_{B\to A})$. This supervision encourages the final cross-attention responses to concentrate on geometrically valid tokens, providing more correspondence-specific relation features $\hat{F}_t$ for the subsequent geometry head.

Sharp Relation Injection. These features encode cross-view relations at $1/32$ resolution, whereas direct interpolation to the $1/8$ coarse-matching scale may blur their spatially localized responses. We therefore introduce two cascaded Sharp Relation Injection (SRI) layers to adaptively propagate them to $1/8$ scale. For either view, an SRI layer takes a lower-resolution interacted feature F_l and a higher-resolution lateral feature F_h as inputs. After a 1×1 Conv-BN block $\mathcal{Q}$, a lightweight kernel predictor $\mathcal{K}$, consisting of a 1×1 Conv-BN-GELU channel compressor, a 3×3 convolution, pixel

shuffle, and softmax normalization, generates a spatially adaptive reassembly kernel:

$$\bar{F}_l = \mathcal{Q}(F_l), \qquad \kappa = \mathcal{K}(\bar{F}_l). \tag{2}$$

The same kernel reassembles both the projected feature and its modulation gate:

$$F_{l\to h} = \mathcal{R}(\bar{F}_l; \kappa), \qquad M_h = \mathcal{R}\big(\mathcal{G}(\bar{F}_l); \kappa\big), \tag{3}$$

where $\mathcal{G}$ denotes a gate predictor composed of a 1×1 convolution, batch normalization, and a sigmoid activation, while $\mathcal{R}$ denotes content-adaptive neighborhood reassembly. The reassembled gate modulates the projected lateral feature, which is then fused with the propagated feature:

$$\hat{F}_h = \Phi\left(\mathcal{P}(F_h) \odot M_h + F_{l\to h}\right), \tag{4}$$

where $\mathcal{P}$ is a 1×1 Conv-BN projection, $\odot$ denotes element-wise multiplication, and Φ comprises a 3×3 depthwise convolution and a 1×1 projection. Cascaded at the $1/16$ and $1/8$ levels, SRI produces the coarse features F_c^A and F_c^B for subsequent matching. See Appendix B for details.

3.3 Coarse Matching

After flattening the enhanced features F_c^A and F_c^B, we compute their pairwise similarity matrix as

$$S(i,j) = \frac{\left\langle F_c^A(i), F_c^B(j)\right\rangle}{\tau}, \tag{5}$$

where τ is the temperature parameter. Following Li et al. (2025a), we adopt an efficient L_1 normalized implementation of dual softmax. The coarse confidence matrix is computed as

$$P_c(i,j) = \frac{Z(i,j)}{\|Z(i,\cdot)\|_1} \cdot \frac{Z(i,j)}{\|Z(\cdot,j)\|_1}, \tag{6}$$

where $Z = \exp(S)$. During inference, inactive tokens are further filtered according to their feature L_2 norms (Choo & Li, 2026). For each retained token in I^A, we select the maximum response in I^B and keep the top-K responses to form the coarse matches $\mathcal{M}_c$.

3.4 Structured Geometry Head

To move beyond independently refined matches driven solely by local appearance, we consolidate confident cross-view correspondences into a compact pair-level representation and decode the camera geometry shared by the image pair. The coarse and fine feature maps F_c and F_f at the $1/8$ resolution are first fused within each image. We retain the top-K_g row-wise matches from P_c as $\mathcal{M}_g$, and gather their fused features F_g. The token for the k-th match is constructed as

$$c_k = \phi_f\left(F_{g,k}\right) + \phi_p\left(r_k\right), \qquad F_{g,k} = \left[F_{g,k}^A; F_{g,k}^B\right], \quad r_k = \left[x_k^A, y_k^A, x_k^B - x_k^A, y_k^B - y_k^A\right], \tag{7}$$

where ϕ_f and ϕ_p denote learnable feature and position encoders, respectively, and (x_k^v, y_k^v) denotes the normalized 2D coordinates of the k-th match in view $v \in \{A, B\}$. Collecting the correspondence tokens as $C = [c_1, \ldots, c_{K_g}]$, we first propagate information among them through self-attention:

$$\widetilde{C} = \mathcal{A}(C, C), \tag{8}$$

where $\mathcal{A}(X, Y)$ uses X as queries and Y as keys and values. Four learnable geometry tokens are then used to aggregate parameter-specific pair-level evidence:

$$\widetilde{Z} = \mathcal{A}(Z, \widetilde{C}), \qquad Z = [z_q, z_t, z_A, z_B], \tag{9}$$

where z_q and z_t correspond to relative rotation and translation, while z_A and z_B encode the intrinsics of the two views. Separate lightweight MLP heads decode the geometry tokens into the relative rotation quaternion q, translation vector t, and per-view fields of view θ^A and θ^B:

$$q = \psi_q(\widetilde{z}_q), \qquad t = \psi_t(\widetilde{z}_t), \qquad \theta^v = \psi_v(\widetilde{z}_v), \quad v \in \{A, B\}. \tag{10}$$

The quaternion is normalized and converted into the relative rotation $\widehat{R}$, while the normalized translation gives its direction $\widehat{t}$. The predicted fields of view are converted into the per-view intrinsic matrices $\widehat{K}^A$ and $\widehat{K}^B$ under the pinhole camera model, assuming that the principal point lies at the image center (Wang et al., 2025). The resulting camera parameters define the fundamental matrix as

$$\widehat{F} = (\widehat{K}^B)^{-T}[\widehat{t}]_\times \widehat{R} (\widehat{K}^A)^{-1}, \tag{11}$$

where $[\widehat{t}]_\times$ denotes the skew-symmetric matrix of $\widehat{t}$. We normalize $\widehat{F}$ by its Frobenius norm to remove its arbitrary scale. To stabilize joint optimization, we stop the refinement gradients from propagating through the predicted fundamental matrix.

3.5 Geometry-Biased Fine Matching

For the k-th coarse match $(p_k^A, p_k^B) \in \mathcal{M}_c$, the indexed fine and coarse features are fused by element-wise addition within each view, independently encoded, and then concatenated:

$$h_k = \phi_m\left(\left[\phi_A(F_{f,k}^A + F_{c,k}^A); \phi_B(F_{f,k}^B + F_{c,k}^B)\right]\right), \tag{12}$$

where ϕ_A, ϕ_B, and ϕ_m denote lightweight token-wise MLPs. For efficient appearance-based localization, h_k is fed into an axis-decoupled coordinate regressor (Li et al., 2025a). The local search window centered at p_k^B is uniformly discretized into $D+1$ candidate offsets along each axis, $u_i = i/D - 1/2$, where $i = 0, \ldots, D$. The two regression heads predict the horizontal and vertical appearance logits $s_{k,i}^x$ and $s_{k,j}^y$. Each offset pair (u_i, u_j) defines a candidate $p_{k,ij}^B$ within the local target window. Its appearance score is formed by combining the two axis-wise responses:

$$S_{k,ij}^{\text{app}} = s_{k,i}^x + s_{k,j}^y. \tag{13}$$

Using the source point p_k^A and the estimated fundamental matrix $\widehat{F}$, we compute the epipolar distance $d_{k,ij}$ for each candidate $p_{k,ij}^B$. The distance is converted into a bounded geometric penalty and injected into the appearance score:

$$S_{k,ij}^{\text{geo}} = S_{k,ij}^{\text{app}} - \lambda g_k \frac{d_{k,ij}^2}{2\sigma^2 + d_{k,ij}^2}, \tag{14}$$

where λ controls the strength of the geometry bias and σ determines its tolerance to epipolar deviations. The gate g_k activates the bias only when the coarse target point p_k^B is close to the predicted epipolar line, preventing unreliable geometry from overwhelming the appearance evidence. The geometry-biased scores are normalized over the candidate grid, and the final fine offset is obtained as the expectation over the candidate coordinates:

$$P_{k,ij} = \text{softmax}_{i,j}\left(S_{k,ij}^{\text{geo}}\right), \qquad \Delta p_k = \sum_{i,j} P_{k,ij} \begin{bmatrix} u_i \\ u_j \end{bmatrix}. \tag{15}$$

The offset Δp_k is added to p_k^B to produce the refined correspondence. For bidirectional refinement, the same procedure is applied in reverse using $\widehat{F}^T$, and the more confident prediction is retained.

3.6 Loss Function

The coarse and fine matching branches are supervised using the standard focal loss $\mathcal{L}_c$ and an RLE loss $\mathcal{L}_f$, respectively (Sun et al., 2021; Li et al., 2025a). In addition to the CGCA loss $\mathcal{L}_{\text{CGCA}}$, we introduce a geometry loss that jointly supervises camera estimation and geometry-biased refinement:

$$\mathcal{L}_{\text{geo}} = \mathcal{L}_{\text{camera}} + \lambda_{\text{ref}} \mathcal{L}_{\text{ref}}. \tag{16}$$

The camera loss $\mathcal{L}_{\text{camera}}$ supervises the predicted rotation, translation direction, and fields of view using geodesic, cosine, and L_1 losses, respectively. It further constrains the resulting $\widehat{F}$ through a robust Sampson loss on ground-truth matches. The refinement loss $\mathcal{L}_{\text{ref}}$ applies a Smooth L_1 loss to the geometry-biased fine offsets and a Sampson loss to the resulting matches under the ground-truth fundamental matrix. The overall training objective is

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_f \mathcal{L}_f + \lambda_a \mathcal{L}_{\text{CGCA}} + \lambda_g \mathcal{L}_{\text{geo}}. \tag{17}$$

Further details of the geometry loss terms are provided in Appendix A.

3.7 Implementation Details

The backbone feature dimensions at the $1/8$, $1/16$, and $1/32$ scales are 128, 256, and 256, respectively. We set the number of stacked Transformer blocks to $L = 2$ and $D = 16$ for fine matching. For coarse matching, the dual-softmax temperature is set to 0.1, and the top-$K_g = 512$ coarse matches are retained for the Structured Geometry Head. The geometry-bias strength λ, tolerance σ, and gating threshold are set to 0.1, 2.0, and 8 pixels, respectively. The matching confidence threshold is set to 0.05 on MegaDepth and 0.2 on ScanNet. The model is trained on MegaDepth at a resolution of 832×832. We use AdamW with an initial learning rate of 1×10^{-3}, a weight decay of 0.1, and a batch size of 16 on four NVIDIA RTX 3090 GPUs. The loss weights are set to $\lambda_c = 1.0$, $\lambda_f = 0.2$, $\lambda_a = 0.02$, $\lambda_g = 0.05$, and $\lambda_{\text{ref}} = 20$.

Table 1: **Relative pose estimation on ScanNet and MegaDepth.** Pose AUCs at different thresholds are reported alongside average runtime, peak allocated GPU memory, and FLOPs per image pair on MegaDepth. The superscript * denotes results from our re-implementation. Bold and underlined values indicate the best and second-best results among semi-dense methods, respectively.

Category	Method	ScanNet-1500 (↑)			MegaDepth-1500 (↑)			Time (ms)(↓)	Memory (GiB)(↓)	FLOPs (T)(↓)
		AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°			
Sparse	SP + SG	16.2	32.8	49.7	49.7	67.1	80.6	104.83	1.02	0.59
	SP + LG	14.8	30.8	47.5	49.9	67.0	80.1	53.96	1.02	0.50
Semi-Dense	LoFTR	16.9	33.6	50.6	52.8	69.2	81.2	376.15	12.19	3.27
	QuadTree	19.0	37.3	53.5	54.6	70.5	82.2	411.74	12.20	3.39
	MatchFormer	15.8	32.0	48.0	53.3	69.7	81.8	658.22	8.58	3.35
	ELoFTR	19.2	37.0	53.6	56.4	72.2	83.5	140.49	8.64	3.51
	JamMa	14.5	29.8	46.2	55.4	70.8	82.1	361.42	5.49	0.96
	EDM	<u>19.8</u>	<u>37.5</u>	<u>54.4</u>	<u>57.5</u>	**73.2**	<u>84.2</u>	83.45	8.24	<u>0.50</u>
	EDM*	19.5	37.1	53.7	55.4	71.2	83.5	<u>83.24</u>	8.24	<u>0.50</u>
	SLiM	18.0	34.7	50.4	**57.9**	<u>72.8</u>	83.5	157.63	<u>5.14</u>	1.86
	GeMiL (Ours)	**20.3**	**38.5**	**55.2**	57.4	**73.2**	**84.3**	**77.47**	**3.59**	**0.31**
Dense	DKM	26.6	47.1	64.2	60.4	74.9	85.1	572.98	9.66	3.91
	RoMa	28.9	50.4	68.3	62.6	76.7	86.3	755.44	6.79	15.30

4 EXPERIMENTS

4.1 RELATIVE POSE ESTIMATION

Datasets and Evaluation Protocol. GeMiL is trained on MegaDepth (Li & Snavely, 2018) with images resized and padded to 832×832, and evaluated on the MegaDepth and ScanNet (Dai et al., 2017) test sets. On MegaDepth, semi-dense and sparse methods use longest-side resolutions of 1152 and 1600 pixels, respectively, while DKM and RoMa follow prior evaluation protocols with input sizes of 880×660 and 672×672. Following ELoFTR, we estimate the essential matrix with RANSAC (Fischler & Bolles, 1981) and report pose AUC at $5°$, $10°$, and $20°$, using the maximum rotation or translation-direction error as the pose error. Runtime, peak GPU memory, and FLOPs are measured on a 1152×1152 MegaDepth pair using an NVIDIA RTX 3090.

Results. As shown in Table 1, GeMiL achieves strong relative pose estimation performance on both MegaDepth and ScanNet. Although it is slightly behind the best semi-dense method at the strictest outdoor threshold, its strong indoor results, together with the qualitative results in Fig. 4 in Appendix D, validate geometry-guided fine matching under weak texture, repetitive patterns, and large viewpoint changes. Meanwhile, L2-norm-based token filtering, direct sparse top-k extraction without reconstructing the full matching matrix, and reduced intermediate tensor construction during geometric fine matching allow GeMiL to retain leading runtime and peak memory efficiency among the compared semi-dense methods.

4.2 HOMOGRAPHY ESTIMATION

Dataset and Evaluation Protocol. We evaluate homography estimation on HPatches (Balntas et al., 2017), which contains 116 sequences with ground-truth homographies. Following the evaluation protocol of ELoFTR (Wang et al., 2024), images are resized such that their shorter side is 480 pixels while preserving the aspect ratio. The homography is estimated from the predicted correspondences using RANSAC. We compute the mean reprojection error between the four image corners warped by the estimated and ground-truth homographies, and report the AUC at thresholds of 3, 5, and 10 pixels. For fairness, we retain up to 1,000 matches for homography estimation.

Results. As shown in the left panel of Table 2, GeMiL achieves the best performance across all thresholds, demonstrating strong generalization to homography estimation.

4.3 VISUAL LOCALIZATION

Datasets and Evaluation Protocol. We evaluate visual localization on the outdoor Aachen Day–Night v1.1 (Sattler et al., 2018) and indoor InLoc (Taira et al., 2018) benchmarks. Results are reported for the day and night queries on Aachen and for DUC1 and DUC2 on InLoc. All matchers are integrated into the open-source hierarchical localization pipeline HLoc (Sarlin et al., 2019).

Table 2: **Evaluation of homography estimation on HPatches and visual localization on Aachen Day-Night v1.1 and InLoc.** Homography AUCs at 3, 5, and 10 pixels and percentages of correctly localized queries at three thresholds are reported. (↑)

Method	HPatches			Aachen Day-Night v1.1		InLoc	
	@3px	@5px	@10px	Day	Night	DUC1	DUC2
				$(0.25\,\mathrm{m}, 2^\circ)$ / $(0.5\,\mathrm{m}, 5^\circ)$ / $(5.0\,\mathrm{m}, 10^\circ)$		$(0.25\,\mathrm{m}, 2^\circ)$ / $(0.5\,\mathrm{m}, 5^\circ)$ / $(1.0\,\mathrm{m}, 10^\circ)$	
SP + SG	37.0	52.6	70.1	89.7 / 96.5 / 99.3	73.8 / 91.1 / 99.5	50.0 / 69.7 / 79.8	47.3 / 77.9 / 80.2
SP + LG	35.7	51.6	70.2	89.2 / 96.5 / 99.3	72.3 / 89.5 / 99.0	48.0 / 68.7 / 79.8	44.3 / 71.0 / 75.6
LoFTR	51.3	63.0	75.5	88.7 / **96.1** / 98.6	**77.0** / 90.6 / **99.5**	49.0 / 71.7 / 84.3	51.1 / 73.3 / 81.7
MatchFormer	51.4	63.1	76.1	**89.4** / 96.0 / 98.8	75.9 / 90.6 / **99.5**	50.0 / 73.7 / **85.4**	58.0 / 80.9 / **87.0**
ELoFTR	53.5	64.6	75.9	88.1 / 95.1 / 98.4	73.8 / 90.6 / 98.4	52.0 / 72.2 / 84.8	**59.5** / **82.4** / **87.0**
JamMa	49.9	61.2	74.0	85.9 / 94.7 / 98.1	72.8 / 90.1 / 97.9	47.5 / 67.2 / 78.3	35.9 / 53.4 / 69.5
EDM	52.7	64.7	77.0	87.4 / 95.6 / 98.2	73.3 / **91.1** / 99.0	47.5 / 70.2 / 81.3	50.4 / 74.8 / 82.4
SLiM	51.2	62.4	75.6	71.8 / 79.4 / 85.7	57.1 / 70.7 / 78.5	40.4 / 58.1 / 68.2	43.5 / 58.0 / 66.4
GeMiL (Ours)	**53.7**	**65.3**	**77.5**	88.1 / **96.1** / **98.9**	76.4 / **91.1** / 99.0	**53.0** / **74.7** / **85.4**	54.2 / 70.2 / 81.7

Localization accuracy is measured by the percentage of queries whose estimated camera poses fall within predefined position and orientation error thresholds. We report $(0.25\,\mathrm{m}, 2^\circ)$, $(0.5\,\mathrm{m}, 5^\circ)$, and $(5.0\,\mathrm{m}, 10^\circ)$ on Aachen and $(0.25\,\mathrm{m}, 2^\circ)$, $(0.5\,\mathrm{m}, 5^\circ)$, and $(1.0\,\mathrm{m}, 10^\circ)$ on InLoc.

Results. As shown in the right panel of Table 2, GeMiL achieves competitive localization accuracy among semi-dense matchers, demonstrating strong generalization to visual localization.

4.4 Multi-view 3D reconstruction

Dataset and Evaluation Protocol. We evaluate on the 13 publicly annotated scenes of the ETH3D high-resolution multi-view benchmark (Schops et al., 2017). All matchers are trained on MegaDepth and tested on ETH3D without fine-tuning. Image pairs and the SfM, PatchMatch Stereo, and depth-fusion settings are fixed across methods, while ground-truth poses and laser scans are used for alignment and evaluation. We report the number of successful reconstructions (*Success*), valid camera registration (*Val. Reg.*), and ETH3D accuracy at 1, 2, and 5 cm. Valid registration and reconstruction accuracy are reported in percentage. Following Hybrid SfM (Liu et al., 2024), estimated cameras are aligned to the ground truth by a similarity transformation and considered valid when their camera-center and rotation errors are within 5 cm and 5°, respectively. Val. Reg. and accuracy are computed per scene and macro-averaged over all 13 scenes, with failed scenes assigned zero. Success requires a non-empty fused point cloud and finite evaluation scores.

Results. Multi-view 3D reconstruction requires accurate and consistent pairwise geometry, as matching errors can propagate through camera registration and reconstruction. As shown in Table 3, GeMiL successfully reconstructs the largest number of scenes and achieves the highest valid registration rate and reconstruction accuracy at all thresholds. Fig. 1(c) further shows that GeMiL produces stable matches on the low-texture walls of the *relief* scene, providing reliable geometric constraints for accurate multi-view reconstruction. This consistent improvement confirms that grounding local matching in image-pair geometry leads to more accurate multi-view reconstruction.

4.5 Understanding GeMiL

Geometry Head Analysis. (1) We directly evaluate the geometry predicted by the Structured Geometry Head and compare it with RANSAC using the same top-K coarse correspondences. The rotation and translation AUCs at 20° evaluate the corresponding angular errors, while the $\widehat{F}$ AUC at 10° and median error measure the discrepancy between the normalized predicted and ground-truth fundamental matrices. We further report the accuracy of the predicted fields of view within 10° and 20°. As shown in Table 4, the Geometry Head substantially outperforms RANSAC in rotation and translation estimation, and improves the $\widehat{F}$ AUC from 81.7 to 84.2 while reducing its median error from 0.48° to 0.30°. Its FoV accuracies of 51.0 and 78.5 further confirm meaningful intrinsic prediction. These results show that the correspondence features encode reliable pairwise geometry beyond that recovered solely from quantized coarse coordinates. Removing the Geometry Head increases the median EPE from 0.727 to 0.755 pixels and degrades PCK@1, median Sampson error,

Table 3: **3D reconstruction on ETH3D. (↑)**

Method	Success	Val. Reg.	Accuracy (1 / 2 / 5 cm)
LoFTR	**12**	66.78	56.80 / 66.92 / 78.12
ELoFTR	8	44.62	48.87 / 54.69 / 59.36
JamMa	11	59.56	57.88 / 67.56 / 78.67
SLiM	8	30.70	37.73 / 45.79 / 55.23
EDM	**12**	77.02	60.64 / 73.93 / 83.85
GeMiL (Ours)	**12**	**78.34**	**64.75 / 75.79 / 86.09**

Table 4: **Geometry analysis on MegaDepth.**

Method	$\widehat{R}$ AUC ↑	$\widehat{t}$ AUC ↑	$\widehat{F}$ AUC ↑	$\widehat{F}$ Med. ↓	FoV Acc.@10°/20° ↑
RANSAC	52.3	22.9	81.7	0.48°	–
w/o SRI	76.8	34.3	83.3	0.33°	50.7 / 77.4
w/o CGCA	75.9	35.0	83.7	0.33°	49.1 / 76.4
Geo. Head	**79.4**	**38.5**	**84.2**	**0.30°**	**51.0 / 78.5**

Method	Med. EPE ↓	PCK@1 ↑	Samp. Med. ↓	Epi.@1 ↑
w/o Geo. Head	0.755	68.07	0.321	90.13
Geo. Head	**0.727**	**69.48**	**0.307**	**92.08**

Table 5: **Structured Geometry Head generalization across semi-dense matchers. (↑)**

Method	AUC@5°	AUC@10°	AUC@20°
ELoFTR	54.9	71.2	82.9
+ Geo. head	55.3 (+0.4)	71.3 (+0.1)	82.8 (−0.1)
+ Geo. head & bias	55.4 (+0.5)	71.5 (+0.3)	83.2 (+0.3)
EDM	55.4	71.2	83.5
+ Geo. head	56.1 (+0.7)	72.1 (+0.9)	83.5 (+0.0)
+ Geo. head & bias	56.5 (+1.1)	72.4 (+1.2)	83.8 (+0.3)
SLiM	55.4	69.2	79.0
+ Geo. head	55.7 (+0.3)	70.1 (+0.9)	80.3 (+1.3)
+ Geo. head & bias	56.7 (+1.3)	70.9 (+1.7)	81.0 (+2.0)

Table 6: **Ablation study on MegaDepth.**

Variant		@5°	@10°	@20°
(a)	(1) w/o geometric bias	56.0	72.1	83.5
	(2) w/o Structured Geometry Head	55.9	71.7	83.7
	(3) All tokens instead of matched tokens	56.5	72.4	83.6
	(4) w/o geometry gate	56.3	71.9	83.4
	(5) w/o 2D coordinate encoding	57.1	72.8	83.8
(b)	(6) CGCA on all cross-attention layers	56.4	72.4	83.5
	(7) w/o CGCA loss	56.1	72.2	83.0
(c)	(8) SRI → bilinear interpolation	56.4	72.5	83.3
	Full model	**57.4**	**73.2**	**84.3**

and epipolar inlier ratio, confirming its benefit to both fine localization and geometric consistency. Translation direction remains more challenging than rotation, consistent with prior learning-based pose estimation and its degeneracy under pure rotational motion (Chen et al., 2021; Tirado-Garín & Civera, 2024). Importantly, GeMiL uses the predicted geometry as a selective local prior rather than a standalone pose estimate. The compatibility gate in Eq. 14 activates the geometric bias only for coarse matches close to the predicted epipolar line, while the bounded penalty limits its influence on the appearance scores. Further details are provided in Appendix C. (2) To evaluate its architectural generality, we integrate the Structured Geometry Head into ELoFTR, EDM, and SLiM, and adapt the predicted epipolar prior to their respective fine localization processes. For a fair comparison, all baseline models are retrained and evaluated on MegaDepth. As shown in Table 5, adding the geometry head improves most metrics across the three matchers, suggesting that direct geometry supervision enhances their cross-view relation features. Further incorporating the predicted geometry into fine localization brings additional gains, showing that the epipolar prior can effectively disambiguate fine-level candidates.

Ablation Studies. We evaluate the main design choices of GeMiL on MegaDepth, with results reported in Table 6. **(a) Geometry Modeling.** (1)–(2) Removing either the geometric bias or the Structured Geometry Head consistently degrades AUC, validating geometry-guided fine localization and explicit pair-level modeling. (3) Using all tokens instead of matched tokens performs worse, indicating that selected correspondences provide cleaner geometry evidence. (4) Removing the geometry gate causes clear degradation across all thresholds, while (5) removing 2D coordinate encoding leads to smaller but consistent drops. **(b) CGCA.** (6) Applying CGCA to all cross-attention layers performs worse than supervising only the final layer. (7) Removing the CGCA loss also lowers performance, validating its correspondence-grounded supervision. **(c) SRI.** (8) Replacing SRI with bilinear interpolation consistently reduces the AUCs, confirming the effectiveness of adaptive cross-scale relation propagation. Their direct geometry ablations in Table 4 further show that CGCA improves all evaluated geometry metrics, while SRI improves most of them, confirming that both modules provide more reliable relation evidence for the Structured Geometry Head.

5 Conclusions

We presented GeMiL, an efficient semi-dense matcher that introduces shared image-pair geometry into fine-stage correspondence localization. Its Structured Geometry Head predicts camera geometry and converts it into an epipolar prior, while CGCA and SRI improve the reliability and cross-scale propagation of relation features. Experiments across four geometric tasks demonstrate strong accuracy, generalization, and efficiency, validating geometry-guided fine matching beyond appearance-only correspondence search.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript and check grammar to improve its readability, as well as to assist in diagnosing software issues during development. They were not used to generate research ideas, develop the proposed method or mathematical formulations, design experiments, write or implement code, process or generate data, or interpret experimental results. All AI-assisted language revisions and diagnostic suggestions were carefully reviewed by the authors. The proposed method, implementation, experimental design, and data processing were carried out by the authors. The authors take full responsibility for the final content of this work, including its text, claims, code, and experimental results.

REFERENCES

Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, pp. 5173–5182, 2017.

Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE TRO*, 37(6):1874–1890, 2021.

Jiahao Chang, Jiahuan Yu, and Tianzhu Zhang. Structured epipolar matcher for local feature matching. In *CVPRW*, pp. 6177–6186, 2023.

Gonglin Chen, Jinsen Wu, Haiwei Chen, Wenbin Teng, Zhiyuan Gao, Andrew Feng, Rongjun Qin, and Yajie Zhao. Geometry-aware feature matching for large-scale structure from motion. In *3DV*, pp. 34–43, 2025.

Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *ECCV*, pp. 20–36, 2022.

Kefan Chen, Noah Snavely, and Ameesh Makadia. Wide-baseline relative camera pose estimation with directional learning. In *CVPR*, pp. 3258–3268, 2021.

Sin Wai Choo and Bo Li. Scalable feature matching via state space modeling and sparse correlation. In *CVPR*, pp. 6685–6694, 2026.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pp. 2432–2443, 2017.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, pp. 224–236, 2018.

Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *CVPR*, pp. 17765–17775, 2023.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *CVPR*, pp. 19790–19800, 2024.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm: Robust and interpretable topic-assisted feature matching. *AAAI*, 37(2):2447–2455, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *CVPR*, pp. 21594–21603, 2024.

Dominik A Kloepfer, João F Henriques, and Dylan Campbell. Scenes: Subpixel correspondence estimation with epipolar supervision. In *3DV*, pp. 21–30, 2024.

Minkyung Kwon, Jinhyeok Choi, Jiho Park, Seonghu Jeon, Jinhyuk Jang, Junyoung Seo, Minseop Kwak, Jin-Hwa Kim, and Seungryong Kim. Correspondence-attention alignment for multi-view diffusion models. In *CVPR*, pp. 2316–2326, 2026.

Xi Li, Tong Rao, and Cihui Pan. Edm: Efficient deep feature matching. In *ICCV*, pp. 26198–26208, 2025a.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pp. 2041–2050, 2018.

Zizhuo Li, Yifan Lu, Linfeng Tang, Shihua Zhang, and Jiayi Ma. Comatch: Dynamic covisibility-aware transformer for bilateral subpixel-level semi-dense image matching. In *ICCV*, pp. 18521–18530, 2025b.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV*, pp. 17581–17592, 2023.

Shaohui Liu, Yidan Gao, Tianyi Zhang, Rémi Pautrat, Johannes L Schönberger, Viktor Larsson, and Marc Pollefeys. Robust incremental structure-from-motion with hybrid features. In *ECCV*, pp. 249–269, 2024.

David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

Xiaoyong Lu and Songlin Du. Jamma: Ultra-lightweight local feature matching with joint mamba. In *CVPR*, pp. 14934–14943, 2025.

Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE TRO*, 31(5):1147–1163, 2015.

Junjie Ni, Guofeng Zhang, Guanglin Li, Yijin Li, Xinyang Liu, Zhaoyang Huang, and Hujun Bao. Eto: Efficient transformer-based local feature matching by organizing multiple homography hypotheses. In *NeurIPS*, volume 37, pp. 60260–60274, 2024.

Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *CVPR*, pp. 2682–2691, 2024.

Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In *NeurIPS*, volume 32, 2019.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, pp. 2564–2571, 2011.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, pp. 12708–12717, 2019.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pp. 4937–4946, 2020.

Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *CVPR*, pp. 3247–3257, 2021.

Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *CVPR*, pp. 8601–8610, 2018.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pp. 4104–4113, 2016.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pp. 3260–3269, 2017.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, pp. 8922–8931, 2021.

Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, pp. 7199–7209, 2018.

Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. In *ICLR*, 2022.

Javier Tirado-Garín and Javier Civera. From correspondences to pose: Non-minimal certifiably optimal relative pose without disambiguation. In *CVPR*, pp. 403–412, 2024.

Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. In *NeurIPS*, volume 33, pp. 14254–14265, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pp. 5294–5306, 2025.

Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *ACCV*, pp. 256–273, 2022.

Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *CVPR*, pp. 21666–21675, 2024.

Shihua Zhang, Zhenjie Zhu, Zizhuo Li, Tao Lu, and Jiayi Ma. Matching while perceiving: Enhance image feature matching with applicable semantic amalgamation. *AAAI*, 39(10):10094–10102, 2025.

Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter CY Chen, and Zhengguo Li. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE TMM*, 25:3101–3112, 2022.

Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *CVPR*, pp. 4667–4676, 2021.

A Additional Loss Details

A.1 Camera Geometry Loss.

We provide the detailed formulation of the camera loss introduced in Sec. 3.6, including supervision for rotation, translation direction, field of view, and the fundamental matrix:

$$\mathcal{L}_{\text{camera}} = \mathcal{L}_R + \lambda_\theta \mathcal{L}_{\text{fov}} + \mathcal{L}_t + \lambda_F \mathcal{L}_F. \tag{18}$$

For the predicted rotation $\widehat{R}$ and ground-truth rotation R^*, we use the geodesic distance on $\mathrm{SO}(3)$:

$$\mathcal{L}_R = \frac{1}{B} \sum_{b=1}^{B} \arccos\left(\frac{\mathrm{tr}(\widehat{R}_b^\top R_b^*) - 1}{2} \right). \tag{19}$$

Since translation scale is unobservable from an image pair, we supervise only its direction:

$$\mathcal{L}_t = \frac{1}{B} \sum_{b=1}^{B} \left(1 - \widehat{t}_b^\top t_b^*\right), \tag{20}$$

where both $\widehat{t}_b$ and t_b^* are unit vectors.

For each image, the ground-truth horizontal and vertical fields of view are derived from the effective intrinsics after resizing:

$$\theta_x^* = 2 \arctan\left(\frac{W_{\text{valid}}}{2 f_x} \right), \qquad \theta_y^* = 2 \arctan\left(\frac{H_{\text{valid}}}{2 f_y} \right), \tag{21}$$

where H_{valid} and W_{valid} exclude padded regions. We apply an L_1 loss to the horizontal and vertical fields of view predicted for both images.

A.2 Epipolar Supervision.

For a correspondence (x, x') and a fundamental matrix F, the Sampson distance is defined as

$$d_{\text{S}}(x, x'; F) = \frac{(x'^\top F x)^2}{(Fx)_1^2 + (Fx)_2^2 + (F^\top x')_1^2 + (F^\top x')_2^2 + \epsilon}. \tag{22}$$

We constrain the predicted fundamental matrix $\widehat{F}$ using ground-truth correspondences:

$$\mathcal{L}_F = \frac{1}{|\mathcal{M}^*|} \sum_{(x,x') \in \mathcal{M}^*} \log\left(1 + \frac{d_{\text{S}}(x, x'; \widehat{F})}{\tau} \right), \tag{23}$$

where the logarithmic robustifier reduces the influence of large geometric residuals.

A.3 Geometry-Biased Refinement Loss.

We further supervise the final correspondence coordinates obtained after injecting the predicted geometry:

$$\mathcal{L}_{\text{ref}} = \lambda_{\text{coord}} \mathcal{L}_{\text{coord}} + \lambda_{\text{epi}} \mathcal{L}_{\text{epi}}^{\text{ref}}. \tag{24}$$

The coordinate term is a Smooth L_1 loss:

$$\mathcal{L}_{\text{coord}} = \frac{1}{|\mathcal{V}|} \sum_{k \in \mathcal{V}} \mathrm{SmoothL1}\left(\widehat{\Delta p_k}^{\text{geo}}, \Delta p_k^* \right), \tag{25}$$

where $\mathcal{V}$ contains correspondences whose ground-truth offsets lie inside the local fine-matching window. The offsets are normalized by the window size before supervision.

The refined coordinates are additionally constrained by the ground-truth fundamental matrix:

$$\mathcal{L}_{\text{epi}}^{\text{ref}} = \frac{1}{|\mathcal{V}|} \sum_{k \in \mathcal{V}} d_{\text{S}}\left(p_k, \widehat{p}_k^{\text{geo}}; F^* \right). \tag{26}$$

Both matching directions are included in the supervision.

A.4 Loss Settings.

Within $\mathcal{L}_{\text{camera}}$, the rotation and translation terms have unit weights, while the field of view and fundamental matrix terms are both weighted by 0.1. The robust Sampson loss uses $\tau = 1$, $\epsilon = 10^{-8}$, and at most 512 ground-truth correspondences for each image pair. The weight of the camera loss is linearly increased during the first training epoch. Within $\mathcal{L}_{\text{ref}}$, the coordinate and epipolar terms are weighted by $\lambda_{\text{coord}} = 0.05$ and $\lambda_{\text{epi}} = 0.01$, respectively. Sampson distances in the refinement loss are clipped at 100, with $\epsilon = 10^{-8}$. The refinement loss is enabled from the beginning of training.

B Additional Methodological Details

B.1 Architecture of Sharp Relation Injection.

Figure 3 illustrates the detailed architecture of SRI. Given the lower-resolution interacted feature F_l and the higher-resolution lateral feature F_h, SRI predicts a content-adaptive kernel κ from $\bar{F}_l = \mathcal{Q}(F_l)$. The shared reassembly operation applies κ to $\bar{F}_l$ and $\mathcal{G}(\bar{F}_l)$ to generate the propagated feature $F_{l\to h}$ and modulation gate M_h, respectively. The projected lateral feature $\mathcal{P}(F_h)$ is modulated by M_h, combined with $F_{l\to h}$, and fused by Φ to produce the enhanced feature $\hat{F}_h$.

B.2 Geometry Gate.

The geometry gate determines whether the predicted fundamental matrix is sufficiently reliable for refining a coarse match. For the forward direction, the source point p_k^A defines an epipolar line in image I^B:

$$\ell_k^B = \widehat{F}\widetilde{p}_k^A, \tag{27}$$

where $\widetilde{p}_k^A$ denotes the homogeneous coordinate of p_k^A. The distance between the coarse target point p_k^B and this line is computed as

$$d_k^c = \frac{\left|\widetilde{p}_k^{B\top}\ell_k^B\right|}{\sqrt{(\ell_{k,1}^B)^2 + (\ell_{k,2}^B)^2 + \epsilon}}. \tag{28}$$

The binary gate is then defined as

$$g_k = \mathbf{1}\left[d_k^c \le \tau_g\right], \tag{29}$$

where $\tau_g = 8$ pixels is the gating threshold measured in the original image coordinates. A single gate is shared by all candidates within the local search window. When $g_k = 1$, the geometric penalty is applied to the complete candidate distribution. When $g_k = 0$, the geometry term is disabled and the original appearance scores are preserved. For refinement in the reverse direction, the same procedure is applied using $\widehat{F}^\top$.

C Direct Geometry Evaluation Protocol

C.1 RANSAC Baseline.

Given a correspondence $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$, we first normalize the image coordinates using the ground-truth intrinsics:

$$\widetilde{\mathbf{x}}_i = K_A^{-1}\mathbf{x}_i, \qquad \widetilde{\mathbf{x}}'_i = K_B^{-1}\mathbf{x}'_i. \tag{30}$$

An essential matrix is then estimated with RANSAC and decomposed into relative rotation and translation candidates. The valid solution is selected through a cheirality check. We compute the rotation and translation-direction errors as

$$e_R = \arccos\left(\frac{\operatorname{tr}\left(\widehat{R}^\top R^*\right) - 1}{2}\right), \tag{31}$$

and

$$e_t = \arccos\left(\frac{\widehat{t}^\top t^*}{\|\widehat{t}\|_2\|t^*\|_2}\right). \tag{32}$$

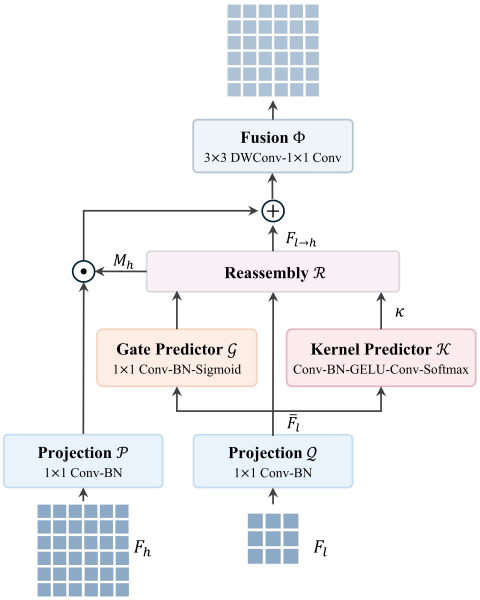


Figure 3: **Overview of Sharp Relation Injection (SRI).**

Since the translation recovered from an essential matrix is defined up to sign, we use

$$e_t \leftarrow \min(e_t, 180^\circ - e_t). \tag{33}$$

The rotation and translation AUCs are computed separately from their respective recall curves.

C.2 Fundamental Matrix Error.

Since the fundamental matrix is defined up to a nonzero scale, we first normalize both the predicted and ground-truth matrices by their Frobenius norms. We then measure their sign-invariant angular discrepancy as

$$e_F = \frac{180}{\pi} \arccos\left(\text{clamp}\left(\frac{\left|\left\langle \text{vec}(\widehat{F}), \text{vec}(F^*)\right\rangle\right|}{\|\widehat{F}\|_F \|F^*\|_F}, 0, 1\right)\right), \tag{34}$$

where the absolute value accounts for the equivalence between F and $-F$. We report the median error and the normalized area under the cumulative recall curve over the range $[0^\circ, 10^\circ]$.

D Additional Qualitative Results

Additional comparisons on MegaDepth and ScanNet demonstrate the robustness of GeMiL under repetitive structures, weak texture, and large viewpoint changes. It retains broad correspondence coverage while producing fewer geometric outliers than LoFTR, ELoFTR, and SLiM, as shown in Fig. 4.

The *terrace* scene in Fig. 5 illustrates how this improvement benefits multi-view reconstruction. The correspondences predicted by GeMiL provide more reliable constraints for camera registration and recover substantially more of the building facade than ELoFTR.

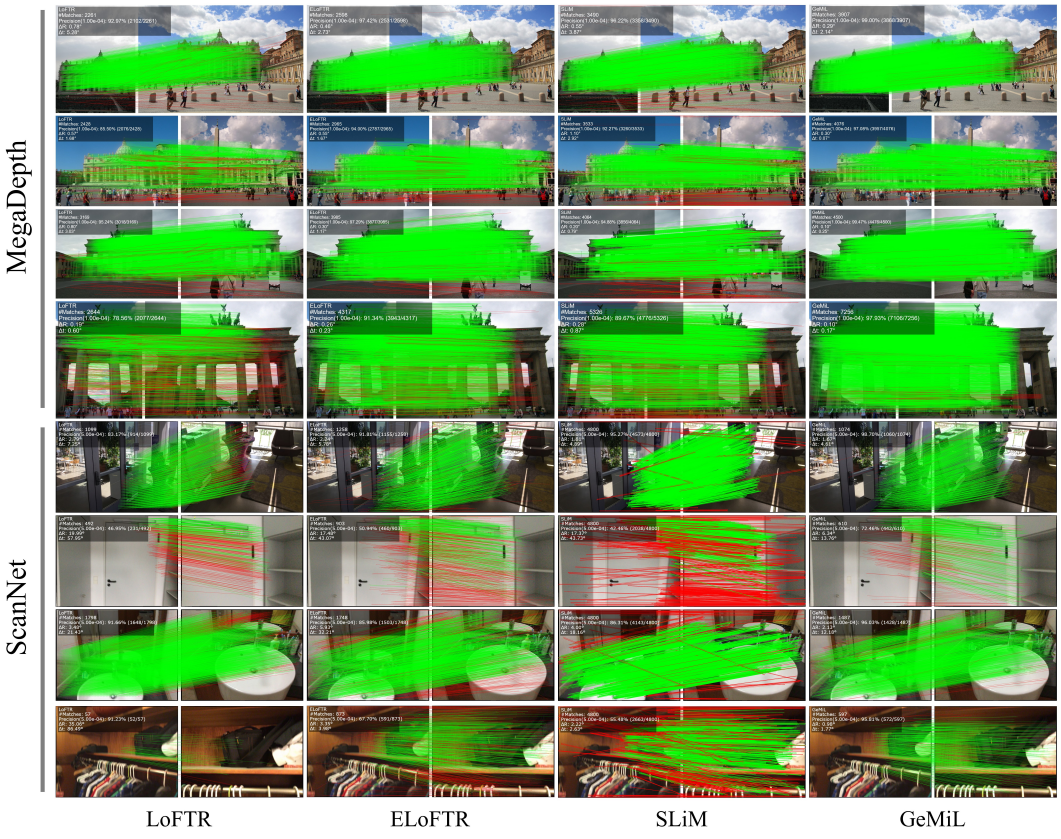


Figure 4: **Qualitative matching comparisons on MegaDepth and ScanNet.** Green and red lines denote matches with epipolar errors below and above 5×10^{-4} in normalized image coordinates, respectively.

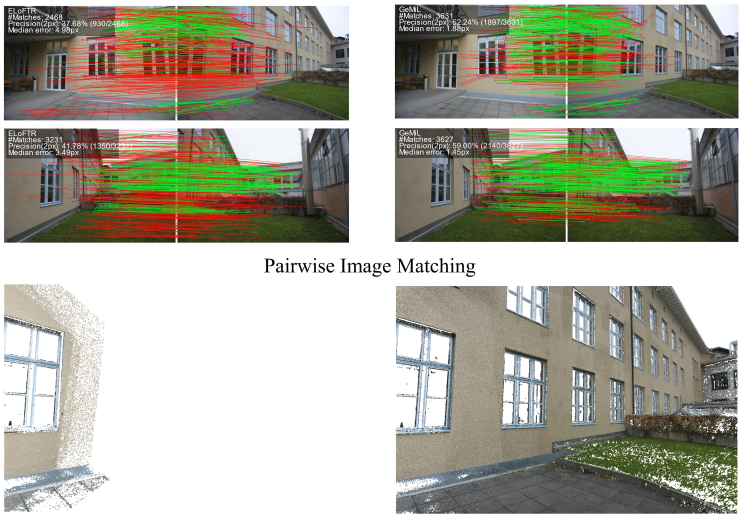


Figure 5: **Qualitative comparison of pairwise matching and 3D reconstruction on ETH3D.** Compared with ELoFTR, GeMiL produces denser and more reliable correspondences, leading to a more complete reconstruction.