

SemLight: Distilled Semantic–Geometric Fusion for Efficient Local Feature Matching

Guanglu Shi, Xiangzeng Liu[†], Yunan Li, Tuo Pang, and Qiguang Miao

Xidian University, Xi'an, China

{guanglushi,tuopang}@stu.xidian.edu.cn

{xzliu,yunanli,qgmiao}@xidian.edu.cn

[†]Corresponding author

Abstract. Robust and efficient local feature matching remains a bottleneck for real-time visual localization, especially in environments with low texture, repetitive structures, or illumination changes. Most prior approaches model either appearance or geometry in isolation, failing to exploit the semantic–geometric interactions crucial for disambiguating challenging matches. To address this limitation, we propose SemLight, built on the insight that compact semantic cues can serve as a lightweight prior for resolving matching ambiguity under challenging conditions. Specifically, high-level semantic priors are first distilled into compact, matching-oriented descriptors via a task-aware student network. These descriptors then conditionally re-weight appearance and geometric channels locally, enabling efficient disambiguation with minimal computational overhead. Extensive experiments on relative pose estimation, homography estimation, and visual localization demonstrate state-of-the-art performance. Under comparable computational cost to XFeat, our method achieves a 17.8% relative improvement in AUC@5 over XFeat (50.2 vs. 42.6).

Keywords: Local feature matching · Semantic—geometric ambiguity · Lightweight network

1 Introduction

Local feature matching remains a cornerstone of many high-level vision systems, such as Structure-from-Motion [12], SLAM [22], visual localization [7], and relative pose estimation [38]. Despite significant progress in learned descriptors and geometry-aware matching, two challenges remain: semantic–geometric ambiguity and deployment cost. Ambiguity arises when visually or geometrically similar patterns correspond to different semantic entities, leading to incorrect matches in repetitive or low-texture regions. Addressing this issue often requires dense semantic representations, which introduce substantial computational overhead and limit practical deployment.

Traditional local feature matching pipelines consist of feature detection, description, and correspondence estimation. Classical methods rely on handcrafted descriptors and nearest-neighbor matching [3, 21, 27], while learning-based methods jointly learn keypoints, descriptors, and matching strategies [10, 17, 24, 29, 32, 34]. Geometry-aware approaches further incorporate structural cues to improve robustness under viewpoint and appearance changes [20, 33]. However, these methods mainly improve either appearance modeling, geometric consistency, or matching inference, and still struggle in cases where geometrically consistent structures correspond to different semantic entities.

To mitigate this limitation, existing semantic-enhanced matching methods incorporate high-level semantics via keypoint selection [13, 35], region-level correspondence guidance [18, 19, 39, 40], or feature-level integration [4, 9, 14, 36]. Keypoint-based methods improve robustness by leveraging semantic saliency, but remain fundamentally sparse and thus insufficient for resolving ambiguities in repetitive or low-texture regions. Region-level approaches introduce semantics at the region level, yet often rely on intermediate segmentation that is category-dependent and may lack fine-grained spatial precision. Feature-level integration methods embed semantic representations into dense descriptors, but typically depend on heavy semantic backbones, leading to significant computational overhead. Overall, these methods either provide indirect or coarse semantic guidance or incur substantial efficiency costs, making it difficult to jointly achieve semantic disambiguation and efficient matching.

To address this trade-off, we propose a lightweight semantic-geometric matching framework that combines compact semantic distillation with adaptive semantic-conditioned feature fusion. Semantic knowledge is first distilled into a compact CNN to produce matching-oriented descriptors, which are subsequently used to dynamically modulate local appearance and geometric representations. This design improves descriptor discriminability while maintaining low computational cost. By providing stable complementarity to geometric normals under challenging conditions (Fig. 2), these semantic cues enable more reliable correspondence estimation via adaptive fusion, ultimately yielding a favorable accuracy–efficiency balance (Fig. 1). The main contributions of this paper are:

- (1) We propose a compact local feature extraction framework that jointly models semantic and geometric consistency to mitigate semantic–geometric ambiguity in repetitive and texture-poor scenes. The resulting descriptor improves discriminability under challenging conditions while remaining suitable for real-time deployment.
- (2) We develop a lightweight semantic feature learner trained via knowledge distillation. This design transfers high-level semantic priors into a compact

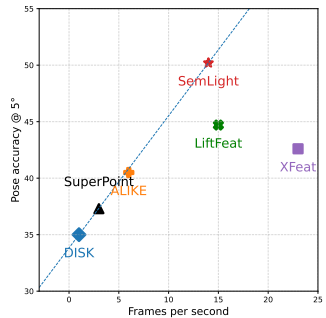


Fig. 1: Accuracy–efficiency trade-off. SemLight achieves a balance between accuracy and runtime.

descriptor tailored for local matching, achieving strong discriminative power with minimal computational overhead.

- (3) We design a lightweight cross-modal fusion module that adaptively integrates semantic, appearance, and 3D geometric cues through conditional re-weighting. The module efficiently enhances descriptor robustness without relying on heavy architectures or expensive computations.

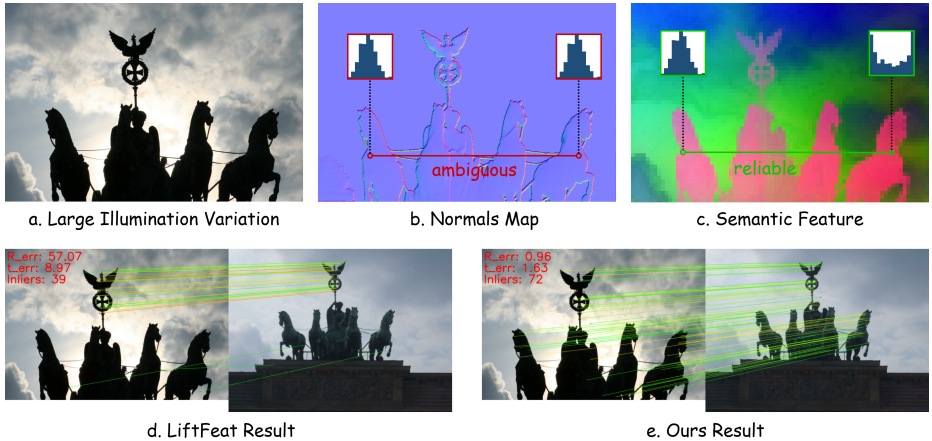


Fig. 2: Qualitative comparison on a scene with large illumination variation. (b) Geometric normals alone are ambiguous under lighting changes. (c) Our semantic features remain stable. (d) LiftFeat struggles due to reliance on geometry. (e) Our method successfully resolves ambiguity via semantic–geometric fusion.

2 Related Work

2.1 Local Feature Matching

Classical methods with handcrafted descriptors such as SIFT [21] and ORB [27] remain widely adopted due to their invariance properties and ease of deployment. Learned feature matching methods have substantially improved robustness and accuracy, ranging from joint detection-description networks [8] to learned matching modules that formulate correspondence estimation as a structured assignment problem using graph neural networks or optimal transport [17, 29]. While these approaches achieve strong performance, they often incur considerable computational overhead and are not explicitly designed for resource-constrained platforms, limiting their applicability in large-scale or real-time robotic systems. To address efficiency concerns, several works have proposed compact descriptors or hybrid pipelines that combine classical detectors with learned representations, enabling practical deployment in visual SLAM. For instance, XFeat [26] introduced an efficient CNN architecture that supports fast sparse and semi-dense

matching. LiftFeat [20] has enhanced robustness in challenging scenes by incorporating 3D geometric cues into 2D feature representations. More recently, dense and semi-dense matching methods often based on transformer architectures have further pushed matching accuracy under challenging conditions [10,32], albeit at the cost of increased runtime and memory consumption. In contrast to methods that primarily emphasize accuracy or efficiency in isolation, our work focuses on a balanced design that integrates lightweight geometric cues with compact semantic priors, achieving robust and discriminative matching under repetitive, low-texture, and illumination-challenged scenarios while maintaining suitability for real-time and embedded applications.

2.2 Semantic-Geometric Matching

The integration of semantic cues to enhance feature invariance has evolved through several research threads. Early works like DEAL [25] combined geometric and texture features for deformation-invariant nonrigid matching, while Sim2Real [42] separated intra- and inter-object features for single-object correspondence. Building on this, SFD2 [35] distilled semantic segmentation features to classify keypoints by stability for outdoor localization, though it remained dependent on fixed semantic classes. More recently, semantic-geometric fusion has progressed along two main directions. The first direction uses semantic priors to perform coarse region matching first, then employs off-the-shelf point matchers for fine-level correspondence within matched regions [18, 19, 39]. The second direction directly integrates semantic and geometric cues into point descriptors for end-to-end matching. Cadar et al. [4] integrated DINOv2 [23] features into descriptors and MATCHA [36] proposed a unified attention-based model for diverse tasks. DeDoDe [9] also incorporated DINOv2 features but without explicit feature aggregation. Collectively, these methods advance from specialized constraints to integrated representations, yet often rely on multi-stage pipelines and heavyweight models at inference, limiting their practicality in resource-constrained scenarios. In contrast, our approach adopts a lightweight design that distills semantics into a compact CNN and enables dynamic pixel-wise fusion in a single forward pass, balancing efficiency and accuracy for mobile and embedded applications.

3 SemLight

In this section, we present the overall network architecture of SemLight, as shown in Fig. 3. First, DistilSem is proposed to transfer the high-level semantic priors from a large self-supervised model to a compact semantic descriptor. Then, RG-Fusion is developed to integrate semantic, appearance, and local 3D surface information into a unified descriptor. Finally, implementation and training details of the proposed framework are provided.

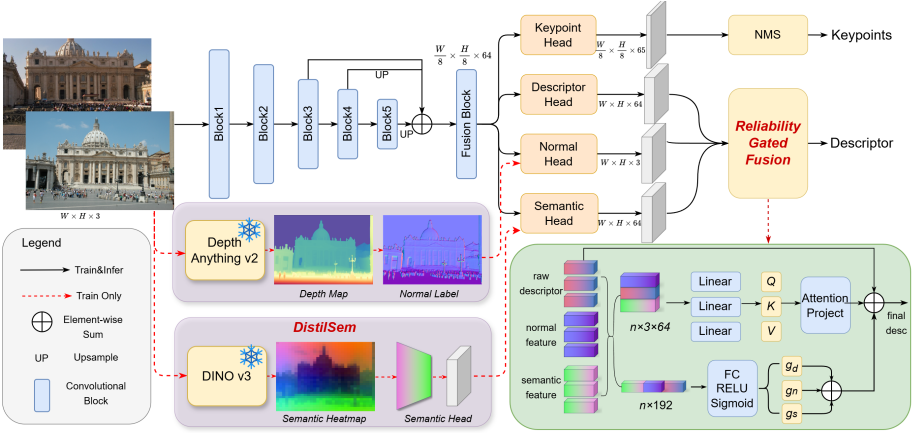


Fig. 3: Overview of the proposed SemLight. Given an input image, the network jointly predicts keypoints, descriptors, surface normals and semantic features via multi-task heads. The depth-based pseudo normals and DistilSem are introduced as auxiliary supervision during training. RGFusion adaptively integrates appearance, geometric, and semantic cues to produce robust local descriptors, while auxiliary branches are removed at inference.

3.1 Network Architecture

Our framework consists of three components: a lightweight CNN backbone for feature extraction, geometry and semantic branches that provide supervisory signals, and a feature-fusion branch that integrates these cues into a unified local descriptor. Let the input image be $I \in \mathbb{R}^{W \times H \times 3}$. Our feature extractor is a compact 5-block convolutional encoder with progressive channel growth; in our implementation the block depths are $\{4, 8, 16, 32, 64\}$ and spatial resolution is reduced by a factor of 2^i after Block i . Features from Block3—5 are channel-normalized via 1×1 convolutions, upsampled to a common $\frac{W}{8} \times \frac{H}{8}$ grid and summed, producing a fused tensor of size $\frac{W}{8} \times \frac{H}{8} \times 64$.

Keypoint detection head. A lightweight 1×1 projection on the fused feature produces per-cell scores of size $\frac{W}{8} \times \frac{H}{8} \times (64+1)$ (the extra channel denotes “no-keypoint”). A channel-wise softmax followed by a simple decoding/upsampling yields a dense keypoint heatmap at the input resolution.

Descriptor head. The fused feature is projected into a 64-dimensional descriptor space with a small convolutional block; the descriptor map is bilinearly upsampled to the input resolution and L_2 -normalized to produce per-pixel descriptors $D(u, v) \in \mathbb{R}^{64}$ suitable for standard similarity metrics.

3D geometry head. A shallow convolutional branch predicts per-pixel surface normals $N(u, v) \in \mathbb{R}^3$, which are unit-normalized and supervised by L_2 regression against pseudo-ground-truth normals derived from Depth Anything v2 [37]. Converting depth to normals reduces sensitivity to global scale and provides complementary geometric cues that improve descriptor discriminability in low-texture or repetitive regions.

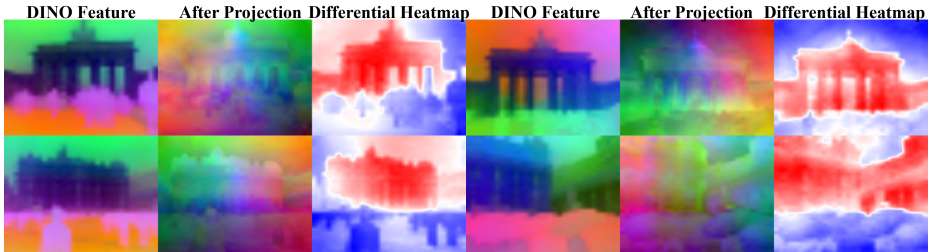


Fig. 4: PCA visualization of semantic projection. The differential heatmap highlights strengthened/suppressed responses in red/blue.

3.2 Distilled Semantic Embedding (DistilSem).

To incorporate high-level semantic priors into local feature matching without sacrificing efficiency, we propose the *Distilled Semantic Embedding (DistilSem)*, a lightweight semantic-injection mechanism based on compact teacher-student distillation. Our key observation is that directly deploying large foundation models at inference is computationally prohibitive, while training an independent semantic student backbone substantially increases model complexity and optimization difficulty. DistilSem resolves this tension by distilling only the task-relevant semantic information into a compact representation that can be seamlessly integrated into the shared matching backbone.

Specifically, a two-stage strategy is adopted. In the first offline stage, a small projection head $P(\cdot)$ is attached to a pre-trained DINOv3 [31] model to compress its high-dimensional token features into a matching-oriented semantic embedding. Then this projection head is frozen as the teacher model. Given an input image I , the frozen DINOv3 ViT-S backbone produces token features $\phi_{\text{DINO}}(I) \in \mathbb{R}^{h \times w \times 384}$, which are projected to a compact 64-dimensional teacher map:

$$T(u, v) = P(\phi_{\text{DINO}}(I)), \quad T(u, v) \in \mathbb{R}^{64}. \quad (1)$$

As illustrated in Fig. 4, the learned projection preserves matching-relevant semantic cues while suppressing responses from noisy or less informative regions.

In the second stage, the compact semantic representation is distilled into the matching network itself. Rather than introducing a separate semantic backbone, a lightweight semantic head $\psi_{\theta}(\cdot)$ is attached to the shared backbone features $F(u, v)$, yielding a student semantic map

$$S(u, v) = \psi_{\theta}(F(u, v)), \quad S(u, v) \in \mathbb{R}^{64}. \quad (2)$$

By distilling semantic knowledge through a compact projection into the shared CNN backbone, DistilSem provides two benefits: (i) efficiency, requiring no additional semantic encoder at inference; (ii) feature compatibility, as appearance and geometry are extracted from the same backbone and semantics are distilled into this shared feature space. DistilSem thus enhances local features with semantics at negligible parameter and runtime overhead.

3.3 Reliability Gated Fusion (RGFusion)

Local feature matching in real-world scenes inherently involves multi-source uncertainty: appearance cues are sensitive to illumination changes and repetitive textures, geometric cues derived from normals may be unreliable near discontinuities or reflective surfaces, and semantic cues vary in discriminative power depending on scene context. A principled fusion mechanism must therefore satisfy two complementary requirements: (i) *selective reliability modeling*, i.e., dynamically modulating the contribution of each modality according to its local reliability; and (ii) *conditional cross-modal reasoning*, i.e., allowing informative modalities to actively compensate ambiguous ones. Guided by these considerations, we introduce the *Reliability Gated Fusion (RGFusion)*, a lightweight yet expressive fusion module that combines channel-wise gated fusion and token-level cross-modal attention to integrate appearance, geometry, and semantics.

Let $\mathbf{d}, \mathbf{n}, \mathbf{s} \in \mathbb{R}^D$ denote the per-keypoint embeddings of the appearance descriptor, surface normal, and distilled semantic feature, respectively, all projected into a shared latent space. We first address modality reliability through a channel-wise gated fusion. Rather than treating all modalities equally, we learn input-adaptive gates that reweight each modality at the channel level. Specifically, we concatenate the three embeddings and pass them through a compact bottleneck network:

$$\mathbf{z} = \text{ReLU}(W_1[\mathbf{d}; \mathbf{n}; \mathbf{s}] + b_1), \quad \mathbf{z} \in \mathbb{R}^H, \quad H \ll 3D, \quad (3)$$

followed by a sigmoid projection producing modality-specific gates

$$\mathbf{g} = [\mathbf{g}_d; \mathbf{g}_n; \mathbf{g}_s] = \sigma(W_2\mathbf{z} + b_2) \in (0, 1)^{3D} \quad (4)$$

The gated fusion output is computed as

$$\mathbf{f}_{\text{gate}} = \mathbf{g}_d \odot \mathbf{d} + \mathbf{g}_n \odot \mathbf{n} + \mathbf{g}_s \odot \mathbf{s}. \quad (5)$$

This operation explicitly models *local modality reliability*: for example, when geometric cues are unstable, the network can suppress $\mathbf{n}$ while relying more on semantic or appearance channels; conversely, in low-texture regions, semantic or geometric cues may be amplified. Importantly, the bottleneck design ensures that this adaptive weighting introduces only marginal computational overhead.

While gated fusion handles how much each modality should contribute, it does not capture *how modalities should interact*. To this end, we introduce a lightweight cross-modal token attention mechanism that explicitly models conditional complementarity. We treat $\mathbf{d}$, $\mathbf{n}$, and $\mathbf{s}$ as a set of three modality tokens, form the token matrix $\mathbf{X} = [\mathbf{d}; \mathbf{n}; \mathbf{s}] \in \mathbb{R}^{3 \times D}$, and apply scaled dot-product attention:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{X}W_Q(\mathbf{X}W_K)^\top}{\sqrt{D}}\right), \quad (6)$$

$$\mathbf{O} = \mathbf{A}(\mathbf{X}W_V) \in \mathbb{R}^{3 \times D}. \quad (7)$$

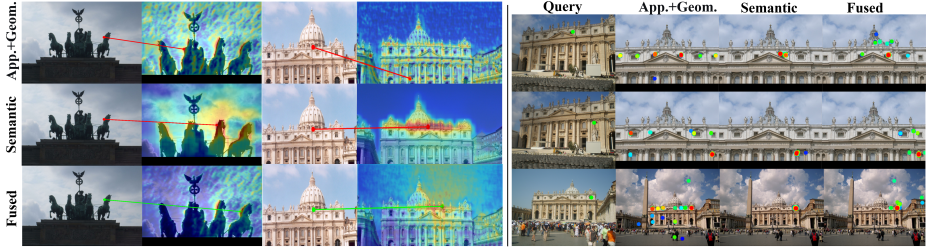


Fig. 5: Effect of semantic fusion on ambiguous correspondences. Heatmaps and top- k matches show correspondence strength, with red/blue denoting high/low relevance.

We take the updated descriptor token $\mathbf{O}_{0,:}$ and form an enhanced appearance representation via a residual connection:

$$\mathbf{f}_{\text{att}} = \mathbf{d} + \text{Proj}(\mathbf{O}_{0,:}). \quad (8)$$

This operation allows the descriptor to selectively absorb discriminative components from semantic or geometric tokens when needed, enabling conditional reasoning such as semantic disambiguation of geometrically similar structures or geometric reinforcement of visually ambiguous patterns. Since attention is performed over only three tokens per keypoint, its computational cost is negligible compared to spatial or graph-level attention.

Finally, the gated and attention-enhanced features are added to produce the final fused descriptor. As shown in Fig. 5, appearance–geometry features alone produce diffuse responses over locally similar structures, whereas incorporating semantic cues through RGFusion suppresses semantically inconsistent candidates and concentrates the responses around the correct correspondence. In summary, RGFusion combines reliability-aware channel weighting with conditional cross-modal interaction, achieving a favorable balance between descriptor discriminability and computational efficiency.

3.4 Supervision Strategy

Below we describe the losses used to supervise each task and the final multi-task objective.

1) **Keypoint Loss:** For keypoint supervision, we adopt the same strategy as SuperPoint [6]. The keypoint head outputs a logits map of size $\frac{W}{8} \times \frac{H}{8} \times 65$, where the last channel corresponds to the “no-keypoint” class. We use keypoints detected by the ALIKE detector [41] as pseudo ground-truth labels. For each cell (i, j) , let $k_{i,j}$ denote the target class index corresponding to the teacher keypoint location. The keypoint loss is computed using the Negative Log-Likelihood (NLL) loss:

$$L_{\text{keypoint}} = - \sum_{i,j} \log \left(\text{softmax}(\mathbf{K}_{i,j})_{k_{i,j}} \right), \quad (9)$$

2) **Descriptor Loss:** Given descriptors sampled from images I_A and I_B , we feed them into the feature fusion module to obtain the final descriptors $d_A \in \mathbb{R}^{m \times 64}$

and $d_B \in \mathbb{R}^{n \times 64}$. We then compute the similarity score matrix $S \in \mathbb{R}^{m \times n}$. Let M_{gt} denote the ground-truth matching matrix. Following [32], we minimize the negative log-likelihood of the predicted similarity matrix with respect to the ground truth:

$$L_{\text{desc}} = - \sum_{i,j} M_{\text{gt}}(i,j) \log S(i,j). \quad (10)$$

3) Normal Loss: To enforce geometric consistency, we supervise the surface normal prediction branch using a cosine similarity loss. For each pixel, the predicted normal vector n_{pred} is compared with the ground-truth normal vector n_{gt} , encouraging alignment of surface orientation. The normal loss is defined as:

$$L_{\text{normal}} = 1 - \frac{n_{\text{pred}} \cdot n_{\text{gt}}}{\|n_{\text{pred}}\| \|n_{\text{gt}}\|}, \quad (11)$$

where n_{pred} and n_{gt} denote the predicted and ground-truth surface normals, respectively.

4) Semantic Loss: To inject high-level semantic priors into local descriptors, we introduce a semantic loss that supervises the student semantic head using a frozen teacher. Let $S(p) \in \mathbb{R}^{64}$ and $T(p) \in \mathbb{R}^{64}$ denote the student and teacher semantic features at spatial location p , respectively. We employ a cosine-distance loss to align the normalized semantic embeddings:

$$L_{\text{sem}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left(1 - \left\langle \frac{S(p)}{\|S(p)\|_2}, \frac{T(p)}{\|T(p)\|_2} \right\rangle \right), \quad (12)$$

5) Total Loss: The total training loss is defined as a weighted sum of the above components:

$$L_{\text{total}} = L_{\text{keypoint}} + \alpha_1 L_{\text{desc}} + \alpha_2 L_{\text{normal}} + \alpha_3 L_{\text{sem}}, \quad (13)$$

where α_1 , α_2 , and α_3 are weighting factors that balance geometric, appearance, and semantic supervision, respectively. In our experiments, we empirically set $\alpha_1 = 2$, $\alpha_2 = \alpha_3 = 1$.

4 Experiments

We evaluate the proposed method on three tasks: relative pose estimation, homography estimation and visual localization. The implementation details, comparative methods, and some qualitative illustrations are given in the following.

4.1 Implementation Details

We train the proposed framework using a two-stage strategy to efficiently incorporate semantic and geometric priors. Pseudo ground-truth surface normals are generated using the pre-trained Depth-Anything v2 model [37]. Training data consists of a mixture of MegaDepth [15] and synthetic COCO [16] image pairs.

All input images are resized to 800×600 pixels. We optimize the network using the Adam optimizer with an initial learning rate of 4×10^{-3} and a batch size of 8. Training process takes 30 hours on single NVIDIA RTX 4090D GPU, no additional fine-tuning is performed during evaluation.

Comparative Methods. To assess both accuracy and efficiency, we compare our method with representative lightweight local feature extractors, including ORB [27], SuperPoint [6], ALIKE [41], SiLK [11], SFD2 [35], Cadar et al. [4], XFeat [26] and LiftFeat [20]. For ALIKE and SiLK, we adopt their smallest available backbones (ALIKE-Tiny and VGG-based variants) to ensure a fair comparison. For all methods, we retain the top 4096 detected keypoints and perform matching using mutual nearest neighbor (MNN) search. All methods are trained on MegaDepth, except SFD2, which is trained on Aachen v1.0 and web images, and Cadar et al., which is trained on ScanNet.

4.2 Relative Pose Estimation

Table 1: Relative pose results on MegaDepth and ScanNet. We report the AUC scores of translation and rotation errors at different thresholds. The best results are in bold, and the second-best are underlined.

Methods	Dim	ScanNet-1500			MegaDepth-1500		
		AUC@5	AUC@10	AUC@20	AUC@5	AUC@10	AUC@20
ORB (ICCV2011)	256-b	9.0	18.5	29.9	17.9	27.6	39.0
SuperPoint (CVPR2018)	256-f	12.5	24.4	36.7	37.3	50.1	61.5
DISK (NIPS2020)	128-f	9.6	19.3	30.4	35.0	51.4	64.9
ALIKE (TMM2022)	64-f	8.0	16.4	25.9	<u>49.4</u>	<u>61.8</u>	<u>71.4</u>
SiLK (ICCV2023)	32-f	15.9	30.1	44.5	39.9	55.1	66.9
ZippyPoint (CVPR2023)	256-b	8.5	15.1	23.7	23.6	34.9	46.3
SFD2 (CVPR2023)	128-f	13.0	25.6	38.3	44.9	58.2	68.3
Cadar et al. (ACCV2023)	256-f	20.3	36.8	<u>52.4</u>	35.7	51.2	64.2
XFeat (CVPR2024)	64-f	16.7	32.6	<u>47.8</u>	42.6	56.4	67.7
LiftFeat (ICRA2025)	64-f	18.5	34.9	51.2	44.7	59.5	70.3
Ours	64-f	<u>19.2</u>	<u>35.8</u>	52.5	50.2	62.5	72.7

Setup. We evaluate relative pose accuracy on two standard benchmarks with viewpoint, illumination variation and repetitive structures: MegaDepth [15] and ScanNet [5]. For fair comparison, MegaDepth images are resized to a max side of 1200 pixels, while ScanNet frames are evaluated at VGA resolution. We report area under the cumulative recall curve (AUC) for pose errors at thresholds $\{5^\circ, 10^\circ, 20^\circ\}$. Poses are recovered from the essential matrix using MAGSAC++ [2]. Descriptor format (floating-point ‘f’ or binary ‘b’) and dimensionality are indicated for each method.

Results. Table 1 reports the relative-pose AUC on ScanNet and MegaDepth. Our method achieves the best performance at all thresholds on MegaDepth,



Fig. 6: Qualitative matching results on the MegaDepth and ScanNet datasets. Our method maintains robust correspondences under challenging conditions involving significant illumination and viewpoint changes, as well as repetitive textures. Correct matches are drawn in green, while incorrect matches are shown in red.

showing a favorable accuracy–efficiency trade-off. The improvements are particularly evident on MegaDepth: $\text{AUC}@5^\circ$ increases from 42.6 (XFeat) to 50.2. Notably, compared to SFD2 [35], which also integrates semantic features via distillation, our method delivers consistently higher accuracy with only half the descriptor dimensionality. This suggests that the distilled semantic priors and adaptive re-weighting effectively mitigate semantic–geometric ambiguity in challenging outdoor scenes.

On ScanNet, our method ranks second at $\text{AUC}@5^\circ$ and $\text{AUC}@10^\circ$, only behind Cadar et al. (which is trained on ScanNet), while achieving the best $\text{AUC}@20^\circ$. This demonstrates strong cross-domain generalization to indoor scenes. Although Cadar et al. [4] also leverages feature-level semantic integration, our approach outperforms it at the coarser threshold and achieves substantially better results on MegaDepth, while maintaining a much more compact descriptor. This further confirms the robustness and cross-domain effectiveness of our method. The compact 64-d floating-point descriptor further reduces memory and matching costs, making the method suitable for large-scale or real-time applications.

Fig. 6 provides qualitative examples that illustrate this gain—our method finds reliable matches under strong illumination changes, large viewpoint variations, and repetitive textures where competing methods often fail.

4.3 Homography Estimation

Table 2: Mean Homography Accuracy (MHA, %) under illumination and viewpoint changes on HPatches. Best results are in **bold**, second-best are underlined.

Method	Illumination (MHA)			Viewpoint (MHA)		
	@3	@5	@7	@3	@5	@7
SiLK	78.5	82.3	83.8	48.6	59.6	62.5
SuperPoint	94.6	<u>98.5</u>	98.8	<u>71.1</u>	79.6	83.9
DISK	94.6	98.8	99.6	66.4	77.5	81.8
ORB	74.6	84.6	85.4	63.2	71.4	78.6
ZippyPoint	94.2	96.9	98.5	66.1	76.8	80.7
ALIKE	94.6	<u>98.5</u>	99.6	68.2	77.5	81.4
XFeat	95.0	98.1	98.8	68.6	81.1	86.1
LiftFeat	<u>95.6</u>	98.8	<u>99.2</u>	<u>71.1</u>	<u>81.7</u>	87.5
Ours	95.8	98.8	<u>99.2</u>	72.1	82.5	<u>87.2</u>

Setup. We evaluate homography estimation on the HPatches [1], which contains planar image sequences with varying viewpoint and illumination changes and provides ground-truth homography annotations for image pairs. For each method, point matches are first estimated and the homography is computed using MAGSAC++. We report the Mean Homography Accuracy (MHA), which measures the percentage of image pairs whose average corner reprojection error falls below pixel thresholds of $\{3, 5, 7\}$.

Results. Table 2 reports Mean Homography Accuracy on HPatches under both illumination and viewpoint changes. Our method obtains the best accuracy at the tighter thresholds across both splits, indicating more precise corner alignment when exact geometric registration is required. Concretely, on the viewpoint split we improve MHA@3 and MHA@5 over the LiftFeat by approximately 1.0 and 0.8 percentage points, respectively. These gains suggest that the proposed semantic priors directly help disambiguate repetitive or low-texture regions. At the more permissive MHA@7 threshold our method remains highly competitive, with only marginal differences against the very best performing baselines; this indicates that the proposed fusion improves the precision tail of the correspondence distribution while maintaining overall robustness.

4.4 Visual Localization

Setup. We evaluate on the Aachen Day-Night v1.1 [30] benchmark, which exhibits severe illumination changes between database and query images. All im-

Table 3: Visual localization accuracy on the Aachen Day-Night v1.1 benchmark. Results are reported as the percentage of correctly localized queries under position thresholds $\{0.25\text{m}, 0.5\text{m}, 5\text{m}\}$ and rotation thresholds $\{2^\circ, 5^\circ, 10^\circ\}$. Best results are shown in bold and second-best are underlined.

Method	Day			Night		
	0.25m / 2°	0.5m / 5°	5m / 10°	0.25m / 2°	0.5m / 5°	5m / 10°
SuperPoint	87.4	<u>93.2</u>	97.0	77.6	85.7	95.9
ORB	66.9	76.1	81.7	10.2	12.2	19.4
ZippyPoint	80.7	88.6	93.7	61.2	70.4	79.6
ALIKE	85.7	92.4	96.7	81.6	88.8	<u>99.0</u>
XFeat	84.7	91.5	96.5	77.6	<u>89.8</u>	98.0
LiftFeat	<u>87.6</u>	93.1	<u>97.1</u>	82.1	89.9	99.1
Ours	88.2	94.8	97.9	<u>81.7</u>	88.5	<u>99.0</u>

ages are resized such that the longer side does not exceed 1024 pixels, and 4,096 keypoints are extracted per image for all methods. We report localization accuracy using the official HLoc [28] evaluation protocol, where a query is considered correctly localized if the estimated pose falls within joint position and orientation error thresholds of $\{0.25\text{ m}, 0.5\text{ m}, 5\text{ m}\}$ and $\{2^\circ, 5^\circ, 10^\circ\}$, respectively.

Results. Table 3 reports visual localization accuracy on the Aachen Day–Night v1.1 benchmark. Our method attains the highest daytime performance at all three thresholds, surpassing LiftFeat by 0.6%, 1.7%, and 0.8%, and outperforming lightweight baselines (e.g., +3.3% vs XFeat and +2.4% vs ALIKE at 0.5m/5°). This demonstrates that distilled semantic cues noticeably improve matching reliability; the fused semantic–geometric descriptors yield both high-precision and coarse-level localization gains. The modest performance drop on nighttime queries can be attributed to a domain gap in the distilled semantic features: since the semantic distillation is performed on MegaDepth consisting predominantly of daytime images, the distilled semantic head tends to produce less stable responses under the very low-light night scenes. This degradation in semantic cue quality weakens the reliability of the adaptive fusion mechanism, leading to a slight decline in localization accuracy.

4.5 Runtime Analysis

We compare the efficiency of our method with SuperPoint, XFeat, and LiftFeat using VGA resolution. CPU runtime is measured on an Intel i5-1135G7 @ 2.4 GHz, and GPU runtime on a mobile NVIDIA MX450. As reported in Table 4, the complexity of our method is close to LiftFeat and substantially lower than SuperPoint. In terms of runtime, our approach achieves similar efficiency to LiftFeat on both CPU and GPU with only a minor overhead introduced by the semantic-enhanced matching pipeline. While XFeat remains faster, our method consistently outperforms it across three downstream tasks. Compared to SuperPoint, our method is significantly faster and more accurate. Overall, these

Table 4: Runtime and model complexity comparison.

Metrics	SuperPoint	XFeat	LiftFeat	Ours
Params (M)	1.30	0.66	0.85	0.92
FLOPs (G)	19.85	1.33	4.96	5.41
Desc. Dimension	256-f	64-f	64-f	64-f
Runtime / CPU (ms)	273	43	65	72
Runtime / GPU (ms)	25.9	3.8	4.9	5.4

results demonstrate that our approach achieves a favorable trade-off between accuracy and efficiency, making it suitable for deployment on resource-constrained platforms.

4.6 Ablation Study

Table 5: Ablation study on MegaDepth for relative pose estimation. All times are GPU inference time.

Method	AUC@5	AUC@10	AUC@20	Time (ms)
Baseline	44.7	59.5	70.3	4.9
BL+DistilSem+Naive Addition	47.8	60.6	70.7	5.1
BL+DistilSem+Reliability Modeling	49.4	61.7	71.9	5.3
BL+DistilSem+Cross-Modal Reasoning	48.9	61.3	71.4	5.2
Full Model	50.2	62.5	72.7	5.4
Replace DistilSem with DINOv3	53.6	65.9	74.6	738.8

Table 5 reports ablation results on MegaDepth-1500, where the baseline is LiftFeat. Integrating DistilSem into the baseline consistently improves pose estimation accuracy with negligible computational overhead, demonstrating its effectiveness as a cost-efficient semantic enhancement. Building upon DistilSem, both Reliability Modeling and Cross-Modal Reasoning further improve performance, indicating that uncertainty-aware weighting and cross-modal interaction provide complementary benefits for robust matching. Compared with Cross-Modal Reasoning, which only models interactions among the three modality features at each keypoint and thus has limited modeling capacity, gated fusion directly predicts three channel-wise reliability gates, making it a more effective and efficient way to modulate modality contributions. The full model achieves the best overall performance with only a small increase in runtime (5.4 ms), highlighting a favorable accuracy–efficiency trade-off. In contrast, replacing DistilSem with a large pretrained model (DINOv3) yields higher accuracy but introduces a dramatic computational cost (738.8 ms), underscoring the efficiency advantage of our distilled semantic representation.

5 Conclusion

This work presents a lightweight framework for robust local feature matching that tackles two practical challenges: the semantic–geometric ambiguity of purely appearance- or geometry-driven descriptors and the high deployment cost of large semantic backbones. By distilling high-level semantic priors into a compact student network and introducing an efficient fusion mechanism that adaptively re-weights semantic, appearance, and geometric cues, our method produces descriptors that are both discriminative and computationally efficient. Extensive experiments on three tasks confirm that the proposed design yields consistent accuracy improvements over lightweight baselines while preserving real-time throughput.

Despite these improvements, the distilled semantic priors are learned primarily from daytime imagery, which may reduce their reliability under extreme nighttime conditions. Future work will explore domain-adaptive distillation and the incorporation of additional complementary modalities to further enhance robustness and generalization across diverse environments.

Acknowledgements

This work was supported by the Natural Science Basic Research Program of Shaanxi (2024JCYBMS-467).

References

1. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 5173–5182 (2017)
2. Baráth, D., Nuskova, J., Ivashechkin, M., Matas, J.: Magsac++, a fast, reliable and accurate robust estimator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 1301–1309 (2020)
3. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). *Comput. Vision Image Understanding* **110**(3), 346–359 (2008), similarity Matching in Computer Vision and Multimedia
4. Cadar, F., Potje, G.A., Martins, R., Demonceaux, C., Nascimento, E.R.: Leveraging semantic cues from foundation vision models for enhanced local feature correspondence. In: Asian Conference on Computer Vision(ACCV). pp. 1268–1283 (2024)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scan-net: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 2432–2443 (2017)
6. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-Supervised Interest Point Detection and Description. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 337–33712 (2018)

7. Dong, S., Wang, S., Liu, S., Cai, L., Fan, Q., Kannala, J., Yang, Y.: Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*. pp. 16739–16752 (2025)
8. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint description and detection of local features. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*. pp. 8084–8093 (2019)
9. Edstedt, J., Bökman, G., Wadenbäck, M., Felsberg, M.: DeDoDe: Detect, Don't Describe — Describe, Don't Detect for Local Feature Matching. In: *International Conference on 3D Vision (3DV)*. pp. 148–157 (2024)
10. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*. pp. 19790–19800 (2024)
11. Gleize, P., Wang, W., Feiszli, M.: Silk: Simple learned keypoints. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICC)*. pp. 22442–22451 (2023)
12. He, X., Sun, J., Wang, Y., Peng, S., Huang, Q., Bao, H., Zhou, X.: Detector-free structure from motion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*. pp. 21594–21603 (2024)
13. Jiang, Q., Lu, X., Liang, D., Du, S.: Semmatcher: Semantic-aware feature matching with neighborhood consensus. *Journal of Visual Communication and Image Representation* p. 104611 (2025)
14. Li, J., Zhang, R., Li, G., Li, T.H.: Sde2d: Semantic-guided discriminability enhancement feature detector and descriptor. *IEEE Transactions on Multimedia* **27**, 275–286 (2025)
15. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*. pp. 2041–2050 (2018)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *European Conference on Computer Vision(ECCV)*. pp. 740–755 (2014)
17. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICC)*. pp. 17581–17592 (2023)
18. Liu, X., Shi, G., Wang, C., Zhang, X., Miao, Q.: Sgcm: Semantic and geometric consistency for robust aerial image matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **18**, 19696–19709 (2025)
19. Liu, X., Wang, C., Shi, G., Zhang, X., Miao, Q., Fan, M.: Sgad: Semantic and geometric-aware descriptor for local feature matching. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC)*. pp. 27095–27104 (October 2025)
20. Liu, Y., Lai, W., Zhao, Z., Xiong, Y., Zhu, J., Cheng, J., Xu, Y.: Liftfeat: 3d geometry-aware local feature matching. In: *IEEE International Conference on Robotics and Automation (ICRA)*. pp. 11714–11720 (2025)
21. Lowe, D.: Distinctive image features from scale-invariant key points. *Int. J. Comput. Vision* **20**, 91–110 (2003)
22. Murai, R., Dexheimer, E., Davison, A.J.: Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*. pp. 16695–16705 (2025)

23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
24. Pautrat, R., Suárez, I., Yu, Y., Pollefeys, M., Larsson, V.: Gluestick: Robust image matching by sticking points and lines together. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9706–9716 (2023)
25. Potje, G., , Cadar, F., Araujo, A., Martins, R., Nascimento, E.R.: Enhancing deformable local features by jointly learning to detect and describe keypoints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1306–1315 (2023)
26. Potje, G., Cadar, F., Araujo, A., Martins, R., Nascimento, E.R.: Xfeat: Accelerated features for lightweight image matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2682–2691 (2024)
27. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2564–2571 (2011)
28. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12716–12725 (2019)
29. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4937–4946 (2020)
30. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T.: Benchmarking 6dof outdoor visual localization in changing conditions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8601–8610 (2018)
31. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
32. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8918–8927 (2021)
33. Sun, P., Guan, B., Yu, Z., Shang, Y., Yu, Q., Barath, D.: Learning affine correspondences by integrating geometric constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27038–27048 (2025)
34. Wang, Y., He, X., Peng, S., Tan, D., Zhou, X.: Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21666–21675 (2024)
35. Xue, F., Budvytis, I., Cipolla, R.: Sfd2: Semantic-guided feature detection and description. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5206–5216 (2023)

36. Xue, F., Elflein, S., Leal-Taixé, L., Zhou, Q.: Matcha: Towards matching anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 27081–27091 (2025)
37. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: Advances in Neural Information Processing Systems(NeurIPS). vol. 37, pp. 21875–21911 (2024)
38. Yu, Y., Liu, S., Pautrat, R., Pollefeys, M., Larsson, V.: Relative pose estimation through affine corrections of monocular depth priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 16706–16716 (2025)
39. Zhang, Y., Shen, S., Zhao, X.: Mesa: Effective matching redundancy reduction by semantic area segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–18 (2025)
40. Zhang, Y., Zhao, X.: Mesa: Matching everything by segmenting anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 20217–20226 (2024)
41. Zhao, X., Wu, X., Miao, J., Chen, W., Chen, P.C.Y., Li, Z.: Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia* **25**, 3101–3112 (2022)
42. Zhong, C., Yang, C., Qi, J., Sun, F., Liu, H., Mu, X., Huang, W.: Sim2real object-centric keypoint detection and description. In: Proceedings of the AAAI Conference on Artificial Intelligence(AAAI). vol. 36, pp. 5440–5449 (2022)