TriCo-Net: Learning Semantically Aware Local Features via Triple Consistency

Longze Zhu[1,2], Li Yan[1,2]*, Hong Xie[1,2], Hao Wu[1,2], Shan Su[1,2], Binbing Wang[1], Xiaoteng Yang[1], Junjie Yuan[1], Aoran Li[3]

[1] Wuhan University, The School of Geodesy and Geomatics, Wuhan 430079, Hubei, China
[2] Hubei Luojia Laboratory, Wuhan 430079, Hubei, China
[3] Henan Normal University, The College of Software, Xinxiang 453000, Henan, China

Keywords: Local Feature Matching, Semantic-Aware Descriptors, Knowledge Distillation, Multi-Scale Consistency, Visual Localization.

Abstract

Local feature matching in complex scenes is hindered by semantic ambiguity, where detectors often latch onto transient or repetitive patterns. We present TriCo-Net, which learns semantically aware and discriminative local features by enforcing a Triple Consistency (TriCo) principle across implicit semantics, scale, and spatial context. During training, an Implicit Semantic Strategy (ISS) distills cues from a segmentation teacher to modulate keypoint reliability and descriptor learning, while introducing no overhead at inference. A Scale-wise Semantic Harmonizer (SSH) aligns and fuses feature-pyramid levels to ensure cross-scale coherence, and a Global Context Propagator (GCP) broadcasts scene-level dependencies to resolve local ambiguities. On Aachen Day–Night v1.1, TriCo-Net achieves strong and consistent gains in visual localization, particularly under night conditions, and exhibits robustness to blur, noise, and large homographies. Ablations show complementary benefits from ISS, SSH, and GCP, with ISS contributing most at tight thresholds and at night. TriCo-Net narrows the day–night performance gap while maintaining mid-range throughput, offering a practical trade-off between robustness and efficiency.

1. Introduction

Robust local feature matching is foundational to autonomous driving and SLAM, where long-term localization and relocalization must remain reliable under day–night shifts, dynamic objects, and adverse imaging conditions (Khatib et al., 2025, Zhu et al., 2022). To meet these requirements, recent work favors learned local features (Sarlin et al., 2019, Sattler et al., 2016), which jointly optimize keypoints and descriptors and consistently outperform handcrafted methods such as SIFT (Arandjelović and Zisserman, 2012, Lowe, 2004). Yet prevailing training pipelines rely almost exclusively on low-level geometric and photometric supervision and provide little high-level semantic guidance (Zhao et al., 2022a). This reliance induces a critical limitation: detectors develop a bias toward locally salient but semantically unstable features. As illustrated in Figure 1, the bias draws detectors to transient objects or repetitive textures, ultimately undermining matching in complex, real-world scenes.

To bridge this semantic gap, prior work has explored two main directions. One line of work employs semantic segmentation to filter keypoints (Zhao et al., 2022b, Long et al., 2025). This approach, however, suffers from two fundamental drawbacks. First, it introduces a dependency on computationally expensive external networks. More critically, it imposes a rigid, binary constraint that discards all keypoints within a masked region. This all-or-nothing strategy is brittle; in feature-sparse areas, it risks a catastrophic loss of all geometric information, rendering entire objects unmatchable. Alternatively, attention mechanisms have been employed to implicitly weigh feature importance (Li et al., 2023). Yet, without explicit semantic supervision, learned attention often regresses to highlighting textural saliency rather than true semantic stability, thereby reinforcing the original bias instead of resolving it (Li et al., 2023).

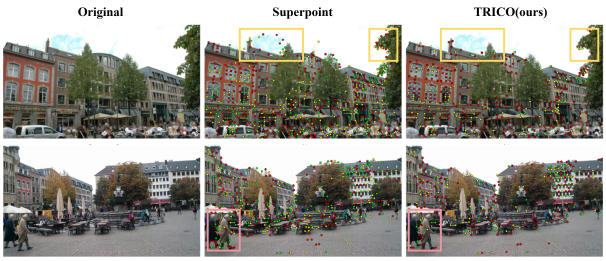

Figure 1. **Semantic bias in learned local feature detectors.** SuperPoint favors transient/repetitive textures—sky, foliage, dynamic objects—yielding sparse, unstable keypoints in low-texture (yellow) and occluded (red) regions. TriCo-Net suppresses these, focusing on semantically stable structures.

In essence, the limitations of prior work are rooted in two intertwined challenges. First, a supervision challenge: how to imbue feature detectors with semantic awareness without resorting to external models or expensive manual annotations. Second, a granularity challenge: how to assess the semantic stability of individual keypoints at the point-level, rather than relying on coarse, region-level heuristics or misleading texture-based cues. Meeting these challenges is fundamental to further advances in local feature matching.

To address these challenges, we introduce TriCo-Net, a novel framework that learns semantically robust local features by enforcing a Triple Consistency (TriCo) principle across the im-

* Corresponding author
Acknowledgments: This work was partially supported by the National Natural Science Foundation of China (Grants 42371451 and 42394061), and the Natural Science Foundation of Wuhan (No.2024040701010028).
Emails: lz.zhu@whu.edu.cn (LZ.Zhu); lyan@sgg.whu.edu.cn (L.Yan); hxie@sgg.whu.edu.cn (H.Xie); whwhuedusgg@whu.edu.cn (H.Wu); sushan@whu.edu.cn (S.Su); whu_wangbb@whu.edu.cn (BB.Wang); xtyang@whu.edu.cn (XT.Yang); 2021302141082@whu.edu.cn (JJ.Yuan); 3938448365@163.com (AR.Li).

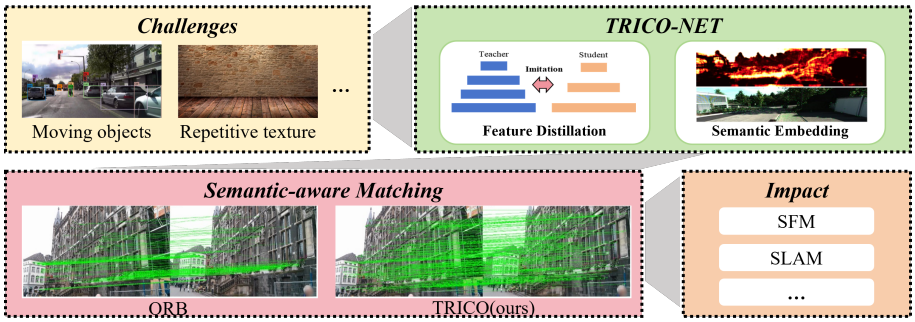

Figure 2. **TriCo-Net overview.** By addressing moving objects and repetitive textures via feature distillation and semantic embedding, TriCo-Net yields semantic-aware correspondences and improves downstream tasks such as SfM and SLAM.

plicit, scale, and spatial dimensions, as depicted in Figure 2. First, for Implicit Consistency, we employ an Implicit Semantic learning Strategy (ISS) that distills explicit semantic knowledge into the network during training, enabling efficient, dependency-free inference. Second, to enforce Scale Consistency, we introduce the Scale-wise Semantic Harmonizer (SSH), a novel module that adaptively fuses and aligns semantic cues from different levels of the feature pyramid. Finally, for Spatial Consistency, we propose the Global Context Propagator (GCP), which leverages a self-attention mechanism to broadcast scene-level context to all local features, ensuring their semantic interpretation is globally coherent. Collectively, these components empower TriCo-Net to produce descriptors that are both highly discriminative and semantically aware, leading to significant performance gains in challenging matching scenarios.

In summary, our main contributions are as follows:

- We propose TriCo-Net, a novel framework that learns semantically robust local features by enforcing a Triple Consistency (TriCo) principle across the implicit, scale, and spatial dimensions.
- We introduce an Implicit Semantic learning Strategy (ISS) that distills high-level knowledge into the feature network during training, enabling it to produce semantic-aware features at inference time with zero additional computational overhead.
- We design a novel Consistency Enhancement Module (CEM) to realize this principle, which consists of two architectural components: a Scale-wise Semantic Harmonizer (SSH) that enforces cross-scale consistency by adaptively fusing and aligning semantic cues, and a Global Context Propagator (GCP) that instills spatial consistency by modeling scene-level dependencies via self-attention.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details our proposed method, TriCo-Net. Section 4 presents the experimental results. Section 5 provides an in-depth discussion, and Section 6 concludes the paper.

2. Related Work

Our work builds on learned local feature extractors and introduces a semantic guidance paradigm. We first review advances in learning to detect and describe local features, and then discuss how semantic signals have been incorporated into feature learning.

2.1 Learned Local Feature Extraction

Local feature extraction has progressed from handcrafted detectors and descriptors—SIFT (Lowe, 2004), SURF (Bay et al., 2006), and ORB (Rublee et al., 2011)—to learned representations. Early deep methods improved descriptors given fixed keypoints (Tian et al., 2017, Mishchuk et al., 2017). Subsequent end-to-end frameworks jointly learn detection and description, yielding prominent sparse extractors such as SuperPoint (DeTone et al., 2018), D2-Net (Dusmanu et al., 2019), R2D2 (Revaud et al., 2019), and ASLFeat (Luo et al., 2020). In parallel, detector-free dense correspondence models learn features and matching within a unified architecture, as exemplified by LoFTR (Sun et al., 2021). Although these systems achieve strong geometric accuracy, their supervision is dominated by low-level cues. As a result, extracted features often emphasize textured or repetitive structures rather than stable, semantically meaningful landmarks. Our approach targets the representation itself: we seek features that encode semantic stability while remaining discriminative for geometry.

2.2 Semantic Guidance for Feature Extraction

To move beyond purely geometric supervision, several lines of work inject semantics at different stages. Post-hoc filtering with external segmentation models (Toft et al., 2018, Shi et al., 2019) removes keypoints on transient regions, but adds computation and offers coarse control. Implicit context modeling via Transformer attention (Jiang et al., 2021) captures long-range cues; however, without explicit semantic objectives the learned features may still prioritize textural salience over category-consistent stability. Knowledge distillation (Sarlin et al., 2019, Zhao et al., 2020) improves efficiency and transfers priors, yet prior studies seldom ensure that distilled semantics remain consistent across scales and spatial neighborhoods. We explicitly regularize this cross-scale and cross-context consistency during feature learning, aiming to produce semantically stable, match-ready features without heavy post-processing.

3. Proposed Method

In this section, we introduce TriCo-Net, our proposed framework for learning semantically robust local features. Our approach is built upon the Triple Consistency (TriCo) principle, which enforces semantic coherence across the implicit, scale, and spatial dimensions. We first present the overall architecture of TriCo-Net. Then, we detail the core components designed to enforce each consistency: the Implicit Semantic learning Strategy (ISS), the Scale-wise Semantic Harmonizer (SSH),

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-61-2026 | © Author(s) 2026. CC BY 4.0 License.

and the Global Context Propagator (GCP). Finally, we describe the joint training objective that integrates these components.

3.1 Overall Architecture

Our method employs a teacher-student framework, termed the Implicit Semantic learning Strategy (ISS), to instill semantic awareness into a Student Network via knowledge distillation, as illustrated in Figure 3. The teacher, a pre-trained ConvNeXt segmentation model (Liu et al., 2022), provides dual-level supervision during training: it supplies robust intermediate feature representations and derives a task-level reliability map from its final segmentation mask to guide the decoding stage of the student. Crucially, the teacher is discarded after training, allowing the student to operate at inference with no additional computational overhead.

The Student Network adopts a standard encoder-decoder architecture designed for efficient feature extraction. At its core is our proposed Consistency Enhancement Module (CEM), which processes multi-scale features from the encoder. The CEM consists of two sequential components: 1) the Scale-wise Semantic Harmonizer (SSH), which aligns semantic cues across feature pyramid levels to enforce scale consistency, and 2) the Global Context Propagator (GCP), which leverages self-attention to resolve local ambiguities by broadcasting global context, thereby ensuring spatial consistency. The decoder then processes these semantically-refined features to generate the final keypoint heatmap and descriptor maps.

3.2 Implicit Semantic learning Strategy

The Implicit Semantic learning Strategy (ISS) is realized via a dual-level knowledge distillation mechanism, comprising feature-level and task-level guidance. Task-level guidance employs the teacher-derived semantic and reliability map to spatially modulate the loss function for the student's final decoding stage. Concurrently, feature-level guidance compels the student's intermediate feature maps to mimic the teacher's more robust representations through a feature imitation loss.

Task-Level Semantic Guidance We define keypoint reliability as a product of Local Textural Saliency (LTS) and Global Semantic Stability (GST). This addresses the unreliability of keypoints on dynamic or non-rigid objects.

The GST map, $\mathcal{S}_{\text{GST}}$, is generated by assigning pre-defined weights to semantic categories obtained from a teacher network's segmentation mask. Table 1 shows these weights.

Table 1. **Stability weights for semantic categories.**

Semantic Category	Stability Type	Weight (ω)
Building, Road	Long-term Static	1.0
Vegetation	Short-term Static	0.5
Pedestrian, Vehicle	Dynamic	0.1
Tree, Sky	Unstructured	0.1

The LTS map, $\mathcal{S}_{\text{LTS}}$, uses detection scores from a pre-trained SuperPoint as pseudo-ground truth, as illustrated in Fig. 4. The final reliability ground-truth map $\mathcal{S}_{\text{gt}}$ is the element-wise product:

$$\mathcal{S}_{\text{gt}} = \mathcal{S}_{\text{LTS}} \odot \mathcal{S}_{\text{GST}} \tag{1}$$

The detection loss $\mathcal{L}_{\text{det}}$ is a binary cross-entropy loss between the student's predicted reliability map $\hat{\mathcal{S}}$ and the ground-truth map $\mathcal{S}_{\text{gt}}$.

$$\mathcal{L}_{\text{det}} = -\frac{1}{N}\sum_{p}\Big(\mathcal{S}_{\text{gt}}(p)\log(\hat{\mathcal{S}}(p)) + (1-\mathcal{S}_{\text{gt}}(p))\log(1-\hat{\mathcal{S}}(p))\Big) \tag{2}$$

The total descriptor loss, $\mathcal{L}_{\text{desc}}$, is a weighted sum of the inter-class and intra-class components:

$$\mathcal{L}_{\text{desc}} = w_{\text{inter}}\mathcal{L}_{\text{inter}} + w_{\text{intra}}\mathcal{L}_{\text{intra}} \tag{3}$$

The inter-class loss is formulated as a triplet loss over descriptors from different semantic categories:

$$\mathcal{L}_{\text{inter}} = \frac{1}{N}\sum\left(\|x_i^{c1} - x_j^{c1}\|^2 - \|x_i^{c1} - x_k^{c2}\|^2 + m\right) \tag{4}$$

where x_i^{c1}, x_j^{c1} are descriptors from class $c1$, and x_k^{c2} is a descriptor from a different class $c2$. The summation is performed over all valid triplets in the batch of size N, and m is the margin.

The intra-class loss is defined using an average precision (AP) based soft ranking objective, operating independently within each semantic class c:

$$\mathcal{L}_{\text{intra}} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i=1}^{N_c}(1 - \mho_{ap}[d_i^c; \mathcal{P}_i^c, \mathcal{N}_i^c]) \tag{5}$$

where d_i^c is the i-th query descriptor of class c, while $\mathcal{P}_i^c$ and $\mathcal{N}_i^c$ represent the sets of its positive (geometrically corresponding) and negative (non-corresponding) samples within the same class, respectively.

Feature-Level Semantic Guidance To ensure the student network learns semantically meaningful representations, we introduce a feature-level guidance mechanism that operates on the backbone encoder. The objective is to align the student's intermediate features with those of a powerful teacher network.

We extract feature maps $\mathbf{F}_{\text{T}}$ from an intermediate layer of pre-trained ConvNeXt encoder and corresponding feature maps $\mathbf{F}_{\text{S}}$ from a homologous layer in the student encoder. A feature imitation loss, $\mathcal{L}_{\text{feat}}$, is defined to minimize the discrepancy between them.

To handle disparities in channel dimensions or spatial resolution, a projection layer ϕ is applied to the student's features. The loss is formulated as the squared L2 norm of the difference:

$$\mathcal{L}_{\text{feat}} = \|\phi(\mathbf{F}_{\text{S}}) - \mathbf{F}_{\text{T}}\|_2^2 \tag{6}$$

The network is trained end-to-end by minimizing a total loss function, $\mathcal{L}_{\text{total}}$, which is a weighted sum of the individual loss components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{desc}}\mathcal{L}_{\text{desc}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} \tag{7}$$

3.3 Consistency Enhancement Module (CEM)

CEM promotes cross-scale and global consistency in the student representation. CEM consists of two components. A Scale-wise Semantic Harmonizer (SSH) first aligns and fuses the multi-scale features, which is then refined by a Global Context Propagator (GCP), which injects long-range dependencies. The overall procedure is summarized in Algorithm 1.

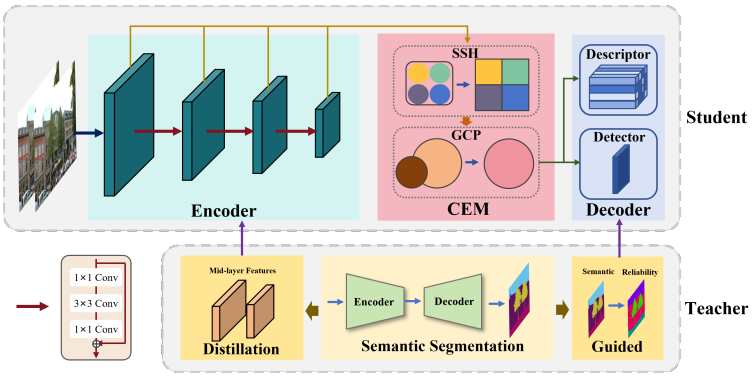

Figure 3. **Overall architecture of TriCo-Net.** ISS distills semantics from a ConvNeXt teacher into a joint detector and descriptor network. Mid-level features supervise the encoder and a segmentation-based reliability map guides the decoder, with no inference overhead. The CEM, comprising SSH and GCP, enforces cross-scale and spatial consistency in keypoints and descriptors.

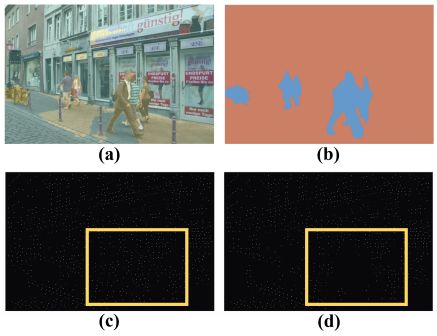

Figure 4. **Semantic-guided feature detection.** **(a)** semantic segmentation mask, **(b)** Global Semantic Stability, **(c)** Local Textural Saliency (LTS), **(d)** keypoint reliability

Scale-wise Semantic Harmonizer (SSH) To reconcile discrepancies across pyramid levels and obtain a unified, scale-consistent representation, we introduce the Scale-wise Semantic Harmonizer (SSH), as illustrated in Fig. 5. Given four-scale student features $\{\mathbf{F}_i^S \mid \mathbf{F}_i^S \in \mathbb{R}^{C_i \times H_i \times W_i}, i = 1, \ldots, 4\}$, SSH proceeds in three stages. We set the unified channel dimension to C=256 and the unified spatial size to $H/8 \times W/8$.

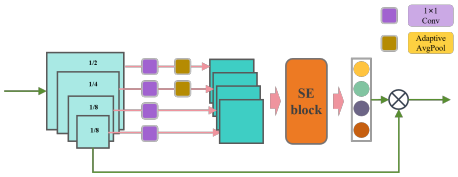

Figure 5. **Scale-wise Semantic Harmonizer.**

First, per-scale alignment standardizes both channels and spatial resolution. For each scale $i \in 1, \ldots, 4$, a 1×1 convolution $\mathcal{C}_i$ maps channels to 256, and adaptive average pooling $\mathcal{P}$ adjusts the spatial size to $H/8 \times W/8$:

$$\tilde{\mathbf{F}}_i \;=\; \mathcal{P}\big(\mathcal{C}_i(\mathbf{F}_i^S)\big) \;\in\; \mathbb{R}^{256\times H/8\times W/8}. \tag{8}$$

Next, we summarize each aligned feature into a compact scale descriptor to gauge cross-scale importance. Specifically, each $\tilde{\mathbf{F}}_i$ is globally averaged across spatial dimensions to obtain a channel descriptor $\mathbf{z}_i \in \mathbb{R}^{256}$ (see Fig. 6). We then concatenate

Algorithm 1: Consistency Enhancement Module (CEM)

Input: $\{F_i^S\}_{i=1}^4$; unified channel C=256; projection dim d=32
Output: F^{GCP}

```
// Scale-wise Semantic Harmonizer (SSH)
```
for $i \leftarrow 1$ **to** 4 **do**
  $\tilde{F}_i \leftarrow \mathrm{P}\big(\mathrm{C}_i(F_i^S)\big)$
$z \leftarrow \mathrm{Concat}\big(\mathrm{GAP}(\tilde{F}_1), \ldots, \mathrm{GAP}(\tilde{F}_4)\big)$;
$s \leftarrow \sigma\big(W_2\,\delta(W_1 z)\big)$; // $s = [s_1, \ldots, s_4]$
$F^{S,\mathrm{SSH}} \leftarrow \sum_{i=1}^4 s_i \cdot \tilde{F}_i$;
```
// Global Context Propagator (GCP)
```
$X \leftarrow \mathrm{reshape}(F^{S,\mathrm{SSH}})$;
$Q \leftarrow W_q X$; $K \leftarrow W_k X$; $V \leftarrow W_v X$;
$A \leftarrow \mathrm{softmax}\big((Q^\top K)/\sqrt{d}\big)$;
$Y \leftarrow V A^\top$;
$\widehat{F} \leftarrow \mathrm{reshape}(W_o Y)$;
$F^{\mathrm{GCP}} \leftarrow F^{S,\mathrm{SSH}} + \widehat{F}$;
return F^{GCP};

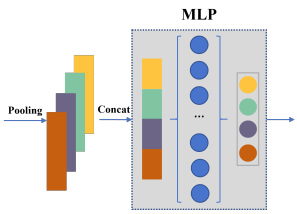

Figure 6. **SE Block.**

the four descriptors to form

$$\mathbf{z} = [\mathbf{z}_1; \mathbf{z}_2; \mathbf{z}_3; \mathbf{z}_4] \in \mathbb{R}^{1024}. \tag{9}$$

A two-layer MLP predicts scale-level attention scores:

$$\mathbf{s} = \sigma\big(\mathbf{W}_2, \delta(\mathbf{W}_1\mathbf{z})\big),\ \mathbf{W}_1 \;\in \mathbb{R}^{256\times1024},\quad \mathbf{W}_2 \in \mathbb{R}^{4\times256}, \tag{10}$$

where $\delta(\cdot)$ and $\sigma(\cdot)$ denote ReLU and Sigmoid, respectively. We write $\mathbf{s} = [s_1, s_2, s_3, s_4]^\top \in \mathbb{R}^4$.

Each attention score modulates its corresponding aligned feature, yielding reweighted scale features:

$$\mathbf{G}_i \;=\; s_i \cdot \tilde{\mathbf{F}}_i, \qquad i = 1, \ldots, 4. \tag{11}$$

The SSH output is then obtained by aggregating these reweighted

features via summation:

$$\mathbf{F}^{S,\mathrm{SSH}} = \sum_{i=1}^{4} \mathbf{G}_i \in \mathbb{R}^{256\times H/8\times W/8}. \tag{12}$$

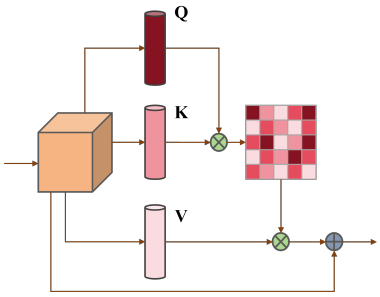

Figure 7. **Global Context Propagator.**

Global Context Propagator (GCP) To inject long-range dependencies on top of the locally enhanced features from SSH, we introduce the Global Context Propagator (GCP), as illustrated in Fig. 7. Let the channel count be C=256 and the number of spatial locations be S=$(H/8)(W/8)$. We first reshape $\mathbf{F}^{S,\mathrm{SSH}}$ into a matrix $\mathbf{X} \in \mathbb{R}^{C\times S}$, and set the projection dimension to d=32 to extract cross-location global relations with manageable computational cost.

Three 1×1 convolutions generate query, key, and value:

$$\begin{aligned}&\mathbf{Q} = \mathbf{W}_q\mathbf{X}, \quad \mathbf{K} = \mathbf{W}_k\mathbf{X}, \quad \mathbf{V} = \mathbf{W}_v\mathbf{X},\\ &\mathbf{W}_q,\ \mathbf{W}_k,\ \mathbf{W}_v \in \mathbb{R}^{d\times C}.\end{aligned} \tag{13}$$

We compute a scaled dot-product affinity across spatial positions and aggregate global context accordingly:

$$\mathbf{A} = \mathrm{softmax}\big(\tfrac{1}{\sqrt{d}}\,\mathbf{Q}^\top\mathbf{K}\big) \in \mathbb{R}^{S\times S}. \tag{14}$$

$$\mathbf{Y} = \mathbf{V}\,\mathbf{A}^\top \in \mathbb{R}^{d\times S}. \tag{15}$$

To align with the original channel space, we project back via a 1×1 convolution $\mathbf{W}_o \in \mathbb{R}^{C\times d}$ and reshape to the feature-map form:

$$\widehat{\mathbf{F}} = \mathcal{R}\big(\mathbf{W}_o\mathbf{Y}\big) \in \mathbb{R}^{C\times H/8\times W/8}, \tag{16}$$

where $\mathcal{R}$ denotes reshape. Finally, GCP is combined with the input in a residual manner to yield the CEM output:

$$\mathbf{F}^{\mathrm{GCP}} = \mathbf{F}^{S,\mathrm{SSH}} + \widehat{\mathbf{F}}. \tag{17}$$

4. Experiments

In this section, we first summarize the datasets and settings, then present four experiments: long-term large-scale localization, patch correspondence, an ablation study of ISS/SSH/GCP, and model inference runtime.

4.1 Datasets and Settings

Our experiments are conducted on the Aachen Day–Night v1.1 benchmark using the official split (Sattler et al., 2018). This version expands upon v1.0 by adding 2,369 reference images and 93 night queries, resulting in a total of 6,697 reference images and 1,015 query images (824 day, 191 night). The entire experimental process, including training and evaluation, was executed on an NVIDIA RTX 3090 GPU.

4.2 Long-term Large-scale Localization

We evaluate visual relocalization with Hierarchical Localization (HLoc) pipeline and report the percentage of successfully localized queries at (2°, 0.25m), (5°, 0.5m), and (10°, 5 m). Day and night are summarized separately for clarity and comparability. For readability, we group baselines into classic (C), semantics-using (S), and learned-feature (L) methods. Results are summarized in Table 2.

Classic methods (C). TriCo achieves the highest accuracy on both day and night. On day queries, it improves over the strongest classic baseline (ASv1.1) by +3.9, +2.6, and +0.8 percentage points at the 2°, 5°, and 10° thresholds, respectively (89.2/94.8/98.7% compared with 85.3/92.2/97.9% for ASv1.1). On night queries, the gains are substantial: +40.4, +42.1, and +34.1 points over ASv1.1 (80.2/91.1/98.4% compared with 39.8/49.0/64.3%), indicating strong robustness under low-light conditions.

Semantics-using methods (S). TriCo also leads overall among methods that incorporate semantics. On day queries, it exceeds SPADesc by +1.6 and +0.9 points at 2° and 5° (89.2/94.8% compared with 87.6/93.9%), and surpasses SFD2 by +1.3 at 10° (98.7% compared with 97.4%). On night queries, TriCo outperforms SPADesc at the tighter thresholds by +3.8 and +0.5 points and ties at 10° (80.2/91.1/98.4% compared with 76.4/90.6/98.4%).

Learned-feature methods (L). TriCo improves over the strongest prior daytime baseline (D2Net) by +6.6, +3.7, and +3.4 points at 2°, 5°, and 10°, respectively. At night, TriCo matches or exceeds the best baselines at medium and coarse thresholds (+1.3 and +2.4 points over D2Net at 5° and 10°), while D2Net remains slightly better at the tightest threshold (TriCo is lower by 3.5 points at 2°).

Overall, TriCo substantially reduces the day–night performance gap, showing a difference of only 3.7 points at (5°, 0.5m) (94.8% day; 91.1% night) compared with 43.2 points for ASv1.1 (92.2% day; 49.0% night). The larger nighttime improvements reflect that our method enhances robustness to illumination and appearance variations.

4.3 Patch Correspondence

We evaluate patch correspondence on 112 images randomly sampled from Aachen Day–Night v1.1 using homography-based synthetic pairs. Mean Matching Accuracy (MMA) measures the proportion of matches with reprojection error below a pixel threshold τ from 0 to 6px. TriCo is compared with SuperPoint, AKAZE, and ORB under four perturbations: normal and large homography amplitude, motion blur, and Gaussian noise (Figure 8).

Figure 8 shows that TriCo is competitive at tight thresholds and dominates at mid-to-loose thresholds across all perturbations. Under normal homographies, TriCo matches or slightly trails the best classical baseline at $\tau = 1$ px but attains the highest matching accuracy when averaged over $\tau \in [0, 6]$ px, and exhibits a clear advantage from $\tau \geq 3$ px. Under large-amplitude transforms, TriCo maintains stable growth in MMA

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-61-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 2. **Visual localization results on Aachen dataset**

Group	Method	Day	Night
		Accuracy @ Thresholds (%) $(2°, 0.25m)/(5°, 0.5m)/(10°, 5m)$	
C	ASv1.1	85.3/92.2/97.9	39.8/49.0/64.3
	SIFT	82.8/88.1/93.1	30.6/43.9/58.2
	CSL	52.3/80.0/94.3	29.6/40.8/56.1
	CPF	76.7/88.6/95.8	33.7/48.0/62.2
	TriCo (ours)	89.2/94.8/98.7	80.2/91.1/98.4
S	SSM	71.8/91.5/96.8	58.2/76.5/90.8
	VLM	62.4/71.8/79.9	35.7/44.9/54.1
	SMC	52.3/80.0/94.3	29.6/40.8/56.1
	SFD2	56.9/81.6/97.4	27.6/66.2/90.2
	SPADesc	87.6/93.9/96.8	76.4/90.6/98.4
	TriCo (ours)	89.2/94.8/98.7	80.2/91.1/98.4
L	Superpoint	80.5/87.4/94.2	42.9/62.2/76.5
	D2Net	82.6/91.1/95.3	83.7/89.8/96.0
	TriCo (ours)	89.2/94.8/98.7	80.2/91.1/98.4

References: ASv1.1 (Sattler et al., 2016); CSL (Svärm et al., 2016); CPF (Cheng et al., 2019); SSM (Shi et al., 2019); VLM (Xin et al., 2019); SMC (Toft et al., 2018); SFD2 (Xue et al., 2023); SPADesc (Meng et al., 2025); SIFT (Lowe, 2004); Superpoint (DeTone et al., 2018); D2Net (Dusmanu et al., 2019).

as τ increases, indicating improved tolerance to geometric distortion. With motion blur and Gaussian noise, TriCo consistently reaches the highest plateau values at $\tau \in [4, 6]$ px compared with SuperPoint (DeTone et al., 2018), AKAZE (Alcantarilla and Solutions, 2011), and ORB (Rublee et al., 2011), reflecting robustness to photometric degradations.

Qualitative examples in Figure 9 further confirm these patterns. TriCo yields denser and more uniformly distributed inliers aligned with scene geometry, maintaining coverage across facades and structural edges while suppressing long-baseline and cross-surface outliers. Under blur, warps, and noise, TriCo retains coherent correspondences where SuperPoint drifts and ORB largely fails.

Overall, TriCo achieves higher recall at practical thresholds without sacrificing precision. The consistent improvements across perturbations demonstrate that integrating SSH, GCP, and ISS enhances both geometric reliability and semantic robustness in visual correspondence.

4.4 Ablation Study

We conduct an ablation on ISS, SSH, and GCP under the Aachen Day–Night v1.1 protocol (Table 3). Disabling all components reduces the pipeline to SuperPoint, we use this configuration as the reference.

Any two-component configuration improves over this reference. Among pairs, ISS+GCP is strongest, reaching 86.8–97.4% (day) and 69.5–92.2% (night); ISS+SSH is close behind, while SSH+GCP is weaker, especially at night. These results indicate that including ISS is particularly beneficial and that ISS and GCP together provide a strong base.

Marginal contributions, computed by adding each module to the corresponding two-component variant, are consistent across thresholds. ISS adds +2.5–4.8 points (day) and +11.6–20.2 points (night); GCP adds +1.4–2.8 (day) and +6.95–12.7 (night); SSH adds +1.3–2.4 (day) and +6.2–10.7 (night). In terms of contribution, ISS ranks first, GCP ranks second, and SSH ranks third, with the advantage of ISS most pronounced at night and at the tightest tolerance.

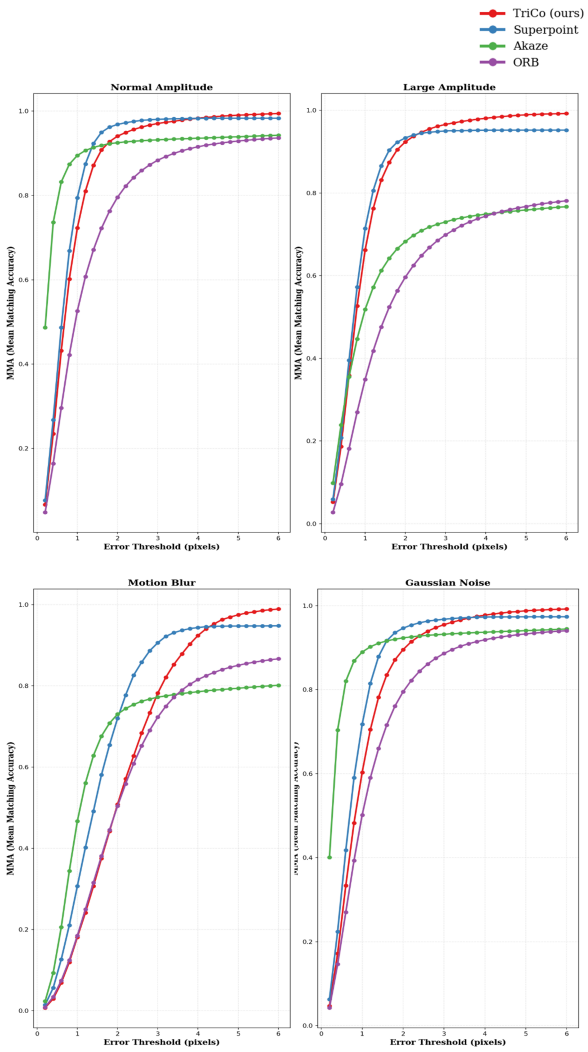

Figure 8. **Comparisons on Aachen dataset with Mean Matching Accuracy.**

4.5 Model Inference

Table 4 compares per-image latency for a fixed input size of 1024×1024. The runtimes range from 15.2 ms for Superpoint to 116.8 ms for ASLFeat. TriCo records 53.7 ms, corresponding to approximately 18.6 frames per second on the evaluated setup.

In relative terms, TriCo runs about 40% faster than R2D2 (77.6 ms) and more than 100% faster than ASLFeat (116.8 ms), while it is roughly 35% slower than SFD2 (39.1 ms). Superpoint remains the fastest method, reaching 15.2 ms, equivalent to 65.8 frames per second. The latency ranking in ascending order is Superpoint, SFD2, TriCo, R2D2, and ASLFeat.

Overall, TriCo offers mid-range throughput at the standard 1 megapixel resolution. It is suitable when sub-60 ms latency is acceptable, providing a clear speed advantage over heavier learned-feature baselines while trading some runtime efficiency against the fastest detectors.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-61-2026 | © Author(s) 2026. CC BY 4.0 License.

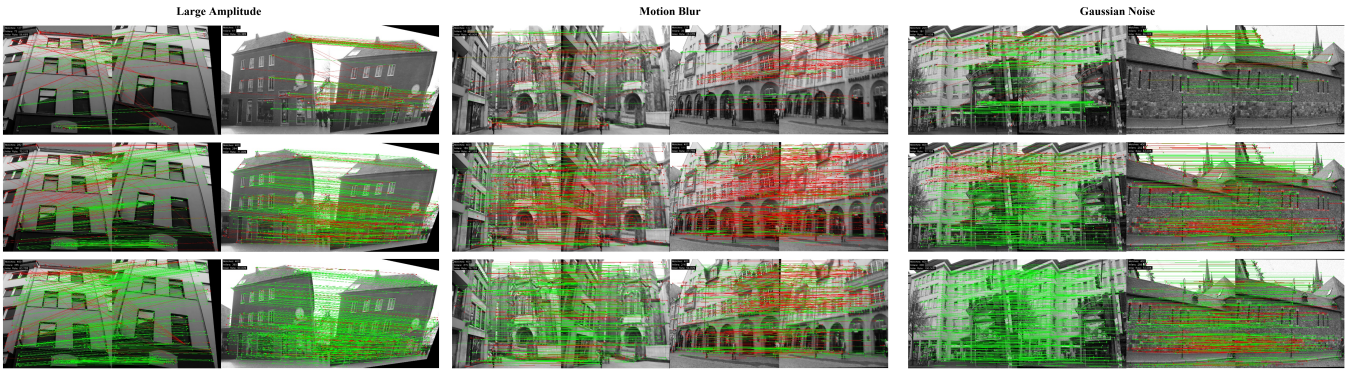

Figure 9. **Qualitative visualization of feature matching results across different methods.** The three rows, from top to bottom, show the results for ORB, SuperPoint, and TriCo (ours), respectively.

Table 3. **Ablation study on Aachen dataset.**

ISS	SSH	GCP	Day	Night
			Accuracy @ Thresholds (%) $(2^\circ, 0.25m)/(5^\circ, 0.5m)/(10^\circ, 5m)$	
×	×	×	80.5/87.4/94.2	42.9/62.2/76.5
✓	✓	×	86.4/92.2/97.3	67.5/82.6/91.4
✓	×	✓	86.8/92.6/97.4	69.5/84.0/92.2
×	✓	✓	84.4/91.0/96.2	60.0/76.0/86.8
✓	✓	✓	89.2/94.8/98.7	80.2/91.1/98.4

Table 4. **Comparison of running time across different models**

Model	Input size	Running time (ms)
Superpoint	1024 × 1024	15.2
R2D2	1024 × 1024	77.6
ASLFeat	1024 × 1024	116.8
SFD2	1024 × 1024	39.1
TriCo (ours)	1024 × 1024	53.7

References: Superpoint (DeTone et al., 2018); R2D2 (Revaud et al., 2019); ASLFeat (Luo et al., 2020); SFD2 (Xue et al., 2023). TriCo is our method.

5. Discussion

TriCo-Net advances local feature matching by infusing semantic awareness via our Triple Consistency (TriCo) principle. Compared to hard semantic masks or purely attention-based models, our soft supervision and consistency constraints maintain a richer, more robust feature set. This directly enhances performance in challenging conditions like day-night transitions and large viewpoint changes.

Nonetheless, analyzing failure cases reveals boundary conditions. As shown in Table 2, our accuracy at the tightest nighttime threshold (2°/0.25m) slightly trails D2Net. This reflects an intentional trade-off: to maintain overall semantic stability in low light, TriCo-Net actively suppresses textural saliency, occasionally discarding fine-grained but noisy geometric corners. Additionally, severe occlusions or motion blur in highly dynamic environments (e.g., dense crowds) can impair the teacher’s semantic parsing, limiting temporal robustness for SLAM applications. Finally, GCP’s full self-attention incurs quadratic complexity, limiting scalability to high-resolution images.

Moving forward, a primary goal is to enhance the inherent generalization of the semantic representations to highly diverse and dynamic scenes. To achieve this, future work will explore self-supervised domain adaptation, efficient attention mechanisms, and data-driven approaches to adaptively learn stability weights rather than relying on fixed priors.

6. Conclusion

In this work, we introduced TriCo-Net, a novel framework that fundamentally enhances the semantic robustness of local features. By enforcing our Triple Consistency principle, TriCo-Net successfully bridges the gap between textural saliency and semantic stability, leading to state-of-the-art performance, particularly under challenging day-night conditions. Our findings demonstrate that modeling semantic consistency is a critical and effective strategy for overcoming the limitations of traditional feature matchers in complex real-world environments.

Looking forward, the principles established by TriCo-Net open several promising avenues. The future of robust feature matching lies not only in refining geometric precision but also in achieving true semantic generalization across diverse domains and modalities. We envision next-generation models that can learn semantic stability in a self-supervised manner, adapt dynamically to any scene context, and integrate seamlessly with downstream tasks like SLAM and robotics. This work represents a significant step towards creating truly universal and context-aware local features for perception systems.

References

Alcantarilla, P. F., Solutions, T., 2011. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell*, 34(7), 1281–1298.

Arandjelović, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval. *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2911–2918.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features. *European conference on computer vision*, Springer, 404–417.

Cheng, W., Lin, W., Chen, K., Zhang, X., 2019. Cascaded parallel filtering for memory-efficient image-based localization. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1032–1041.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 224–236.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T., 2019. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*.

Jiang, W., Trulls, E., Hosang, J., Tagliasacchi, A., Yi, K. M., 2021. Cotr: Correspondence transformer for matching across images. *Proceedings of the IEEE/CVF international conference on computer vision*, 6207–6217.

Khatib, F., Moran, D., Trostianetsky, G., Kasten, Y., Galun, M., Basri, R., 2025. Generalizable visual localization for gaussian splatting scene representations. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 178–189.

Li, Y., Wang, J., Dai, X., Wang, L., Yeh, C.-C. M., Zheng, Y., Zhang, W., Ma, K.-L., 2023. How does attention work in vision transformers? A visual analytics attempt. *IEEE transactions on visualization and computer graphics*, 29(6), 2888–2900.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Long, J., Wang, F., Liu, M., Wang, Y., Zou, Q., 2025. DOG-SLAM: Enhancing dynamic visual SLAM precision through GMM-based dynamic object removal and ORB-boost. *IEEE Transactions on Instrumentation and Measurement*.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91–110.

Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L., 2020. Aslfeat: Learning local features of accurate shape and localization. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6589–6598.

Meng, H., Lu, H., Ding, B., Wang, Q., 2025. SPADesc: Semantic and parallel attention with feature description. *Neurocomputing*, 625, 129567.

Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J., 2017. Working hard to know your neighbor’s margins: Local descriptor learning loss. *Advances in neural information processing systems*, 30.

Revaud, J., De Souza, C., Humenberger, M., Weinzaepfel, P., 2019. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. Orb: An efficient alternative to sift or surf. *2011 International conference on computer vision*, Ieee, 2564–2571.

Sarlin, P.-E., Cadena, C., Siegwart, R., Dymczyk, M., 2019. From coarse to fine: Robust hierarchical localization at large scale. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12716–12725.

Sattler, T., Leibe, B., Kobbelt, L., 2016. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9), 1744–1756.

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J. et al., 2018. Benchmarking 6dof outdoor visual localization in changing conditions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8601–8610.

Shi, T., Shen, S., Gao, X., Zhu, L., 2019. Visual localization using sparse semantic 3d map. *2019 IEEE international conference on image processing (ICIP)*, IEEE, 315–319.

Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X., 2021. Loftr: Detector-free local feature matching with transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8922–8931.

Svärm, L., Enqvist, O., Kahl, F., Oskarsson, M., 2016. City-scale localization for cameras with known vertical direction. *IEEE transactions on pattern analysis and machine intelligence*, 39(7), 1455–1461.

Tian, Y., Fan, B., Wu, F., 2017. L2-net: Deep learning of discriminative patch descriptor in euclidean space. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 661–669.

Toft, C., Stenborg, E., Hammarstrand, L., Brynte, L., Pollefeys, M., Sattler, T., Kahl, F., 2018. Semantic match consistency for long-term visual localization. *Proceedings of the European Conference on Computer Vision (ECCV)*, 383–399.

Xin, Z., Cai, Y., Lu, T., Xing, X., Cai, S., Zhang, J., Yang, Y., Wang, Y., 2019. Localizing discriminative visual landmarks for place recognition. *2019 International conference on robotics and automation (ICRA)*, IEEE, 5979–5985.

Xue, F., Budvytis, I., Cipolla, R., 2023. Sfd2: Semantic-guided feature detection and description. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5206–5216.

Zhao, L., Peng, X., Chen, Y., Kapadia, M., Metaxas, D. N., 2020. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6528–6537.

Zhao, X., Wu, X., Miao, J., Chen, W., Chen, P. C., Li, Z., 2022a. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia*, 25, 3101–3112.

Zhao, Y., Xiong, Z., Zhou, S., Peng, Z., Campoy, P., Zhang, L., 2022b. KSF-SLAM: a key segmentation frame based semantic SLAM in dynamic environments. *Journal of Intelligent & Robotic Systems*, 105(1), 3.

Zhu, L., Kang, Z., Zhou, M., Yang, X., Wang, Z., Cao, Z., Ye, C., 2022. CMANet: Cross-modality attention network for indoor-scene semantic segmentation. *Sensors*, 22(21), 8520.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-61-2026