

ViMatch: Video-Driven Self-Supervised Local Feature Tracking for Visual Navigation

Yuqing Wang¹, Graduate Student Member, IEEE, Yan Wang², Hailiang Tang², Tisheng Zhang³,
Member, IEEE, Yilong Yuan⁴, Pan Jiang⁴, and Xiaoji Niu*⁵, Member, IEEE

Abstract—Local feature tracking is an essential problem in computer vision, aiming to estimate stable keypoint trajectories across consecutive frames for downstream tasks such as Visual Navigation and Structure-From-Motion (SfM) systems. Although deep learning-based matching methods have achieved remarkable results, their performance may degrade in unseen real-world environments, as they are typically trained on specific benchmark datasets. In this paper, we propose ViMatch, a lightweight local feature tracking framework designed to incorporate unlabeled video sequences into the learning process. Unlike traditional pipelines that rely on complex 3D reconstruction, ViMatch leverages local consistency between adjacent video frames to learn stable feature correspondences. Specifically, the model is first pretrained on standard datasets and then adapted using consecutive frames from SpatialVID without manual correspondence annotations. Unlike adaptation strategies based on static correspondence supervision or large-scale global matching, ViMatch formulates local feature tracking adaptation as a video-driven consistency optimization process, where teacher-guided correspondence priors are progressively refined through intra-frame spatial stability and inter-frame temporal coherence. By extending the descriptor invariance of keypoints from single-frame random perturbations to continuous video streams, and by further filtering out outliers through robust fitting and pixel-level consistency checks, ViMatch enables teacher-initialized self-supervised adaptation from unlabeled videos. Experimental results demonstrate that ViMatch significantly improves long-term feature tracking performance, achieving a 2–8× improvement in inference speed compared with existing methods under the local tracking setting. When integrated into the OpenVINS visual-inertial odometry system as a feature tracking component, ViMatch effectively reduces the absolute translation error by up to 70% on the evaluated datasets.

Index Terms—Index Terms—Local feature tracking, self-supervised adaptation, video-based correspondence estimation.

I. INTRODUCTION

LOCAL feature tracking aims to estimate the temporal trajectories of detected keypoints across consecutive frames within constrained observation regions, which is an essential component of visual navigation systems. Reliable feature trajectories provide fundamental measurements for downstream geometric estimation, including Structure from Motion (SfM) [1], visual localization [2], [3], and Visual Navigation [4]. In contrast to general image matching methods that estimate correspondences between image pairs without predefined local search constraints, local feature tracking assumes that initial keypoints are available and focuses on accurate and stable correspondence estimation within limited observation regions.

Although handcrafted descriptors such as ORB [5] have been widely adopted, they often yield limited matching recall in challenging scenarios, including illumination changes and extreme weather conditions. In recent years, learning-based methods have emerged as promising alternatives, offering higher accuracy [6]–[9] and denser correspondences. However, due to the lack of ground-truth correspondences in in-the-wild scenarios, existing approaches are typically trained on domain-specific datasets (e.g., indoor or outdoor) in isolation. This domain-specific training limits their generalization to unseen environments and restricts practical deployment in diverse real-world applications. Although GIM [10] improves cross-domain correspondence estimation through large-scale video-based learning, it is primarily designed for general image matching with unconstrained correspondence estimation. Its inference speed remains limited by the complexity of the original network architecture.

To address these limitations, we propose ViMatch, a lightweight, video-driven local feature tracking framework designed to exploit unlabeled video sequences for improving feature trajectory stability. As illustrated in Fig. 1, ViMatch improves local descriptor invariance by refining keypoint trajectories with self-supervised signals from consecutive video frames, targeting local feature tracking within bounded observation regions. Inspired by recent advances in video-based representation learning, ViMatch differs from general video-based correspondence learning frameworks by focusing on local trajectory refinement around detected keypoints and ex-

Copyright (c) 2026 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFB3909200. (Corresponding author: Xiaoji Niu.)

Yuqing Wang is with the GNSS Research Center, Wuhan University, Wuhan 430079, China, and also with the Electronic Information School, Wuhan University, Wuhan 430079, China (e-mail: m18071357432@163.com).

Yan Wang, Hailiang Tang, are with the GNSS Research Center, Wuhan University, Wuhan 430079, China, also with the Hubei Technology Innovation Center for Spatiotemporal Information and Positioning Navigation (e-mail: thl@whu.edu.cn; wstephen@whu.edu.cn).

Tisheng Zhang and Xiaoji Niu are with the GNSS Research Center, Wuhan University, Wuhan 430079, China, also with the Electronic Information School, Wuhan University, Wuhan 430079, China, also with Hubei Technology Innovation Center for Spatiotemporal Information and Positioning Navigation, Wuhan 430079, China, and also with Hubei Luojia Laboratory, Wuhan 430079, China (e-mail: zts@whu.edu.cn; xjniu@whu.edu.cn).

Yilong Yuan and Pan Jiang are with the Tencent Technology, Beijing 100000, China (e-mail: yosonyuan@tencent.com; mambajiang@tencent.com).

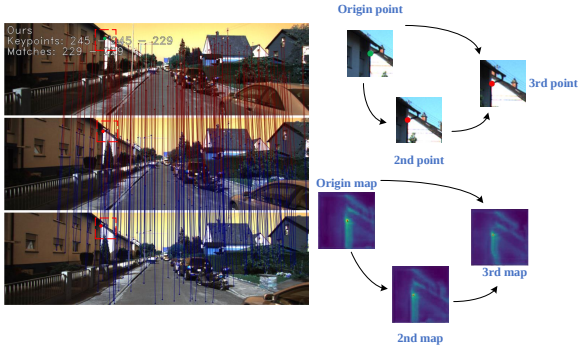


Fig. 1. **From sparse to dense.** We improve the local invariance of feature descriptors by aligning and refining 2D keypoints using self-supervised signals from consecutive video frames. This formulation is designed for local feature tracking and visual navigation scenarios, where correspondences are estimated within bounded local regions.

exploiting intrinsic spatiotemporal consistency rather than large-scale correspondence propagation.

Unlike video-based self-training strategies that primarily emphasize correspondence propagation, ViMatch introduces explicit spatial and temporal consistency constraints to improve local feature tracking stability under bounded observation regions. Therefore, the key contribution of ViMatch lies in formulating local feature tracking adaptation as a measurement consistency optimization framework combined with a lightweight local tracking architecture, rather than simply applying video-based adaptation.

ViMatch generates candidate trajectories by applying a pretrained model to consecutive frames from video streams. Outliers are then filtered via robust cross-frame fitting, and the model is refined by enforcing single-frame consistency, which improves local spatial invariance, and multi-frame consistency, which maintains stable keypoint trajectories across frames. This design allows ViMatch to handle challenging visual conditions, including motion blur, dynamic blur, and illumination variations.

Since ViMatch relies on local observation windows, it is primarily designed for short-to-moderate displacement feature tracking rather than unconstrained global image matching. When the true correspondence moves beyond the predefined search region due to large motion, severe occlusion, or rapid viewpoint changes, tracking failures may occur. In practical visual navigation systems, this limitation can be mitigated through coarse-to-fine pyramid inference and periodic keypoint re-detection.

The resulting framework is lightweight, teacher-initialized, and self-supervisedly refined using unlabeled videos, while remaining compatible with existing visual navigation pipelines. It provides an effective balance between computational efficiency, robustness, and long-term tracking accuracy.

The main contributions of the proposed video-driven local feature tracking method are summarized as follows:

- We formulate local feature tracking adaptation as a video-driven measurement consistency optimization problem, enabling continuous refinement of feature trajectories from unlabeled video streams without requiring explicit

correspondence annotations.

- We introduce a hybrid consistency learning strategy that jointly exploits intra-frame spatial stability and inter-frame temporal coherence, improving long-term feature tracking robustness under local observation constraints.
- Extensive experiments are conducted to evaluate the proposed framework on visual navigation, cross-dataset feature tracking, efficiency, and robustness benchmarks under diverse visual conditions.

The remainder of this paper is organized as follows. Section II reviews deep learning-based image tracking methods. Section III presents the lightweight network architecture, the proposed hybrid self-supervised adaptation module, and the associated loss functions. Section IV describes the training details and evaluates the method across various visual navigation and feature tracking tasks. Finally, Section V concludes the paper.

II. RELATED WORK

Recent advances in image matching have focused on learning robust features from domain-specific datasets. Early works [11]–[14] explored improving single-image descriptors using supervised or self-supervised signals, demonstrating improved accuracy under specific domain conditions. More recently, transformer-based architectures have been applied to image matching: Sarlin et al. [15] proposed a model that takes a pair of images as input, achieving significant improvements in matching accuracy and robustness, particularly under challenging conditions such as viewpoint and illumination changes. Semi-dense and dense matching methods [6], [7] further increase the density of correspondences, benefiting downstream tasks such as 3D reconstruction, visual localization, and visual odometry (VO). Despite these advances, existing learning-based methods are generally trained separately on indoor datasets (e.g., ScanNet [16]) and outdoor datasets (e.g., MegaDepth [17]), which limits their generalization to complex, dynamic real-world scenarios.

To address cross-domain generalization, methods such as GIM attempt to propagate pseudo-labels across diverse domains to extend the inference capability of the network to unseen environments. However, GIM primarily focuses on general image matching through large-scale video-based learning and global correspondence estimation, which requires relatively complex network architectures. In contrast, ViMatch targets lightweight local feature tracking for visual navigation, where correspondences are refined around detected keypoints within bounded local observation regions. Rather than relying solely on large-scale correspondence propagation, ViMatch introduces explicit spatial and temporal consistency constraints to improve long-term feature tracking stability.

In addition to generalization, computational efficiency is a critical concern in practical applications. Global feature methods, such as SEKD [18] and DALF [19], extract holistic representations from entire images, which can support sparse 3D reconstruction but incur high computational cost, restricting their use in high-resolution or real-time settings. In contrast, local feature methods focus on patch-level descriptors to estimate correspondences within restricted regions

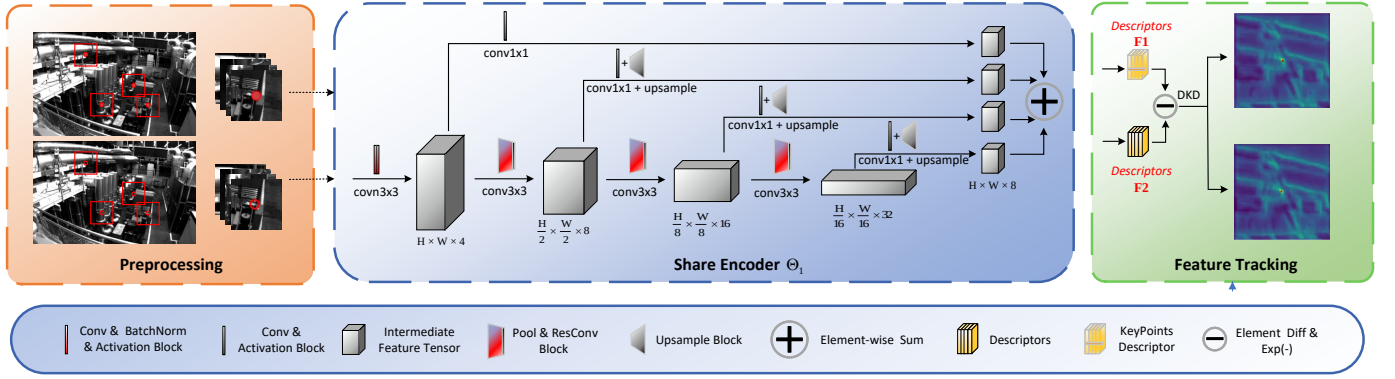


Fig. 2. **The pipeline of the proposed method.** In the **preprocessing** stage, fixed-size image patches surrounding the keypoints are extracted. The method achieves exceptional speed through shallow convolutional operations, followed by the generation of a compact dense descriptor map $\mathbf{D}$ in the subsequent encoding phase. Sub-pixel feature locations are obtained through similarity computation and differentiable feature extraction.

[20]–[22], achieving competitive performance while reducing computational overhead.

Building on these observations, ViMatch aims to achieve an effective balance between adaptation capability and computational efficiency for local feature tracking. It integrates patch-level teacher-initialized self-supervised adaptation with single-frame and multi-frame consistency constraints from unlabeled video sequences, enabling robust feature trajectory estimation for visual navigation. By relying on local observation regions rather than global correspondence search, ViMatch reduces network complexity while maintaining tracking accuracy and real-time performance for practical visual perception tasks.

III. METHODOLOGY

A. Overview

For clarity, we first define the notation used throughout this section. Let $\mathbf{I}_t$ denote the image at time step t in an ordered image sequence. The i -th keypoint location in frame t is represented as $\mathbf{x}_i^t$, and its corresponding descriptor is denoted as $\mathbf{d}_i^t$. The dense descriptor map extracted from the network is represented as $\mathbf{D}_t$, and $\mathbf{S}_i^t(u, v)$ denotes the similarity response map between a query descriptor and candidate locations.

Given an ordered image sequence $\{\mathbf{I}_t\}_{t=1}^T$, the objective is to estimate the spatial trajectories of a set of feature points across frames.

Let $\{\mathbf{x}_i^1\}_{i=1}^K$ denote the initial keypoints detected in the reference frame $\mathbf{I}_1$.

The task is to recover their corresponding states $\mathbf{x}_i^t, t = 1, \dots, T$, such that spatial consistency and temporal coherence are jointly preserved.

From a measurement-theoretic standpoint, feature tracking can be interpreted as a sequential state estimation problem. Each keypoint location $\mathbf{x}_i^t$ represents a latent spatial state, while the surrounding image patch provides an observation. Descriptor similarity serves as a likelihood function that links the latent state to image measurements. The overall framework therefore seeks to maximize correspondence likelihood while minimizing spatial variance and temporal drift.

Stage I: Teacher-Guided Initialization via Knowledge Distillation

In the early training stage, a teacher–student paradigm is adopted to establish a structured feature embedding space. A

pretrained teacher network provides supervision signals derived from ground-truth correspondences available in standard datasets. The student network, termed *ViMatch*, is trained to mimic the alignment behavior of the teacher through knowledge distillation.

ViMatch constructs similarity maps using an exponential element-wise differencing formulation. Given a query descriptor $\mathbf{d}_i^t$ extracted from keypoint $\mathbf{x}_i^t$ and a dense descriptor map $\mathbf{D}_{t+1}$ from the target frame, the similarity response is defined as

$$\mathbf{S}_i^{t+1}(u, v) = \exp(-\|\mathbf{D}_{t+1}(u, v) - \mathbf{d}_i^t\|_2).$$

This formulation produces bounded similarity responses that decay smoothly with descriptor distance. This similarity response provides a continuous correspondence confidence distribution in feature space, where larger values indicate more likely correspondence locations. Peak responses correspond to maximum-likelihood correspondence estimates. Through distillation, ViMatch learns to generate well-structured similarity distributions that form sharp and spatially localized peaks for true matches.

Stage II: Teacher-Initialized Self-Supervised Adaptation on Unlabeled Videos

After supervised initialization, the model enters an extended self-training stage using unlabeled video streams containing diverse scene variations. In this phase, explicit ground-truth correspondences are no longer required. Instead, the model continuously tracks the initial keypoints $\mathbf{x}_i^1$ across frames and optimizes its parameters using intrinsic temporal coherence.

Two complementary consistency constraints are imposed:

- 1) Intra-frame consistency, which stabilizes similarity distributions within local regions and suppresses spatial ambiguity;
- 2) Inter-frame consistency, which constrains accumulated drift by enforcing agreement between recursive tracking and direct cross-frame propagation (see Sec. III-C).

Together, these constraints transform feature tracking into a closed-loop self-regularized optimization process. The model progressively refines its descriptor representation from raw video streams, improving tracking robustness under diverse visual conditions and mitigating feature degradation caused by appearance changes.

To ensure scalability for high-resolution imagery, correspondence estimation is performed within local observation windows. For each keypoint $\mathbf{x}_i^t$, a fixed-size local image patch $\mathbf{H}_i^t \in \mathbb{R}^{W \times H \times C}$ centered at the keypoint is extracted as network input.

Restricting similarity computation to local patches serves both computational and statistical purposes. Computationally, it reduces redundant evaluations and ensures nearly constant complexity per keypoint, independent of global image resolution. Statistically, it defines a bounded spatial prior that limits correspondence search to physically plausible displacement ranges, thereby reducing the probability of spurious high-similarity responses.

In summary, the proposed framework combines distilled supervised initialization with autonomous self-supervised temporal regularization under a localized measurement model. This design enables stable and scalable feature trajectory estimation across long in-the-wild sequences without requiring explicit annotations during deployment.

B. Network Architecture

To achieve a lightweight design, our framework consists of a fully convolutional network Θ_1 with a shared encoder and a feature tracking module that employs Differentiable Keypoint Detection (DKD) [23], as illustrated in Fig. 2. Specifically, the DKD module takes the similarity difference map generated from dense descriptor comparison as input and predicts the sub-pixel correspondence location by detecting reliable response peaks. The input of DKD is denoted as $\mathbf{S}_i^{t+1}$, while its output is the estimated correspondence location $\hat{\mathbf{x}}_i^{t+1}$. The network Θ_1 comprises three residual convolutional blocks, multiple normalization layers, and pooling layers. It outputs compact feature maps composed of low-dimensional descriptors, which serve as the foundation for similarity computation and subsequent feature tracking.

(a) **Shared Encoder** The image feature encoder transforms an input image patch $\mathbf{H}_i^t \in \mathbb{R}^{H \times W \times 3}$ into a dense descriptor map of size $W \times H \times 32$. Four 3×3 convolutional kernels are successively applied to reduce spatial resolution while increasing channel depth, thereby accelerating feature extraction. Each intermediate feature map is further processed by a 1×1 convolution followed by upsampling to restore the original resolution, ensuring channel consistency for concatenation. ReLU activations [24] are applied after each convolution. This design enables efficient inference while preserving fine-grained information at the original resolution for generating compact image descriptors.

(b) Feature Tracking

Given a keypoint $\mathbf{x}_i^t$ in the reference frame, the corresponding dense descriptor map is represented as $\mathbf{D}_{t+1} \in \mathbb{R}^{H \times W \times 32}$ in the target frame. Bilinear interpolation is applied to obtain the descriptor at the keypoint location, yielding the descriptor $\mathbf{d}_i^t \in \mathbb{R}^{32}$. A similarity difference map is then constructed by comparing the query descriptor with the dense descriptor map in the target frame:

$$\mathbf{S}_i^{t+1}(u, v) = \exp(-\|\mathbf{D}_{t+1}(u, v) - \mathbf{d}_i^t\|_2).$$

The similarity difference map $\mathbf{S}_i^{t+1}$ provides a continuous correspondence confidence distribution, where high responses indicate potential correspondence locations. DKD is subsequently applied to $\mathbf{S}_i^{t+1}$ to detect reliable response peaks and obtain the sub-pixel correspondence location $\hat{\mathbf{x}}_i^{t+1}$. The obtained correspondences are further used for feature tracking and self-supervised optimization (see Sec. III-C).

C. Hybrid Self-Supervised Loss

To extend ViMatch to in-the-wild scenarios, we design a hybrid self-supervised module that captures local spatial consistency within video streams. The first component, single-frame consistency, applies random perturbations to local image patches to enhance the network's invariance to small spatial displacements. The second component, multi-frame consistency, enforces descriptor similarity for the same feature points within a sliding window, thereby improving spatial stability and feature tracking robustness across frames.

1) *Single-frame consistency loss*: When the tracking region is restricted to local patches, the effective receptive field of the network becomes significantly constrained. Under this setting, the feature extraction process relies heavily on the precise spatial alignment of the sampled patch. Even slight deviations in the bounding location around the same candidate keypoint may result in noticeable variations in descriptor responses and tracking confidence. In practice, such deviations are inevitable due to detection noise, image blur, illumination changes, and quantization effects during patch cropping. As a consequence, feature localization should be regarded as a stochastic measurement process rather than a deterministic operation.

Let the true keypoint location be denoted as $\mathbf{x}_i^t$. During patch extraction, the observed sampling center can be expressed as

$$\mathbf{x}_{i,obs}^t = \mathbf{x}_i^t + \Delta\mathbf{x} \quad (1)$$

where $\Delta\mathbf{x}$ represents spatial perturbation induced by localization uncertainty. In the absence of ground-truth correspondences during training—since the teacher model cannot be directly generalized to in-the-wild scenarios—such perturbations cannot be explicitly corrected. Therefore, the learning objective must account for potential deviations in real-world feature localization and ensure descriptor stability under small spatial displacements.

To address this issue, we introduce a single-frame consistency constraint that explicitly minimizes intra-frame prediction variance. Instead of extracting a single patch centered at $\mathbf{x}_i^t$, multiple patches are sampled around the same candidate keypoint within a single image, as illustrated in Fig. 3. By exposing the network to controlled spatial perturbations, the proposed strategy encourages the learned descriptor to exhibit local smoothness and spatial repeatability.

Specifically, let $\mathbf{x}_i^t$ denote the initial tracked point in image $\mathbf{I}_t$. Around $\mathbf{x}_i^t$, we extract N perturbed patches $\{\mathbf{H}_i^{t,(n)}\}_{n=1}^N$ from $\mathbf{I}_t$ by introducing small random spatial offsets sampled from a zero-mean distribution. These offsets simulate intra-frame localization uncertainty and appearance variation. In

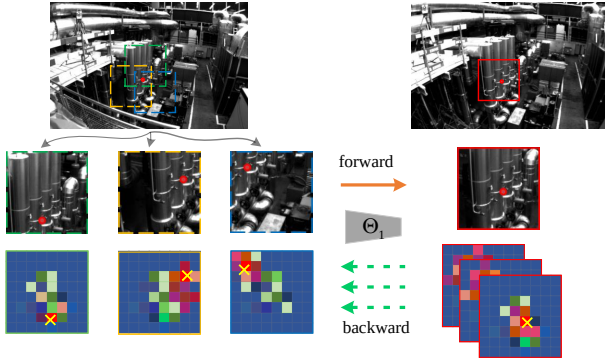


Fig. 3. **Single-frame consistency loss.** The red dots represent the matched features. For each tracked point in the first frame, multiple randomly shifted image patches are generated around the original keypoint location, and the corresponding patch in the second frame is used as the tracking reference. A consistency loss is constructed using the predicted coordinates $\hat{\mathbf{x}}_i^{t+1,(n)}$ from these perturbed patches.

contrast, only one candidate patch $\mathbf{H}_i^{t+1}$ is sampled from the corresponding location in the subsequent frame $\mathbf{I}_{t+1}$, serving as a spatial reference.

For each perturbed pair $(\mathbf{H}_i^{t,(n)}, \mathbf{H}_i^{t+1})$, the predicted correspondence locations are denoted as $(\mathbf{x}_i^{t,(n)}, \hat{\mathbf{x}}_i^{t+1,(n)})$. Ideally, if the descriptor representation is spatially stable, all $\hat{\mathbf{x}}_i^{t+1,(n)}$ should converge to the same location despite perturbations in the input patch.

The single-frame consistency loss therefore enforces spatial alignment among these predicted correspondences, effectively minimizing the intra-frame spatial variance of $\hat{\mathbf{x}}_i^{t+1}$. This constraint improves local measurement repeatability and suppresses amplification of small sampling errors, thereby enhancing stability under appearance perturbations.

Inspired by both sparse-to-sparse and sparse-to-dense correspondence paradigms, tracking is formulated as a descriptor similarity estimation problem. Given a keypoint $\mathbf{x}_i^t$ in image $\mathbf{I}_t$, let $\mathbf{d}_i^t \in \mathbb{R}^C$ denote its C -dimensional descriptor. Let $\mathbf{D}_{t+1} \in \mathbb{R}^{H \times W \times C}$ denote the dense descriptor map extracted from the corresponding patch in image $\mathbf{I}_{t+1}$.

Instead of performing direct nearest-neighbor search, we construct a continuous similarity response map by measuring descriptor distances at every spatial location. The similarity map is defined as

$$\mathbf{S}_i^{t+1}(u, v) = \exp(-\|\mathbf{D}_{t+1}(u, v) - \mathbf{d}_i^t\|_2) \quad (2)$$

where $(u, v) \in \mathbf{H}_i^{t+1}$.

This exponential transformation ensures that similarity responses are bounded within $(0, 1]$ and decay smoothly with increasing descriptor distance. The resulting similarity map can be interpreted as a continuous correspondence confidence distribution, where high responses indicate more likely correspondences. Therefore, peak responses in $\mathbf{S}_i^{t+1}$ correspond to maximum-likelihood correspondence estimates.

By applying DKD to the similarity map, we obtain the predicted correspondence location $\hat{\mathbf{x}}_i^{t+1}$ within $\mathbf{H}_i^{t+1}$. The predicted location is given by the spatial maximum (or soft-

argmax) of the similarity distribution, yielding sub-pixel localization accuracy.

To enforce spatial stability under perturbations, we define the single-frame reprojection consistency loss as

$$\mathcal{L}_{srp} = \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^N \text{dist}_{th}(\hat{\mathbf{x}}_i^{t+1,(i)}, \hat{\mathbf{x}}_i^{t+1,(j)}) \quad (3)$$

where the Euclidean distance metric is defined as

$$\text{dist}(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2) = \|\hat{\mathbf{x}}_1 - \hat{\mathbf{x}}_2\|_2. \quad (4)$$

The operator $(\cdot)_{th}$ discards predictions whose pairwise distances exceed a predefined spatial tolerance, thereby preventing unstable or ambiguous correspondences from influencing optimization.

From a statistical perspective, $\mathcal{L}_{srp}$ minimizes the intra-frame spatial variance of predicted correspondences,

$$\mathcal{L}_{srp} \propto \text{Var}(\hat{\mathbf{x}}_i^{t+1}), \quad (5)$$

thus improving localization repeatability and reducing uncertainty propagation to subsequent frames.

All feature coordinates are rescaled to native image resolution before distance computation, ensuring that the loss is evaluated in metric-consistent pixel space and preserving geometric accuracy during optimization.

2) *Multi-Frame Consistency Loss:* Although restricting correspondence search to local retrieval regions effectively suppresses gross mismatches, it inevitably introduces cumulative pixel drift in long sequences. This phenomenon is illustrated in Fig.4. The underlying cause is not solely local ambiguity, but the progressive accumulation of small descriptor deviations across frames. When feature descriptors are learned from domain-specific datasets, their generalization to cross-domain, in-the-wild video streams is limited. Under recursive tracking, slight variations induced by viewpoint change, illumination fluctuation, motion blur, or occlusion lead to gradual spatial bias propagation. Such bias accumulates temporally and manifests as drift at the pixel level.

From a measurement perspective, this issue can be interpreted as error propagation in sequential state estimation. Let $\mathbf{x}_i^t$ denote the spatial state of the i -th feature point at frame t . Recursive propagation implicitly assumes local descriptor consistency between adjacent frames, but does not explicitly constrain long-term global consistency. Consequently, small per-frame estimation errors compound over time.

To suppress this drift, we introduce a self-supervised multi-frame consistency formulation that enforces agreement between short-term recursive tracking and long-term direct propagation. Given a video sequence $\{\mathbf{I}_t\}_{t=1}^T$, we select a set of feature points $\{\mathbf{x}_i^1\}$ that remain trackable for at least three frames. For each point, two propagation paths are constructed:

1) Recursive tracking:

$$\vec{\mathbf{x}}_i = \mathbf{x}_i^1 \rightarrow \mathbf{x}_i^2 \rightarrow \dots \rightarrow \mathbf{x}_i^T$$

2) Direct cross-frame propagation:

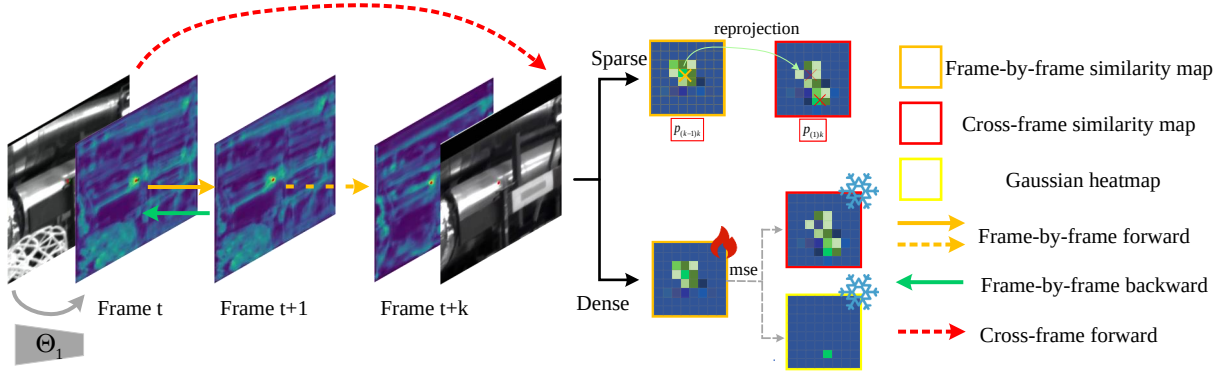


Fig. 4. **Multi-frame consistency loss.** The red dots denote the matched features. For each feature point in the first frame, two propagation paths are constructed: recursive tracking through consecutive frames and direct propagation from the first frame to the target frame. The discrepancy between the predicted coordinates from the two paths provides a self-supervised consistency signal for multi-frame tracking.

$$\vec{\mathbf{x}}_i = \mathbf{x}_i^1 \rightarrow \mathbf{x}_i^T$$

The recursive estimate accumulates local uncertainties, while the direct estimate relies on a single long-baseline similarity evaluation and is therefore less sensitive to intermediate perturbations. Their discrepancy provides a natural self-supervised error signal.

The sparse multi-frame reprojection loss is defined as

$$\mathcal{L}_{mrp} = \frac{1}{M} \sum_{i=1}^M \text{dist}_{th}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_i), \quad (6)$$

where M is the number of valid feature tracks and $\text{dist}_{th}(\cdot)$ discards correspondences whose deviation exceeds a predefined tolerance. Minimizing $\mathcal{L}_{mrp}$ reduces temporal inconsistency and constrains accumulated drift, which can be interpreted as minimizing the long-term state deviation in a sequential measurement model.

While $\mathcal{L}_{mrp}$ provides sparse spatial supervision, it does not regulate the internal structure of similarity distributions. Therefore, a dense consistency constraint is further introduced in descriptor space.

Let $\mathbf{D}_t$ denote the dense descriptor map at frame t . Following Sec. III-C1, the similarity map generated by feature point $\mathbf{x}_i^t$ is denoted as $\mathbf{S}_i^{(t+1)}$.

Two similarity maps are obtained for each feature point:

$$\mathbf{S}_i^{t+1}(\vec{\mathbf{x}}_i) \quad \text{and} \quad \mathbf{S}_i^{t+1}(\vec{\vec{\mathbf{x}}}_i).$$

Since cross-frame propagation avoids intermediate recursive accumulation, its similarity map provides a relatively reliable reference after confidence filtering. However, direct propagation may still become unreliable under large viewpoint changes, severe occlusion, repetitive textures, or violations of the local search range. Therefore, bidirectional verification and threshold-based filtering are applied before using these correspondences for optimization.

The dense similarity consistency loss is defined as

$$\mathcal{L}_{sim} = \left\| \left(\mathbf{S}_i^{t+1}(\vec{\mathbf{x}}_i) - (\mathbf{S}_i^{t+1}(\vec{\vec{\mathbf{x}}}_i))_{\text{stop}} \right)_{th} \right\|^2, \quad (7)$$

where $(\cdot)_{\text{stop}}$ indicates that gradients are not propagated through the reference branch. This formulation enforces descriptor-level distribution alignment rather than merely coordinate agreement.

However, mean squared error minimization alone may encourage globally elevated similarity responses instead of well-localized peaks. To ensure unimodality and spatial concentration, a Gaussian target distribution centered at $\vec{\vec{\mathbf{x}}}_i$ is introduced:

$$\mathbf{G}(u, v) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(u - u_i)^2 + (v - v_i)^2}{2\sigma^2}\right), \quad (8)$$

where (u_i, v_i) denotes the coordinate of $\vec{\vec{\mathbf{x}}}_i$.

The unimodality regularization is defined as

$$\mathcal{L}_{hot} = \left\| \left(\mathbf{S}_i^{t+1}(\vec{\mathbf{x}}_i) - \mathbf{G} \right)_{th} \right\|^2. \quad (9)$$

This term enforces that similarity responses approximate a Gaussian-shaped target distribution, thereby reducing ambiguity and improving sub-pixel localization stability.

The overall dense supervision term becomes

$$\mathcal{L}_{dense} = \mathcal{L}_{sim} + \mathcal{L}_{hot}. \quad (10)$$

Finally, combining single-frame stability, multi-frame spatial consistency, and descriptor-level regularization, the complete unsupervised objective is formulated as

$$\mathcal{L}_{unsup} = \epsilon \mathcal{L}_{srp} + \zeta \mathcal{L}_{mrp} + \eta \mathcal{L}_{dense}, \quad (11)$$

where ϵ , ζ , and η balance spatial variance suppression, long-term drift reduction, and similarity distribution regularization, respectively. In our implementation, they are empirically set to $\epsilon = 1$, $\zeta = 1$, and $\eta = 2$ based on the validation performance on SpatialVID, and remain fixed during all cross-dataset evaluations.

To avoid degenerate solutions in which the network concentrates responses near patch centers (a common shortcut under static scenes), random spatial perturbations are applied to initial keypoint sampling during training. Moreover, bidirectional verification is employed: backward propagation from frame T

to frame 1 must satisfy identical threshold constraints before contributing to the loss. This symmetric constraint further reduces bias and enhances estimation reliability.

From a statistical viewpoint, the proposed multi-frame framework jointly minimizes spatial variance, temporal drift, and descriptor distribution entropy, forming a self-consistent measurement regularization mechanism suitable for in-the-wild video training without external annotations.

IV. EXPERIMENTS AND RESULTS

A. Evaluation Setup

We employ ALIKED as the teacher model. To ensure a fair comparison with existing approaches, teacher-guided initialization is performed on the widely used MegaDepth dataset [17], which provides supervised correspondence information for feature representation learning. After initialization, ViMatch is self-supervisedly adapted exclusively using unlabeled video sequences from SpatialVID [31]. The complete training and evaluation protocol is summarized in Table I.

Following this protocol, all evaluation datasets are excluded from the adaptation stage. We first evaluate the visual navigation performance by integrating the SpatialVID-adapted

model into OpenVINS and testing on EuRoC and UMA without additional fine-tuning. We then evaluate computational efficiency and conduct ablation studies to analyze the influence of teacher supervision, keypoint detectors, and individual consistency components. Finally, cross-dataset feature tracking is evaluated on KITTI and TartanAir, while HPatches is used as an independent local correspondence benchmark.

These datasets cover diverse visual conditions, including synthetic environments, autonomous driving scenes, indoor/outdoor MAV sequences, and local correspondence evaluation scenarios.

For all experiments, the training window size is fixed at 30 frames. The network is implemented in PyTorch and trained using the Adam optimizer with an initial learning rate of 2×10^{-4} and a batch size of 2. Training is performed on an NVIDIA GeForce RTX 4070 GPU for 50 epochs. Unless otherwise specified, ALIKED is used as the default teacher model, and all training configurations are fixed during evaluation.

B. Visual Odometry Evaluation

We integrate the proposed video-driven self-supervised training framework into the state-of-the-art visual-inertial odometry (VIO) system OpenVINS, replacing the conventional optical-flow module as a feature tracking component for evaluating the downstream visual navigation performance. For evaluation, ALIKED is employed as the teacher model in the ViMatch initialization training, while LET-NET, XFeat, and SuperPoint+LightGlue are included as additional baselines.

We conducted comparative experiments on the widely used EuRoC dataset and the challenging UMA dataset (Table IV). ViMatch undergoes self-supervised adaptation exclusively on SpatialVID and is directly evaluated on EuRoC and UMA

TABLE I
TRAINING AND EVALUATION PROTOCOL OF ViMATCH.

Stage	Dataset	Purpose
Initialization	MegaDepth	Teacher-guided initialization
Adaptation	SpatialVID	Self-supervised adaptation
VIO Evaluation	EuRoC, UMA	Unseen-dataset VIO evaluation
Tracking Evaluation	KITTI	Long-term feature tracking evaluation
Cross-domain Evaluation	TartanAir	Synthetic-to-real transfer evaluation
Tracking Benchmark	HPatches	Local matching benchmark

TABLE II
ABSOLUTE TRAJECTORY ERROR (ATE) AND ABSOLUTE ROTATION ERROR (ARE). TRANSLATION ERRORS ARE REPORTED IN METERS, AND ROTATION ERRORS ARE REPORTED IN DEGREES. RED DENOTES THE BEST, WHILE BLUE DENOTES THE SECOND BEST.

Method	MH_03		MH_05		V1_02		V1_03		V2_02		V2_03		Mean	
	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)
OpenVINS [25]	0.1509	1.8826	0.4936	2.2101	0.0583	2.3035	0.0601	3.0555	0.0578	2.0987	0.2001	2.7169	0.1701	2.3779
Sup. [26]+Lig. [27]	0.1647	2.0279	0.4339	1.2753	0.0709	2.8605	0.0920	2.5322	0.0737	2.3279	0.2077	2.2267	0.1738	2.2084
LET-NET [28]	0.1958	2.2749	0.3481	1.2943	0.0875	2.4295	0.0847	3.4756	0.0657	2.2186	0.5872	2.2181	0.2282	2.3290
ALIKED [29]	0.1534	2.4034	0.2705	1.3029	0.0553	2.5894	0.0749	3.0517	0.0670	2.0453	0.1266	2.2436	0.1246	2.2727
XFeat [30]	fail	fail	0.6939	3.6493	0.0767	2.7285	0.1427	2.9422	0.2137	4.7513	fail	fail	0.2818	3.5178
ViMatch	0.1418	1.9025	0.3072	1.1584	0.0551	2.4216	0.0597	1.9460	0.0482	2.3658	0.1304	1.8012	0.1237	1.9326

Method	class_csc1		class_eng		conference_csc1		fantasy_csc1		gattaca_csc1		lab_module_csc		Mean	
	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)	t_ape (m)	r_ape (°)
OpenVINS [25]	0.0782	17.0680	0.1592	46.8445	0.2920	43.1039	1.9482	142.5431	0.1937	17.7539	0.0134	17.8477	0.4475	47.5269
Sup. [26]+Lig. [27]	0.0468	17.2325	0.0520	7.3361	0.1898	40.4545	0.3700	41.2528	0.0767	5.6705	0.1470	17.1108	0.1471	21.5095
LET-NET [28]	0.2134	59.5626	fail	fail	0.4265	73.8894	fail	fail	0.2614	21.9927	0.5293	36.5864	0.3576	48.0078
ALIKED [29]	0.0281	8.0499	0.0628	24.5949	0.2749	47.5270	1.3561	55.7266	0.1025	3.6646	0.0112	1.7753	0.3059	23.5564
XFeat [30]	0.2105	76.9492	fail	fail	fail	fail	fail	fail	14.9470	161.7195	fail	fail	7.5788	119.3344
ViMatch	0.0371	12.6247	0.0954	4.0126	0.1778	25.1042	0.2317	29.1365	0.1921	5.0376	0.0623	6.7803	0.1327	13.7827

* Sup.+Lig. is an abbreviation for Superpoint + Lightglue.

* t_ape(m) represents the absolute translation error, r_ape(°) represents the absolute rotation error, and Mean represents the average value across all sequences.

TABLE III

THE NUMBER OF NETWORK PARAMETERS, GFLOPS ON IMAGES OF VARYING RESOLUTIONS, INFERENCE TIME, AND END-TO-END FPS, WITH THE TOP THREE PERFORMING RESULTS HIGHLIGHTED IN RED, BLUE, AND GREEN, RESPECTIVELY.

Models	Params/M	Low(480p)			Medium (720p)			High Resolution(1080p)			High Resolution(2k)		
		GFLOPs	CPU/ms	FPS	GFLOPs	CPU/ms	FPS	GFLOPs	CPU/ms	FPS	GFLOPs	CPU/ms	FPS
ALiKE(N) [23]	0.346	13.02	210.22	4.64	39.94	669.66	1.47	88.57	1502.13	0.65	156.29	2928.25	0.34
Sup. [26]+Lig. [27]	3.237	5.90	567.14	1.51	5.45	570.76	1.52	6.06	562.05	1.53	6.05	578.66	1.54
LET-NET [28]	0.084	2.02	80.62	10.36	5.92	296.41	3.15	13.32	581.86	1.46	25.22	1102.77	0.75
ALiKED(N16) [29]	0.192	3.48	254.55	3.85	9.07	601.12	1.58	23.51	1689.19	0.59	40.58	3693.15	0.27
XFeat [30]	0.658	1.33	58.12	17.01	3.99	212.58	4.70	8.96	507.88	1.97	15.91	932.22	1.07
ViMatch(single patch)	0.036	0.51	34.46	26.00	2.02	152.09	6.25	8.14	586.20	1.65	32.39	2857.42	0.34
ViMatch(pyramid patch)	0.036	0.51	34.46	26.00	0.51	35.12*2	13.53	0.51	35.14*2	12.97	0.51	35.10*2	13.48

without any target-domain fine-tuning or adaptation. No sequence from either evaluation dataset is involved in self-supervised training, ensuring a strict separation between the adaptation and evaluation datasets.

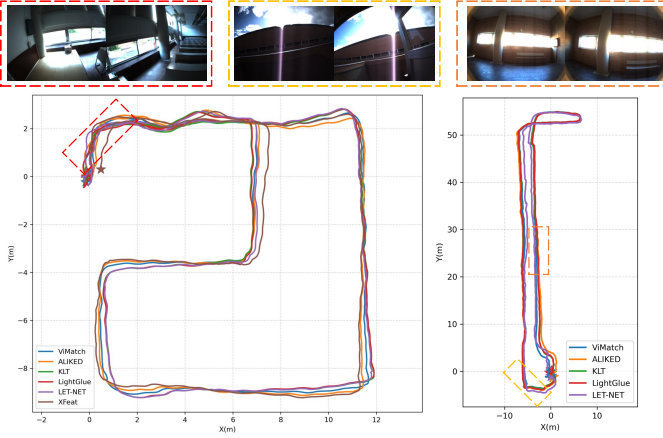


Fig. 5. Comparison of estimated trajectories on the UMA-VI [32] dataset for the *class_csc1* (left) and *lab_module_csc* (right) sequences. The proposed method demonstrates robustness to highly dynamic illumination changes and weakly textured regions, leading to improved trajectory accuracy.

On EuRoC, ViMatch achieves the best overall mean accuracy (0.1237 m / 1.9326°), reducing translation and rotation errors by 27.3% and 18.7%, respectively, compared with the original OpenVINS front-end. Compared with the ALiKED baseline, ViMatch further reduces translation and rotation errors by 0.7% and 15.0%, respectively.

On the more challenging UMA dataset, the performance gap becomes more pronounced. ViMatch attains the lowest mean error (0.1327 m / 13.7827°), outperforming SuperPoint+LightGlue by 9.8% in translation and 35.9% in rotation, and reducing translation and rotation errors by 56.6% and 41.5%, respectively, compared with ALiKED. Unlike LET-NET and XFeat, which exhibit multiple failure cases, ViMatch completes all sequences successfully.

Although SuperPoint+LightGlue achieves lower ATE in *class_eng* and *gattaca_csc1*, its performance degrades in *lab_module_csc*. In contrast, the proposed method maintains more consistent accuracy across environments. Despite being adapted exclusively on SpatialVID, ViMatch maintains competitive performance across different evaluation environments without target-domain adaptation. These results demonstrate that the learned representation can effectively transfer across

datasets with different visual characteristics and acquisition conditions. Fig. 5 illustrates representative outdoor and indoor trajectories from *class_csc1* and *lab_module_csc*, respectively.

C. Efficiency Comparisons with the State-of-the-Arts

To analyze the computational scalability of ViMatch, we evaluate its efficiency across multiple input resolutions under a controlled CPU-only setting. All experiments are conducted in a PyTorch environment on an Intel Core i7-13700KF processor, with hardware acceleration (GPU/TPU) and deployment-level optimizations (e.g., NCNN, ONNX) disabled to ensure reproducibility and platform consistency.

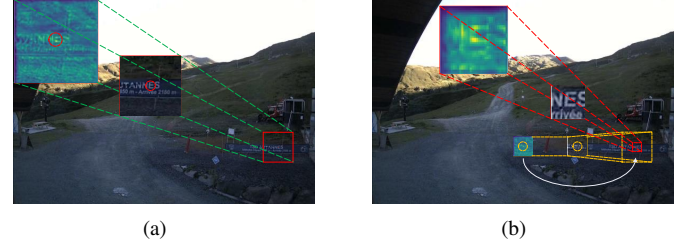


Fig. 6. Different network reasoning architectures. The left side(a) is single-patch direct reasoning, and the right side(b) is pyramid-patches coarse-to-fine reasoning.

The efficiency advantage of ViMatch originates from both the lightweight network architecture and the localized correspondence estimation strategy. Unlike global matching approaches that search arbitrary correspondences over the entire image, ViMatch estimates correspondences within bounded local observation regions around detected keypoints. Therefore, the efficiency comparison should be interpreted under the local feature tracking setting rather than unconstrained global image matching.

For high-resolution inputs, we employ a pyramid-based inference strategy (Fig. 6) to regulate computational growth. Specifically, coarse correspondences are first estimated on downsampled image blocks, followed by local refinement using original-resolution patches. This hierarchical design reduces resolution-dependent complexity while preserving sub-pixel localization accuracy.

We evaluate both single-level and two-level pyramid configurations. In the single-level setting, block size scales proportionally with image resolution (32×32 for 480p, 64×64 for 720p, 128×128 for 1080p, and 256×256 for 2K), resulting

TABLE IV
ROBUSTNESS OF ViMATCH TO DIFFERENT TEACHER SUPERVISION QUALITY.

Teacher	Teacher Corr. Acc. (%) $\uparrow$		Initial ViMatch (%) $\uparrow$		+ SpatialVID Adaptation (%) $\uparrow$		Absolute Gain (%) $\uparrow$	
	MMA@1	MMA@3	MMA@1	MMA@3	MMA@1	MMA@3	MMA@1	MMA@3
XFeat	44.94	73.35	43.02	72.56	53.27	74.38	+10.25	+1.82
Sup.+Lig.	47.78	70.38	49.11	69.31	58.72	76.14	+9.61	+6.83
ALIKED	49.05	73.35	51.96	72.48	55.99	77.79	+4.03	+5.31

in quadratic growth of computation. In contrast, the two-level configuration constrains the first-stage processing to fixed 32×32 blocks and performs refinement only in localized regions, effectively limiting computational expansion.

Table III summarizes parameter count, GFLOPs, inference latency, and end-to-end FPS, including preprocessing (keypoint detection), network inference, and post-processing (similarity computation). Although ViMatch primarily operates as the tracking module, the reported parameters correspond only to the ViMatch network itself. The external detector parameters are not included because ViMatch is detector-agnostic and can be combined with different lightweight detectors. The results show that the pyramid configuration substantially improves efficiency at high resolutions, maintaining approximately 10 FPS at 2K resolution while reducing computational cost by over an order of magnitude compared to direct large-block processing.

The reported speed improvements should therefore be understood within the supported local tracking range. When the true correspondence exceeds the predefined local search region due to large displacement or severe appearance changes, additional strategies such as pyramid inference or keypoint re-detection are required.

Finally, the current evaluation focuses on algorithmic efficiency under a controlled CPU environment rather than complete edge deployment validation. Although the lightweight design indicates potential for resource-limited platforms, dedicated hardware evaluation involving memory usage, energy consumption, and hardware-optimized inference latency remains an important direction for future work.

D. Ablation and Robustness Analysis

1) *Robustness to Teacher Models*: To investigate the influence of initial teacher supervision and the effectiveness of SpatialVID adaptation, we evaluate ViMatch with different teacher models, as summarized in Table V. For each teacher configuration, keypoints are detected using the corresponding detector, and all reported MMA values are averaged over all sequences of HPatches. SpatialVID is used exclusively for self-supervised adaptation, while HPatches is used only for evaluation. The results show that SpatialVID adaptation consistently improves ViMatch after different teacher-based initializations, indicating that the proposed video adaptation strategy provides additional benefits beyond the initial supervised model. Moreover, the improvements obtained with different teacher models demonstrate that ViMatch does not critically depend on a specific teacher configuration and remains robust to variations in initial supervision quality.

TABLE V
MEAN MATCHING ACCURACY (MMA %) ON HPATCHES.

Features + Tracker	Illumination			Viewpoint		
	@1	@3	@5	@1	@3	@5
Fast+LK	37.13	48.13	51.53	5.20	10.17	11.26
Fast+LET-NET	53.93	72.13	81.47	41.44	57.68	62.79
ALIKE (N)	43.17	73.91	77.84	50.34	76.18	80.97
Sup.+Lig	49.46	71.19	75.49	46.10	69.57	73.66
LET-NET	56.98	77.06	85.15	44.53	58.94	63.88
ALIKED (N16)	45.62	70.73	74.96	52.47	72.29	78.11
XFeat	41.25	76.22	82.83	48.62	70.47	71.63
Fast+ViMatch	55.39	75.26	84.77	51.06	73.73	77.18
ALIKED+ViMatch	58.79	79.33	84.95	53.19	76.24	78.84

2) *Robustness to Keypoint Detectors*: We further investigate the dependence of ViMatch on the keypoint detector by combining the same trained ViMatch model with FAST and ALIKED during inference. All configurations are evaluated on HPatches without target-specific adaptation. As shown in Table VI, both Fast+ViMatch and ALIKED+ViMatch achieve competitive accuracy under illumination and viewpoint variations, while ALIKED+ViMatch performs better at stricter thresholds. This demonstrates that ViMatch is compatible with different keypoint detectors rather than being tied to the detector used during initialization.

3) *Ablation of Consistency Components*: To distinguish the contribution of the proposed consistency objective from generic video adaptation, we conduct a detailed ablation study on individual consistency components using the KITTI dataset. Specifically, we progressively introduce the spatial reprojection consistency loss L_{srp} , multi-frame reprojection consistency loss L_{mrp} , similarity consistency loss L_{sim} , and unimodal regularization loss L_{hot} to analyze their individual contributions to long-term feature tracking.

In the first experiment, corresponding image patches were extracted from two consecutive frames, and artificial spatial displacements of varying magnitudes were introduced in the second frame to simulate realistic perturbations encountered during tracking. Without the spatial reprojection consistency loss L_{srp} , the similarity maps corresponding to the same candidate keypoint exhibit displacement-dependent variations. The response peaks shifted irregularly within local patches, and the similarity distributions became spatially dispersed, indicating that descriptor responses were highly sensitive to small spatial perturbations. As a result, the maximum-likelihood localization was unstable, and the pixel error ranged from 1 to 2 pixels.

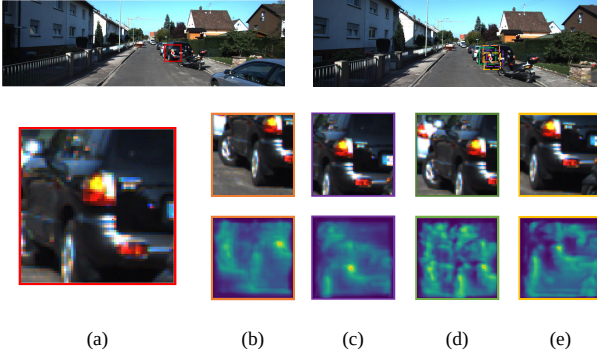


Fig. 7. (a) The image patches to be matched; (b) and (c) similarity maps guided only by the teacher model; (d) and (e) similarity maps with single-frame consistency added.

After introducing L_{srp} , the similarity maps converge to spatially concentrated response distributions with well-aligned peaks across different displacement magnitudes as illustrated in Fig. 7. The learned descriptors become less sensitive to local perturbations, demonstrating that L_{srp} effectively suppresses spatial variance and improves sub-pixel localization precision.

To further evaluate the contribution of temporal consistency, we selected images containing highly repetitive local patterns and compared the model optimized with only spatial consistency (L_{srp}) and the model additionally incorporating multi-frame consistency (L_{mrp}) (Fig. 8). Under only spatial consistency, the optimization objective enforces local agreement within each frame but lacks temporal constraints, resulting in competing high-response regions and ambiguous localization in repetitive scenes.

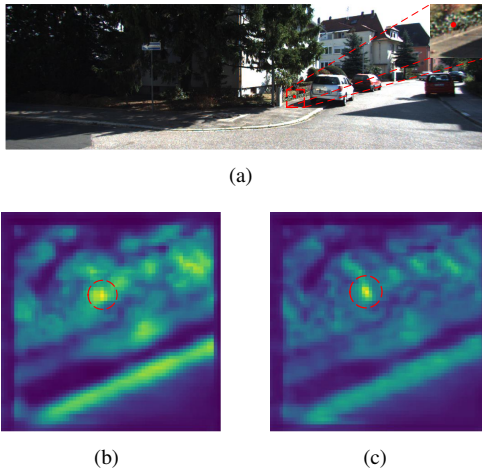


Fig. 8. (a) original image patches (b) similarity map with L_{srp} only (c) similarity map after adding temporal consistency.

By incorporating L_{mrp} , the correspondence estimation is further constrained by temporal consistency across frames. This additional constraint reduces ambiguity caused by repetitive textures and improves the stability of long-term feature trajectories.

Finally, we analyze the effect of dense similarity regularization, including L_{sim} and L_{hot} , using a 15-frame sequence (Fig. 9). Without dense consistency regularization, recursive tracking gradually causes descriptor divergence in latent space due to the accumulation of small alignment errors. Although

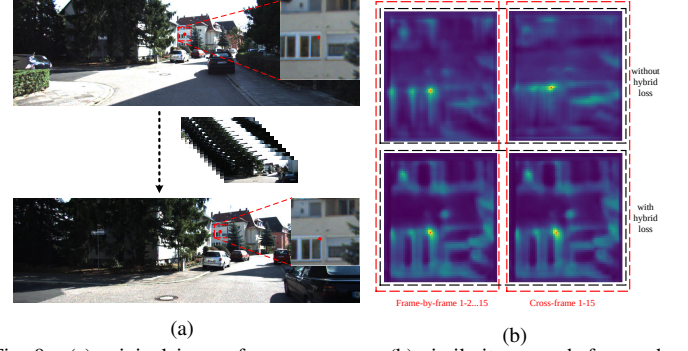


Fig. 9. (a) original image frame sequence; (b) similarity maps before and after adding dense consistency regularization.

short-term correspondences may appear visually reasonable, feature embeddings progressively drift over time, resulting in degraded long-term alignment.

After introducing L_{sim} and L_{hot} , similarity distributions are jointly regularized in the descriptor space. The response maps become more compact and concentrated around the correct locations, producing more reliable correspondence estimates under long-term tracking.

TABLE VI
ABLATION STUDY OF INDIVIDUAL CONSISTENCY COMPONENTS. THE TRACKING RATE IS EVALUATED UNDER DIFFERENT TEMPORAL HORIZONS ON THE KITTI DATASET.

Method	8	16	24	32	40
Initial ViMatch	0.54	0.43	0.23	0.10	0.00
+ L_{srp}	0.65	0.52	0.37	0.18	0.06
+ $L_{srp} + L_{mrp}$	0.78	0.65	0.50	0.31	0.18
+ $L_{srp} + L_{mrp} + L_{sim}$ (w/o L_{hot})	0.76	0.66	0.42	0.25	0.16
+ $L_{srp} + L_{mrp} + L_{hot}$ (w/o L_{sim})	0.81	0.70	0.54	0.38	0.27
+ $L_{srp} + L_{mrp} + L_{sim} + L_{hot}$	0.89	0.78	0.65	0.51	0.44

The quantitative results in Table VII demonstrate the complementary effects of different consistency components. Starting from the initialized model, introducing L_{srp} improves short-term tracking stability by reducing spatial variance caused by local perturbations. Adding L_{mrp} further enhances long-term tracking persistence by constraining temporal drift across frames. The dense consistency components require joint optimization to avoid degenerate similarity distributions. When only L_{sim} is introduced, although the reference branch is detached by stop-gradient, the network may gradually increase similarity responses over the entire local region to minimize the distribution discrepancy, resulting in degraded localization reliability. In contrast, when only the unimodal regularization term L_{hot} is applied without L_{sim} , the network can produce inconsistent response magnitudes at the peak locations, which does not satisfy the desired descriptor invariance. By combining L_{sim} and L_{hot} , the similarity distribution is simultaneously constrained in both consistency and concentration, leading to more stable and discriminative correspondence responses. The complete consistency objective achieves the best tracking performance across all temporal horizons, demonstrating that each component plays a complementary role in improving long-term feature tracking.

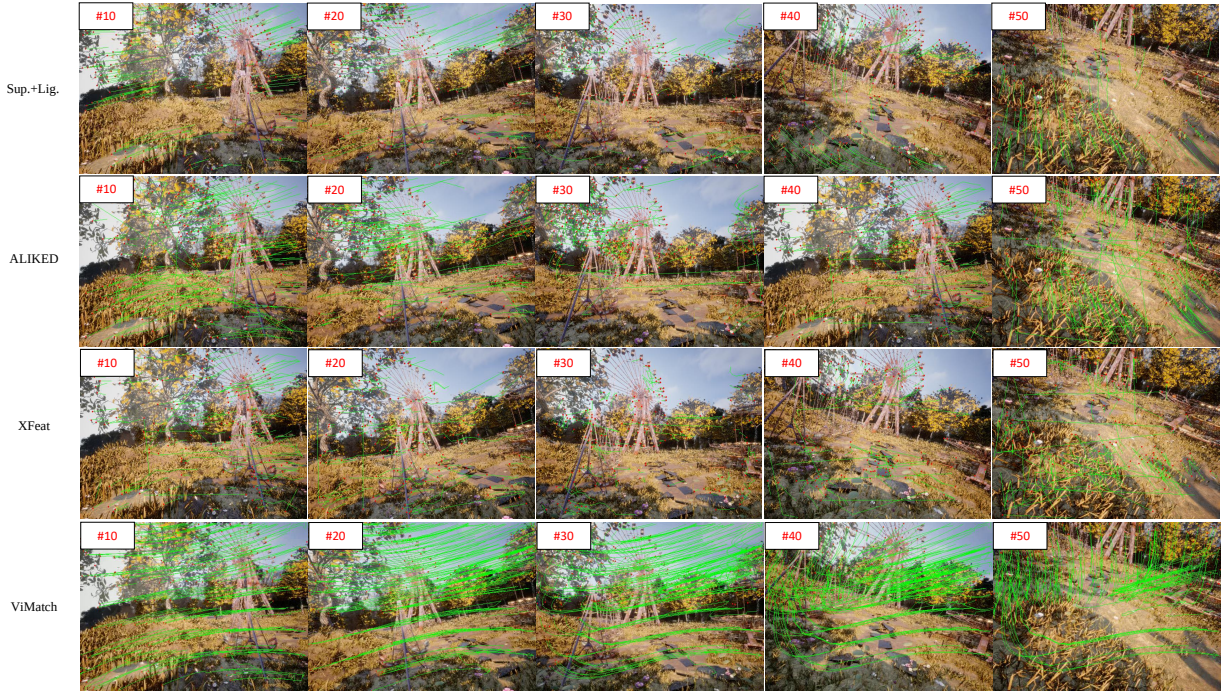


Fig. 10. Qualitative comparison of long-term feature tracking on the TartanAir dataset. ViMatch maintains more continuous feature trajectories under the synthetic-to-real domain gap between SpatialVID and TartanAir. Keypoints in the current frame are drawn in red, and historical trajectories are drawn in green.

TABLE VII
PROPORTION OF TRACKING LENGTHS FOR DIFFERENT MODELS ON THE KITTI ODOMETRY DATASET.

Frame Numbers	10	20	30	40	50
Sup.+Lig.	0.89	0.74	0.53	0.44	0.19
ALIKED	0.82	0.66	0.45	0.31	0.17
XFEAT	0.75	0.62	0.36	0.22	0.09
ViMatch	0.94	0.87	0.73	0.57	0.38

E. Cross-Dataset Feature Tracking Evaluation

The long-term keypoint tracking performance is evaluated on the KITTI Odometry dataset 00-03 and summarized in Table VIII. In each sequence, 300 keypoints are detected in the initial frame and tracked without introducing new points. The reference pixel trajectories are generated using the KITTI camera poses and calibration information. A keypoint is considered successfully tracked if the pixel reprojection error with respect to the reference trajectory remains within 3 pixels. Points that leave the image boundary, become occluded, or cannot be projected to valid image locations are regarded as failed tracks. The reported tracking ratios are averaged over all evaluated sequences.

Compared with competing methods, ViMatch achieves higher tracking persistence across all temporal spans. At 50 frames, ViMatch maintains a tracking ratio of 0.38, which is approximately twice that of Sup.+Lig. and more than three times that of XFeat. These results demonstrate that the proposed hybrid self-supervised loss improves temporal correspondence stability and preserves longer feature trajectories during long-term tracking. Although the tracking-rate metric evaluates temporal correspondence persistence rather than ge-

ometric validity directly, persistent feature trajectories provide more continuous candidate observations for downstream visual estimation. In the OpenVINS pipeline, geometric verification and bundle adjustment are further applied to reject unreliable observations before pose optimization, which is consistent with the improvements observed in the VIO evaluation.

To further examine the transferability of the learned representation, we provide qualitative tracking results on the TartanAir dataset, which is excluded from both initialization and self-supervised adaptation. As shown in Fig. 10, ViMatch produces more continuous feature trajectories under the synthetic-to-real domain gap between SpatialVID and TartanAir, providing additional evidence of its ability to transfer to unseen visual environments.

V. CONCLUSION

In this paper, we introduce ViMatch, a lightweight video-driven local feature tracking framework for visual navigation. The core idea of this work is to exploit local spatial and temporal consistency from unlabeled video sequences to improve long-term feature tracking stability. To this end, we propose a teacher-initialized self-supervised adaptation framework that jointly enforces invariance to small spatial perturbations and temporal consistency across consecutive video frames. The proposed approach achieves high computational efficiency under different image resolutions and demonstrates practical effectiveness when integrated into OpenVINS for visual-inertial odometry evaluation. Furthermore, detailed ablation studies validate the complementary contributions of individual consistency components. Since ViMatch relies on bounded local observation regions, its performance may degrade under extremely large displacements, severe occlusions, or cases

requiring long-term feature re-identification. Future work will investigate more comprehensive geometric constraints and recovery mechanisms to further improve robustness in challenging scenarios.

REFERENCES

- [1] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4104–4113, 2016.
- [2] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(3):611–625, March 2018.
- [3] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *ICRA 2014*, pages 15–22, 2014.
- [4] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.*, 34(4):1004–1020, 2018.
- [5] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV 2011*, pages 2564–2571, 2011.
- [6] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *CVPR*, pages 17765–17775, June 2023.
- [7] Jiaming Sun, Zehong Shen, Yang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021.
- [8] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(8):10247–10266, 2023.
- [9] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 13610–13626. Curran Associates, Inc., 2022.
- [10] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. *ArXiv*, abs/2402.11095, 2024.
- [11] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor’s margins: Local descriptor learning loss, 2018.
- [12] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *CVPR 2019*, pages 8084–8093, 2019.
- [13] Jerome Revaud, Philippe Weinzaepfel, César De Souza, and Martin Humenberger. *R2D2: repeatable and reliable detector and descriptor*. Curran Associates Inc., Red Hook, NY, USA, 2019.
- [14] Michał J. Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient, 2020.
- [15] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, June 2020.
- [16] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017.
- [17] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR 2018*, pages 2041–2050, 2018.
- [18] Yafei Song, Ling Cai, Jia Li, Yonghong Tian, and Mingyang Li. SEKD: Self-evolving keypoint detection and description. *arXiv preprint arXiv:2006.05077*, 2020.
- [19] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R. Nascimento. Enhancing deformable local features by jointly learning to detect and describe keypoints. In *CVPR 2023*, pages 1306–1315, 2023.
- [20] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *CVPR 2015*, pages 3279–3286, 2015.
- [21] Daniel Ponsa Vassileios Balntas, Edgar Riba and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *BMVC*, pages 119.1–119.11. BMVA Press, September 2016.
- [22] Xi Li, Tong Rao, and Cihui Pan. Edm: Efficient deep feature matching, 2025.
- [23] Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter C. Y. Chen, and Zhengguo Li. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Trans. Multimed.*, 25:3101–3112, 2023.
- [24] Lie Xu, Chiu-Sing Choy, and Yi-Wen Li. Deep sparse rectifier neural networks for speech denoising. In *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5, 2016.
- [25] Patrick Geneva, Kevin Eickenhoff, Woosik Lee, Yulin Yang, and Guoquan Huang. OpenVINS: A research platform for visual-inertial estimation. In *Proc. IEEE ICRA*, Paris, France, 2020.
- [26] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. In *CVPRW 2018*, pages 337–33712, Los Alamitos, CA, USA, June 2018. IEEE Computer Society.
- [27] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV 2023*, pages 17581–17592, 2023.
- [28] Yicheng Lin, Shuo Wang, Yunlong Jiang, and Bin Han. Breaking of brightness consistency in optical flow with a lightweight cnn network. *IEEE Robot. Autom. Lett.*, 9(8):6840–6847, 2024.
- [29] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter C. Y. Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023.
- [30] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R. Nascimento. Xfeat: Accelerated features for lightweight image matching. In *CVPR 2024*, pages 2682–2691, 2024.
- [31] Jiahao Wang, Yufeng Yuan, Ruijie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. Spatialvid: A large-scale video dataset with spatial annotations, 2025.
- [32] David Zuñiga-Noël, Alberto Jaenal, Ruben Gomez-Ojeda, and Javier Gonzalez-Jimenez. The uma-vi dataset: Visual-inertial odometry in low-textured and dynamic illumination environments. *The International Journal of Robotics Research*, 39(9):1052–1060, 2020.



Yuqing Wang received the B.Eng. degree in engineering and the M.Eng. degree in engineering from Wuhan University of Technology, Wuhan, China, in 2016 and 2020, respectively. He is currently pursuing the Ph.D. degree with the GNSS Research Center, Wuhan University, Wuhan, China. His primary research interests include visual-inertial SLAM and multi-sensor fusion navigation systems.



Yan Wang received the M.E. degree in computer applied technology from China University of Mining and Technology, Xuzhou, China, in 2019, and the Ph.D. degree from the Global Navigation Satellite System (GNSS) Research Center, Wuhan University, Wuhan, China, in 2023. He is currently a Post-Doctoral Fellow with the GNSS Research Center, Wuhan University. His research interests focus on indoor navigation, sensor fusion algorithms, and computer vision.



Hailiang Tang received the M.E. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2020 and 2023, respectively. He is currently a Post-Doctoral Fellow with the Global Navigation Satellite System (GNSS) Research Center, Wuhan University. His research interests include autonomous robotics systems, visual and light detection and ranging (LiDAR) simultaneous localization and mapping (SLAM), GNSS/Inertial Navigation System (INS) integration, and deep learning.



Tisheng Zhang (Member, IEEE) received the B.Sc. and Ph.D. degrees in communication and information system from Wuhan University, Wuhan, China, in 2008 and 2013, respectively. From 2018 to 2019, he was a Post-Doctoral Researcher at The Hong Kong Polytechnic University. He is a Professor at the GNSS Research Center, Wuhan University. His research interests focus on the fields of GNSS receiver, multi-sensor deep integration, and robot navigation.



Yilong Yuan earned his Master's degree from the Institute of Network Technology, Beijing University of Posts and Telecommunications. Currently, he works at Tencent as Expert Researcher and Lead of Positioning Algorithm R&D at the Positioning IT Center. His work focuses on the research and development of multi-sensor fusion positioning, high-precision satellite positioning and high-precision indoor positioning technologies..



Pan Jiang received his Ph.D. degree from Harbin Institute of Technology, China in 2021. His main research direction is high-precision inertial navigation algorithms. Now he works at Tencent (Beijing) Co., Ltd., his main research direction is multi-source fusion algorithm..



Xiaoji Niu (Member, IEEE) is currently a Professor with the GNSS Research Center, Wuhan University, China. He has published more than 90 academic papers and owns 28 patents. He leads a multi-sensor navigation group focusing on GNSS/INS integration, low-cost navigation sensor fusion, and its new applications.