Leveraging Consistent Spatio-Temporal Correspondence for Robust Visual Odometry

Zhaoxing Zhang[1]*, Junda Cheng[1]*, Gangwei Xu[1], Xiaoxiang Wang[1], Can Zhang[1], Xin Yang[1,2]†

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology
[2]Hubei Key Laboratory of Smart Internet, Huazhong University of Science and Technology
{zzx, jundacheng, gwxu, xiaoxiang0012, can_zhang, xinyang2014}@hust.edu.cn

Abstract

Recent approaches to VO have significantly improved performance by using deep networks to predict optical flow between video frames. However, existing methods still suffer from noisy and inconsistent flow matching, making it difficult to handle challenging scenarios and long-sequence estimation. To overcome these challenges, we introduce **S**patio-**T**emporal **V**isual **O**dometry (**STVO**), a novel deep network architecture that effectively leverages inherent spatio-temporal cues to enhance the accuracy and consistency of multi-frame flow matching. With more accurate and consistent flow matching, STVO can achieve better pose estimation through the bundle adjustment (BA). Specifically, STVO introduces two innovative components: 1) the Temporal Propagation Module that utilizes multi-frame information to extract and propagate temporal cues across adjacent frames, maintaining temporal consistency; 2) the Spatial Activation Module that utilizes geometric priors from the depth maps to enhance spatial consistency while filtering out excessive noise and incorrect matches. Our STVO achieves state-of-the-art performance on TUM-RGBD, EuRoc MAV, ETH3D and KITTI Odometry benchmarks. Notably, it improves accuracy by 77.8% on ETH3D benchmark and 38.9% on KITTI Odometry benchmark over the previous best methods.

Introduction

Visual Odometry (VO) is a pivotal technology for estimating a robot's position and orientation by analyzing data from visual sensors. Classical VO methods utilize optimization techniques to either minimize the photometric error of pixel intensities (Forster, Pizzoli, and Scaramuzza 2014; Engel, Koltun, and Cremers 2017) or minimize the reprojection error of correspondences (Mur-Artal, Montiel, and Tardos 2015; Mur-Artal and Tardós 2017; Campos et al. 2021; Yuan et al. 2022, 2024, 2023) to estimate trajectories. However, due to limited matching capabilities, classical methods frequently encounter issues with robustness, particularly when the image lacks distinct feature points or the assumption of photometric consistency is not met. In the last decade, there has been a shift towards employing end-to-end deep learning methods (Wang et al. 2017; Teed and Deng 2018; Tang and Tan 2018; Wang, Hu, and Scherer 2021), which map visual inputs directly to poses. End-to-end deep learning methods can achieve more stable matching, but neural networks often struggle to predict 6-degree-of-freedom poses from high-dimensional features directly. As a result, their accuracy often falls short compared to classical optimization backends which have strong geometric constraints.

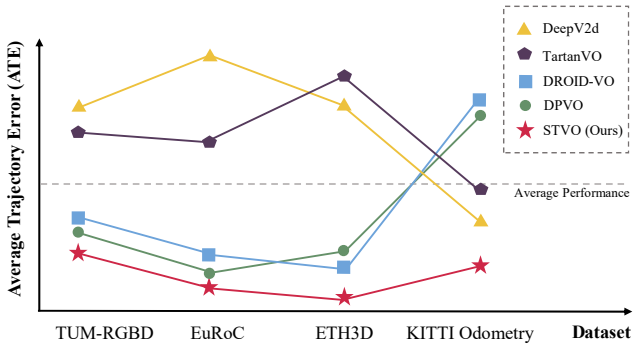


Figure 1: Comparison of STVO with other influential visual odometry methods. Our STVO, denoted by red stars, achieves state-of-the-art performance in all benchmarks.

To address these limitations, a hybrid VO framework (Teed and Deng 2021; Teed, Lipson, and Deng 2024; Chen et al. 2024) has emerged, integrating the strengths of classical methods with deep learning techniques. These frameworks initially employ neural networks to predict flow correspondences and subsequently use correspondences for pose estimation through geometric optimization. By combining the powerful matching capabilities of deep learning with classical geometric constraints, hybrid VO frameworks significantly reduce overfitting and enhance robustness and accuracy. However, since correspondence estimation is pivotal to the robustness of hybrid VO systems, the persistent issues of noisy and inconsistent flow matching in existing hybrid methods result in significant performance degradation, particularly in challenging scenarios. Furthermore, VO systems commonly experience cumulative trajectory drift, which is exacerbated by these matching problems, resulting in increasingly inaccurate estimations over long sequences.

Existing hybrid methods rely solely on the information between two frames to obtain correspondences. However,

*These authors contributed equally.
†Corresponding author.

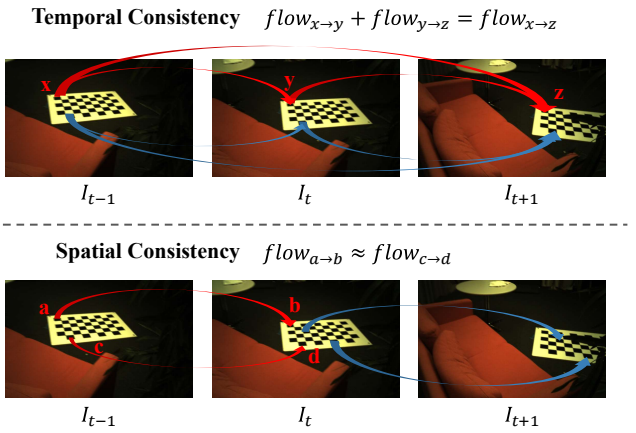


Figure 2: Diagram of Temporal Consistency and Spatial Consistency Across Multiple Frames

flow estimated from just two frames is prone to noise and inconsistencies, especially in challenging regions. These methods overlook the potential of VO systems to jointly estimate poses across multiple frames in local windows, which could also be leveraged for multi-frame joint optical flow estimation. This approach can significantly enhance the accuracy and robustness of optical flow by optimally integrating spatio-temporal cues. With more consistent flow matching in both spatial and temporal domains, VO can achieve better pose estimation through the bundle adjustment (BA) step. This insight is driven by two key observations, as illustrated in Figure 2. First, there is temporal consistency across multiple frames, where adjacent frames share similar motion trends and are constrained by flow consistency. Second, each frame exhibits spatial consistency, where different points on the same object maintain highly uniform motion patterns.

Building on these observations, we introduce Spatio-Temporal Visual Odometry (STVO), a novel deep network architecture that integrates spatio-temporal cues to optimize multi-frame optical flow matching. Firstly, we introduce a Temporal Propagation Module which leverages multi-frame information to extract and propagate temporal cues across adjacent frames. The Temporal Propagation Module maintains a motion state for each source frame and iteratively uses currently predicted optical flows to warp motion state across all adjacent frames, enabling the accurate reflection of motion dynamics. Finally, the temporal information obtained from adjacent frames is propagated back to the source frame and updates the motion state for the next iteration. Secondly, we introduce the Spatial Activation Module. This module innovatively utilizes depth information and geometric priors to enhance spatial consistency in flow estimation. It leverages geometric information from depth maps to create a Spatial Attention Matrix, which employs a attention-based approach to model spatial cues. The Spatial Attention Matrix is then used to activate the coarse context and correlation features, enabling comprehensive spatial understanding and filtering out noise and incorrect matches.

Extensive experiments demonstrate that our STVO outperforms all prior works across all four real-world benchmarks. Specifically, on the challenging ETH3D dataset, STVO shows a 77.8% improvement over the previous best method. Additionally, on the long-sequence KITTI Odometry benchmark, it achieves a 38.9% improvement over the previous best method.

In summary, our main contributions are as follows:

- To the best of our knowledge, we are the first to highlight the significance of both spatial and temporal consistency for matching in Visual Odometry. We introduce Spatio-Temporal Visual Odometry (STVO), a novel deep network architecture that harnesses spatio-temporal cues to mitigate performance degradation in challenging scenarios and long-sequence estimations.
- We propose the Temporal Propagation Module, which capitalizes on the high correlation between neighboring frames to enhance temporal consistency.
- We propose the Spatial Activation Module, which utilizes depth information and geometric priors to maintain spatial consistency and filter out noise and incorrect matches.
- Our STVO achieves state-of-the-art accuracy on four benchmarks: TUM-RGBD, EuRoC MAV, ETH3D, and KITTI Odometry, while also showcasing outstanding performance under challenging conditions and excellent resistance to drift in long sequences.

Related Work

Visual odometry methods can be broadly categorized into classical methods, deep learning-based methods and Hybrid VO framework.

Classical Methods can be further divided into direct and indirect approaches. *Direct* approaches operate based on the photometric constancy assumption, obtain pose by minimizing the photometric error of pixel intensities across the images (Forster, Pizzoli, and Scaramuzza 2014; Engel, Koltun, and Cremers 2017; Engel, Schöps, and Cremers 2014). *Indirect* approaches use hand-crafted features to find correspondences and then minimize the reprojection error on the correspondences (Mur-Artal, Montiel, and Tardos 2015; Mur-Artal and Tardós 2017; Campos et al. 2021). A common issue with classical methods is that handcrafted features often lack robustness, making VO systems prone to failure.

Deep Learning-Based Methods have been proposed to enhance robustness. Early methods replaced traditional handcrafted feature extraction with neural networks for feature extraction (DeTone, Malisiewicz, and Rabinovich 2018; Ono et al. 2018) or adopted end-to-end approaches to predict the pose (Wang et al. 2017; Teed and Deng 2018; Wang, Hu, and Scherer 2021). Although learning-based VO has shown to improve accuracy and robustness compared to classical approaches, they still face the challenge of significant performance degradation when dealing with data that differs from training distributions (Zhan et al. 2021; Zhang et al. 2020a,b; Wang et al. 2024; Cheng et al. 2024b,c).

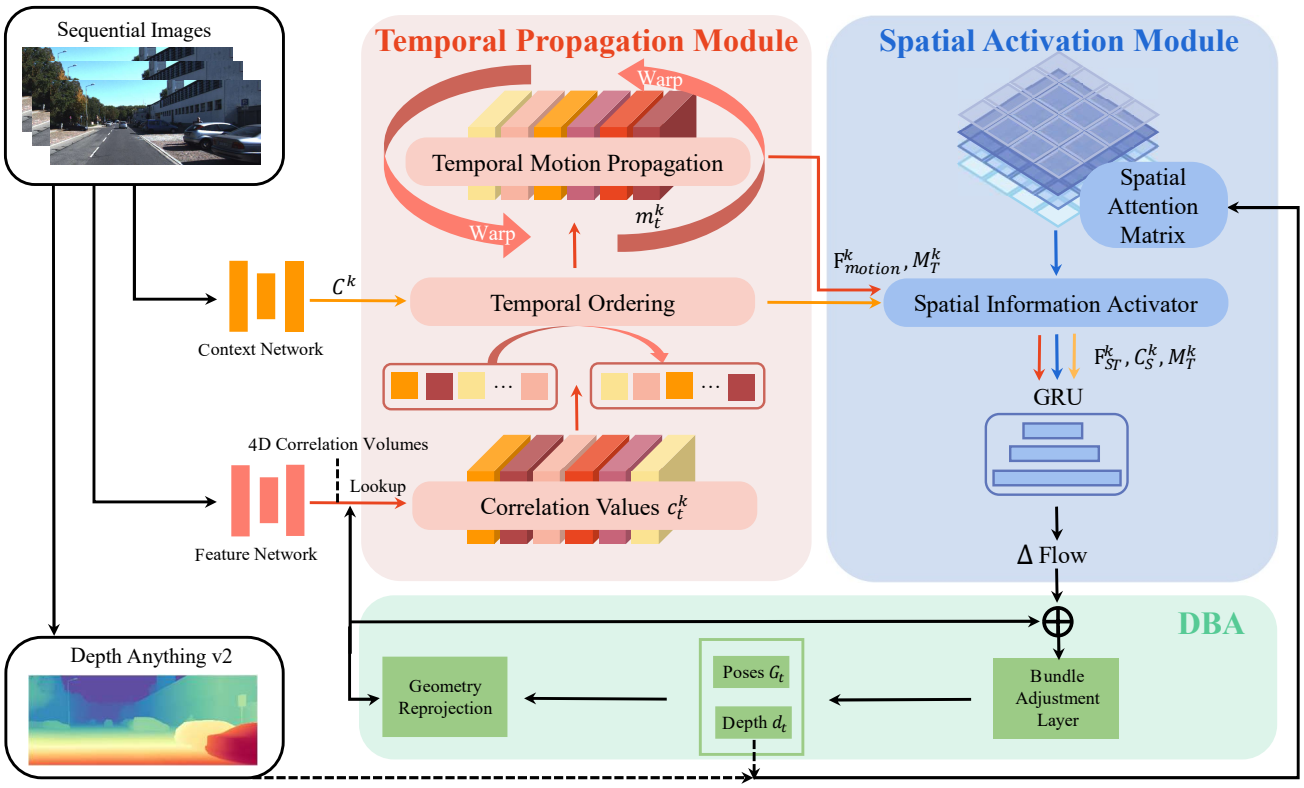


Figure 3: Overview of STVO. The architecture consists of three key modules: 1) Temporal Propagation Module, which enhances temporal consistency; 2) Spatial Activation Module, which maintains spatial consistency and filters out incorrect matches; 3) Differentiate Bundle Adjustment (DBA) Module, which updates poses and depths using optical flow estimates. The dashed lines in STVO indicate that the depth map input can be flexibly chosen between the depth generated by Depth Anything V2 and the output depth map of Bundle Adjustment. Both input options have demonstrated significant effectiveness.

Hybrid VO Framework was proposed to combine the strengths of classical methods and deep learning methods. There has already been a series of outstanding works in this field such as (Teed and Deng 2021; Teed, Lipson, and Deng 2024; Chen et al. 2024; Gurumurthy et al. 2024; Klenk et al. 2024; Cheng et al. 2024a), among them, two particularly notable and closely related to our work are DROID-SLAM and DPVO. DROID-SLAM integrates RAFT for iterative dense optical flow prediction to obtain a more robust correspondence, which greatly improves the robustness of VO system. To improve the efficiency of VO, DPVO (Teed, Lipson, and Deng 2024) replaces dense optical flow tracking with patch-based tracking. However, this patch-based matching approach can exacerbate instability issues in extreme scenarios, such as motion blur or highly dynamic scenes.

Optical Flow is the task of estimating dense 2D pixel-level motion between a pair of frames. Many deep learning-based optical flow methods (Ranjan and Black 2017; Sun et al. 2018; Jiang et al. 2021; Luo et al. 2022; Shi et al. 2023; Xu et al. 2022, 2023a,b, 2024) have shown outstanding performance. A recent standout work, RAFT (Teed and Deng 2020), employs a multi-scale search and recurrent manner to estimate flow, effectively balancing both accuracy and efficiency. Many hybrid VO systems use RAFT-style networks to obtain optical flow, thereby improving VO performance. However, these systems typically use RAFT merely as a module for obtaining two-frame correspondences, without fully integrating the characteristics of VO. In fact, the local sliding window in VO contains rich temporal and motion information, which can be effectively utilized to enhance the robustness of optical flow estimation, rather than relying on isolated pairwise two-frame matching. This is precisely what STVO aims to achieve.

Background

We choose DROID-VO as our baseline (the VO front-end of DROID-SLAM (Teed and Deng 2021)). In this section, we provide a brief overview of the relevant components of DROID-VO to help readers better understand our approach.

Representation. Our network operates on an ordered collection of images $\{I_t\}_{t=0}^N$, camera poses $G_t \in SE(3)$, and inverse depths $d_t \in \mathbb{R}_+^{H \times W}$ follow DROID-VO. We adopt a frame-graph $(\mathcal{V}, \mathcal{E})$ to represent co-visibility between keyframes. And the frame-graph is built dynamically during training and inference to ensure that, within the local window, only the r reference frames closest to the target keyframe are retained.

Update Operator. The key to DROID-VO's outstanding

performance lies in using the RAFT-based update operator to iteratively refine the optical flow. This approach projects the estimates of depth d_t and pose G_t in each Bundle Adjustment (BA) iteration to obtain initial optical flow f_t^k and uses the GRU module to refine the flow. To be specific, the correlation features and context features are injected into the GRU, which maintains a hidden state h^k during the update process and outputs the revision flow field $r_{ij} \in \mathbb{R}^{H\times W\times 2}$ and the associated confidence map $w_{ij} \in \mathbb{R}_+^{H\times W\times 2}$. The improved optical flow is then used to optimize more accurate depth and pose in the BA module, thereby forming a feedback loop that enhances the overall accuracy.

Differentiable Bundle Adjustment. After obtaining the predicted flow revisions and the flow confidence, the differentiate bundle adjustment (DBA) Layer is applied to get the updated poses and pixel-wise depths. The DBA objective is as follows:

$$T^*, d^* = \arg\min_{T,d} \sum_{(i,j)\in\mathcal{E}} \left\| \hat{P}_{ij} - \tilde{P}_{ij} \right\|_{\Sigma_{ij}}^2 \tag{1}$$

where $\hat{P}_{ij}$, $\tilde{P}_{ij}$ and Σ_{ij} denote the reprojected position, the estimated flow, and the confidence weights from the I_i to the I_j, respectively.

Method

Overview

Figure 3 presents the overall architecture of our network. STVO consists of four main steps: 1) *Feature Extraction Network*, which extracts features used to compute the optical flow cost volume and generates correlation values c_t^k. 2) *Temporal Propagation Module*, by imposing temporal consistency constraints on the correlation values, a more robust matching cost is obtained. 3) *Spatial Activation Module*, which applied the spatial consistency constraints to the correlation values to derive the final matching cost volume, then through a GRU module to obtain the revision flow. 4) *Differentiable Bundle Adjustment (DBA) Module*, the refined optical flow is utilized by the DBA module to optimize the depth and pose. The optimized depth d_t and pose G_t are then used for geometric reprojection, providing an initial optical flow estimate for the next iteration, thereby creating a positive feedback loop. The key contributions of STVO lie in two main components: the Temporal Propagation Module and the Spatial Activation Module, which can complement each other to achieve a more stable optical flow, further enhancing the potential of the positive feedback loop and improving the accuracy of pose estimation.

Temporal Propagation Module

Temporal Ordering. As shown in the first row of Figure 3, the dynamic addition and removal of keyframes within the local window causes the edges $\mathcal{E}$ in the frame graph to lose their chronological order, thereby disrupting the temporal relationships between frames. This disruption makes it challenging to leverage temporal consistency for constraining the optical flow. To address this, we applied temporal ordering to the correlation values, sorting them according to

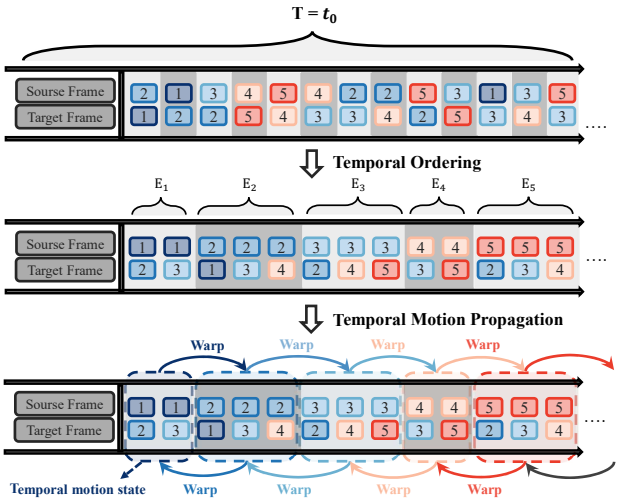


Figure 4: Diagram of Temporal Propagation Module.

the sequence of their source frames. For simplicity, we denote the set of edges in the frame-graph that originate from the same source frame t as the Source Edge Set E_t in the following text, where $E_t = \{(t, j) \mid t \in \mathcal{V} \text{ and } (t, j) \in \mathcal{E}\}$.

Temporal Motion Propagation. To leverage the benefits of temporal ordering, allowing the system to utilize the sequential information and enhance temporal consistency, STVO maintains a motion state $m_t^k \in R^{H\times W\times D_m}$ for each Source Edge Set E_t, where m_t^0 is randomly initialized. The motion state m_t^k contains motion information at the k-th iteration for I_t and will be updated in each iteration.

In iteration k, for each edge (m, n) in the Source Edge Set E_m, the motion state m_m^k is warped using the predicted optical flow $f_{m\to n}$. This warping produces the corresponding dynamic motion state $M_{m\to n}^k \in R^{H\times W\times D_m}$, enabling the target frame to leverage the motion information from the source frame. We then concatenate the source frame's motion state m_m^k, the target frame's motion state m_n^k, and the acquired dynamic motion state $M_{m\to n}^k$ to obtain the temporal motion state $M_T^k \in R^{H\times W\times 3D_m}$ as below, which providing a comprehensive representation of the temporal motion across this frame edge (m, n).

$$\begin{aligned} M_{m\to n}^k &= \text{Warp}(m_m^k; f_{m\to n}^k), \\ M_T^k &= \text{Concat}(m_m^k, M_{m\to n}^k, m_n^k) \end{aligned} \tag{2}$$

The we concatenate the correlation feature F_{corr}^k and our temporal motion state M_T^k to generate the temporal motion feature $F_{motion}^k \in R^{H\times W\times D_M}$ and local motion state $m_{m\to n}^{k+1}$ by *TemporalEncoder*, which is a cascaded two-layer 2D convolution. Subsequently, STVO propagates all the collected local motion states back to I_m to update the motion state for the next iteration.

$$\begin{aligned} F_{motion}^k, m_{m\to n}^{k+1} &= \text{TemporalEncoder}(F_{corr}^k, M_T^k), \\ m_m^{k+1} &= \frac{1}{|T_m|} \sum_{n\in T_m} m_{m\to n}^{k+1} \end{aligned} \tag{3}$$

where $T_m = \{n \mid (m, n) \in E_m\}$ represents the collection of all target frames for each frame image I_m, and $|T_m|$ is the

	360	desk	desk2	floor	plant	room	rpy	teddy	xyz	Avg
ORB-SLAM3	-	**0.017**	0.210	-	0.034	-	-	-	**0.009**	-
DSO	0.173	0.567	0.916	0.080	0.121	0.379	0.058	-	0.036	-
DeepV2d	**0.144**	0.105	0.321	0.628	0.217	**0.215**	0.046	0.294	0.051	0.225
TartanVO	0.178	0.125	0.122	0.349	0.297	0.333	0.049	0.339	0.062	0.206
DROID-VO	0.161	0.028	0.099	**0.033**	0.028	0.327	0.028	0.169	0.013	0.098
DPVO	0.165	0.034	**0.042**	0.050	0.036	0.388	0.034	**0.057**	0.012	0.091
STVO(Ours)	0.171	0.031	0.068	0.035	**0.027**	0.242	**0.027**	0.111	0.015	**0.080**

Table 1: Performance comparisons on the TUM-RGBD Dataset on ATE[m]. $(-)$ indicates that the method failed to track. We use **bold** and _ to highlight the methods that rank 1st and 2nd.

number of frame edges in E_m. The temporal motion feature F^k_{motion} and the temporal motion state M^k_T are then transferred to the Spatial Activation Module to further enhance spatial consistency.

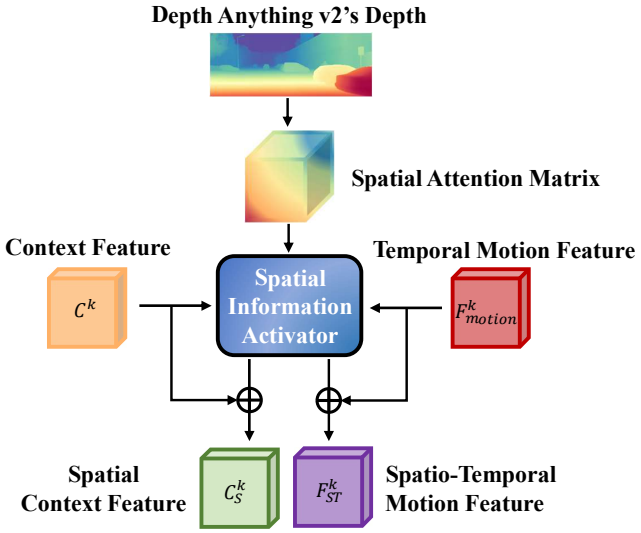


Figure 5: Diagram of Spatial Activation Module

Spatial Activation Module

Spatial Attention Matrix. As shown in Figure 3, for each image $I_m \in R^{H_1 \times W_1 \times 3}$, we generate a corresponding depth map $D_m \in R^{H \times W}$ by using the Depth Anything v2 small model or using our BA depth (output of DBA module), where $H = H_1/8$, $W = W_1/8$. To enhance flow consistency, we exploit the spatial consistency information embedded in the depth map through spatial attention. Specifically, the query feature q_m and key feature k_m are derived by projecting the depth map D_m into a higher-dimensional feature space. These features are used to compute the Spatial Attention Matrix (SAM), which captures the spatial relationships and context dependencies within the depth map D_m:

$$\begin{aligned} q_m = D_m W_q, k_m = D_m W_k \\ SAM = \sigma\left(q_m k_m^\top\right) \end{aligned} \tag{4}$$

where W_q,$W_k \in R^{1 \times D_{in}}$ are projection matrices, and $q_m, k_m \in R^{HW \times D_{in}}$ denotes query, key features. $\sigma(\cdot)$ denotes the softmax function.

Spatial Information Activator. As shown in Figure 5, the Spatial Attention Matrix is utilized in the Spatial Information Activator to aggregate object consistency information and filter out irrelevant information, thereby making the matching features more robust. Specifically, we apply attention aggregation to both the temporal motion feature F^k_{motion} and context feature C^k, resulting in our spatio-temporal motion Feature F^k_{ST} and spatial context feature C^k_S respectively as below:

$$\begin{aligned} C^k_S = C^k + \alpha_c \cdot SAM \cdot C^k \\ F^k_{ST} = F^k_{motion} + \alpha_f \cdot SAM \cdot F^k_{motion} \end{aligned} \tag{5}$$

where α_c and α_f are learned scalar parameters. Once the matching features have been optimized through our Temporal Propagation Module and Spatial Activation Module, they are combined with the temporal motion state M^k_T and fed into the GRU module to obtain the refined optical flow for the current iteration.

Implementation Details

STVO is implemented using PyTorch (Paszke et al. 2019) and C++. The training strategy for STVO is almost consistent with that of DROID-VO, which is conducted on monocular images from the synthetic TartanAir dataset (Wang et al. 2020). We train our network for 250k steps with a batch size of 4, resolution 384 × 512, and 7 frame clips, and unroll 15 update iterations.

Experiments

In the experimental section, we validated the effectiveness of the STVO design through comprehensive quantitative and qualitative comparisons. Additionally, we performed ablation studies to analyze the contributions of each component.

Quantitative Comparison

In this section, we conduct experiments on popular real-world VO benchmarks to quantitatively evaluate the effectiveness of our approach. Following prior works, We assess the estimated trajectory on Absolute Trajectory Error (ATE) using scale alignment with evo (Grupp 2017). We compare STVO with classical methods, such as ORB-SLAM3 (Campos et al. 2021), DSO (Engel, Koltun, and Cremers 2017),

	cables	camera_shake	ceiling	desk_changing	einstein	mannequin	plant_scene	sfm_lab_room	Avg
ORB-SLAM3	0.210	-	-	-	0.138	-	-	-	-
DSO	0.174	-	-	1.406	0.301	0.805	-	0.827	-
DeepV2d	0.229	0.100	1.974	0.964	0.069	0.559	0.262	0.057	0.526
TartanVO	0.364	0.107	1.889	0.851	0.120	0.227	0.651	0.277	0.561
DROID-VO	0.030	**0.034**	0.193	0.186	**0.002**	**0.009**	**0.005**	1.066	0.190
DPVO	**0.020**	0.071	0.398	1.405	0.007	0.020	0.017	0.060	0.249
STVO(Ours)	0.028	0.038	**0.183**	**0.045**	0.003	0.013	0.009	**0.015**	**0.042**

Table 2: Performance comparisons on the ETH3D Dataset on ATE[m]. $(-)$ indicates that the method failed to track.

	00	03	04	05	06	07	08	09	10	Avg
ORB-SLAM3	74.69	**0.64**	1.81	33.22	47.07	16.20	54.46	46.61	**7.13**	31.31
DSO	48.04	0.80	**0.36**	48.45	57.59	53.67	113.02	92.19	11.03	47.23
DeepV2d	101.65	7.15	4.08	27.05	**7.39**	8.70	**18.91**	**10.13**	14.77	22.20
TartanVO	63.84	7.71	2.89	54.61	24.67	19.29	59.55	32.61	25.04	32.25
DROID-VO	109.00	5.57	1.05	60.37	38.03	21.41	105.64	73.04	13.77	47.53
DPVO	108.88	1.60	1.55	56.60	59.37	17.64	97.48	62.94	10.03	46.23
STVO(Ours)	**16.13**	2.15	1.93	**19.49**	14.07	**7.72**	29.18	17.28	14.16	**13.56**

Table 3: Performance comparisons on the KITTI Odometry Dataset on ATE[m].

and influential deep learning methods like DeepV2d (Teed and Deng 2018), TartanVO (Wang, Hu, and Scherer 2021), DROID-VO (Teed and Deng 2021), and DPVO (Teed, Lipson, and Deng 2024) across four public datasets: TUM-RGBD (Sturm et al. 2012), EuRoC MAV (Burri et al. 2016), ETH3D (Schops, Sattler, and Pollefeys 2019) and KITTI Odometry (Geiger, Lenz, and Urtasun 2012). For a fair comparison, we present the results for DPVO as the average of five runs, instead of using the median of five runs as reported in (Teed, Lipson, and Deng 2024). This choice is made because the median method tends to exclude outliers and does not fully represent the model's average performance.

TUM-RGBD. The TUM-RGBD dataset (Sturm et al. 2012) consists of indoor scene data captured using handheld cameras, which introduces challenges like rolling shutter artifacts, motion blur, and significant rotations. In Table 1, we benchmark on TUM-RGBD and compare to SOTA. STVO achieves the best or second-best results in 6 out of 9 sequences and delivers the best overall average performance. Compared to DPVO and DROID-VO, it shows improvements of 12% and 18%, respectively.

EuRoC MAV. In Table 4, we present our benchmarking results on the EuRoC MAV(Burri et al. 2016) dataset. EuRoC MAV is a widely used benchmark for evaluating VO systems, collected using Micro Aerial Vehicles (MAV). The dataset includes fast and dynamic camera movements that pose significant challenges for VO systems. To align with the practices of similar studies such as DROID-VO and DPVO, we process every other frame, effectively doubling the system's frame rate to 40 FPS. STVO outperforms previous methods on the majority of sequences on EuRoC MAV. The average error of STVO is 5% lower than the state-of-the-art DPVO method and 21% lower than DROID-VO.

	V101	V102	V103	V201	V202	V203	Avg
ORB-SLAM3	**0.035**	0.139	0.713	1.352	0.056	0.632	0.487
DSO	0.089	0.107	0.903	**0.044**	0.132	1.152	0.404
DeepV2d	0.717	0.695	1.483	0.839	1.052	0.591	1.173
TartanVO	0.447	0.389	0.622	0.433	0.749	1.152	0.632
DROID-VO	0.103	0.165	0.158	0.102	0.115	**0.204**	0.141
DPVO	0.050	0.148	0.093	0.086	**0.049**	0.282	0.118
STVO(Ours)	0.055	**0.098**	**0.078**	0.064	0.125	0.248	**0.111**

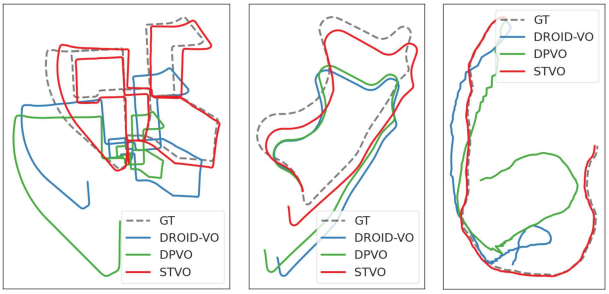
Table 4: Performance comparisons on EuRoC MAV dataset.

ETH3D. In Table 2, we use the SLAM benchmark of ETH3D (Schops, Sattler, and Pollefeys 2019). It includes eight distinct scenes, each comprising multiple sequences. We consistently select the first sequence from each scene for a fair comparison. The ETH3D dataset covers a wide range of scenarios, with many sequences being particularly challenging due to low lighting and significant motion blur. This diversity and difficulty allow for a comprehensive evaluation of VO systems' performance. For instance, in the desk_changing sequence, where objects move quickly, both classical methods and the DPVO algorithm, which relies on randomly selecting image patches between only two frames, struggle to achieve stable matching, leading to failure. In contrast, STVO demonstrated the best performance with an error of only 0.045 meters, showcasing its robustness and precision. Notably, our approach also achieved the lowest average error across all test sequences. Compared to DPVO, we reduced the average error from 0.249 meters to 0.042 meters, representing an improvement of 83.1%.

KITTI Odometry. In Table 3, we benchmark on the KITTI Odometry (Geiger, Lenz, and Urtasun 2012) dataset. KITTI Odometry is designed for autonomous driving applications. Among existing methods, performance on KITTI Odometry has been generally unsatisfactory. Apart from STVO, DeepV2d demonstrates the best performance due to being the only method trained on the KITTI dataset. However, STVO not only exceeds all other methods, but also achieves a notable 38.9% improvement over DeepV2d, without any fine-tuning on KITTI. This highlights STVO's excellent ability to mitigate trajectory drift in long sequences and its strong generalization capability.

Qualitative Comparison

In this section, we qualitatively compare trajectories and optical flow to validate the effectiveness of STVO.



KITTI Odometry:00 KITTI Odometry:09 ETH3D:sfm_room_lab

Figure 6: Visualization comparison of trajectory.

Trajectory Comparison. As shown in Figure 6, both DROID-VO and DPVO exhibit significant trajectory drift in the KITTI Odometry long sequences 00 and 09, which cover distances of 3724m and 1705m, respectively. In contrast, STVO shows remarkable resistance to drift, highlighting its superior performance in long-sequence estimation. For ETH3D sequence sfm_room_lab, large area repeated texture increases the difficulty of matching, resulting in serious trajectory drift of DPVO and DROID-VO algorithms that only rely on two frames of information for matching. However, STVO consistently aligns with the ground truth, demonstrating its ability to maintain accurate trajectory estimation even in challenging situations.

Flow Comparison. As shown in Figure 7, compared to DROID-VO, STVO has a sharper optical flow and a more accurate structure due to the efficient integration of the Spatial Activation Module with spatial information. Even when handling complex situations, STVO still demonstrates superior stability and accuracy in optical flow estimation.

Ablations

We conduct ablation studies in Table 5 to validate each component of STVO on the TUM-RGBD dataset by comparing the model's average ATE, GPU memory usage and average frames per second (FPS) under different settings. DROID-VO serves as our baseline. By individually adding our Spatial Activation Module and the Temporal Propagation Module to the baseline (denoted as Base + SAM and Base + TPM), we achieve an improvement in ATE by 9.2% and 13.3% respectively without introducing significant memory cost. Moreover, their combined use yields even better results, i.e. our final STVO model. Specifically, to evaluate the impact of Depth Anything V2's geometry priors, we replace Depth Anything V2's depth maps with those from Bundle Adjustment (denoted as DepAny → BA depth). The result shows that using BA depth alone can also significantly improve VO performance with fewer memory costs compared to STVO. This indicates that the SAM's improvement is not solely derived from the robust priors provided by Depth Anything V2, but rather from our design that leverages depth information to enforce motion consistency. This highlights the superiority of our Spatial Activation Module.

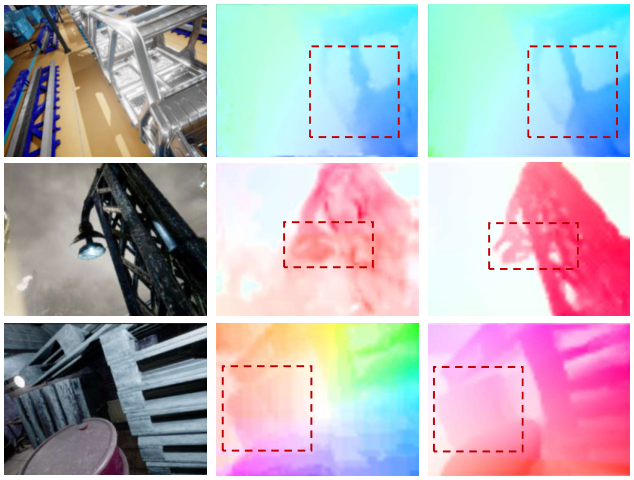


Image DROID-VO STVO

Figure 7: Visual comparison of optical flow in challenging conditions: highlights, fog, and repetitive structures.

Ablation	SAM	TPM	ATE	GPU	FPS
DROID-VO (Baseline)			0.098	4.2	9.8
Base+SAM	✓		0.089	5.6	7.8
Base+TPM		✓	0.085	4.8	9.1
Full model (STVO)	✓	✓	0.080	6.0	7.3
DepAny → BA depth	✓	✓	0.082	5.1	8.0

Table 5: Ablation Study on the Designs of STVO.

Conclusion

We present STVO, a novel deep Visual Odometry architecture that integrates spatio-temporal cues to enhance both spatial and temporal consistency in multi-frame matching, effectively mitigating performance degradation in challenging scenarios and long-sequence estimations. STVO outperforms all previous methods on TUM-RGBD, EuRoC MAV, ETH3D, and KITTI Odometry. STVO is the first to highlight the significance of spatial and temporal consistency in multi-frame flow matching for VO.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62122029, 62472184), the Fundamental Research Funds for the Central Universities, and the National Natural Science Foundation of China (623B2036).

References

Burri, M.; Nikolic, J.; Gohl, P.; Schneider, T.; Rehder, J.; Omari, S.; Achtelik, M. W.; and Siegwart, R. 2016. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10): 1157–1163.

Campos, C.; Elvira, R.; Rodríguez, J. J. G.; Montiel, J. M.; and Tardós, J. D. 2021. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6): 1874–1890.

Chen, S.; Liu, K.; Wang, C.; Yuan, S.; Yang, J.; and Xie, L. 2024. Salient Sparse Visual Odometry With Pose-Only Supervision. *IEEE Robotics and Automation Letters*.

Cheng, J.; Cai, Z.; Zhang, Z.; Yin, W.; Muller, M.; Paulitsch, M.; and Yang, X. 2024a. RoMeO: Robust Metric Visual Odometry. arXiv:2412.11530.

Cheng, J.; Xu, G.; Guo, P.; and Yang, X. 2024b. Coatrsnet: Fully exploiting convolution and attention for stereo matching by region separation. *International Journal of Computer Vision*, 132(1): 56–73.

Cheng, J.; Yin, W.; Wang, K.; Chen, X.; Wang, S.; and Yang, X. 2024c. Adaptive fusion of single-view and multi-view depth for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10138–10147.

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

Engel, J.; Koltun, V.; and Cremers, D. 2017. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3): 611–625.

Engel, J.; Schöps, T.; and Cremers, D. 2014. LSD-SLAM: Large-scale direct monocular SLAM. In *European conference on computer vision*, 834–849. Springer.

Forster, C.; Pizzoli, M.; and Scaramuzza, D. 2014. SVO: Fast semi-direct monocular visual odometry. In *2014 IEEE international conference on robotics and automation (ICRA)*, 15–22. IEEE.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Grupp, M. 2017. evo: Python package for the evaluation of odometry and SLAM. https://github.com/MichaelGrupp/evo.

Gurumurthy, S.; Ram, K.; Chen, B.; Manchester, Z.; and Kolter, Z. 2024. From Variance to Veracity: Unbundling and Mitigating Gradient Variance in Differentiable Bundle Adjustment Layers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27507–27516.

Jiang, S.; Campbell, D.; Lu, Y.; Li, H.; and Hartley, R. 2021. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9772–9781.

Klenk, S.; Motzet, M.; Koestler, L.; and Cremers, D. 2024. Deep event visual odometry. In *2024 International Conference on 3D Vision (3DV)*, 739–749. IEEE.

Luo, A.; Yang, F.; Li, X.; and Liu, S. 2022. Learning Optical Flow With Kernel Patch Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8906–8915.

Mur-Artal, R.; Montiel, J. M. M.; and Tardos, J. D. 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, 31(5): 1147–1163.

Mur-Artal, R.; and Tardós, J. D. 2017. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5): 1255–1262.

Ono, Y.; Trulls, E.; Fua, P.; and Yi, K. M. 2018. LF-Net: Learning local features from images. *Advances in neural information processing systems*, 31.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4161–4170.

Schops, T.; Sattler, T.; and Pollefeys, M. 2019. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 134–144.

Shi, X.; Huang, Z.; Bian, W.; Li, D.; Zhang, M.; Cheung, K. C.; See, S.; Qin, H.; Dai, J.; and Li, H. 2023. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12469–12480.

Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; and Cremers, D. 2012. A benchmark for the evaluation of RGB-D SLAM systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, 573–580. IEEE.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8934–8943.

Tang, C.; and Tan, P. 2018. Ba-net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807*.

Teed, Z.; and Deng, J. 2018. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 402–419. Springer.

Teed, Z.; and Deng, J. 2021. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34: 16558–16569.

Teed, Z.; Lipson, L.; and Deng, J. 2024. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36.

Wang, S.; Clark, R.; Wen, H.; and Trigoni, N. 2017. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE international conference on robotics and automation (ICRA)*, 2043–2050. IEEE.

Wang, W.; Hu, Y.; and Scherer, S. 2021. Tartanvo: A generalizable learning-based vo. In *Conference on Robot Learning*, 1761–1772. PMLR.

Wang, W.; Zhu, D.; Wang, X.; Hu, Y.; Qiu, Y.; Wang, C.; Hu, Y.; Kapoor, A.; and Scherer, S. 2020. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4909–4916. IEEE.

Wang, X.; Liu, J.; Feng, M.; Zhang, Z.; and Yang, X. 2024. 3D Multi-Object Tracking with Semi-Supervised GRU-Kalman Filter. *arXiv preprint arXiv:2411.08433*.

Xu, G.; Cheng, J.; Guo, P.; and Yang, X. 2022. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12981–12990.

Xu, G.; Wang, X.; Ding, X.; and Yang, X. 2023a. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21919–21928.

Xu, G.; Wang, X.; Zhang, Z.; Cheng, J.; Liao, C.; and Yang, X. 2024. Igev++: Iterative multi-range geometry encoding volumes for stereo matching. *arXiv preprint arXiv:2409.00638*.

Xu, G.; Wang, Y.; Cheng, J.; Tang, J.; and Yang, X. 2023b. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuan, Z.; Deng, J.; Ming, R.; Lang, F.; and Yang, X. 2024. SR-LIVO: LiDAR-Inertial-Visual Odometry and Mapping With Sweep Reconstruction. *IEEE Robotics and Automation Letters*.

Yuan, Z.; Lang, F.; Xu, T.; and Yang, X. 2022. Sr-lio: Lidar-inertial odometry with sweep reconstruction. *arXiv preprint arXiv:2210.10424*.

Yuan, Z.; Wang, Q.; Cheng, K.; Hao, T.; and Yang, X. 2023. SDV-LOAM: semi-direct visual–LiDAR Odometry and mapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 11203–11220.

Zhan, H.; Weerasekera, C. S.; Bian, J.-W.; Garg, R.; and Reid, I. 2021. DF-VO: What should be learnt for visual odometry? *arXiv preprint arXiv:2103.00933*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020a. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; and Sun, Q. 2020b. Feature pyramid transformer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 323–339. Springer.