2025 International Conference on Advanced Robotics and Intelligent Systems (ARIS 2025)
College of EECS, NCHU, Taichung 40227, Taiwan, August 17-19, 2025

VSTFusion-VO: Monocular Visual Odometry with Video Swin Transformer Multimodal Fusion

Chia-Hsuan Hsu, Hsin-Chun Lin, Sin-Ye Jhong, Hui-Che Hsu, Ming-Xian Hong, and Yung-Yao Chen*, *Senior Member, IEEE*

***Abstract*—This paper presents a learning-based monocular visual odometry (VO) framework that leverages a Video Swin Transformer for hierarchical 3D spatiotemporal modeling. Beyond incorporating pseudo-depth, our method employs early multimodal fusion and 3D patch embedding to jointly encode RGB and geometric information before transformer-based processing, enabling effective spatiotemporal representation learning without relying on ground-truth depth. Trained end-to-end to predict 6-DoF poses, the model captures both local motion patterns and long-range dependencies. Experiments on the KITTI Odometry dataset demonstrate superior performance compared to prior learning-based VO methods in both translation and rotation accuracy. The code is available at: `https://github.com/tongyu0924/VSTFusion-VO`**

***Index Terms*—visual odometry, Swin Transformer, multimodal fusion, KITTI dataset**

I. Introduction

Visual Odometry (VO) is a core technology in robotic perception and autonomous navigation, with widespread applications in fields such as autonomous driving, indoor robotics, and virtual reality. By analyzing a series of consecutive images captured by a camera, VO calculates the real-time pose change—encompassing both rotation and translation—of a device relative to its starting position to estimate its trajectory. A key advantage of VO is its independence from external positioning systems like GPS. This makes it an efficient and valuable solution for GPS-denied environments such as indoor spaces, underground tunnels, or dense urban canyons, all while relying on relatively low-cost sensors.

Traditional VO algorithms are primarily categorized into two main approaches: feature-based and direct methods. Feature-based approaches, such as the well-known ORB-SLAM2 [1], follow a modular pipeline that involves detecting and matching sparse keypoints (e.g., SIFT [2], ORB [3]) across frames for geometric optimization. In contrast, direct methods utilize dense pixel intensity information to estimate motion, which can be faster in some scenarios but is often more sensitive to illumination changes. Despite their successes, these traditional methods' heavy reliance on handcrafted features and fine-tuning makes them vulnerable under dynamic conditions and in low-texture scenes.

To overcome these limitations, learning-based VO approaches [4], [5] directly estimate poses from raw RGB inputs, reducing dependence on manual design and improving adaptability. Nevertheless, they still struggle with long-term temporal consistency, generalization to unseen environments, and scale ambiguity in monocular settings [6], [7].

Transformer-based architectures have recently shown strong potential for capturing spatiotemporal dependencies [8], [9]. VO methods like TSformer-VO [10] and SWFormer-VO [11] enhance motion encoding using attention mechanisms. However, many existing models are misaligned with video-native transformer designs and often incorporate depth information only at late stages, which limits their ability to mitigate scale drift [6], [7], [12].

To address these challenges, we propose **VSTFusion-VO**, a transformer-based monocular VO framework featuring two key innovations: (1) early fusion of RGB and pseudo-depth inputs using 3D patch embedding to integrate geometric information from the outset, and (2) a video-native hierarchical transformer backbone based on the Video Swin Transformer [13], enabling multi-scale spatiotemporal modeling. This design preserves temporal continuity, mitigates scale drift without requiring external depth sensors, and strengthens motion encoding via multimodal attention.

The main contributions of this paper are summarized as follows:

- We propose a novel **early-fusion mechanism** that integrates RGB and pseudo-depth information at the initial input stage. By jointly encoding appearance and geometric cues into a unified representation using 3D patch embedding, our model effectively mitigates the scale ambiguity inherent in monocular systems.
- We design a **hierarchical temporal modeling backbone** based on the Video Swin Transformer, an architecture specifically tailored for video data. This allows our model to efficiently capture both local motion patterns and long-range spatiotemporal dependencies, which is critical for robust trajectory estimation.
- We conduct extensive experiments on the KITTI Odometry benchmark [14], demonstrating that VSTFusion-VO achieves **state-of-the-art performance**. Our method surpasses previous learning-based approaches in both translation and rotation accuracy, validating the effectiveness of combining early multimodal fusion with hierarchical temporal modeling.

II. Related Work

As introduced previously, learning-based approaches have emerged to overcome the limitations of traditional Visual Odometry (VO). This section reviews the key research streams that inform our proposed method. We first provide a broader

* Correspondence: yungyaochen@gapps.ntust.edu.tw

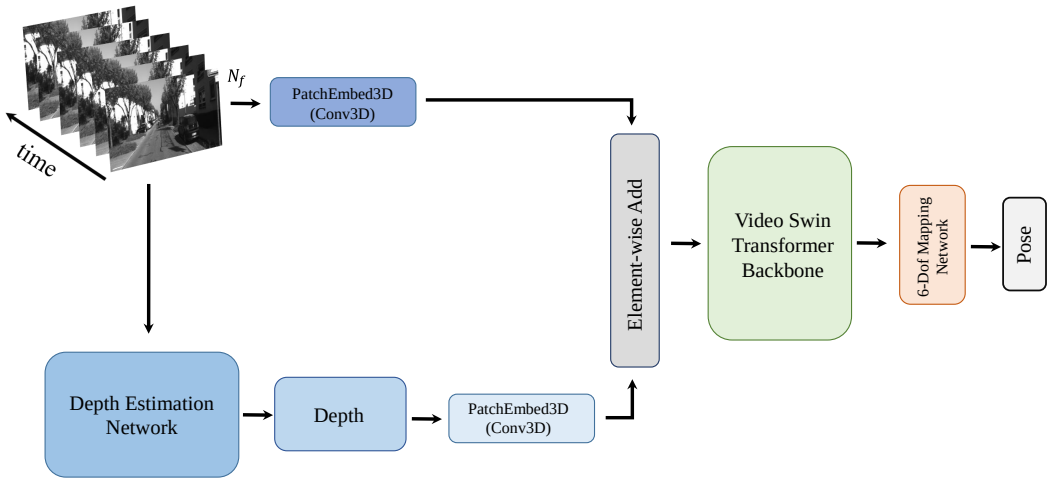


Fig. 1: The proposed VO architecture fuses RGB and pseudo-depth (inferred from monocular input) via 3D patch embedding, enabling early-stage geometric integration. A hierarchical Video Swin Transformer captures spatiotemporal dependencies, followed by a regression head that estimates 6-DoF camera motion.

overview of deep learning advancements in monocular VO. We then delve into two areas critical to our contributions: strategies for RGB-depth fusion to address scale ambiguity, and the evolution of transformer-based models for temporal modeling.

A. Deep Learning for Monocular Visual Odometry

Deep learning-based methods have reshaped the VO landscape by learning representations directly from data, thus reducing the reliance on handcrafted pipelines. Early pioneering works can be broadly categorized into supervised and unsupervised paradigms. Supervised methods, such as DeepVO [4], were among the first to successfully apply an end-to-end learning approach. They typically use a Convolutional Neural Network (CNN) to extract visual features, followed by a Recurrent Neural Network (RNN) to model temporal dynamics and directly regress 6-DoF poses. In parallel, unsupervised methods, notably SfMLearner [5], introduced an innovative self-supervised paradigm. By jointly training a depth network and a pose network to minimize the photometric reprojection error between consecutive frames, these models eliminated the need for ground-truth pose labels.

While these foundational works demonstrated the potential of deep learning, recent approaches have focused on further improving robustness and scalability. This includes leveraging spatiotemporal transformers [15], developing patch-level refinement techniques [16], and incorporating attention-based motion encoding [17] to address persistent challenges like scale ambiguity and generalization.

B. RGB-Depth Fusion for VO

Depth information is crucial in monocular VO for mitigating scale drift and enhancing geometric consistency. Early works, such as GeoNet [6] and UnDeepVO [18], incorporate depth supervision but typically fuse depth features only in later stages of the pipeline. Depth-VO-Feat [19] integrates depth through auxiliary heads or decoders.

More recent methods, including DVSO [20], adopt modality-specific encoders with separate branches for RGB and depth. RAFT-Stereo [21] further introduces stereo-specific correlation volumes to improve depth matching. In contrast, our method performs early fusion of RGB and pseudo-depth via shared 3D convolutional embeddings [22], allowing unified feature encoding prior to temporal modeling.

C. Transformer-based Temporal Modeling

Transformer models have recently shown strong potential in temporal visual perception. TSformer-VO [10] adopts a ViT-based architecture with divided space-time attention to model spatiotemporal dependencies from split-frame sequences. SWFormer-VO [11] extends this by leveraging hierarchical self-attention in Swin Transformers to model long-range temporal relations.

However, these models rely solely on RGB input, neglecting geometric information from depth, which can limit performance under low-texture or scale-sensitive conditions [23], [24]. Our method addresses this by integrating pseudo-depth features and employing a video-native temporal backbone to enhance geometric reasoning in VO tasks.

III. Proposed Method

Our proposed method, VSTFusion-VO, is designed to address the key challenges of scale ambiguity and insufficient temporal modeling identified in prior works. To achieve this, our framework introduces two core innovations that directly correspond to our main contributions: an early-fusion mechanism for integrating geometric cues, and a hierarchical temporal backbone for robust motion encoding. The overall architecture is illustrated in Fig. 1.

A. Spatial-guided Early Fusion via Joint Embedding

A primary challenge in monocular VO is the inherent scale ambiguity. While incorporating depth information can mitigate this, many existing methods perform late fusion, where RGB and depth features are processed in separate streams and only merged in later stages. This approach can be suboptimal, as the initial feature extraction lacks geometric context.

To overcome this, we introduce a joint embedding module that performs **early fusion** of RGB and pseudo-depth inputs. By integrating geometric information at the very beginning of the pipeline, we enable the network to learn a unified spatiotemporal representation where appearance and geometry are jointly encoded. This allows the subsequent transformer layers to reason about motion and scene layout more effectively from the outset.

Given an RGB frame I_t and its corresponding pseudo-depth map D_t at time t, we first project each modality into 3D patches and then merge them via element-wise addition to form a unified feature tensor $\mathbf{F}_t$:

$$\mathbf{F}_t = \text{PatchEmbed3D}(I_t) + \text{PatchEmbed3D}(D_t) \quad (1)$$

This fusion in the patch space ensures that the model can leverage multimodal cues from the earliest stage, enhancing temporal consistency and robustness in challenging scenes with low texture or dynamic lighting.

B. Temporal Fusion with Video Swin Transformer

Accurate VO requires a model that can capture complex motion patterns across a sequence of frames. Many previous transformer-based VO models treat video as a simple bag of frames, which can disrupt the crucial temporal continuity. We address this by employing a **hierarchical temporal modeling backbone** (Fig. 2) based on the first three stages of the Video Swin Transformer [13], an architecture natively designed for video data.

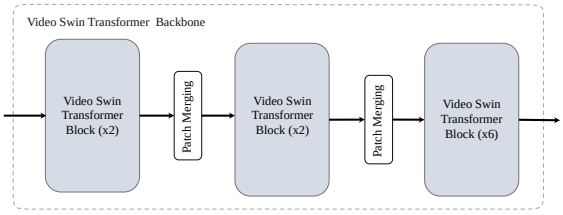


Fig. 2: Temporal dependencies are captured by the first three stages of the Video Swin Transformer.

The Video Swin Transformer's design is particularly well-suited for VO. It operates by progressively reducing spatial resolution through patch merging layers while expanding the temporal receptive field. This hierarchical process is highly effective, as it enables the model to capture both fine-grained, local motions (e.g., pixel displacements) in the early stages and long-range, global temporal patterns (e.g., sustained camera movements) in the later stages.

Formally, each Swin Transformer block in the temporal encoder applies 3D window-based multi-head self-attention (3DW-MSA), followed by feed-forward networks (FFN) and residual connections. The shifted window (3DSW-MSA) mechanism ensures information exchange between neighboring non-overlapping windows, enhancing spatiotemporal continuity. The computation in a pair of consecutive blocks is defined as:

$$\hat{\mathbf{z}}^l = \text{3DW-MSA}(\text{LN}(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1}, \quad (2)$$

$$\mathbf{z}^l = \text{FFN}(\text{LN}(\hat{\mathbf{z}}^l)) + \hat{\mathbf{z}}^l, \quad (3)$$

$$\hat{\mathbf{z}}^{l+1} = \text{3DSW-MSA}(\text{LN}(\mathbf{z}^l)) + \mathbf{z}^l, \quad (4)$$

$$\mathbf{z}^{l+1} = \text{FFN}(\text{LN}(\hat{\mathbf{z}}^{l+1})) + \hat{\mathbf{z}}^{l+1} \quad (5)$$

Here, $\mathbf{z}^l$ denotes the output of the l-th block and LN refers to Layer Normalization. This powerful temporal fusion module enables the model to effectively track motion over time, which is crucial for accurate and stable pose estimation.

C. Pose Estimation

After temporal encoding, the output feature tensor, which is rich with multimodal spatiotemporal information, is passed to a 6-DoF mapping network for pose regression. Each relative pose is represented as a 6-dimensional vector comprising 3 translation parameters (x, y, z) and 3 rotation parameters, typically expressed using axis-angle or Euler angles.

The mapping network operates on the temporally enriched features, enabling the model to preserve both spatial context and temporal continuity for accurate and stable relative pose estimation across consecutive frames.

To supervise the end-to-end training, we adopt a Mean Squared Error (MSE) loss between the predicted poses and the ground-truth poses:

$$\mathcal{L} = \frac{1}{6B} \sum_{n=1}^{B} \sum_{i=1}^{6} (y_{i,n} - \hat{y}_{i,n})^2 \quad (6)$$

where B is the batch size, and $y_{i,n}$ and $\hat{y}_{i,n}$ denote the i-th ground-truth and predicted pose components for the n-th sample, respectively. This formulation jointly optimizes translation and rotation, promoting robust and stable motion estimation.

IV. Experiments

The effectiveness of VSTFusion-VO is demonstrated through quantitative and qualitative experiments on the KITTI Odometry benchmark, evaluating accuracy, robustness to dynamic motion, and trajectory consistency. Details of the dataset, evaluation metrics, training protocol, and baseline comparisons are provided below.

A. Dataset

We evaluate our model on the KITTI Odometry benchmark [14], which consists of 22 real-world driving sequences recorded at 10 FPS. Among these, 11 sequences (00–10) include GPS-based ground-truth poses for evaluation. The dataset captures urban and highway scenes with diverse dynamics, including varying speeds up to 90 km/h and sharp turns.

TABLE I: Accuracy Comparison on KITTI Sequences

Metric	Method	01	03	04	05	06	07	10
Translational error (%)	Deep-VO	156.389	73.552	10.803	56.184	64.397	71.790	128.732
	TSformer-VO2	**23.671**	18.344	9.035	9.437	17.101	13.998	14.913
	SWFormer-VO2	24.688	**11.661**	5.655	11.448	**10.308**	11.349	9.267
	VSTFusion-VO	25.112	14.754	**4.836**	**9.386**	10.404	**8.204**	**8.647**
Rotational error (deg/100m)	Deep-VO	10.036	15.671	3.849	29.898	31.395	50.821	41.465
	TSformer-VO2	5.855	11.644	4.874	**3.907**	5.395	7.440	4.381
	SWFormer-VO2	**3.648**	**6.731**	**1.596**	4.501	**2.924**	7.708	**3.043**
	VSTFusion-VO	5.772	9.197	2.550	4.092	3.688	**6.440**	3.446
ATE (m)	Deep-VO	**19.981**	**11.744**	3.850	123.298	107.995	22.831	57.901
	TSformer-VO2	101.699	20.123	6.005	**37.679**	46.788	23.141	23.141
	SWFormer-VO2	82.743	14.834	4.373	50.151	**24.255**	28.124	16.970
	VSTFusion-VO	76.283	20.341	**3.294**	42.306	25.968	**19.529**	**14.331**
RPE (m)	Deep-VO	3.577	0.553	0.261	0.808	1.152	0.741	1.135
	TSformer-VO2	**0.542**	0.128	0.141	0.137	0.181	0.123	0.159
	SWFormer-VO2	0.723	**0.095**	0.104	**0.100**	0.139	**0.092**	0.126
	VSTFusion-VO	0.703	0.101	**0.085**	0.104	**0.133**	0.102	**0.117**
RPE (°)	Deep-VO	0.440	0.438	0.137	0.535	0.476	0.703	0.580
	TSformer-VO2	0.310	0.284	0.174	0.263	0.251	0.282	0.322
	SWFormer-VO2	**0.238**	0.222	**0.125**	**0.197**	0.189	**0.206**	**0.238**
	VSTFusion-VO	0.260	**0.221**	0.129	0.201	**0.179**	0.214	0.241

We evaluated multiple variants of TSformer-VO and SWFormer-VO, and report only TSformer-VO2 and SWFormer-VO2, the strongest in their series, to ensure fair comparison. TSformer-VO was reproduced using official weights and evaluated under the same protocol as SWFormer-VO [25], with results matching the original reports. Only VSTFusion-VO is newly implemented in this work.

In our experiments, we train on sequences 00, 02, 03, and 09, and test on 01, 04, 05, 06, 07, 08, and 10. Pseudo-depth maps are generated using Monodepth2 [7] to enhance geometric perception in monocular settings. All frames are resized to 192×640 before being input to the model. To address scale ambiguity, predicted poses are aligned with ground truth using 7-DoF similarity transformation.

B. Evaluation Metrics

We evaluate the performance of our visual odometry system using the following standard metrics commonly used in the KITTI benchmark:

- **Translational Error** (T_{err}): Measures the average relative translation error between predicted and ground truth poses over the full trajectory:

$$T_{\text{err}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\left\| \mathbf{t}_i^{\text{gt}} - \mathbf{t}_i \right\|_2}{\left\| \mathbf{t}_i^{\text{gt}} \right\|_2} \times 100 \tag{7}$$

  where $\mathbf{t}_i^{\text{gt}}$ and $\mathbf{t}_i$ are the ground truth and predicted translations at frame i.
- **Rotational Error** (R_{err}): Computes the average relative rotation error using the Frobenius norm, normalized by traveled distance:

$$R_{\text{err}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\left\| \mathbf{R}_i^{\text{gt}} - \mathbf{R}_i \right\|_F}{d_i} \times 100 \tag{8}$$

  where $\mathbf{R}_i^{\text{gt}}$ and $\mathbf{R}_i$ are the ground truth and predicted rotation matrices, and d_i is the traveled distance at frame i.
- **Absolute Trajectory Error (ATE)**: Evaluates the global trajectory alignment using RMSE:

$$\text{ATE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{t}_i^{\text{gt}} - \mathbf{t}_i \right\|_2^2} \tag{9}$$

  ATE quantifies the overall discrepancy between the predicted and ground truth trajectories.
- **Relative Pose Error (RPE)**: Assesses local motion consistency between adjacent frames.
  Translation RPE:

$$\text{RPE}_{\text{trans}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| \Delta \mathbf{t}_i^{\text{gt}} - \Delta \mathbf{t}_i \right\|_2 \tag{10}$$

  Rotation RPE:

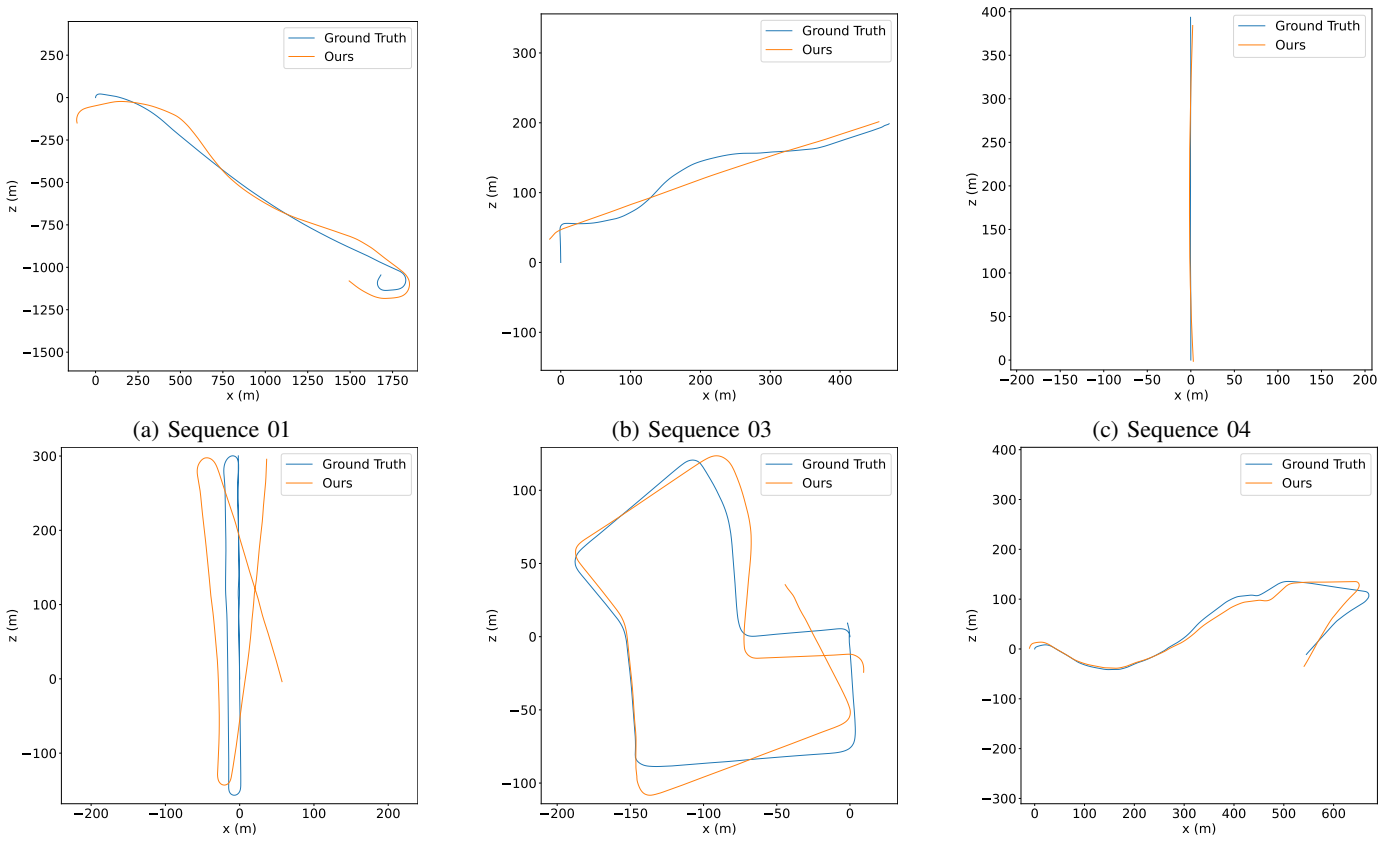
$$\text{RPE}_{\text{rot}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| \mathbf{R}_i^{\text{gt}} - \mathbf{R}_i \right\|_F \tag{11}$$

  where $\Delta \mathbf{t}_i = \mathbf{t}_{i+1} - \mathbf{t}_i$ denotes the frame-to-frame relative translation.

To address the scale ambiguity in monocular VO, a 7-DoF similarity transformation is applied during evaluation to align predicted trajectories with ground truth.

C. Comparison with State-of-the-Art

Table I compares our proposed VSTFusion-VO with several representative visual odometry (VO) methods on the KITTI Odometry benchmark [14], including the recurrent Deep-VO [4], and transformer-based TSformer-VO2 [26] and SWFormer-VO2 [25]. Among the TSformer and SWFormer



(a) Sequence 01 (b) Sequence 03 (c) Sequence 04

(d) Sequence 06 (e) Sequence 07 (f) Sequence 10

Fig. 3: Visual comparison of predicted and ground-truth trajectories on KITTI sequences 01, 03, 04, 06, 07, and 10. These sequences cover diverse scenarios including long-range navigation, sharp turns, and complex urban environments. All trajectories are aligned using a 7-DoF similarity transformation.

series, we report only VO2 variants, as they demonstrate the strongest performance in their respective families.

VSTFusion-VO achieves consistently strong results across five standard metrics: translational error, rotational error, absolute trajectory error (ATE), and relative pose error (RPE) in both translation and rotation. Compared to the transformer-based baselines, our model yields lower ATE and RPE, particularly in challenging sequences involving dynamic motion or trajectory curvature.

Overall, VSTFusion-VO ranks among the top two methods across all metrics, and notably outperforms TSformer-VO2 and SWFormer-VO2 in sequences with sharp turns or rapid motion. While Deep-VO performs competitively in some metrics, it suffers from unstable performance and large translational drift in complex scenes.

As summarized in Table II, VSTFusion-VO achieves a 3.59% reduction in translational error, 8.76% lower ATE, and 2.54% lower translational RPE compared to SWFormer-VO2. These improvements highlight the benefits of integrating early-stage RGB-depth fusion and hierarchical video transformer modeling.

These results confirm that our method effectively balances rotational and translational accuracy, delivering robust performance in dynamic environments and outperforming state-of-the-art learning-based baselines.

TABLE II: Comparison of Average ATE, RPE, and Translational Error between SWFormer-VO2 and VSTFusion-VO.

Metric	**SWFormer-VO2**	**VSTFusion-VO**	**Improvement (%)**
Trans. error (%)	12.053	11.620	↓ 3.59%
ATE (m)	31.636	28.865	↓ 8.76%
RPE (m)	0.197	0.192	↓ 2.54%

Note: ↓ indicates improvement (lower error is better).

D. Component-wise Analysis

Although our method incorporates pseudo-depth, all models operate under the same monocular and RGB-only setting without access to external depth sensors, ensuring a fair comparison. TSformer-VO lacks both depth fusion and video-specific temporal modeling, while SWFormer-VO uses a hierarchical Swin Transformer without depth fusion or temporal attention. VSTFusion-VO integrates pseudo-depth with a video-native backbone, leading to consistent improvements across KITTI sequences as shown in Table I, especially in long-range trajectories.

These results suggest that the main performance gains stem from hierarchical temporal modeling and multimodal integration, rather than access to ground-truth depth.

E. Trajectory Visualization on KITTI Sequences

Figure 3 shows qualitative comparisons between predicted trajectories and ground truth on selected KITTI sequences. All trajectories are aligned using a 7-DoF similarity transformation to remove global scale and orientation mismatches. These visualizations highlight the accuracy of our method under diverse and challenging conditions.

The sequences cover various driving scenarios. Sequence 01 features long-range highway motion, prone to scale drift. Sequences 03 and 04 are mostly linear and serve as consistency references. Sequence 06 includes high-curvature turns, while Sequence 07 captures urban scenes with occlusions and abrupt direction changes. Sequence 10 reflects dense urban layouts with frequent motion discontinuities.

Our method closely matches the ground truth across all sequences. It effectively suppresses drift in Sequence 01 and preserves trajectory curvature and orientation in Sequences 06 and 07. These results demonstrate the robustness of our depth-aware transformer in maintaining trajectory accuracy across different environments.

V. Conclusion and Future Work

We presented a monocular VO framework, VSTFusion-VO, that integrates early-stage RGB–depth fusion with a hierarchical Video Swin Transformer. The proposed model captures spatiotemporal features effectively and enables accurate pose estimation without relying on ground-truth depth. Experiments on the KITTI benchmark show consistent improvements over both traditional and transformer-based VO baselines. Particularly in challenging scenarios involving sharp turns and rapid motion, our model exhibits exceptional stability and trajectory accuracy, validating the synergistic effect of early geometric integration and video-native temporal modeling.

Future work will advance along several promising directions. In addition to exploring lightweight model variants for real-time deployment and integrating semantic or inertial information to enhance robustness, we also plan to investigate the following key areas: model generalization, unsupervised learning frameworks, and optimization of fusion mechanisms.

References

[1] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 2564–2571.

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

[5] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[6] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1983–1992.

[7] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[8] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 2, no. 3, 2021, p. 4.

[9] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6836–6846.

[10] A. O. Françani and M. R. Maximo, "Transformer-based model for monocular visual odometry: a video understanding approach," *IEEE Access*, 2025.

[11] Z. Wu and Y. Zhu, "Swformer-vo: A monocular visual odometry model based on swin transformer," *IEEE Robotics and Automation Letters*, 2024.

[12] X. Wang, F. Ma, H. Zhao, J. Costeira, K. Daniilidis, and R. Urtasun, "Unsupervised learning of upscaled depth via coarse-to-fine matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[13] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[14] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[15] A. Francani and M. Maximo, "Transformer-based model for monocular visual odometry: A video understanding approach," in *IEEE Access*, 2023.

[16] Z. Teed and J. Deng, "Deep patch visual odometry," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: https://openreview.net/forum?id=X0OVY57tG7

[17] O. B. Dufour, A. Mohebbi, and S. Achiche, "An attention-based deep learning architecture for real-time monocular visual odometry: Applications to gps-free drone navigation," *arXiv preprint arXiv:2404.17745*, 2024.

[18] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 7286–7291.

[19] H. Zhan, R. Garg, C. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Depth-vo-feat: Deep depth and visual odometry from sparse and noisy depth inputs with rgb guidance," in *ECCV*, 2018, pp. 696–712.

[20] X. Wang, G. Yang, B. Shi, Y.-W. Tai, and C.-K. Tang, "Dvso: Deep visual slam with online depth fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2510–2520.

[21] Z. Teed and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6644–6653.

[22] Y. Lyu, S. Song, J. Tao, and J. Yi, "Hrvo: Towards high-resolution visual odometry," in *European Conference on Computer Vision*, 2022, pp. 635–653.

[23] B. Su and T. Zang, "A global pose and relative pose fusion network for monocular visual odometry," *IEEE Access*, vol. 12, pp. 106 238–106 251, 2024.

[24] H. Pu, J. Luo, G. Wang, T. Huang, and H. Liu, "Visual slam integration with semantic segmentation and deep learning: A review," *IEEE Sensors Journal*, 2023, available on ResearchGate.

[25] W. Zhang, X. Luo, and Q. Huang, "Swformer-vo: Sliding window transformer for visual odometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

[26] Y. Lin, C. Liu, and Y. Wang, "Tsformer-vo: Transformer-based spatiotemporal attention for visual odometry," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.