2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

# Diff-VIO: A Diffusion Model-Based Pose Optimizer for Visual Inertial Odometry

Wenyuan Qin[1,2], Xiangxi Kong[1], Sizhuo Zhang[1], Hao Xu[1] and Xiwang Dong[1], *Senior Member, IEEE*

***Abstract*— Visual inertial odometry (VIO) serves as a cornerstone of environmental perception and spatial localization, with broad applications in autonomous driving, robotic navigation, and embodied intelligence. Although recent deep learning based VIO methods have achieved impressive accuracy and computational efficiency, most approaches optimize errors within a maximum a posteriori (MAP) framework, often overlooking explicit prior modeling which constrains the upper bounds of achievable performance. To address this challenge, Diff-VIO is introduced, which is a VIO optimization framework grounded in diffusion models. An end-to-end coarse pose generator is first employed. It outputs an initial pose estimate and supplies priors for the diffusion refinement. To constrain the solution space, a diffusion-based refinement module injects pose priors during generation. This process is supported by a global context transformer encoder and a conditional decoder, which model long-range dependencies and predict residual noise for precise pose refinement. Experiments conducted on the KITTI benchmark demonstrate that the proposed method outperforms state-of-the-art VIO techniques in both accuracy and robustness. Additional evaluations on a dataset collected with an Intel RealSense D435i further validate the strong generalization capability of the proposed method across diverse hardware platforms. As the first diffusion-based VIO framework, Diff-VIO introduces a novel optimization paradigm for learning-based visual-inertial odometry systems.**

## I. Introduction

Obtaining reliable motion estimation in unknown environments is essential for a wide range of robotic applications, including autonomous driving [1], augmented reality [2], and unmanned aerial vehicle (UAV) navigation [3]. Visual inertial odometry (VIO) serves as a critical component in these fields. By integrating data from cameras and inertial measurement units (IMUs), VIO enables the estimation of six degrees of freedom (6-DoF) poses, effectively addressing the challenges of perception and localization in environments without access to GPS. Due to its cost-effectiveness and ease of deployment, VIO has attracted considerable research interest [4]–[6].

Most existing VIO methods are predominantly geometry-based, integrating manually designed initialization, feature tracking, corresponding keyframe selection, and joint optimization [7], [8] or filter [9], [10], as depicted in Fig. 1a. These approaches rely on 2D-3D mapping of feature points to establish correspondences and infer coordinate transformations. These correspondences are then fed into a joint nonlinear optimization framework to achieve accurate robot pose estimation. However, due to the intricate and often computationally intensive nature of their optimization processes, these methods typically incur significant time and memory costs. In contrast, some methods adopt an end-to-end framework for pose estimation, with a representative architecture shown in Fig. 1b. These approaches typically fuse data from both modalities at the front-end, followed by a pose regression network that directly estimates the global pose. Such end-to-end methods have gained prominence as a preferred choice, primarily because they eliminate the need for laborious hand-engineered features and delegate the entire optimization process to the neural network.

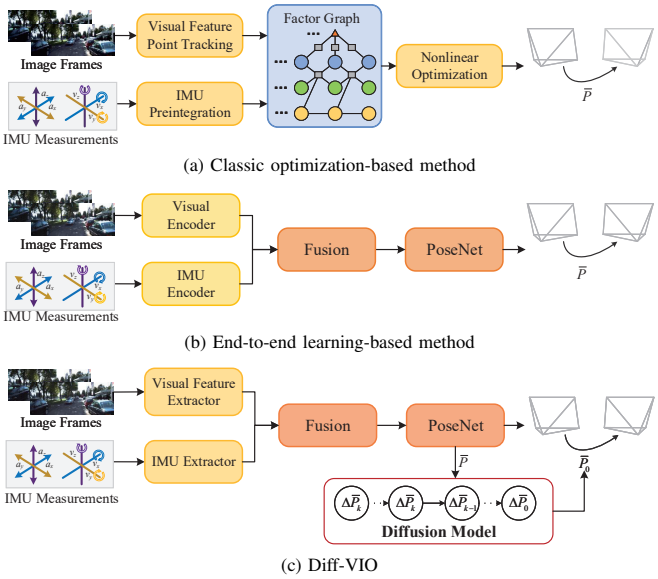



Fig. 1. Frameworks of different VIO methods

Despite recent advances, end-to-end methods continue to face challenges in achieving high localization accuracy [11], [12]. One of the earliest efforts, proposed by [13], introduced an end-to-end framework that directly regresses poses from raw image and IMU inputs. Subsequently, [14] improves accuracy by introducing a decision strategy that quantifies the contribution of visual information throughout the inference process. Further refinement is achieved by [15], who designs a multi-stage fusion strategy to enhance modality interaction by encoding image and IMU data into embedding vectors.

This work was supported by the National Key R&D Plan of China under Grant No. 2024YFB4708300, the National Natural Science Foundation of China under Grants U2241217, 62473027, 62473029, 62403038, 62203032, and 62388101, and the Beijing Natural Science Foundation under Grants JQ23019 and 4232046. (Corresponding author: Hao Xu.)

[1]The authors are with the Science and Technology on Aircraft Control Laboratory, Beihang University, Beijing, 100191 (e-mail: wyqin@buaa.edu.cn; sy2242116@buaa.edu.cn; iszsz@buaa.edu.cn; xuhao3e8@buaa.edu.cn; xwdong@buaa.edu.cn)

[2]Zhongguancun Laboratory, Beijing, 100094, China

Although these methods have demonstrated encouraging results, they often fail to explicitly model the inherent noise present in the raw data. This noise frequently exhibits characteristics of a random walk, making it difficult for conventional supervised learning approaches to capture effectively. Accurately modeling and compensating for such noise remains a significant challenge.

To address these challenges, we propose Diff-VIO, a novel visual-inertial odometry framework based on diffusion models. Unlike existing methods that focus on feature matching or fusion strategies, Diff-VIO formulates pose estimation as a generative task in translation and rotation spaces, allowing the integration of prior pose knowledge for more robust predictions. However, the generative diversity of diffusion models may hinder the consistency required for accurate pose estimation. To overcome this, strong conditional cues are introduced to guide generation. The framework first produces coarse pose estimates and latent features through a preliminary network. These are then used as conditions in a diffusion-based refinement module, which employs an encoder-decoder structure to model global pose distributions and iteratively refine the estimates through denoising.

The main contributions of this paper are summarized as follows:

- Introduction of Diff-VIO, a pioneering VIO system that leverages diffusion models for iterative pose refinement. This work marks the successful application of diffusion models to the VIO task, leading to significantly enhanced localization accuracy.
- To enable precise local-to-global pose refinement, a novel methodology is developed within Diff-VIO. This involves establishing a robustly constrained generative pose model that jointly captures likelihood and prior distributions, meticulously integrating coarse pose guidance features to control generative diversity. Furthermore, it incorporates the design of a global-context transformer encoder and a conditional transformer decoder to facilitate effective and intelligent information exchange between global contextual cues and local pose estimates, leading to highly accurate pose estimation through an iterative denoising process.
- Extensive experiments on the challenging KITTI Odometry benchmark demonstrate the effectiveness of the proposed framework. The proposed method not only surpasses state-of-the-art end-to-end VIO methods but also achieves performance levels that push beyond the typical capabilities of traditional discriminative models.

## II. Related work

### A. Traditional Visual Inertial Odometry

Traditional VIO methods are typically classified into filter-based and optimization-based approaches, depending on their state estimation strategies. Filter-based methods, such as MSCKF [10], estimate states by fusing visual information and inertial measurements through an Extended Kalman Filter (EKF). ROVIO [16] improves localization accuracy by incorporating image intensity errors into the EKF update process. Although these methods offer high speed and low computational overhead, their performance is often limited by the accumulation of estimation errors over time. In contrast, optimization-based approaches generally achieve higher accuracy by leveraging complementary multi-modal information across a sliding temporal window. For instance, OKVINS [17] utilizes a BRISK-based visual front-end combined with a back-end that performs pose optimization within a sliding window. While optimization-based approaches deliver improved accuracy, they often entail increased computational complexity and may exhibit reduced robustness in dynamic or unstructured environments.

### B. Learning-based Visual Inertial Odometry

With the widespread success of deep learning across various domains, learning-based methods have seen rapid advancement in the field of VIO, achieving notable performance gains. Several approaches follow a hybrid paradigm, combining deep learning techniques with traditional VIO pipelines. For example, CodeVIO [18] integrates a learnable depth estimation network with the traditional MSCKF framework to simultaneously perform localization and scene reconstruction. DynaDepth [19] addresses the recovery of absolute pose scale by incorporating visual and inertial data within an EKF framework. More recently, end-to-end learning frameworks have emerged as a dominant direction. Adaptive VIO [20] employs geometric supervision to dynamically update network parameters via online learning, thereby improving adaptability during deployment. Despite achieving high accuracy and real-time performance, many of these learning-based methods remain limited by the noise characteristics inherent in raw sensory data.

### C. Diffusion Model

Diffusion models constitute a prominent class of deep generative models that iteratively reconstruct data samples from random noise through a denoising process. During training, Gaussian noise is progressively added to clean data samples, while the inference phase reverses this process to generate samples that approximate the original data distribution. These models have demonstrated outstanding performance across a wide range of tasks, including 3D point cloud generation [21], [22], point cloud registration [23], human pose estimation [24], and scene flow estimation [25]. For instance, DIT-3D [21] leverages a diffusion-based framework for 3D point cloud generation, while DifFlow3D [25] proposes a scene flow refinement network to capture motion uncertainty. Due to their iterative refinement mechanism, diffusion models are particularly well-suited for modeling the pose refinement process in VIO, offering the potential to mitigate inherent biases in end-to-end VIO frameworks. Nonetheless, their application in this domain remains largely unexplored.

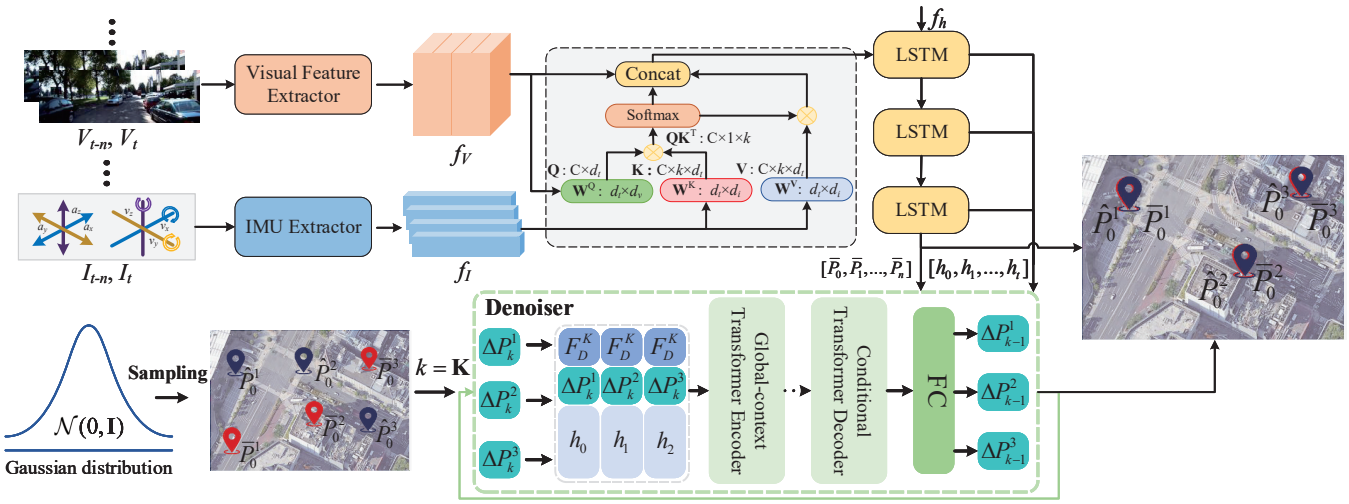



Fig. 2. **Overview of the Proposed Diff-VIO Framework.** The input consists of a $T$-frame image sequence and corresponding IMU measurements, from which modality-specific features are independently extracted. After cross-modal fusion, the combined features are processed by an LSTM for coarse pose estimation. In the refinement stage, noise is added to the initial pose residual $\Delta P_0$ via forward diffusion. A global-context transformer encoder encodes the conditional information, which is then decoded by a conditional transformer to estimate the noise $\bar{Z}$. The refined pose is obtained by adding the predicted residual to the coarse estimate.

## III. DIFF-VIO

This paper introduces a novel two-stage pose refinement framework, termed Diff-VIO. The first stage employs a coarse pose generator to produce an initial pose estimate along with corresponding conditional information. This estimate benefits from adaptive modality fusion enabled by specially designed weighting factors. In the second stage, an iterative denoising strategy based on a conditional diffusion model is applied to mitigate the inherent errors typically observed in end-to-end VIO framework. This module features a novel encoder–decoder architecture that facilitates efficient information flow from global context to local pose. Through iterative learning, the model recovers the deviation distribution between the estimated and ground-truth poses from noisy inputs, thereby effectively compensating for localization errors.

*1) VIO:* Given a sequence of consecutive images $V = \{V_i\}_{i=1}^{2}$ and corresponding IMU measurements $I = \{I_i\}_{i=1}^{2n}$, the goal of VIO is to estimate the pose transformation $P$ between the two frames. This task can be formulated as a probabilistic inference problem, where the objective is to obtain the optimal pose transformation $P^*$ that maximizes the posterior probability conditioned on the observed data.

According to Bayes' theorem, the posterior can be decomposed into the product of the likelihood and the prior. The optimal pose transformation $P^*$ can thus be obtained via Maximum A Posteriori (MAP) estimation:

$$\begin{aligned} P^* &= \arg\max_P p(P \mid [V_1, I_1^n], [V_2, I_2^n]) \\ &= \arg\max_P p([V_1, I_1^n], [V_2, I_2^n] \mid P) \cdot p(P) \\ &= \arg\max_P \left\{ \log p([V_1, I_1^n], [V_2, I_2^n] \mid P) + \log p(P) \right\}, \end{aligned} \tag{1}$$

where $p([V_1, I_1^n], [V_2, I_2^n] \mid P)$ denotes the likelihood term, representing the observational evidence from the image and IMU measurements given the pose transformation $P$. The term $p(P)$ represents the prior, encoding any prior knowledge or assumptions about $P$.

*2) Conditional Diffusion Model:* Diffusion model are a type of generative framework capable of learning the underlying data distribution by progressively transforming noise into structured data. When the model is conditioned on auxiliary information $C$, it is referred to as a *conditional diffusion model*. This model consists of two processes: a forward diffusion process and a reverse denoising process.

In the forward diffusion process, Gaussian noise is incrementally added to an initial data sample $X_0$, resulting in a sequence of noisy data points $\{X_k\}_{k=1}^{K}$ through a Markov chain defined as:

$$q(X_k \mid X_{k-1}) = \mathcal{N}(X_k; \sqrt{1-\beta_k} X_{k-1}, \beta_k \mathbf{I}), \tag{2}$$

where $k \in \{1, \ldots, K\}$, $K$ denotes the number of diffusion steps, $\beta_k \in [0, 1)$ refers to the hyperparameters, and $\mathbf{I}$ represents the identity matrix. The notation $\mathcal{N}(x; \mu, \sigma)$ denotes a Gaussian distribution with mean $\mu$ and covariance $\sigma$ [26]. This forward process also admits a closed-form expression:

$$X_k = \sqrt{\alpha_k} X_0 + \sqrt{1-\alpha_k} Z, \quad Z \sim \mathcal{N}(0, \mathbf{I}), \tag{3}$$

where $\alpha_k = \prod_{i=1}^{k} (1-\beta_i)$.

The reverse process involves training a neural network $\mathcal{F}_\theta(X_k; k; C)$ to denoise the noisy input $X_k$ and estimate the original sample $X_0$, denoted as $\hat{X}_{k\to 0}$. This is accomplished by minimizing the negative log-likelihood during the training process:

$$\begin{aligned} \mathbb{E}_{X_0 \sim p_{\text{data}}} [\log \mathcal{F}_\theta(X_0; k; C)] \geq \mathbb{E}_q \Bigg[ & \log \mathcal{F}_\theta(X_0; k; X_1; C) \\ & - \sum_{k>1} D_{\text{KL}} \left( q(X_{k-1} \mid X_k, X_0) \,\|\, \mathcal{F}_\theta(X_{k-1}; k; X_k; C) \right) \Bigg], \end{aligned} \tag{4}$$

where $p_{\text{data}}$ is the distribution of the training data. According to Bayes' rule, the reverse denoising step at each timestep $k$

can be expressed as:

$$X_{k-1} = \sqrt{\alpha_{k-1}}\mathscr{F}_\theta(X_k;k;C) + \frac{\sqrt{1-\alpha_{k-1}-\sigma_k^2}}{\sqrt{1-\alpha_k}} \\ (X_k - \sqrt{\alpha_k}\mathscr{F}_\theta(X_k;k;C)) + \sigma_k Z, \tag{5}$$

where $\sigma_t$ denotes the variance at diffusion step $t$.

### A. Coarse Pose Generator

*1) Feature Extractor:* To perform pose estimation, a coarse pose generator is constructed, adopting an architectural design similar to that proposed in [14]. The model receives as input a video sequence $\{V_i\}_{i=1}^{T}$ and corresponding IMU measurements $\{I_i\}_{i=1}^{nT}$, where $n$ represents the sampling rate ratio of the IMU relative to the video frame rate, and $T$ is the number of input frames. Each image $V_i \in \mathbb{R}^{3\times H\times W}$, with $H$ and $W$ denoting the image height and width, respectively. Each IMU measurement $I_i \in \mathbb{R}^6$ comprises linear acceleration and angular velocity along the three spatial axes. The network produces a sequence of estimated relative poses $\{P_i\}_{i=1}^{T-1}$ as output.

Following most end-to-end VIO pipelines [15], the proposed coarse pose generator first processes data from the visual and inertial modalities independently through a visual feature extractor $E_\text{v}$ and an inertial feature extractor $E_\text{i}$, respectively. Since the first frame serves as the reference pose, the input video sequence of length $T$ is divided into overlapping frame pairs for relative pose estimation. Correspondingly, IMU data between each pair of visual frames is jointly encoded. Thus, for each time step $t$, the visual and inertial features are computed as $x_t^v = E_\text{v}(V_{t\to t+1})$ and $x_t^i = E_\text{i}(I_{t\to t+1})$, where $x_t^v \in \mathbb{R}^{d_v}$ and $x_t^i \in \mathbb{R}^{d_i}$ represent the visual and inertial features at time step $t$, respectively. These features are then stacked to form the sequences $X^v \in \mathbb{R}^{(T-1)\times d_v}$ and $X^i \in \mathbb{R}^{(T-1)\times d_i}$.

*2) Cross-modal Fusion:* Existing VIO methods often employ simple concatenation or basic fusion strategies, which fail to model the intricate dependencies between visual and inertial features. As noted in [14], such approaches are sensitive to transient noise in individual modalities, resulting in degraded localization performance. To address this limitation, we propose a cross-modal attention mechanism that adaptively emphasizes salient information for improved odometry accuracy.

As illustrated in Fig. 2, the attention module treats the visual feature $f_v$ as the *Query*, and the inertial feature $f_I$ as both the *Key* and *Value*. This design enables the visual modality to selectively attend to and refine the inertial representation. Attention weights are computed via linear projections and dot-product operations, and used to modulate $f_I$, effectively capturing fine-grained inter-modal correlations. The refined inertial features are then concatenated with $f_v$ to form a unified representation $f_F$, which is passed to an LSTM for relative pose estimation

### B. Diffusion Refinement Module

To improve pose estimation accuracy, we employ a conditional diffusion model [27] to learn the prior distribution of pose residuals and iteratively refine coarse predictions. Given the coarse pose estimate $\bar{P}$ and the ground truth pose $\hat{P}$, the initial pose residual $\Delta P_0 = \hat{P}\bar{P}^{-1}$.

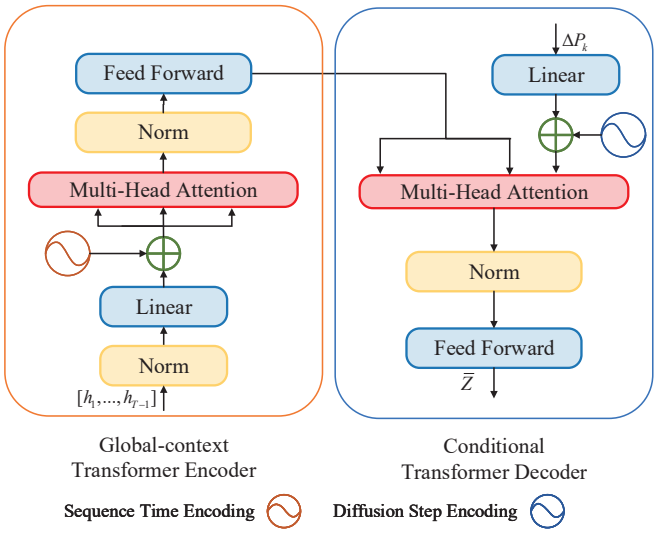



Fig. 3. Architecture of the diffusion refinement module

The forward diffusion process corrupts the clean residual $\Delta P_0$ with Gaussian noise, producing a noisy sequence $\Delta P_k$ over diffusion steps $k$:

$$\Delta P_k = \sqrt{\alpha_k}\Delta P_0 + \sqrt{1-\alpha_k}Z, \quad Z \sim \mathscr{N}(0,\mathbf{I}), \tag{6}$$

where $\alpha_k = \prod_{i=1}^{k}(1-\beta_i)$ and $Z$ is standard Gaussian noise.

The reverse denoising process is parameterized by a neural network $\mathscr{F}_\theta$, which estimates the clean residual from the noisy input as:

$$\Delta P_{k-1} = \sqrt{\alpha_{k-1}}\mathscr{F}_\theta(\Delta P_k;k;C) + \frac{\sqrt{1-\alpha_{k-1}-\sigma_k^2}}{\sqrt{1-\alpha_k}} \\ (\Delta P_k - \sqrt{\alpha_k}\mathscr{F}_\theta(\Delta P_k;k;C)) + \sigma_k Z, \tag{7}$$

where $C$ denotes global contextual conditioning and $\sigma_k$ is the variance at step $k$.

To represent the pose within a Gaussian space amenable to diffusion processes, normalization of its components is essential. The translation component is normalized through standard scaling, whereas the rotation is parameterized using rotation vectors. Euler angles are excluded due to their susceptibility to singularities and the pronounced nonlinearities inherent in spatial rotation representations, both of which can adversely affect optimization convergence. Although quaternions are extensively employed in robotic applications, their unit-norm constraint complicates the modeling of rotational errors. Moreover, the intrinsic requirement of diffusion models for unconstrained iterative sampling in Euclidean space imposes further limitations on the choice of rotation representation. Rotation vectors, encoding rotations as axis-angle parameters where the vector magnitude corresponds to the rotation angle, inherently support additive residual modeling. This property renders rotation vectors particularly suitable for pose residual optimization within the diffusion model framework.

*1) Global-Context Transformer Encoder:* We leverage the hidden features $h_t$, generated by the first-stage pose regression network, as conditional inputs to the diffusion-based refinement module. These features, obtained through supervised learning, encode rich cross-modal representations. However, due to the unidirectional nature of standard pose regression architectures, they lack global temporal interactions, which are crucial for capturing comprehensive motion dynamics.

To overcome this limitation, we introduce a Global-Context Transformer Encoder designed to model temporal dependencies across the entire sequence. Specifically, hidden features from multiple time steps are aggregated as input, and fixed positional encodings are incorporated to guide the attention mechanism in learning temporal relationships along the trajectory. The fixed positional encoding, following the formulation in [28]. We adopt fixed positional encoding since the relative temporal order is preserved during training, providing consistent structural guidance for the encoder to effectively capture sequential dependencies in pose representations.

*2) Conditional Transformer Decoder:* To refine the pose estimates, we propose a Conditional Transformer Decoder that accepts a noisy pose residual and global contextual features as input, and predicts the corresponding noise $\bar{Z}$ at a given diffusion timestep $k$.

Within this decoder, we employ a learnable temporal encoding strategy to capture fine-grained temporal cues, enhancing both the denoising performance and the trajectory consistency of the generated poses. Specifically, for each diffusion step $k$, a learnable temporal embedding is produced via a *TimeEmbedding* module, implemented as a multi-layer perceptron (MLP). Unlike conventional sinusoidal encodings, this MLP-based formulation enables the model to generate trainable, nonlinear, and expressive embeddings that better reflect the noise scale and statistical properties of the current diffusion stage. This adaptability allows the decoder to dynamically adjust its denoising behavior throughout the pose refinement process.

The learnable temporal embedding is computed as:

$$k_{\text{emb}} = \text{MLP}\left(\text{SiLU}\left(\text{MLP}(k/K)\right)\right), \tag{8}$$

where $k$ is the current diffusion timestep, $K$ is the total number of steps, and SiLU denotes the sigmoid-weighted linear unit activation. The resulting embedding $k_{\text{emb}}$ is subsequently injected into the Transformer decoder, modulating the prediction process based on the temporal noise characteristics associated with each diffusion stage.

### C. Loss Functions

*1) Coarse Pose Estimation Loss:* During the coarse pose estimation stage, we supervise the network using a standard mean squared error (MSE) loss, which jointly penalizes both translation and rotation errors. The loss is defined as:

$$\mathscr{L}_{\text{pose}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(\|\hat{p}_t - p_t\|_2^2 + \alpha\left\|\hat{\phi}_t - \phi_t\right\|_2^2\right), \tag{9}$$

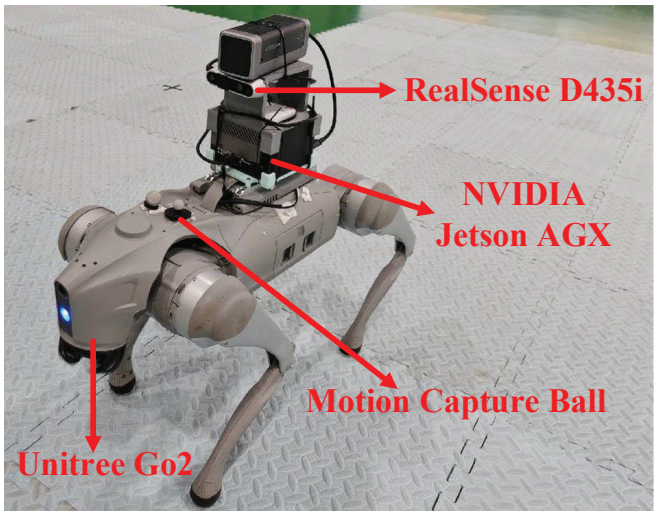



Fig. 4. Data collection platform

where $T$ is the input sequence length, $p_t$ and $\phi_t$ denote the ground truth translation and rotation at time step $t$, and $\hat{p}_t$, $\hat{\phi}_t$ are their corresponding predictions. The weighting factor $\alpha$ balances the translation and rotation terms; following prior work [14], [15], we set $\alpha = 100$ to ensure appropriate scaling between modalities.

*2) Diffusion Denoising Loss:* In the diffusion-based refinement stage, we employ a standard noise prediction objective commonly used in denoising diffusion models. At each diffusion step $k \in \{1,\ldots,K\}$, Gaussian noise $Z \sim \mathcal{N}(0,\mathbf{I})$ is added to the clean pose residual $\Delta P_0$, producing a noisy state $\Delta P_k$. The goal of the denoising network $\mathscr{F}_\theta$ is to predict the added noise $Z$, conditioned on $\Delta P_k$, timestep $k$, and a latent context representation $C$. The diffusion loss is thus formulated as:

$$\mathscr{L}_{\text{diffusion}} = \mathbb{E}_{\Delta P_0,k,Z}\left[\|\mathscr{F}_\theta(\Delta P_k;k;C) - Z\|^2\right], \tag{10}$$

where $\mathscr{F}_\theta(\Delta P_k;k;C)$ denotes the network's predicted noise.

By jointly optimizing both the coarse pose estimation loss and the diffusion denoising loss, our framework progressively refines initial pose estimates into accurate, fine-grained predictions.

## IV. EXPERIMENTS

### A. Dataset and Implementation Details

*1) Dataset:* We evaluate the proposed framework on the KITTI Odometry dataset [35], a widely adopted benchmark in large-scale autonomous driving scenarios. This dataset comprises 22 driving sequences collected across diverse environments, including urban, rural, and highway scenes. RGB images and ground-truth poses are recorded at 10 Hz, while IMU measurements are sampled at 100 Hz. Additionally, the dataset provides synchronized RGB, LiDAR, and inertial data streams. Following the experimental protocol in [14], we use sequences {00, 01, 02, 04, 06, 08, 09} for training, and sequences {05, 07, 10} for testing. Sequence 03 is omitted due to the absence of raw IMU data, and sequences {11–22} are excluded as they lack ground-truth trajectories. Given the imperfect

TABLE I

ALGORITHM PERFORMANCE COMPARISON

| Method | Mode | Seq.05 | | Seq.07 | | Seq.10 | | Ave | |
|---|---|---|---|---|---|---|---|---|---|
| | | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ |
| ORB-SLAM2 [29] | Geo. | **9.12** | **0.2** | 10.34 | **0.3** | **4.04** | **0.3** | **7.83** | **0.27** |
| VINS-Mono [7] | Geo. | 11.6 | 1.26 | **10.0** | 1.72 | 16.5 | 2.34 | 12.7 | 1.77 |
| VIOLeander [30] | Self-Sup. | 3.00 | **1.40** | 3.60 | 2.06 | 2.04 | 1.37 | 2.88 | **1.61** |
| DeepVIO [31] | Self-Sup. | **2.86** | 2.32 | **2.71** | **1.66** | **0.85** | **1.03** | **2.14** | 1.67 |
| Hard Fusion-VIO [32] | Sup. | 4.25 | 1.67 | 4.46 | 2.17 | 5.81 | 1.55 | 4.84 | 1.80 |
| VS-VIO [14] | Sup. | 2.61 | 1.06 | 1.83 | 1.35 | 3.11 | 1.12 | 2.52 | 1.18 |
| Gravity-Shift-VIO [33] | Sup. | 2.98 | 1.40 | 2.76 | 1.87 | 4.45 | 1.21 | 3.39 | 1.49 |
| CAP-VIO [34] | Sup. | 4.76 | 1.39 | 5.18 | 2.55 | 6.55 | 1.83 | 5.49 | 1.92 |
| CMIF-VIO [15] | Sup. | 2.65 | 0.95 | 1.76 | 0.98 | 3.45 | 1.04 | 2.62 | 0.99 |
| Ours | Sup. | 2.84 | 1.09 | 1.82 | 0.99 | 2.82 | 0.84 | 2.49 | 0.97 |
| Ours | Sup.&Diff | **2.28** | **0.76** | **1.63** | **0.58** | **2.68** | **0.59** | **2.20** | **0.64** |
| Delta | | 19.7% | 30.3% | 10.4% | 41.4% | 5% | 29.8% | 11.6% | 34% |
| Ours | Sup. | 2.49 | 0.90 | 1.77 | 0.84 | 3.01 | 0.82 | 2.42 | 0.85 |
| Ours | Sup.&Diff | **2.46** | **0.83** | **1.66** | **0.54** | **2.64** | **0.63** | **2.25** | **0.67** |
| Delta | | 1.2% | 7.8% | 6.2% | 35.7% | 12.3% | 23.2% | 7% | 21.2% |

Green indicates that the activation function in the coarse pose module is Softmax, while pink indicates Sigmoid.

synchronization between image and IMU streams, we apply interpolation to align the raw IMU data temporally with the image timestamps.

To further assess the generalization capability of our model to different platforms, we conduct additional experiments on a quadruped robot equipped with an Intel RealSense D435i camera. Ground-truth poses are acquired via a high-precision motion capture system. The camera captures RGB images at 30 Hz and IMU data at 200 Hz, as illustrated in Fig. 4.

*2) Implement Details:* Our implementation is based on PyTorch 2.4.0 and runs on an NVIDIA RTX 3090 GPU with an Intel(R) Core(TM) i9-12900KS CPU. During training, all input images are resized to $512 \times 256$. To ensure fair comparison with existing methods, we adopt the same sequence length $T = 11$ used in prior work, where each image pair corresponds to 11 IMU measurements. Both the visual and inertial feature extractors are configured with an output dimensionality of 512, and the batch size is set to 32. The latent dimensionality of the diffusion model is set to $D = 64$. Both the coarse pose generator and the diffusion-based refinement module are trained for 30 epochs. To stabilize training, we adopt a staged learning rate schedule, with learning rates of $5 \times 10^{-4}$, $5 \times 10^{-5}$, and $1 \times 10^{-6}$ applied every 10 epochs. For the diffusion-based pose refinement stage, we set the number of diffusion steps to $K = 50$.

To enhance the robustness of visual feature extraction, a pre-trained FlowNet is employed as the visual encoder. Trained on the FlyingChairs dataset [36], it effectively captures fine-grained pixel-level motion between image pairs, facilitating downstream pose estimation.

*3) Metrics:* To assess localization accuracy, we adopt the relative translation error $t_{\text{rel}}$ and relative rotation error $r_{\text{rel}}$ computed over trajectory segments of varying lengths, as commonly used in the KITTI benchmark [35]. The translation error $t_{\text{rel}}$ is reported as a percentage (%), while the rotation error $r_{\text{rel}}$ is expressed in degrees per 100 meters (°/100 m).

### B. Performance Evaluation

As summarized in Table I, the proposed method outperforms several state-of-the-art baselines. Compared to the classic geometry-based approach [7], our model achieves an 82.7% improvement in average translation accuracy and a 56.6% improvement in rotation accuracy, demonstrating superior pose estimation capability. Although the overall accuracy is slightly lower than that of the self-supervised method [31], this discrepancy primarily stems from the performance gap in Sequence 10. Across the other two sequences, our method exhibits a clear advantage.

In comparison to the state-of-the-art supervised approach [15], our model improves average translation accuracy by 16% and rotation accuracy by 35.4%, further validating the effectiveness of our algorithm. We also report the results of each stage in our two-stage pipeline. With the integration of the diffusion-based refinement module, the average translation accuracy improves by 11.6%, and the rotation accuracy increases by 34%, indicating that the proposed refinement mechanism significantly boosts pose estimation performance.

TABLE II

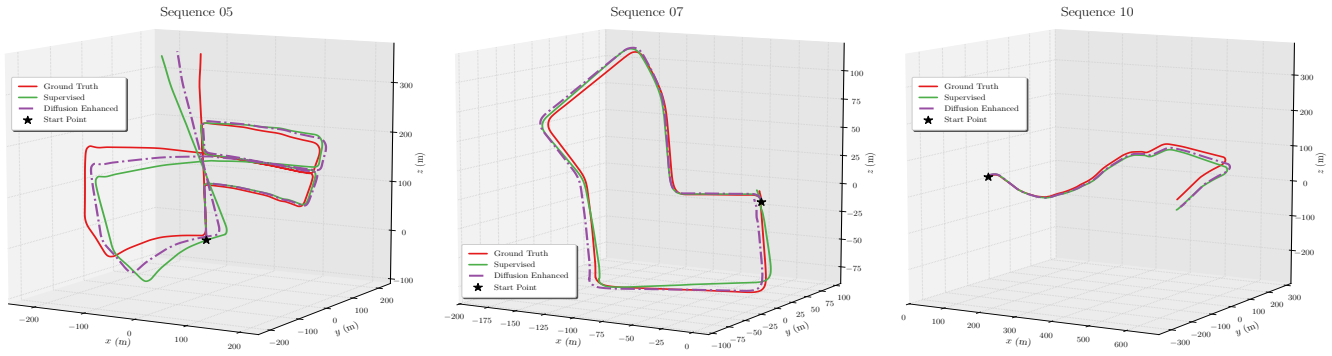
PERFORMANCE COMPARISON OF DIFFERENT FUSION STRATEGIES

| Method | Seq.05 | | Seq.07 | | Seq.10 | |
|---|---|---|---|---|---|---|
| | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ |
| Soft | 3.13 | 1.19 | 2.73 | 1.11 | 2.99 | 0.67 |
| Hard | 3.34 | 1.35 | 2.23 | 1.34 | 2.39 | 0.97 |
| Cat | 3.36 | 1.24 | 2.17 | 1.01 | **2.17** | **0.67** |
| Atte. | 2.84 | 1.09 | 1.82 | 0.99 | 2.82 | 0.84 |
| Sigm. | **2.49** | **0.90** | **1.77** | **0.84** | 3.01 | 0.82 |

Cat denotes simple feature concatenation; Soft applies a linear transformation after concatenation; Hard employs Gumbel-Softmax to adaptively weight each feature dimension; Atte. represents the proposed attention-based fusion strategy. Sigm. refers to sigmoid function.

### C. Ablation Study

*Effect of Fusion Strategy:* To validate the effectiveness of the proposed fusion strategy, we evaluate the model performance under different fusion schemes, as shown in Table II. Compared with the simple concatenation baseline, our method achieves significantly higher accuracy on Sequences 5 and 7. Although a slight accuracy drop is observed on Sequence 10, we attribute this to the relatively simple motion trajectory in that sequence, where our fusion strategy may introduce minor information loss. In such cases,




(a) Sequence 05 (b) Sequence 07 (c) Sequence 10

Fig. 5. Visualization of KITTI testing sequences

TABLE III

PERFORMANCE OF THE DIFFUSION REFINEMENT MODULE UNDER DIFFERENT CONDITION COMBINATIONS

| Glo. | Cross. | C | Seg.05 | | Seg.07 | | Seg.10 | | Ave | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ | $t_{rel}\downarrow$ | $r_{rel}\downarrow$ |
| ✗ | ✗ | ✗ | 2.39 | 0.85 | 1.73 | 0.63 | 3.26 | 0.75 | 2.46 | 0.74 |
| ✗ | ✗ | ✓ | 2.38 | 0.85 | 1.75 | 0.62 | 3.20 | 0.72 | 2.44 | 0.73 |
| ✗ | ✓ | ✓ | 2.36 | 0.82 | 1.64 | 0.59 | 2.88 | 0.63 | 2.29 | 0.68 |
| ✓ | ✗ | ✓ | 2.66 | 0.83 | 1.71 | 0.71 | 2.82 | 0.74 | 2.40 | 0.76 |
| ✓ | ✓ | ✓ | **2.28** | **0.76** | **1.63** | **0.58** | **2.68** | **0.59** | **2.20** | **0.64** |

Glo. denotes the global context transformer encoder, Cross. represents the cross-attention mechanism, and C refers to the hidden features.

TABLE IV

ALGORITHM RUNNING TIME ANALYSIS.

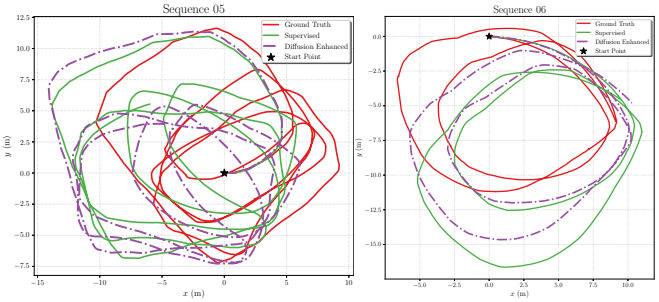
| | Data | Coarse | Diffusion |
|---|---|---|---|
| Time | 145∼150ms | 4∼6ms | 114∼118ms |

Data refers to the data loading process, Coarse denotes the Coarse Pose Generator, and Diffusion corresponds to the Diffusion Refinement Module.

the naive concatenation approach may preserve more relevant features, thereby demonstrating a marginal advantage.

*Effect of Transformer Structure:* To evaluate the contribution of each component, ablation studies are conducted and summarized in Table III. The complete model achieves the highest overall performance. Compared to the diffusion-only baseline, the full model improves translation and rotation accuracy by 10.5% and 14.1%, respectively. Removing the cross-attention mechanism leads to performance drops of 8.3% in translation and 15.8% in rotation. Similarly, excluding the global context module results in declines of 3.9% and 5.8% in translation and rotation accuracy, respectively.

*Runtime Analysis:* Real-time performance is a critical consideration for VIO systems. To evaluate the computational efficiency of the proposed method, we report the runtime breakdown of each module in Table IV. The data loading stage in our pipeline takes approximately 145∼150 ms, while the Coarse Pose Generator operates within 4∼6 ms. The diffusion model, due to its inherently iterative nature, incurs higher computational overhead. Specifically, with the number of diffusion steps set to $K = 50$, the refinement module requires approximately 114∼118 ms. Although this introduces additional latency, the overall runtime remains acceptable for scenarios that permit near real-time performance.




(a) Sequence 05 (b) Sequence 06

Fig. 6. Visualization of the practical testing sequences.

### D. Practical Experiment

To evaluate the performance of the algorithm on different platforms, six ROS bags were collected, with Sequences 01–04 used for training and Sequences 05–06 for testing. As shown in Fig. 6, the proposed method demonstrates strong generalization to real-world environments, although certain deviations are observed

Following data analysis, two primary factors are identified as the main contributors to the observed deviations. First, the Intel RealSense D435i employs a consumer-grade IMU, which introduces substantial measurement noise. Second, mechanical vibrations from the quadruped robot during locomotion further degrade performance. In particular, deviations along the gravity axis are primarily caused by the limited vertical variation in the dataset, which consists mainly of planar trajectories. The absence of elevation changes such as slopes or stairs significantly constrains the accuracy of vertical pose estimation.

Nevertheless, the method achieves high accuracy and real-time performance in planar motion scenarios, demonstrating its potential as a foundational module for embodied intelligence in robotic systems.

## V. CONCLUSION AND FUTURE WORK

This paper proposed Diff-VIO, a diffusion model-based pose optimization framework for visual-inertial odometry. The method began with a coarse pose estimator that

generated initial pose estimates along with conditional features. These features were subsequently processed by a global context transformer encoder and a conditional transformer decoder, which learned the distribution of pose residuals to enable refined predictions under prior constraints. Experimental results demonstrate that Diff-VIO achieves high localization accuracy while maintaining real-time performance, underscoring its potential to enhance the spatial memory capabilities of embodied intelligence. Future research will aim to expand the diversity of training data across different robotic platforms to further improve generalization and facilitate broader deployment in intelligent systems.

## REFERENCES


[1] J. Wu, X. Cheng, F. Liu, and X. Tang, "LiDAR-aided object visual-inertial odometry using anchored residual in dynamic scene," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 7, pp. 10146–10159, 2025.

[2] Y. Dai, Y. Lin, X. Lin, C. Wen, L. Xu, H. Yi, S. Shen, Y. Ma, and C. Wang, "SLOPER4D: A scene-aware dataset for global 4D human pose estimation in urban environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 682–692.

[3] W. Wei, Y. Zhou, Y. Hu, Z. Li, S. Wang, X. Liu, and J. Li, "BotVIO: A lightweight transformer-based visual-inertial odometry for robotics," *IEEE Trans. Robot.*, vol. 41, pp. 3760–3778, 2025.

[4] S. Song, H. Lim, A. J. Lee, and H. Myung, "DynaVINS: A visual-inertial SLAM for dynamic environments," *IEEE Rob. Autom. Lett.*, vol. 7, no. 4, pp. 11523–11530, 2022.

[5] J. Liu, X. Li, Y. Liu, and H. Chen, "RGB-D inertial odometry for a resource-restricted robot in dynamic environments," *IEEE Rob. Autom. Lett.*, vol. 7, no. 4, pp. 9573–9580, 2022.

[6] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, 2021.

[7] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, 2018.

[8] L. Von Stumberg, V. Usenko, and D. Cremers, "Direct sparse visual-inertial odometry using dynamic marginalization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 2510–2517.

[9] M. Li and A. I. Mourikis, "High-precision, consistent EKF-Based visual-inertial odometry," *Int. J. Robot. Res.*, vol. 32, no. 6, p. 690–711, 2013.

[10] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2007, pp. 3565–3572.

[11] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "VINet: Visual-inertial odometry as a sequence-to-sequence learning problem," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4995–5001.

[12] E. J. Shamwell, K. Lindgren, S. Leung, and W. D. Nothwang, "Unsupervised deep visual-inertial odometry with online error correction for RGB-D imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2478–2493, 2020.

[13] C. Chen, S. Rosa, Y. Miao, C. X. Lu, W. Wu, A. Markham, and N. Trigoni, "Selective sensor fusion for neural visual-inertial odometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 10534–10543.

[14] M. Yang, Y. Chen, and H.-S. Kim, "Efficient deep visual and inertial odometry with adaptive visual modality selection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 233–250.

[15] Z. Wang, Y. Zhang, X. Xu, M. Xiong, X. Su, and F. Meng, "CMIF-VIO: A novel cross modal interaction framework for visual inertial odometry," *IEEE Rob. Autom. Lett.*, vol. 10, no. 2, pp. 875–882, 2025.

[16] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-Based approach," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2015, pp. 298–304.

[17] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[18] X. Zuo, N. Merrill, W. Li, Y. Liu, M. Pollefeys, and G. Huang, "CodeVIO: Visual-inertial odometry with learned optimizable dense depth," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2021, pp. 14382–14388.

[19] S. Zhang, J. Zhang, and D. Tao, "Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating IMU motion dynamics," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 143–160.

[20] Y. Pan, W. Zhou, Y. Cao, and H. Zha, "Adaptive VIO: Deep visual-inertial odometry with online continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 18019–18028.

[21] S. Mo, E. Xie, R. Chu, L. Yao, L. Hong, M. Nießner, and Z. Li, "DiT-3D: Exploring plain diffusion transformers for 3D shape generation," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 67960–67971.

[22] Z. Wu, Y. Wang, M. Feng, H. Xie, and A. Mian, "Sketch and text guided diffusion model for colored point cloud generation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 8895–8905.

[23] W. Li, Y. Yang, S. Yu, G. Hu, C. Wen, M. Cheng, and C. Wang, "DiffLoc: Diffusion model for outdoor LiDAR localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 15045–15054.

[24] W. Shan, Z. Liu, X. Zhang, Z. Wang, K. Han, S. Wang, S. Ma, and W. Gao, "Diffusion-based 3D human pose estimation with multi-hypothesis aggregation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 14715–14725.

[25] J. Liu, G. Wang, W. Ye, C. Jiang, J. Han, Z. Liu, G. Zhang, D. Du, and H. Wang, "DifFlow3D: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 15109–15119.

[26] S. Lu, G. Zhuo, L. Zheng, J. Zhu, and J. Bai, "A generative hierarchical optimization framework for LiDAR odometry using conditional diffusion models," *IEEE Sens. J.*, pp. 1–1, 2025.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 574–585.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6000–6010.

[29] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, 2017.

[30] E. J. Shamwell, S. Leung, and W. D. Nothwang, "Vision-aided absolute trajectory estimation using an unsupervised deep network with online error correction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2018, pp. 2524–2531.

[31] L. Han, Y. Lin, G. Du, and S. Lian, "DeepVIO: Self-supervised deep learning of monocular visual inertial odometry using 3D geometric constraints," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2019, pp. 6906–6913.

[32] C. Chen, S. Rosa, C. X. Lu, B. Wang, N. Trigoni, and A. Markham, "Learning selective sensor fusion for state estimation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 3, pp. 4103–4117, 2025.

[33] J. Chen, S. Zhang, Z. Li, and X. Jin, "Gravity-Shift-VIO: Adaptive acceleration shift and multi-modal fusion with transformer in visual-inertial odometry," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2023, pp. 1–8.

[34] I. Pajula, N. Joswig, A. Morrison, N. Sokolova, and L. Ruotsalainen, "A novel cross-attention-based pedestrian visual–inertial odometry with analyses demonstrating challenges in dense optical flow," *IEEE J. Indoor Seamless Position. Navig.*, vol. 2, pp. 25–35, 2024.

[35] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 3354–3361.

[36] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2758–2766.