Analogy-Augmented Uncertainty-aware Monocular Visual Odometry

Jituo Li, *Member, IEEE,* Shunwang Sun, Tingxi Xue, Xinqi Liu, Jialu Zhang, Huixu Dong, *Member, IEEE,* Guodong Lu, *Member, IEEE*

***Abstract*—Visual odometry (VO) is a critical component of autonomous robot systems, enabling precise pose estimation from visual inputs. Learning-based VO methods are increasingly recognized for their robustness in challenging scenarios, including dynamic environments, motion blur, and low-light conditions. However, their performance is constrained by both the diversity of the data and its utilization rate. To overcome these limitations, we propose an end-to-end monocular VO system incorporating a novel learning-based end-to-end VO framework and multiple analogy augmentation strategies. We introduce the Context Attention Uncertainty-aware VO Network (CUVO), which prioritizes semantically rich regions and mitigating interference from high-uncertainty areas to enhance attentional focus and pose estimation accuracy. Furthermore, our analogy augmentation methods—temporal reversal, random rotation, and geometric mirroring—enhance image pairs and compute corresponding true pose transformations, significantly increasing training data quantity and diversity. Simultaneously, an analogous loss is applied to ensure consistency between the original and augmented data. Extensive experiments demonstrate that CUVO significantly enhances VO performance, outperforming previous end-to-end VO methods on TartanAir and KITTI datasets. By leveraging analogy augmentation strategy to expand training data under limited data conditions (27k), zero-shot capability of CUVO degrades by up to 29.5% on TartanAir and 23.3% on KITTI. Our work introduces the first image-to-pose data augmentation method tailored for VO and establishes CUVO as a robust system for advancing learning-based visual odometry.**

***Index Terms*—Visual Odometry, Analogy Augmentation, Pose Estimation, End-to-End.**

I. Introduction

VISUAL SLAM is a critical component of autonomous robotic systems, with visual odometry (VO) serving as a core subtask that typically functions independently of the SLAM backend [1]. VO is categorized into geometric [1]–[3] and learning-based approaches. Classical geometric methods, such as ORB-SLAM3 [47], achieve remarkable robustness through multi-frame bundle adjustment and loop closure, but suffer from heavy dependence on hand-crafted features and degraded performance in textureless or dynamic settings.

This work was supported in part by the Pioneer & Leading Goose R&D Program of Zhejiang Province, China under Grant 2023C01067, and in part by the State Key Laboratory of Digital-Intelligent Modeling and Simulation, and Funding of Zhejiang University Robotics Institute. (Corresponding author: Jituo Li.)

Jituo Li, Shunwang Sun, Jialu Zhang, Huixu Dong and Guodong Lu are with State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University and with Zhejiang Key Laboratory of Industrial Big Data and Robot Intelligent Systems. They are also with the School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China, and with Robotics Institute, Zhejiang University, Hangzhou 310027, China (e-mail: jituo_li@zju.edu.cn).

Tingxi Xue is with the College of Engineering, Zhejiang University, Hangzhou 310027, China.

Xinqi Liu is with the School of Artificial Intelligence and Robotics, Hunan University, Changsha, 410012, China.

Recent studies demonstrate that learning-based methods offer superior robustness over geometric methods in challenging scenarios, including moving objects, motion blur, and low-light conditions. As a result, increasing research focuses on leveraging deep learning networks for pose estimation [6], [8], [9]. The performance of learning-based VO depends on two key factors: one is the model architecture and training settings, which define the upper performance limit, the other is the quality and diversity of training data, which establish the baseline performance. Model architectures primarily include RNNs [11], CNNs [9], and RCNNs [12], [13]. Training data is mainly made up of real sensor data [14] and synthetic data generated by virtual simulation.

Improving VO performance primarily focuses on enhancing model architectures to better utilize inter-frame information for pose estimation. To effectively leverage this information, models prioritize semantically rich regions. Currently, SimVODIS++ [16], building upon the original SimVODIS [17], proposes using CBAM module, enabling the model to focus on salient regions and filter dynamic objects. However, since it is trained and tested solely on KITTI [18], it lacks reliable generalization capability. TartanVO [21] innovatively introduces flow and camera intrinsics as intermediate quantities for pose estimation. Prioritizing pixel motion as the primary input, this method overlooks original image features and texture, consequently limiting its semantic understanding. For the widely used DROID-SLAM [23], the model combines correlation volumes with optimization-based backends to push performance boundaries. It can infer the confidence of every pixel, which increases focus on specific regions, but its complex architecture results in slow training and more training resource consumption. Therefore, there is a need for a simpler, end-to-end model architecture that can effectively utilize reliable inter-frame regions. Our work focuses on lightweight and end-to-end frame-to-frame visual odometry.

Beyond architecture optimization, training data scarcity remains a critical bottleneck due to high acquisition costs [18], [25], [26]. Although data augmentation is a proven remedy, conventional techniques like brightness adjustment [16] or cropping [21] are decoupled from camera pose transformations. Consequently, they yield limited benefits as they cannot generate valid image-pose pairs.

Addressing the aforementioned model deficiencies and data bottleneck, we propose a novel monocular VO system. This system incorporates a robust VO network and an image-to-pose analogy augmentation module, designed to filter unreliable regions while addressing data scarcity.

To enhance the utilization efficiency of individual data, we

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3664426

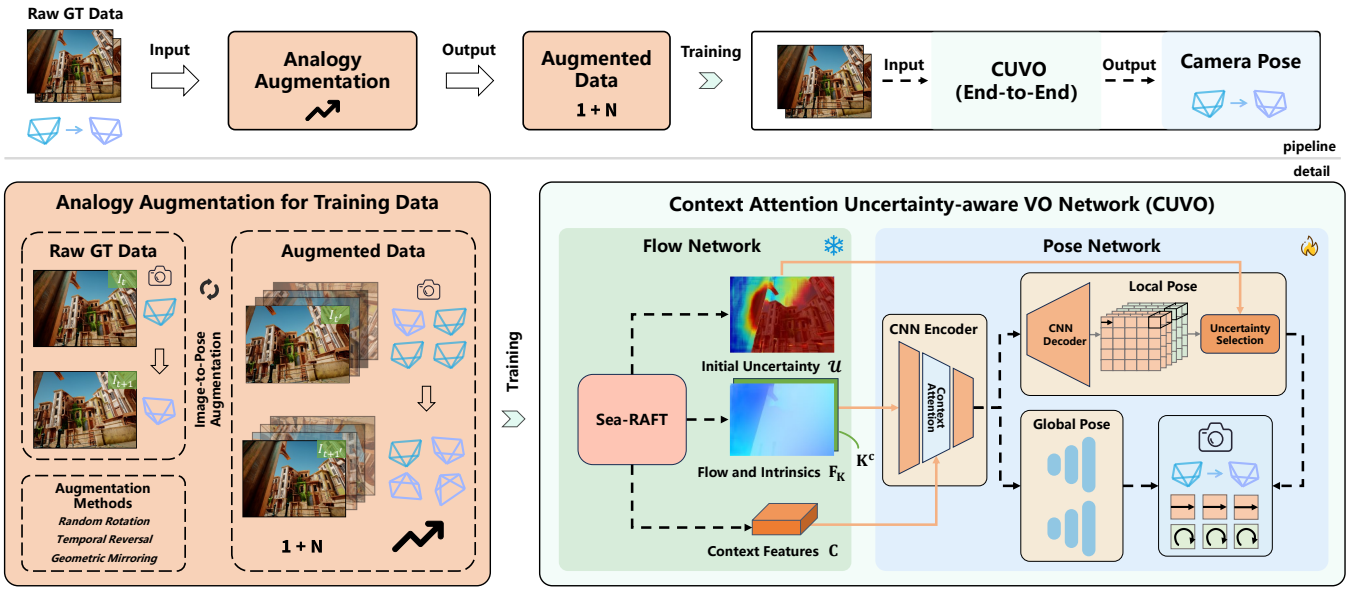

Fig. 1. Overview of the proposed Analogy Augmented VO framework. The top panel illustrates the overall pipeline, where original ground-truth (GT) images and poses are processed by the Analogy Augmentation module to produce augmented data for CUVO training. The system comprises two core components: the **Analogy Augmentation module** and the **CUVO** network. The bottom panel details the specific sub-modules: **Orange**: The Analogy Augmentation module, which employs an image-to-pose analogy strategy to enrich the diversity of the training dataset. **Green**: The Flow Network, which supplies the downstream pose network with optical flow, uncertainty maps, semantic features, and camera intrinsics $\mathbf{K^c}$. **Blue**: The Pose Network, which encodes fused inputs from the intrinsic layer and optical flow. It subsequently computes the final camera pose via a weighted aggregation of local and global pose network outputs. **Black dashed lines** indicate internal data flow, while **orange solid lines** denote the three specific inputs fed into the Pose Network.

propose the **Context Attention Uncertainty-aware VO Network (CUVO)** (as shown in Fig. 1). It consists of a forward optical flow network coupled with a parallel pose estimation network. The flow network leverages texture-rich information, which is integrated as a Context Attention mechanism into the pose network. Simultaneously, an Uncertainty Selection module is incorporated to filter out interference from noise and dynamic objects. This maximizes utilization of reliable data regions for accurate pose estimation in complex environments.

To overcome the constrained data bottleneck, we propose the **Analogy Augmentation Strategy** which enhances training data diversity via two complementary components: analogy augmentation and analogous loss. Unlike prior approaches that preserve camera poses, our augmentation methods transform image pairs while computing corresponding ground truth pose transformations. These methods include: (a) Temporal reversal swaps the order of time-series image pairs; (b) Random rotation applies random angles and extracts the largest inscribed rectangle [27]; (c) Geometric mirroring flips pairs along the horizontal or vertical axes. Their pose transformation is also synchronized. The analogous loss ensures that the VO network learns pose inference for both original and augmented image pairs, thereby enforcing consistency across different views. Experiments demonstrate significant performance improvements in data-limited scenarios, with sustained gains in data-rich conditions and robust cross-dataset zero-shot capability.

The contributions of our works are as follows:

- We propose the Context Attention Uncertainty-aware VO Network (CUVO). It prioritizes semantically informative regions to mitigate noise and dynamic objects. This improves the effective utilization of single-frame data.
- We propose an image-to-pose analogy augmentation method tailored for VO. By applying synchronized and corresponding transformations to both images and poses, we increase the quantity and diversity of training data.
- We apply the analogous loss function together with the analogy augmentation method to the CUVO network. This ensures consistency in pose inference between original and augmented image pairs. This joint learning approach enables CUVO to overcome existing performance limitations, outperforming previous end-to-end VO methods by achieving substantial improvements in accuracy and cross-dataset generalization.

II. Related Work

Learning-based Visual Odometry. Visual odometry is broadly classified into geometric-based methods [1]–[3] and deep learning-based approaches. Geometry-based VO methods are primarily categorized into direct methods [48], [50] that minimize photometric error and indirect feature-matching methods [2], [47]. These approaches fail readily in extreme environments, such as low-texture or dynamic scenes. DYO-SLAM [4], [7] addresses dynamic scene reconstruction, while PAS-SLAM [5] targets ambiguous planar scenes. However,

these approaches require significant manual effort for feature design and parameter tuning. Their complex architectures also lack generalization capabilities.

Recent studies highlight the superior robustness of learning-based methods in challenging scenarios, such as dynamic environments, motion blur, and low-light conditions [21], [28], [29], prompting increased focus on deep learning for camera pose estimation [6], [8]–[10]. Learning-based VO can be categorized by encoding strategy: implicit or explicit. Implicit encoding, exemplified by DeepVO [12], maps images to high-dimensional features for end-to-end pose estimation without explicit feature engineering. However, immature frameworks [23] and limited training data [21] restrict its generalization and accuracy compared to geometric methods.

To address this, Wang et al. [21] introduced explicit encoding in TartanVO, leveraging flow network and camera intrinsics as intermediate representations, trained on the TartanAir dataset [25]. While achieving performance comparable to geometric methods, TartanVO's pose network suffers from insufficient utilization of scene semantics, which restricts its focus to textural features. SwinFVO [22] proposes a joint estimation of depth and flow. However, its reliance on same-domain training and testing (on KITTI) results in unsatisfactory generalization performance. Teed et al. [23] advanced explicit encoding with DROID-VO. They incorporate per-pixel confidence to mitigate noise and dynamic objects, while employing a dense bundle adjustment for enhanced accuracy. Similarly, Ye et al. [33] proposed a panoramic VO, integrating semantics and iterative optimization to enable mutual reinforcement between the VO and segmentation tasks. VO algorithms have evolved from single-task to multi-task frameworks [34]–[36]. However, the aforementioned frameworks suffer from excessive complexity, rendering them challenging to train, and are not end-to-end.

To this end, we introduce a more generalized and robust CUVO, which extends explicit optical flow methods by deeply coupling the semantic and uncertainty information from the flow network with the pose network, thereby realizing a more tightly integrated and streamlined end-to-end VO network, while simultaneously addressing inaccurate pose estimation caused by dynamic scenes and drastic lighting changes.

Data Augmentation for Visual Odometry. Data augmentation is essential for VO, addressing data scarcity by generating diverse training samples to enhance model robustness and cross-dataset generalization [16], [21]. For instance, Zhang et al. [15] employ GANs for panoptic-Level image-to-image translation, which enriches features to facilitate robust feature point matching in traditional methods. Wang et al. [21] employed geometric transformations based on images only to substantially improve the robustness of the VO model in complex environments. Kim et al. [16] leveraged photometric adjustments, including modifications to brightness, contrast, and hue, to train field-of-view regression models, achieving improved performance across environmental variations.

Despite these advancements, existing traditional VO augmentation methods focus on image-level transformations, neglecting pose-consistent augmentation critical for image-pose pairs. Our analogy-based framework addresses this gap by jointly augmenting images and their corresponding pose transformations, significantly advancing VO performance.

Analogical Reasoning and Learning. In cognitive science and robotics, analogical reasoning is recognized as a core mechanism for intelligent generalization. According to the foundational Structure-Mapping Theory [30], analogy involves mapping relational structures from a source domain to a target domain, rather than merely matching surface attributes. In robotics, this paradigm enables agents to transfer learned policies to novel environments by preserving the underlying task structure despite environmental changes [31]. Drawing inspiration from these cognitive foundations, recent studies demonstrate that deep neural networks (DNNs) can similarly enhance performance by learning related tasks alongside a primary task through analogy-based strategies [32], [37], [38]. Essentially, this involves applying data augmentation strategies in conjunction with an augmentation function on DNNs, to achieve results that surpass traditional data augmentation.

In optical flow estimation, Liu et al. [38] pioneered analogy-based learning within an unsupervised framework, validating its effectiveness. Luo et al. [39] further advanced this approach by integrating analogy-driven loss into larger-scale flow network models, significantly improving accuracy and cross-dataset generalization. In contrast, analogy-based learning in VO remains underexplored. Current augmentation methods rely solely on geometric transformations and image noise, without incorporating joint loss functions. Unlike the analogy learning strategies [38] for optical flow mentioned above, VO task cannot be performed entirely at the image matrix level.

Inspired by these challenges, our work integrates camera-specific 6D pose augmentations and redesign the analogous loss to effectively incorporate pose information. Therefore, our work proposes an analogy-based learning framework for VO, including analogy augmentation, which ensures valid pose correspondence, and an analogous loss, which integrates multiple augmentation strategies into a joint loss function to ensure consistency in analogy learning between augmented and original data. Unlike conventional augmentation, our framework expands training data while preserving the data-pose relationship critical for VO tasks.

III. Method

This section introduces the details of our analogy-based learning VO system, as shown in Fig. 1. It primarily comprises two major modules: the CUVO model serves as the main algorithmic framework, while analogy augmentation expands training data and provides the loss function for training.

Firstly, we propose the CUVO (Section III-A), a coupled architecture that enhances the utilization efficiency of data samples and improves semantic understanding and interference suppression. At the data level, we introduce analogy augmentation method (Section III-B) to increase training data diversity and overcome data availability constraints. Based on the analogy augmentation method, we design the analogous loss (Section III-C) to ensure consistency between original and augmented data during training.

A. Context Attention Uncertainty-aware VO Network

Current VO methods encounter limitations in complex scenes, such as dynamic objects and illumination changes. Prior approaches [21], [23], [33] based on optical flow often underutilize semantic information and lack uncertainty modeling, constraining cross-dataset generalization. For example, DROID-SLAM [23] relies on complex paradigms, requiring continuous frames for simultaneous localization and mapping. Analogy-driven augmentation disrupts its 3D point consistency, and without dense bundle adjustment, DROID-VO struggles to accurately estimate camera poses. In contrast, TartanVO [21], despite utilizing only adjacent frames and an extremely streamlined framework, achieves competitive performance. Hence, we adopt TartanVO as our base framework.

As shown in Fig. 1, we propose the Context Attention Uncertainty-aware VO Network (CUVO), which comprises two core components: a flow network and a pose network. The flow network utilizes the SEA-RAFT [41] network to generate context features and initial uncertainty, thereby providing rich prior information. For the pose network, we propose two core components: a context attention module and an uncertainty selection module with local region awareness. These modules leverage flow network features and uncertainty to prioritize texture-rich, static regions for pose estimation. The following subsections detail the flow network and pose network modules.

1) Flow Network: The flow network estimates optical flow while generating semantic context features and initial uncertainty for the pose network in the CUVO. TartanVO's reliance on the outdated PWC-Net [40] yields suboptimal accuracy and performance, prompting us to adopt the pre-trained SEA-RAFT [41]. SEA-RAFT processes two input frames I_t, I_{t+1} using a ResNetFPN backbone to extract features, followed by n iterations through an RNN-based architecture (Decoder Cell) to output optical flow $\mathbf{F}_{t,t+1}$. Additionally, we utilize the final output feature as semantic context features $\mathbf{C} \in \mathbb{R}^{128\times H/8\times W/8}$, rich in texture and scene information.

We model the initial uncertainty map $\mathcal{U} \in \mathbb{R}^{1\times H\times W}$ using the variance of the MoL distribution proposed in SEA-RAFT [41],

$$\sigma(\omega_1, \omega_2, \beta_1, \beta_2) = \frac{2\exp(\omega_1 + 2\beta_1) + \exp(\omega_2 + 2\beta_2)}{\exp(\omega_1) + \exp(\omega_2)} \quad (1)$$

where $\omega_1 + \omega_2 = 1.0$ are mixing coefficients, β_1, β_2 are logarithmic scale parameters. The test [41] reveals $\beta_1 = 0$ as optimal, yielding $\mathcal{U} = \sigma(\omega_1, \omega_2, 0, \beta_2)$.

This MoL-derived variance serves as a statistically principled measure of aleatoric uncertainty, reliably indicating regions of high prediction ambiguity—such as low-texture areas and dynamic objects.

2) Pose Network: As shown in Fig. 1, the pose network comprises three main modules: the CNN Encoder, Local Pose, and Global Pose.

The CNN Encoder primarily encodes matching features from the flow network (Flow and Intrinsics $\mathbf{F_K}$) and simultaneously integrates context features $\mathbf{C}$ using the main context attention module. Specifically, $\mathbf{F_K}$ are primarily formed as follows: Camera intrinsics $\mathbf{K} = [f_x, f_y, c_x, c_y]$ are transformed into an intrinsics layer $\mathbf{K^c} \in \mathbb{R}^{2\times H\times W}$ [21]. This is then concatenated with the flow network's optical flow $\mathbf{F}_{t,t+1}$ to form $\mathbf{F_K} = [\mathbf{F}_{t,t+1}, \mathbf{K^c}]$. This composite input is subsequently downscaled to $(H/4 \times W/4)$ for input to the CNN Encoder.

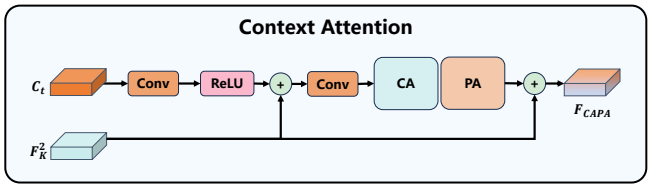

Fig. 2. Context attention module. The inputs to this module are context features $\mathbf{C}$ and the coder's second-layer features $\mathbf{F}^2_\mathbf{K}$. These inputs pass through two convolutional layers, the CAPA module, and a weighting operation to produce $\mathbf{F}_{PACA}$.

The Local Pose consists of two main sub-modules: the CNN Decoder and the uncertainty selection module. The CNN Decoder forms a U-shaped network to decode pixel-level poses with the CNN Encoder. Subsequently, based on the initial uncertainty $\mathcal{U}$ from the flow network, the uncertainty selection module is used to filter low-uncertainty regions and integrate them to obtain the local camera motion pose.

The Global Pose contains two fully connected layers that predict translation $\hat{\mathbf{T}}_g$ and rotation $\hat{\mathbf{R}}_g$ poses. Its input is the global features encoded by the CNN Encoder, which are used to calculate the global pose. Finally, the local and global poses are weighted and fused to obtain the final camera pose.

We introduce two core modules to the pose network: the Context Attention module, for leveraging semantic features, and the Uncertainty Selection module, for refining pose estimates by prioritizing reliable regions.

Context Attention. The module integrates semantic context features $\mathbf{C}$ from the flow network with flow and camera intrinsics. Using a ResNet32 backbone [44], it encodes the flow network's flow $\mathbf{F}_{t,t+1}$ and intrinsics layer $\mathbf{K^c}$. Given the semantic richness of $\mathbf{C}$, we place the Context Attention module in the encoder's second layer, fusing $\mathbf{C}$ with second-layer features $\mathbf{F}^2_\mathbf{K}$ extracted from the flow and intrinsics.

As depicted in Fig. 2, the fusion process proceeds as follows: context features $\mathbf{C}$ are processed through a convolutional layer and ReLU activation, ensuring dimensional consistency, then added element-wise to $\mathbf{F}^2_\mathbf{K}$ for initial semantic-motion integration; subsequently, the added features are processed via global pooling and attention weight computation in the Channel Attention (CA) module [42], which adjusts channel weights to produce refined channel features, computed as

$$\mathbf{F}_{CA} = CA(\mathbf{F}^2_\mathbf{K} + \mathrm{ReLU}(\mathrm{Conv}(\mathbf{C}))) \quad (2)$$

The F_{CA} features are input to the Pixel Attention (PA) module [42], generating a pixel-level attention map. This map is element-wise multiplied with F_{CA}, and the result is added to $\mathbf{F}^2_\mathbf{K}$ to yield:

$$\mathbf{F}_{PACA} = \mathbf{F}^2_\mathbf{K} + \mathrm{PA}(\mathbf{F}_{CA}) \quad (3)$$

The context attention module fuses semantic and motion features to improve optical flow representation in challenging scenes. Element-wise addition ensures stable gradient flow,

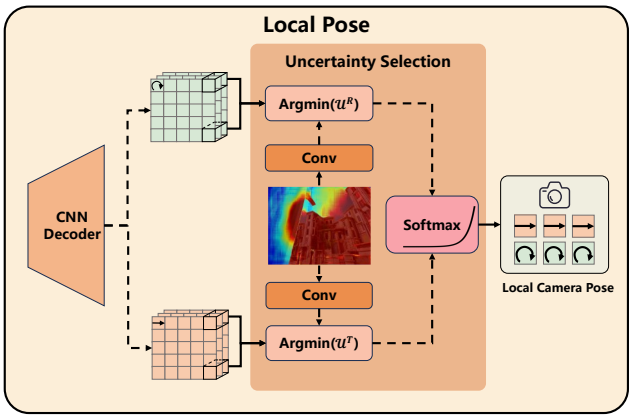

Fig. 3. Local Pose module (Uncertainty Selection module). This module uses CNN to form the right half of the U-shaped network, where the CNN decoder outputs pixel-level translation and rotation pose estimates. Based on the input initial uncertainty $\mathcal{U}$, it selects regions with minimal uncertainty, and local poses $(\hat{\mathbf{T}}_l, \hat{\mathbf{R}}_l)$ are computed via softmax aggregation.

improving training stability. The resulting F_{PACA} features provide high-quality input for the local and global pose decoders, advancing CUVO's pose estimation accuracy.

Uncertainty Selection. Integrated with Context Attention within the U-shaped architecture, this module optimizes pose estimation at a high-dimensional multi-channel level. Instead of relying on conventional single-channel or black-box uncertainty estimation, we establish our framework upon the Mixture-of-Laplace (MoL) distribution.

As depicted in Fig. 3, the CNN decoder predicts pixel-level translation $\hat{\mathbf{T}}_P \in \mathbb{R}^{3\times H/4\times W/4}$ and rotation $\hat{\mathbf{R}}_P \in \mathbb{R}^{3\times H/4\times W/4}$. These are fused with the flow network's initial uncertainty $\mathcal{U} \in \mathbb{R}^{1\times H/4\times W/4}$ to derive robust local poses $\hat{\mathbf{T}}_l \in \mathbb{R}^3$ and $\hat{\mathbf{R}}_l \in \mathfrak{so}(3)$.

The CNN decoder adopts a U-Net architecture [45]. It uses skip connections to fuse encoder features with upsampled decoder features. This balances low-level spatial details and high-level semantics, thereby enhancing the accuracy of $\hat{\mathbf{T}}_P$ and $\hat{\mathbf{R}}_P$. The initial pixel-level uncertainty $\mathcal{U}$ from SEA-RAFT is further processed through two learnable convolutional layers. These layers generate task-aligned 3-channel uncertainty maps $\mathcal{U}^T \in \mathbb{R}^{3\times H/4\times W/4}$ and $\mathcal{U}^R \in \mathbb{R}^{3\times H/4\times W/4}$. The maps represent per-dimension uncertainty for translation and rotation, respectively, and align with $\hat{\mathbf{T}}_P$ and $\hat{\mathbf{R}}_P$. This refinement preserves the probabilistic interpretability of the MoL signal. At the same time, it adapts the signal to the semantic level needed for pose estimation.

The selection module, inspired by PWVO's hierarchical weighting [43], computes local poses by dividing $\hat{\mathbf{T}}_P$, $\hat{\mathbf{R}}_P$, $\mathcal{U}^{\mathbf{T}}$, and $\mathcal{U}^{\mathbf{R}}$ into N patches of size $k \times k$ ($k = 8$ in experiments). For each patch, it extracts the translation t_n and rotation r_n with the minimum uncertainty value, and then computes weights for translation and rotation via:

$$W_n^{\delta} = \mathrm{Softmax}(\mathcal{U}_n^{\delta}), \quad \delta \in \{\mathbf{T}, \mathbf{R}\} \tag{4}$$

followed by deriving local poses as:

$$\hat{\mathbf{T}}_l = \sum_{n=1}^{N} W_n^{\mathbf{T}} t_n, \hat{\mathbf{R}}_l = \sum_{n=1}^{N} W_n^{\mathbf{R}} r_n \tag{5}$$

Concurrently, a parallel global pose module, leveraging high-level CNN encoder features, predicts global poses $\hat{\mathbf{T}}_g \in \mathbb{R}^3$ and $\hat{\mathbf{R}}_g \in so(3)$. The final camera pose $\hat{\mathbf{P}}_{t,t+1} = [\hat{\mathbf{T}}, \hat{\mathbf{R}}]$ is obtained by weighted fusion of local and global poses. The global pose module processes final-layer encoder features through two fully connected layers to predict $\hat{\mathbf{T}}_g$ and $\hat{\mathbf{R}}_g$. Local and global poses are fused as:

$$\hat{\mathbf{P}}_{t,t+1} = (\hat{\mathbf{T}}, \hat{\mathbf{R}}), \quad \text{where} \quad \begin{cases} \hat{\mathbf{T}} = \alpha_g \hat{\mathbf{T}}_g + \alpha_l \hat{\mathbf{T}}_l, \\ \hat{\mathbf{R}} = \alpha_g \hat{\mathbf{R}}_g + \alpha_l \hat{\mathbf{R}}_l \end{cases} \tag{6}$$

with $\alpha_g = \alpha_l = 0.5$ in experiments.

Collectively, the modules detailed above constitute the pose network component of CUVO. As a result, we obtain richer, dimension-specific uncertainty modeling compared to scalar confidence maps used in prior methods [23], [24]. By selecting low-uncertainty predictions within local patches (Eq. 4–5), the module effectively filters unreliable regions and provides a theoretically grounded mechanism for improving robustness and reproducibility. Finally, the global pose module, fused similarly to ResNet's residual structure, stabilizes gradients and reduces pose drift.

B. Analogy Augmentation

Learning-based VO methods heavily rely on the quantity and diversity of training data. Conventional data augmentation techniques, such as cropping and scaling, only modify images without updating the corresponding ground-truth poses. Consequently, these methods fail to enrich the diversity of image-pose correspondences, as the pose distribution remains static regardless of the applied visual variations.

To overcome this, we propose an analogy augmentation method to enable more varied motion patterns, as shown in Fig. 1. Specifically, as 2D transformations are applied to the image pairs, the associated poses are correspondingly transformed in 3D space. For instance, temporal reversal converts forward motion to backward, geometric mirroring flips left-turn motion to right-turn, and random rotation introduces rotational transformations to datasets lacking such patterns. The following subsection provide a detailed description.

As shown in Fig. 4-5, the ground-truth pose transformation is defined in the Camera Coordinate System at time t, denoted CSS_t. The x-axis, y-axis, and z-axis visualize the coordinate system. For images I_t and I_{t+1}, the pose transformation vector from CSS_t to CSS_{t+1} is given by:

$$\mathbf{P}_{t,t+1} = [\mathbf{T}_x, \mathbf{T}_y, \mathbf{T}_z, \mathbf{R}_x, \mathbf{R}_y, \mathbf{R}_z]^{\top} \tag{7}$$

$$\mathbf{P}^n_{t',t+1'} = \mathcal{T}_n \mathbf{P}_{t,t+1} \tag{8}$$

where $\mathbf{P}_{t,t+1}$ is the original ground-truth pose, consisting of the translation $(\mathbf{T}_x, \mathbf{T}_y, \mathbf{T}_z)$ from CSS_t to CSS_{t+1} and the rotation components $(\mathbf{R}_x, \mathbf{R}_y, \mathbf{R}_z)$. For each augmentation type indexed by n, the transformed pose $\mathbf{P}^n_{t',t+1'}$ is obtained via a transformation matrix $\mathcal{T}_n$.

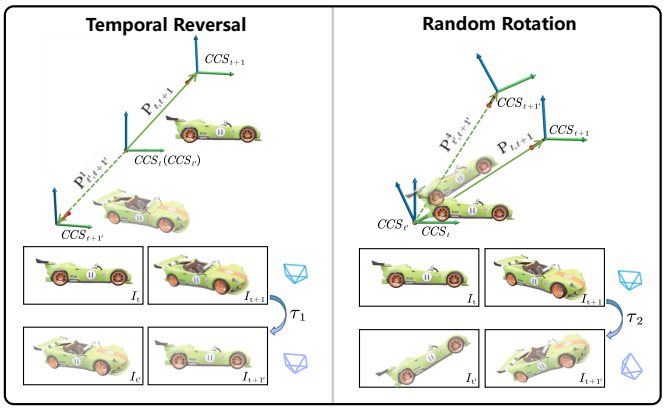

Fig. 4. Temporal Reversal and Random Rotation augmentation strategies. Temporal Reversal constructs backward trajectories to ensure kinematic reversibility, while Random Rotation introduces spatial variations to enhance robustness against aggressive rotational maneuvers.

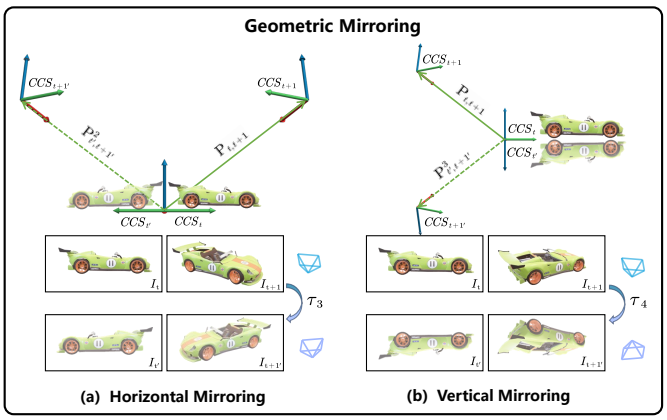

Fig. 5. Geometric Mirroring augmentation strategies. (a) Horizontal Mirroring: Flips the input frames horizontally (left-to-right) to simulate laterally inverted views. (b) Vertical Mirroring: Flips the frames vertically (up-down). The corresponding ground-truth poses (or depth maps) are transformed accordingly to maintain validity.

To support first stage training with ground-truth optical flow, as in subsection IV-B, we provide the corresponding optical flow augmentation strategies. The ground-truth optical flow $\mathbf{F}_{t,t+1} \in \mathbb{R}^{2\times H\times W}$ represents pixel motion $(\mathbf{u}, \mathbf{v})$,

$$\mathbf{F}_{t,t+1}(x,y) = [\mathbf{u}(x,y), \mathbf{v}(x,y)]^\top \tag{9}$$

$$\mathbf{F}^n_{t',t+1'}(x,y) = \mathcal{M}_n \mathbf{F}_{t,t+1}(x,y) \tag{10}$$

where $\mathbf{u}$ is the horizontal displacement and $\mathbf{v}$ is the vertical displacement in image coordinates, and $H \times W$ is the image resolution. The augmented optical flow is denoted as $\mathbf{F}^n_{t',t+1'}(x,y)$, derived via the transformation function $\mathcal{M}_n$.

Temporal Reversal ($n = 1$)**.** As illustrated in Fig. 4, we swap the input image pair such that $I'_{t'} = I_{t+1}$ and $I'_{t+1} = I_t$. In the visualization, the opaque car denotes the projection in the original coordinate systems (CSS_t, CSS_{t+1}), whereas the semitransparent car indicates the projection in the reversed timeline $(CSS_{t'}, CSS_{t+1'})$. Consequently, the relative pose transformation and the augmented flow are given by:

$$\mathbf{P}^1_{t',t+1'} = \mathcal{T}_1 \mathbf{P}_{t,t+1} \tag{11}$$

$$\mathbf{F}^1_{t',t+1'}(x,y) = \mathcal{M}_1 \mathbf{F}_{t,t+1}(x,y) \tag{12}$$

where $\mathcal{T}_1 = -\mathbf{I}_2$ and $\mathcal{M}_1 = -\mathbf{I}_6$, with the subscripts indicating the dimensionality of the identity matrices.

Random Rotation ($n = 2$)**.** As shown in Fig. 4, a random angle $\theta \in [0, 360°)$ is sampled, and images I_t and I_{t+1} are rotated about their centers by θ to yield I'_t and I'_{t+1}. To remove invalid borders, the images are then cropped to their largest inscribed rectangle. The resulting augmented pose and flow are formulated as follows:

$$\mathbf{P}^2_{t',t+1'} = \mathcal{T}_2 \mathbf{P}_{t,t+1} \tag{13}$$

$$\mathbf{F}^2_{t',t+1'} = \mathcal{C}_{rect}(\mathcal{M}_2 \mathbf{F}_{t,t+1}) \tag{14}$$

where $\mathcal{T}_2 = \mathbf{R}_\theta$ and $\mathcal{M}_2 = \mathbf{R}^F_\theta$ denote the transformation matrices for the pose and flow, respectively. To mitigate the influence of invalid boundary regions caused by rotation, the maximum inscribed rectangle $\mathcal{C}_{rect}(\cdot)$ is cropped from the rotated image and flow inputs.

Geometric Mirroring ($n = 3, 4$)**.** Geometric mirroring is divided into horizontal mirroring and vertical mirroring.

Images I_t and I_{t+1} are mirrored horizontally to produce I'_t and I'_{t+1} (Fig. 5 (a)). This implies a pose transformation from I'_t to I'_{t+1}, with the augmented flow defined as:

$$\mathbf{P}^3_{t',t+1'} = \mathcal{T}_3 \mathbf{P}_{t,t+1} \tag{15}$$

$$\mathbf{F}^3_{t',t+1'}(W-x-1, y) = \mathcal{M}_3 \mathbf{F}_{t,t+1}(x,y) \tag{16}$$

where the pose transformation matrix is given by $\mathcal{T}_3 = \text{diag}(1,-1,1,-1,1,-1)$, and the flow transformation matrix is defined as $\mathcal{M}_3 = \text{diag}(-1,1)$.

Vertical mirroring (Fig. 5 (b)) is similar to horizontal mirroring strategy:

$$\mathbf{P}^4_{t',t+1'} = \mathcal{T}_4 \mathbf{P}_{t,t+1}, \tag{17}$$

$$\mathbf{F}^4_{t',t+1'}(x, H-y-1) = \mathcal{M}_4 \mathbf{F}_{t,t+1}(x,y) \tag{18}$$

where the corresponding diagonal transformation matrices are $\mathcal{T}_4 = \text{diag}(1,1,-1,-1,-1,1)$. for the pose and $\mathcal{M}_3 = \text{diag}(1,-1)$ for the flow.

C. Analogous Loss

In order to enhance the model's learning ability for different transformations, we propose that the VO network should jointly learn pose estimation tasks for both original and augmented image pairs through analogy-based learning. Specifically, the network learns pose transformation vectors ($\hat{\mathbf{P}}_{t,t+1}$) from frame I_t to I_{t+1}, as well as $\hat{\mathbf{P}}^n_{t',t+1'}$ from augmented frames $I_{t'}$ to $I_{t+1'}$, alongside transformations for multiple augmentation strategies. This joint learning motivates our analogous loss as shown in Fig. 6.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3664426

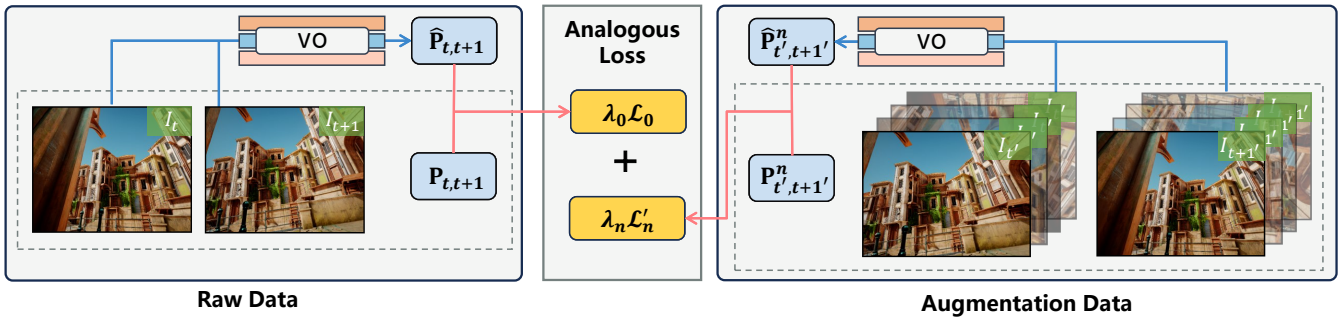

Fig. 6. Pipeline of the Analogous Loss. The training loss from the original input (left) is combined with losses from n types of analogy augmentations (right). The central module computes the final objective via a weighted summation of these components.

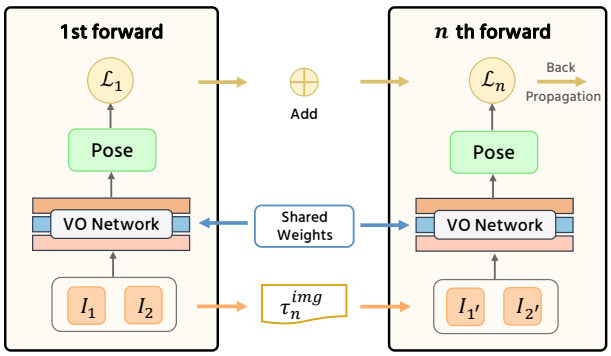

Fig. 7. Analogy Augmentation Strategy Pipeline. A complete training step involves n forward passes: the left side illustrates the model forward process for original samples (without augmentation). Then, we apply transformations to the images and poses to construct n augmented samples, which respectively undergo the additional forward process shown on the right side. Finally, the composite loss is calculated from the n processes, followed by backpropagation to optimize the model.

First, we present the basic loss function. Monocular VO inherently struggles to recover absolute motion scale, and overfitting to specific scales yields poor generalization [21]. To address this, we normalize the camera translation $\mathbf{T}$ between frames to ensure scale invariance, as follows:

$$\mathbf{T}_{norm} = \frac{\mathbf{T}}{\max(|\mathbf{T}|, \epsilon)} \tag{19}$$

where $\epsilon = 10^{-6}$. This normalization enhances the robustness of pose estimation across varied datasets.

Second, we define the predicted pose transformation as $\hat{\mathbf{P}}_{t,t+1} = [\hat{\mathbf{T}}, \hat{\mathbf{R}}]$ and the ground truth as $\mathbf{P}_{t,t+1} = [\mathbf{T}, \mathbf{R}]$. The base loss for original image pairs is:

$$\mathcal{L}(\hat{\mathbf{P}}_{t,t+1}, \mathbf{P}_{t,t+1}) = |\hat{\mathbf{T}}_{norm} - \mathbf{T}_{norm}|_1 + |\hat{\mathbf{R}} - \mathbf{R}|_1 \tag{20}$$

where $\hat{\mathbf{T}}_{norm}$ and $\mathbf{T}_{norm}$ are normalized translations, and $|\cdot|_1$ denotes the L1 norm.

For original data, this loss is:

$$\mathcal{L}_0 = \mathcal{L}(\hat{\mathbf{P}}_{t,t+1}, \mathbf{P}_{t,t+1}) \tag{21}$$

For augmented pairs, we compute:

$$\mathcal{L}'_n = \mathcal{L}(\hat{\mathbf{P}}^n_{t',t+1'}, \mathbf{P}^n_{t',t+1'}) \tag{22}$$

where $\hat{\mathbf{P}}^n_{t',t+1'}$ and $\mathbf{P}^n_{t',t+1'}$ are predicted and ground truth pose transformations for augmented frames. Updating each $\mathcal{L}'_n$ with separate gradients leverages correlations between augmented image-pose pairs but fails to enforce consistency across tasks.

Thus, we propose an analogous loss for different analogy augmented data:

$$\mathcal{L}_{analogous} = \sum_{n=1}^{N} \lambda_n \mathcal{L}'_n \tag{23}$$

where N is the number of augmentation types, and λ_n are weights balancing each loss term.

The total joint loss combines original and analogous losses:

$$\mathcal{L}_{total} = \lambda_0 \mathcal{L}_0 + \mathcal{L}_{analogous} \tag{24}$$

where λ_0 weights the original loss $\mathcal{L}_0$.

Integrating the aforementioned methods, we outline the pipeline of the analogy augmentation strategy, as illustrated in Fig. 7. The process begins with a first forward pass to compute poses for original image pairs, followed by n forward passes for n analogy augmentation methods, with shared weight parameters in the VO network. We then compute a weighted composite loss, combining losses from original and augmented data, and update model parameters via backpropagation. Experimental results (Subsection IV-B1) further demonstrate that the analogous loss significantly enhances VO performance.

IV. Experiments

A. Implementation Details

Training. Our network architecture adopts TartanVO as the baseline. As TartanVO's training code, including data processing, is unavailable, we reimplemented it in PyTorch with enhancements to data processing and model architecture. The validation of our reproduction is available in Appendix A. Following the TartanVO training methodology, our training process proceeds in two distinct stages as follows:

In the first stage, as shown in Fig. 8, we pre-train the Global Pose network using ground-truth optical flow with the aim of eliminating the impact of predicted flow errors. We train the network for 200 epochs with a batch size of 200 and a learning rate of 10^{-4}, utilizing ground-truth flow and pose labels alongside our flow-based analogy augmentation.

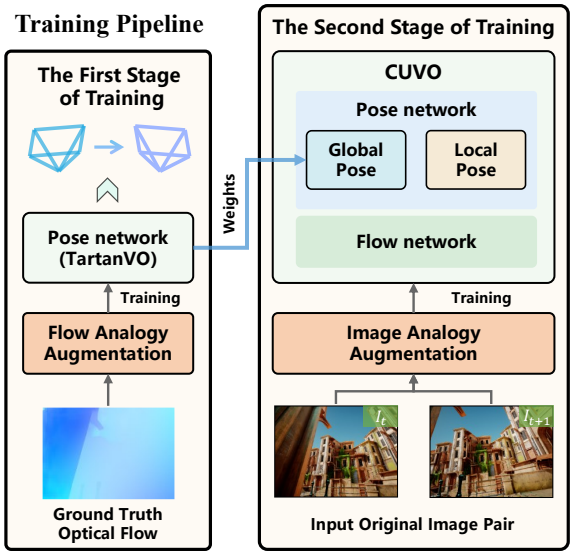

Fig. 8. Training Pipeline. In the first stage, we train TartanVO's pose network using ground truth optical flow combined with the optical flow analogy augmentation strategy. In the second stage, we initialize UCVO's global pose network with the pre-trained pose network weights, and train UCVO using images combined with the image analogy augmentation strategy.

In the second stage, we initialize CUVO's Global Pose network with the first stage pretrained weights, and perform joint optimization for 100 epochs using image-based analogy augmentation with a batch size of 32 and a learning rate of 10^{-5}. The flow network processes input images of size 448×640, while the pose network uses 112×160, minimizing GPU memory usage with minimal accuracy loss.

We employ the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a MultiStepLR scheduler, decaying the learning rate by 0.2 at epochs 60 and 100 in the first stage and at epoch 30 and 50 in the second stage. Training is conducted on four NVIDIA RTX 4090 GPUs using PyTorch.

Evaluation Setup. We evaluate the Absolute Trajectory Error (ATE [m]) in meters on the KITTI [18], TartanAir [25], TUM-RGBD [19], and ETH3D-SLAM [20] datasets. We benchmark CUVO against a comprehensive set of baselines. Our evaluation primarily targets frame-to-frame (end-to-end) baselines for fair comparison, including TartanVO [21] and DytanVO [28]. Furthermore, we report results from state-of-the-art multi-frame methods to position our approach within the broader VO landscape. These additional comparisons span multi-frame optimization-based methods (e.g., DROID-VO [23], DPVO [49], MambaVO [52]) and full SLAM systems (e.g., ORB-SLAM3 [47], LDSO [50], DSO [48], DROID-SLAM [23], MambaVO++ [52]).

To ensure consistency, all learning-based models are trained on the separate TartanAir training set. Specifically, CUVO-Full denotes our model trained with the full analogy augmentation strategy. Traditional approaches (ORB-SLAM3, LDSO, DSO) utilize their official publicly available models.

Overview of Ablation Studies. Our ablation analysis is conducted at two levels. At the model level, we evaluate the contribution of individual components by comparing ATE results on the KITTI and TartanAir, while also reporting their computational efficiency (FPS and GPU memory usage). At the strategy level, the first training stage is designed to evaluate the **scalability** of the Analogy Augmentation strategy. The second training stage assesses the **effectiveness** of the UCVO when combined with the Analogy Augmentation strategy.

B. Result Analysis

In this section, we demonstrate CUVO's competitive performance in inferring poses from continuous image frames through extensive quantitative and qualitative comparative experiments. In these two comparative experiments, we present the experimental results of CUVO both with and without the analogy augmentation strategy. The full analogy augmentation strategy includes temporal reversal, horizontal mirroring, and random rotation, as vertical mirroring has limitations that will be discussed in subsection IV-D3.

*1) **Quantitative Evaluation**:* **KITTI dataset**, which was captured by a car-mounted camera, encompasses a range of urban, rural, and highway environments. Table I presents the ATE comparison on KITTI. Despite some SLAM systems incorporating loop closure and global bundle adjustment, our full analogy augmentation method (CUVO-Full) achieves the lowest average ATE across all VO/SLAM methods. Among VO methods, CUVO reduces ATE by 24.63% compared to TartanVO, while the full analogy augmentation method excluding vertical mirroring (CUVO-Full) lowers ATE by 30.05%. As revealed by ablation studies (Table IX), CUVO with horizontal mirroring is the most effective method on KITTI, yielding a 39.16% ATE reduction compared to TartanVO. This performance surpasses that of the best Multi-Frame VO method, MambaVO [52], achieving an significant 58.87% ATE reduction.

CUVO demonstrates varying degrees of improvement across all seven sequences. However, we observe distinct performance characteristics in specific scenarios. Multi-frame VO methods excel on sequences 01, 03, and 04, primarily benefiting from the long-distance, quasi-linear trajectories and the absence of dynamic objects during initialization. Regarding the performance on sequences 01, 03, and 06, the performance gap (or residual error) compared to the fine-tuned TartanVO baseline is attributed to our use of a frozen optical flow network. The lack of fine-tuning leads to slightly less precise flow estimation in these specific scenes (see Appendix A). We plan to address this limitation by incorporating efficient flow fine-tuning in future work. In sequence 08, which contains numerous dynamic objects (e.g., pedestrians, bicycles, vehicles), CUVO's integration of semantic and pixel-level uncertainty information significantly enhances performance as shown in Fig. 9, achieving up to an 85.01% ATE reduction (CUVO-Full vs. DROID-SLAM), demonstrating substantial improvements in robustness and generalization.

In Table III, we compare the RMSE of translational drift (t_{rel}) and rotational drift (r_{rel}) against various methods on the standard KITTI sequences 06, 07, 09, and 10. CUVO significantly outperforms the baseline, improving t_{rel} by 13.68%

TABLE I
ATE↓ RESULTS ON KITTI ODOMETRY 0-10, COMPARING WITH OTHER SLAM/VO METHODS

Methods		00	01	02	03	04	05	06	07	08	09	10	Avg
Multi-Frame (SLAM/VO)	ORB-SLAM3 [47]	6.77	✗	30.50	1.04	0.93	5.54	16.61	9.70	60.69	7.89	8.65	-
	LDSO [50]	9.32	11.68	31.98	2.85	1.22	5.10	13.55	2.96	129.02	21.64	17.36	22.43
	DROID-SLAM [23]	92.10	344.60	✗	2.38	1.00	118.5	62.47	21.78	161.60	✗	118.70	-
	MambaVO++ [52]	6.19	8.04	27.73	1.94	0.59	3.05	11.79	1.70	105.42	63.24	10.51	21.84
	DROID-VO [23]	**98.43**	84.20	108.80	2.58	0.93	59.27	64.40	24.20	**64.55**	**71.80**	16.91	54.19
	DPVO [49]	113.21	12.69	123.40	2.09	0.68	58.96	**54.78**	19.26	115.90	75.10	**13.63**	53.61
	MambaVO [52]	112.39	**8.16**	**93.78**	**1.80**	**0.66**	**56.51**	57.19	**17.90**	116.01	73.56	14.37	**50.21**
Frame-to-Frame (End-to-End VO)	TartanVO [21]	69.11	53.19	78.78	**2.70**	1.99	55.18	**10.50**	13.87	48.16	27.93	11.90	33.94
	DytanVO [28]	43.14	**30.83**	64.94	4.36	**1.05**	33.83	21.85	23.51	30.43	28.87	10.50	26.66
	CUVO (Ours)	**37.61**	82.54	**44.90**	5.08	1.41	**16.38**	26.93	**4.70**	38.56	17.29	5.96	25.58
	CUVO-Full (Ours)	44.73	69.92	52.41	2.82	1.16	23.90	17.16	6.07	**24.22**	**14.38**	**4.39**	**23.74**

CUVO-Full signifies training with the full analogy augmentation method.
Baseline method is shaded in gray. The best and second-best Frame-to-Frame (End-to-End) VO results are highlighted in **orange bold** and green. The best multi-frame VO results are highlighted in **bold**.

TABLE II
ATE↓ RESULTS ON THE TARTANAIR TEST DATASET.

Methods		ME 000	ME 001	ME 002	ME 003	ME 004	ME 005	ME 006	ME 007	MH 000	MH 001	MH 002	MH 003	MH 004	MH 005	MH 006	MH 007	Avg
Multi-Frame (SLAM/VO)	ORB-SLAM3 [47]	13.61	16.86	20.57	16.00	22.27	9.28	21.61	7.74	15.44	2.92	13.51	8.18	2.59	21.91	11.70	25.88	14.38
	DSO [48]	9.65	3.84	12.20	8.17	9.27	2.94	8.15	5.43	9.92	0.35	7.96	3.46	✗	12.58	8.42	7.50	-
	DROID-SLAM [23]	0.17	0.06	0.36	0.87	1.14	0.13	1.13	0.06	0.08	0.05	0.04	0.02	0.01	0.68	0.30	0.07	0.33
	MambaVO++ [52]	-	-	-	-	-	-	-	-	0.12	0.04	0.02	0.02	0.37	0.14	0.05	0.05	-
	DROID-VO [23]	0.22	0.15	0.24	1.27	1.04	**0.14**	1.32	0.77	0.32	0.13	0.08	0.09	1.52	0.69	0.39	0.97	0.58
	DPVO [49]	**0.16**	**0.11**	**0.11**	**0.66**	**0.31**	**0.14**	**0.30**	**0.13**	**0.21**	0.04	0.04	0.08	0.58	**0.17**	**0.11**	0.15	**0.21**
	MambaVO [52]	-	-	-	-	-	-	-	-	0.24	**0.02**	**0.03**	**0.02**	**0.46**	0.18	0.13	**0.05**	-
Frame-to-Frame (End-to-End VO)	TartanVO [21]	27.30	0.86	**0.64**	7.18	2.02	**0.58**	4.12	**0.42**	**2.12**	0.31	1.28	1.09	0.99	1.40	1.74	**1.42**	3.34
	DytanVO [28]	25.95	1.36	1.17	6.94	2.75	0.96	4.46	0.89	5.10	0.22	1.62	0.79	1.29	4.46	2.06	2.36	3.90
	CUVO (Ours)	**12.20**	**0.42**	1.16	6.12	1.94	0.61	1.81	0.44	2.34	**0.17**	1.23	**0.59**	0.86	2.37	1.53	1.67	2.22
	CUVO-Full (Ours)	12.94	0.52	2.00	**4.55**	**1.48**	0.59	**0.63**	0.65	2.95	0.62	**0.79**	0.84	**0.40**	**0.69**	**0.75**	1.46	**1.99**

TABLE III
PERFORMANCE COMPARISON ON THE KITTI DATASET: RMSE FOR TRANSLATIONAL DRIFT (t_{rel}) AND ROTATIONAL DRIFT (r_{rel}).

Methods		Metric	06	07	09	10	Avg
Multi-Frame (SLAM/VO)	ORB-SLAM3 [47]	t_{rel}	18.68	10.96	15.30	3.71	12.16
		r_{rel}	0.26	0.37	0.26	0.30	0.30
	MASt3R-SLAM [55]	t_{rel}	125.99	37.35	84.14	68.89	79.09
		r_{rel}	12.89	39.13	54.79	47.09	38.48
	DROID-SLAM [23]	t_{rel}	1.99	7.46	2.14	3.09	3.67
		r_{rel}	2.96	7.34	3.78	7.15	5.31
	DPVO [49]	t_{rel}	23.93	10.64	17.23	6.61	14.60
		r_{rel}	2.50	3.11	3.36	3.86	3.21
Frame-to-Frame (End-to-End VO)	TartanVO [21]	t_{rel}	4.72	4.32	6.00	6.89	5.48
		r_{rel}	2.95	3.41	3.11	2.73	3.05
	DytanVO [28]	t_{rel}	**4.70**	8.25	6.97	7.77	6.92
		r_{rel}	**1.98**	3.24	**2.00**	1.88	2.28
	CUVO (Ours)	t_{rel}	5.40	**2.87**	5.39	5.27	4.73
		r_{rel}	2.84	**1.66**	2.38	1.79	**2.17**
	CUVO-Full (Ours)	t_{rel}	7.17	3.58	**3.75**	**3.09**	**4.40**
		r_{rel}	3.12	1.87	2.28	**1.59**	2.21

t_{rel} (%) and r_{rel} (deg/100m) are the average translational and rotational RMSE drift on various segments with length 100-800 m.

and r_{rel} by 40.55%. Additionally, the analogy augmentation strategy alone contributes to improvements of 19.71% (t_{rel}) and 27.54% (r_{rel}) over the baseline.

TartanAir test dataset, used in the CVPR Visual SLAM challenge, presents significant fitting challenges. Table II compares Frame-to-Frame and Multi-Frame VO methods on the TartanAir test split, which features virtual environments with diverse transformations (e.g., lighting, fog, rain, varying motion amplitudes), posing challenges for traditional methods [47], [48]. Learning-based methods benefit from significant viewpoint variations, similar structural features, and shorter camera motion distances compared to KITTI, giving Multi-Frame methods like DROID-VO a significant advantage through inter-frame matching and local optimization. As a Frame-to-Frame VO method without Multi-Frame optimization, CUVO has room for improvement.

Our method significantly advances frame-to-frame VO, with CUVO and CUVO-Full reducing the baseline ATE by 33.53% and a further 6.89%, respectively, and achieving best-in-class results on 10 sequences. Notably, CUVO-Full outperforms multi-frame methods in the texture-poor MH004 sequence. On sequences ME002, MH000, and MH001, CUVO lags behind the baseline, which is primarily attributed to the instability of the frozen flow network in these scenes. Additionally, in certain sequences, CUVO-Full exhibits higher error rates than CUVO. This suggests that applying multiple augmentation strategies simultaneously may introduce excessive noise. As evidenced in Table XVI (Appendix G), individual augmentations remain highly competitive: CUVO coupled solely with Temporal Reversal achieves an ATE of 0.57 on ME002, while Horizontal Mirroring excels on ME007 (0.45) and MH000 (1.48). Consequently, we suggest that adaptively selecting augmentation strategies based on scene characteristics is a promising direction for future optimization.

TUM-RGBD and ETH3D-SLAM datasets. To assess gen-

TABLE IV
ATE↓ RESULTS ON TUM-RGBD AND ETH3D DATASETS.

Methods		TUM-RGBD										ETH3D-SLAM
		360	desk	desk2	floor	plant	room	rpy	teddy	xyz	Avg	Avg
Multi-Frame (SLAM/VO)	ORB-SLAM3 [47]	✗	0.016	✗	✗	0.038	✗	✗	✗	0.005	-	-
	DSO [48]	✗	0.405	0.322	0.041	0.108	0.800	✗	✗	0.058	-	-
	DROID-SLAM [23]	0.111	0.018	0.042	0.021	0.016	0.049	0.026	0.048	0.012	0.038	0.010
	MambaVO++ [52]	0.085	0.016	0.032	0.027	0.025	0.056	0.026	0.029	0.009	0.034	-
	DROID-VO [23]	0.141	0.064	0.078	0.063	0.041	0.393	**0.030**	0.221	0.017	0.116	0.238
	DPVO [49]	0.156	0.034	0.050	0.183	0.034	0.383	0.038	0.073	**0.012**	0.107	**0.203**
	MambaVO [52]	**0.108**	**0.021**	**0.037**	**0.034**	**0.022**	**0.372**	0.031	**0.048**	0.013	**0.076**	-
Frame-to-Frame (End-to-End VO)	TartanVO [21]	**0.160**	0.478	0.539	0.348	0.395	0.417	0.050	**0.329**	0.160	0.320	0.421
	DytanVO [28]	0.188	0.159	0.224	**0.191**	0.343	0.530	0.053	0.508	0.131	0.259	0.364
	CUVO (Ours)	0.206	**0.078**	0.106	0.340	0.207	0.331	0.075	0.334	**0.061**	**0.193**	0.375
	CUVO-Full (Ours)	0.186	0.139	**0.092**	0.319	**0.126**	**0.309**	**0.046**	0.502	0.071	0.199	**0.215**

eralization in indoor environments, we test on TUM-RGBD sequences with "fr1" intrinsics and ETH3D-SLAM training sequences from set "1'

Quantitative results are presented in Table IV. On the TUM-RGBD benchmark, which features cluttered indoor scenes with subtle camera motions, CUVO achieves a remarkable performance gain, reducing the average ATE by 39.69% compared to the TartanVO. CUVO-Full also performs well with a 37.81% reduction, though slightly lower than CUVO. We attribute this to the fact that in already highly cluttered scenes, aggressive analogy augmentation may introduce excessive visual noise.

In stark contrast, on the ETH3D-SLAM dataset, characterized by cleaner and more structured environments, the benefit of augmentation is pronounced. While CUVO reduces the baseline (TartanVO) error by 10.93%, incorporating the full analogy augmentation (CUVO-Full) leads to a substantial 38% reduction. These diverging results underscore the critical importance of adaptively selecting augmentation strategies based on scene complexity and texture characteristics.

*2) **Qualitative Evaluation**:* As depicted in Fig. 9, we performed qualitative analysis on KITTI sequences 08 and 02 and TartanAir's MH006, comparing TartanVO [21], CUVO, and its variant CUVO-Full.

As TartanVO's pose network shares CUVO's first-layer convolution, we use TartanVO as the primary baseline. In columns 3–5, CUVO demonstrates enhanced feature extraction, improving the capture of static object features (e.g., trees, road signs, and buildings in KITTI; doors, pipes, and desks in TartanAir) and refining dynamic object features (e.g., cyclists and moving vehicles in KITTI; dynamic fog in TartanAir). CUVO-Full exhibits improved feature extraction accuracy over CUVO with increased data volume. In the final column, CUVO-Full's uncertainty visualization accurately identifies high-uncertainty regions, including low-texture areas and small distant dynamic objects. And we observed that uncertainty module can filter out some stationary vehicles. In TartanAir's dynamic environments, the Uncertainty Selection module precisely distinguishes high-exposure, low-light, and dynamic gas regions.

TABLE V
RESOURCE ANALYSIS COMPARISON OF OTHER METHODS.

Methods		Params (M)	GPU (GB)	FLOPs (G)	FPS
DROID-VO		4.00	21.22	89.44	6.36
TartanVO		24.51	0.22	53.86	80.03
DytanVO		163.04	3.20	238.69	6.18
CUVO	Flow	8.80	0.46	144.42	62.74
	Pose	19.92	0.26	1.59	
	Total	**28.72**	**0.72**	**146.01**	

C. Time and Memory

We evaluate the computational efficiency of CUVO against other methods on TartanAir and KITTI datasets using a NVIDIA RTX 4090 GPU and Intel Core i7-14700KF CPU. The quantitative comparison of model parameters, GPU memory, FLOPs, and inference speed is detailed in Table V.

Resource Efficiency. As shown in the table, CUVO maintains a highly competitive spatial footprint. With a total parameter count of 28.72M, it is significantly more compact than heavy transformer-based methods like DytanVO (163.04M). More importantly, our method is extremely memory-efficient, requiring only 0.72 GB of GPU memory. This stands in stark contrast to optimization-based approaches like DROID-VO, which demands a prohibitive 21.22 GB, making our method far more suitable for resource-constrained platforms.

Real-Time Performance. CUVO delivers an average speed of 62.74 FPS (15.94ms latency) on the KITTI and TartanAir benchmarks. This throughput is sufficient to support cameras operating at high sampling frequencies of up to 60 Hz. For standard benchmarks like KITTI, which operates at 10Hz (100ms interval), UCVO runs at over 6× the input frequency. This capability not only guarantees robust real-time performance but also leaves ample computational margin for concurrent downstream tasks.

D. Ablation Study

This section offers a detailed methodological analysis, comprising ablation studies on CUVO components (subsection

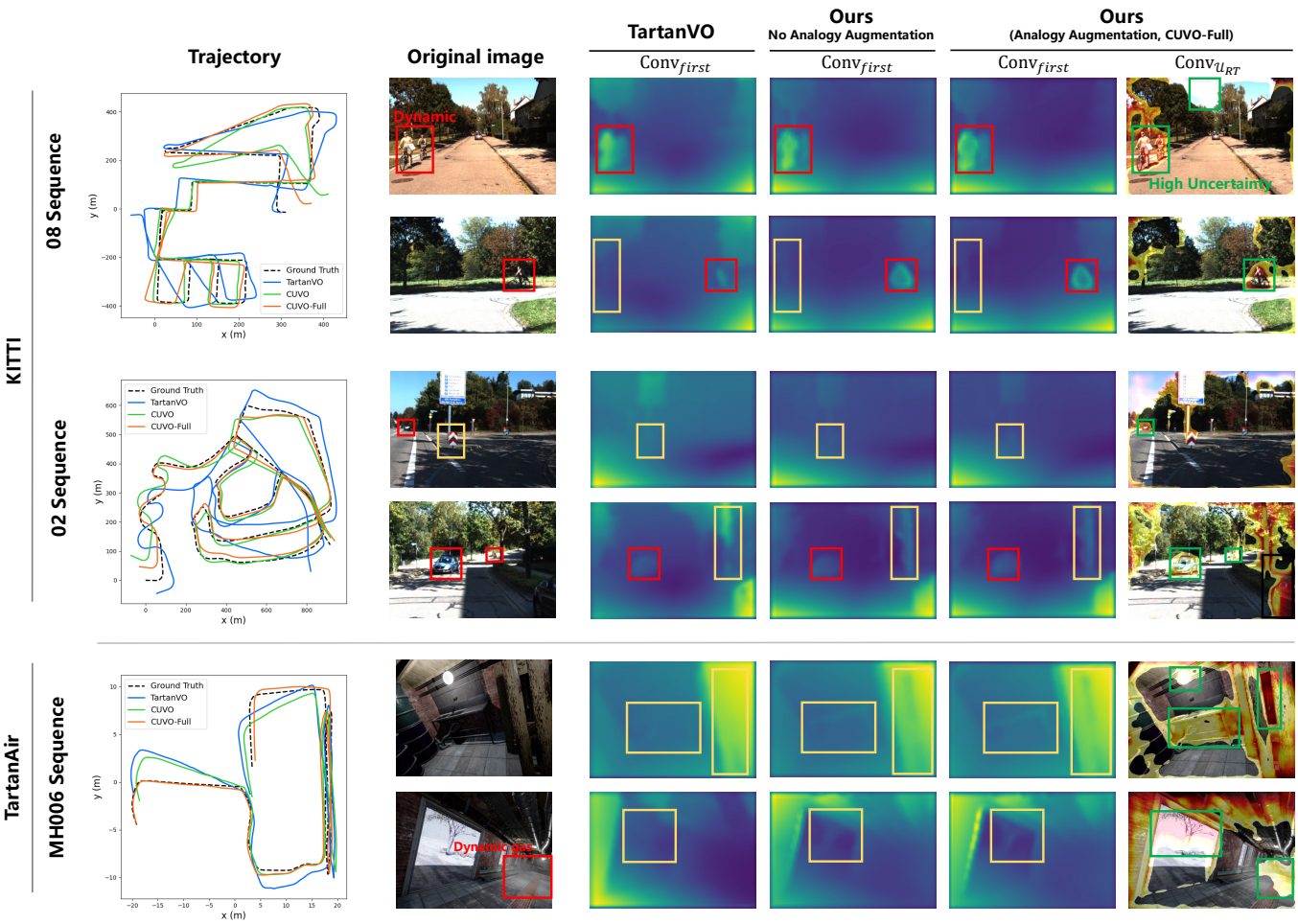

Fig. 9. Qualitative results. The first column shows sequence trajectories, the second column is the images extracted from the sequence, columns 3–5 visualize first-layer convolution outputs of the pose network, and the final column displays pose uncertainty maps integrating translation and rotation. **Annotations include:** red boxes for dynamic objects, yellow for feature details, and green for high uncertain regions (e.g., dynamic, high-brightness, low-light, and distant regions like sky).

TABLE VI
ABLATION EXPERIMENTS ON MODULES IN CUVO

	Methods	KITTI												TartanAir	FPS	GPU
		00	01	02	03	04	05	06	07	08	09	10	Avg	Avg	Avg	(GB)
CUVO	Baseline	76.91	144.58	56.44	3.61	3.48	22.61	57.92	7.32	39.97	24.74	9.78	40.67	2.75	70.10	0.59
	+C	93.00	**77.47**	61.81	**3.06**	1.58	37.51	**26.29**	11.29	48.95	20.69	11.41	35.73	**2.14**	63.88	0.71
	+U	66.20	217.34	98.83	3.96	**1.14**	26.15	38.44	11.13	**31.12**	**17.20**	6.04	47.05	2.60	69.04	0.70
	Ours	**37.61**	82.54	**44.90**	5.08	1.41	**16.38**	26.93	**4.70**	38.56	17.29	**5.96**	**25.58**	2.22	62.74	0.72

The Baseline is built upon TartanVO by replacing the original PWC-Net with Sea-RAFT. Note that Sea-RAFT is kept frozen without fine-tuning.
+ C denotes the Baseline incorporating only with the Context Attention, while + U represents the Baseline incorporating only the Uncertainty Selection. Ours (the full CUVO method) is shaded in gray.

IV-D1) and an examination of various analogy augmentation strategies (subsection IV-D3).

*1) **Ablation Study of CUVO Components**:* To further analyze the contributions of individual components in CUVO, we conducted a comprehensive ablation study evaluating the impact of the Context Attention (+C) and Uncertainty Selection (+U) modules on ATE performance, as detailed in Table VI. We evaluated different variants on the KITTI and TartanAir benchmarks. Our Baseline is established by adapting the TartanVO architecture, replacing its original PWC-Net backbone with a frozen Sea-RAFT network. This Baseline relies solely on optical flow for pose regression, excluding our proposed modules.

Impact of Context Attention (+C). The variant incorporating only Context Attention reduces the ATE by 12.15% on KITTI and 22.18% on TartanAir compared to the Baseline. This module excels in sequences rich in structural details but lacking in dynamic interference. For instance, in KITTI-06, the +C variant achieves the best performance (ATE 26.29), outper-

TABLE VII
ABLATION EXPERIMENTS IN TRAINING STAGE

Methods	KITTI												TartanAir
	00	01	02	03	04	05	06	07	08	09	10	Avg	Avg
Single-Stage	53.26	98.07	125.06	5.42	3.28	35.66	39.34	10.84	**32.07**	17.55	11.26	39.26	2.77
Two-Stage (Ours)	**37.61**	**82.54**	**44.90**	**5.08**	**1.41**	**16.38**	**26.93**	**4.70**	38.56	**17.29**	**5.96**	**25.58**	**2.22**

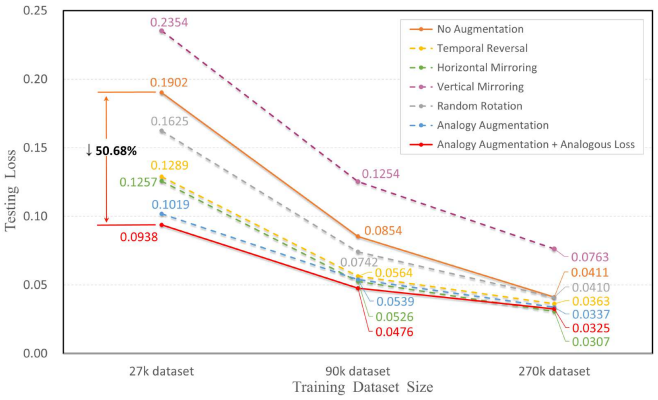

Fig. 10. Comparison of the first stage training final testing loss after 200 epochs using different Analogy Augmentation methods versus No Augmentation data across dataset sizes of 27k, 90k, and 270k.

forming the Baseline (57.92). This suggests that in relatively static environments, leveraging semantic texture information is the primary driver for accurate pose estimation.

Impact of Uncertainty Selection (+U). Conversely, the Uncertainty Selection (+U) variant excels in dynamic scenarios. In sequences featuring moving objects (e.g., KITTI-04, 08, 09), +U outperforms +C; notably, on KITTI-08, +U achieves an ATE of 31.12 compared to 48.95 for +C. This confirms that filtering unreliable optical flow via uncertainty estimation is critical in dynamic regions.

Synergy in the Full Model. The complete CUVO model integrates both modules to handle complex environments. While individual modules may lead in specific scenarios, the Full model achieves the lowest Average ATE on KITTI, significantly outperforming +C and +U. On TartanAir, which lacks dynamic elements, the Full model performs comparably to the +C variant. Thus, the Full model retains the necessary robustness for unpredictable real-world deployment.

Computational Efficiency. As detailed in Table VI, our proposed modules incur minimal computational overhead. The Uncertainty Selection (+U) module is lightweight, causing a negligible drop from 70.10 FPS (Baseline) to 69.04 FPS, while the Context Attention (+C) module reduces speed slightly to 63.88 FPS. Overall, the full CUVO model maintains a real-time speed of 62.74 FPS with only a marginal increase in GPU memory (0.59GB $\rightarrow$ 0.72GB), ensuring efficient deployment without imposing a heavy hardware burden.

*2) **Ablation Study in Training Stage**:* In a single-stage end-to-end pipeline, the pose network is forced to learn from estimated optical flow, which inherently contains noise and errors during the early training phases. This error propagation from the flow network can lead to suboptimal convergence of the pose estimator.

To address this, our two-stage strategy acts as a form of curriculum learning. We first utilize ground-truth optical flow to decouple pose learning from flow estimation errors, ensuring the pose network learns optimal feature extraction from accurate motion cues. To validate the necessity of our two-stage training strategy, we conducted a comparative ablation study as presented in Table VII. We compared this against a single-stage baseline where the model is trained directly from image sequences. The results confirm the superiority of our approach, yielding a substantial ATE reduction of 34.84% on KITTI and 19.86% on TartanAir, demonstrating that stable initialization with ground-truth flow is critical for robustness.

*3) **Ablation Study of Analogy Augmentation Strategy**:* Firstly, to verify the **scalability** of analogy augmentation methods, we demonstrate the effect of applying the ground truth optical flow analogy augmentation strategy to a general pose network, which corresponds to the first stage of training. Concurrently, to comprehensively validate the **effectiveness** of image-to-pose analogy augmentation strategy, we present detailed results of CUVO combined with different augmentation strategies, which corresponds to the second stage of training.

In the first stage, the pose network is trained exclusively on ground-truth optical flow from TartanAir training scenes 1-32 (see Appendix F for setup), with various augmentation strategies applied. It is important to note that the pose network accepts optical flow directly as input. However, given the absence of ground-truth flow in standard benchmarks, we utilize a frozen SEA-RAFT [41] model to generate the necessary optical flow inputs for each image pair during evaluation.

Comparison of First-stage Testing Loss. We present testing performance of the general pose network under different data volumes, evaluating various analogy augmentation strategies. This evaluation also includes testing full analogy augmentation strategies with and without the analogous loss, where both strategies use the same test set and loss function, enabling direct comparison.

Fig. 10 compares testing losses against the No Augmentation baseline (orange solid line). Temporal Reversal (yellow dashed line) significantly lowers loss by enhancing robustness to temporal dynamics. Horizontal Mirroring (green dashed line) also performs well by preserving physical consistency. Conversely, Vertical Mirroring (purple dashed line) yields the highest loss, as it disrupts semantic priors (e.g., gravity) and causes distribution mismatch. Random Rotation (gray dashed line) offers only slight improvements. One possible explanation is that the random rotation operation introduced

TABLE VIII
ATE RESULTS FOR POSE NETWORK TRAINED ON TARTANAIR WITH VARIOUS AUGMENTATION METHODS, EVALUATED ON KITTI, TARTANAIR TEST DATASET, AND EUROC.

Augmentation Methods	TartanAir			KITTI			EuRoC
	27k	90k	270k	27k	90k	270k	270k
No Augmentation	7.63	5.18	4.23	117.20	69.30	40.58	1.27
Temporal Reversal	6.49	4.47	**4.02**	102.66	40.73	**26.87**	1.23
Horizontal Mirroring	6.09	4.60	4.23	107.95	37.50	39.29	**1.12**
Vertical Mirroring	7.84	6.15	5.10	103.79	42.61	54.53	2.10
Random Rotation	7.38	4.72	4.21	80.40	44.82	36.65	1.22
Analogy Augmentation	5.29	**4.34**	4.41	74.00	34.70	28.59	1.78
Analogy Augmentation + Analogous Loss	**5.28**	4.70	4.15	**66.63**	**30.40**	29.66	1.18

TABLE IX
ATE RESULTS FOR CUVO TRAINED ON TARTANAIR WITH VARIOUS AUGMENTATION METHODS, EVALUATED ON KITTI, TARTANAIR TEST DATASET, AND EUROC.

	Augmentation Methods	TartanAir			KITTI			EuRoC
		27k	90k	270k	27k	90k	270k	270k
CUVO	No Augmentation	5.16	2.82	2.22	110.13	52.60	25.58	0.96
	Temporal Reversal	4.24	**2.48**	2.05	**46.14**	**29.83**	24.01	0.91
	Horizontal Mirroring	**3.50**	2.60	**1.85**	121.22	35.98	**20.65**	**0.71**
	Random Rotation	4.32	2.89	2.40	128.92	75.29	56.31	0.83
	Full	3.64	2.80	1.99	84.45	34.63	23.74	0.97

excessive noise, which hindered model convergence during the initial training phase. This effect is particularly prominent with smaller datasets. Another possibility is that the 360° rotation introduced data samples that are similar to existing data domains, such as vertical mirroring, thus diminishing the overall augmentation benefit.

Consequently, our analogy augmentation integrates Temporal Reversal, Horizontal Mirroring, and Random Rotation (excluding Vertical Mirroring). Using No Augmentation as the baseline method (orange solid line), this approach substantially reduces testing loss (blue dashed line), achieving a 46.4% reduction for the 27k dataset, demonstrating enhanced model performance in data-limited conditions. Combining analogy augmentation with analogous loss yields the best testing loss performance (gray solid line), with a 50.7% reduction for the 27k dataset as shown in Fig. 10, an additional 8% improvement over analogy augmentation alone (46.4%), confirming the incremental benefit of analogous loss. Despite increased training volatility from multiple objectives, this combination significantly enhances generalization in data-limited conditions.

Comparison of First-stage ATE Result. We evaluate the Absolute Trajectory Error (ATE) of the first-stage pose network on the KITTI [18], TartanAir [25], and EuRoC [46] datasets, as reported in Table VIII.

For the KITTI dataset, camera motion is primarily translational, and the long road segments exhibit strong temporal dependencies. Therefore, Temporal Reversal demonstrated superior zero-shot performance compared to other geometric transformation methods on this dataset, as detailed in Table VIII. Despite the varied and complex environments of the TartanAir test dataset, our Analogy Augmentation and Analogical Loss method consistently exhibits strong performance and competitiveness. However, performance on the EuRoC dataset is suboptimal. This limitation is primarily attributed to the grayscale nature of the imagery, which lacks the chromatic information relied upon by our network (see Appendix C for a detailed analysis). Overall, when compared to the "No Augmentation" baseline, our various analogy augmentation strategies demonstrated superior zero-shot performance with the sole exception of Vertical Mirroring, providing a robust pre-training foundation for the second stage.

In the second stage, Given the suboptimal performance of vertical mirroring in the first stage, we employed temporal reversal, horizontal mirroring, and random rotation combined with analogous loss. We continued training CUVO on TartanAir datasets of 27k, 90k, and 270k images, using pre-trained weights from the first-stage pose network. We evaluated the ATE on the TartanAir test, KITTI, and Euroc datasets to assess the generalization capabilities of various augmentation methods.

Table IX presents ATE results across dataset sizes and augmentation methods. Horizontally, ATE decreases as dataset size increases (27k $\rightarrow$ 90k $\rightarrow$ 270k), indicating that larger training datasets enhance model performance by learning more robust feature representations, reducing overfitting, and improving generalization. Vertically, we compare augmentation methods against a no-augmentation baseline (No Augmentation). Temporal reversal significantly enhances generalization in data-limited conditions, achieving a 58.1% ATE reduction on KITTI at 27k, excelling in long-sequence tasks. Horizontal mirroring demonstrates optimal performance across all three datasets at 270k. Random rotation performs suboptimally, as 360° random rotation for image pair optical flow estimation introduces excessive noise compared to first-stage optical flow augmentation. However, random rotation improves performance on Euroc, indicating its effectiveness for datasets with significant rotational motion, such as drone data. To mitigate noise, we recommend fixed-angle rotation augmentation with a limited number of angles.

Training with three methods excluding vertical mirroring (Full) significantly improves performance. Specifically, on the 270k dataset, ATE is reduced by 10.36% on TartanAir and 7.19% on KITTI. Furthermore, on the data-limited 27k dataset, ATE is substantially reduced by 29.5% on TartanAir and 23.3% on KITTI. However, Full shows no improvement on EuRoC, where grayscale images lead to suboptimal optical flow inference by CUVO. Excessive augmentation introduces noise, reducing optical flow accuracy and degrading performance. These findings indicate the strong robustness of our analogy augmentation strategy, while also highlighting the necessity of selecting augmentation methods in accordance with each dataset's unique domain properties.

V. CONCLUSION

This paper proposes a novel analogy augmentation framework to enhance the performance of learning-based end-to-end visual odometry. First, we present the Context Attention Uncertainty-aware Visual Odometry Network (CUVO), enhancing the accuracy and robustness of pose estimation in complex dynamic environments. Additionally, we improve data consistency and diversity through an image-pair-to-pose

analogy augmentation strategy and an analogous loss function, addressing challenges of limited training data, data scarcity, and homogeneity, thereby optimizing model performance.

Extensive experiments demonstrate that our framework outperforms most visual odometry and even SLAM algorithms, representing the first image-to-pose augmentation method tailored for visual odometry. By tackling data utilization and diversity challenges, our approach establishes a foundation for more robust and accurate visual odometry systems in autonomous robot applications, paving the way for future advancements in dynamic and diverse environments.

Limitations. Regarding the network model, CUVO primarily relies on inter-frame optical flow for pose estimation, which is inherently less accurate compared to multi-frame methods. Regarding the analogy augmentation strategy, the full-scale analogy augmentation joint method may introduce noise in certain scenarios, necessitating careful domain-specific data selection to mitigate potential artifacts.

Future Work. We plan to address these challenges by integrating multi-modal sensors (e.g., IMU) to recover true metric scale and extending the architecture to incorporate multi-frame temporal optimization for improved consistency. We also intend to explore adaptive selection mechanisms of the augmentation strategy to further reduce potential artifacts in diverse domains.

REFERENCES

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.

[3] C. Wang, G. Zhang, Z. Cheng, and W. Zhou, "Kpdepth-vo: Self-supervised learning of scale-consistent visual odometry and depth with keypoint features from monocular video," *IEEE Trans. Circuits Syst. Video Technol.*, 2025.

[4] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "Dynaslam: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4076–4083, 2018.

[5] X. Hu, Y. Wu, M. Zhao, L. Yang, X. Zhang, and X. Ji, "Pas-slam: A visual slam system for planar-ambiguous scenes," *IEEE Trans. Circuits Syst. Video Technol.*, 2024.

[6] R. Roberts, H. Nguyen, N. Krishnamurthi, and T. Balch, "Memory-based learning for visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 47–52, 2008.

[7] X. Hu, Y. Wu, M. Zhao, Z. Cao, X. Zhang, and X. Ji, "Dyo-slam: Visual localization and object mapping in dynamic scenes," *IEEE Trans. Circuits Syst. Video Technol.*, 2025.

[8] T. A. Ciarfuglia, G. Costante, P. Valigi, and E. Ricci, "Evaluation of non-geometric methods for visual odometry," *Robot. Auton. Syst.*, vol. 62, no. 12, pp. 1717–1730, 2014.

[9] K. Konda and R. Memisevic, "Learning visual odometry with a convolutional network," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, vol. 2, pp. 486–490, 2015.

[10] Y. Cao, X.-S. Zhang, F. Luo, C. Lin, and Y.-J. Li, "Unsupervised visual odometry and action integration for pointgoal navigation in indoor environment,"*IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 6173–6184, 2023.

[11] R. Wang, S. M. Pizer, and J.-M. Frahm, "Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5555–5564, 2019.

[12] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 2043–2050, 2017.

[13] Y. Huang, B. Zhao, C. Gao, and X. Hu, "Learning optical flow with r-cnn for visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 14,410–14,416, 2021.

[14] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pp. 969–977, 2018.

[15] L. Zhang, P. Ratsamee, Z. Luo, Y. Uranishi, M. Higashida, and H. Takemura, "Panoptic-level image-to-image translation for object recognition and visual odometry enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 938–954, 2023.

[16] U.-H. Kim, S.-H. Kim, and J.-H. Kim, "Simvodis++: Neural semantic visual odometry in dynamic environments," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4244–4251, 2022.

[17] U.-H. Kim, S.-H. Kim, and J.-H. Kim, "Simvodis: Simultaneous visual odometry, object detection, and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 428–441, 2020.

[18] 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)*, pp. 3354–3361, 2012.

[19] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2012, pp. 573–580.

[20] T. Schops, T. Sattler, and M. Pollefeys, "Bad slam: Bundle adjusted direct rgb-d slam," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 134–144.

[21] W. Wang, Y. Hu, and S. Scherer, "Tartanvo: A generalizable learning-based vo," in *Proc. Conf. Robot Learn. (CoRL)*, pp. 1761–1772, 2021.

[22] R. Song, R. Li, Z. Xiao, and B. Yan, "Swinfvo: Self-supervised visual odometry with enhanced global spatiotemporal perception," *IEEE Trans. Circuits Syst. Video Technol.*, 2025.

[23] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 16,558–16,569, 2021.

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

[25] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pp. 4909–4916, 2020.

[26] P. Kirsanov, A. Gaskarov, F. Konokhov, K. Sofiiuk, A. Vorontsova, I. Slinko, D. Zhukov, S. Bykov, O. Barinova, and A. Konushin, "Discoman: Dataset of indoor scenes for odometry, mapping and navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pp. 2470–2477, 2019.

[27] K. Alomar, H. I. Aysel, and X. Cai, "Data augmentation in classification and segmentation: A survey and new strategies," *J. Imaging*, vol. 9, no. 2, p. 46, 2023.

[28] S. Shen, Y. Cai, W. Wang, and S. Scherer, "Dytanvo: Joint refinement of visual odometry and motion segmentation in dynamic environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 4048–4055, 2023.

[29] J. Zhang, J. Li, J. Li, Y. Sun, X. Liu, Z. Zheng, and G. Lu, "Mbrvo: A blur robust visual odometry based on motion blurred artifact prior," *IEEE Robot. Autom. Lett.*, 2024.

[30] D. Gentner, "Structure-mapping: A theoretical framework for analogy," *Cognit. Sci.*, vol. 7, no. 2, pp. 155–170, 1983.

[31] M. M. Veloso and J. G. Carbonell, "Derivational analogy in prodigy: Automating case acquisition, storage, and utilization," *Mach. Learn.*, vol. 10, no. 3, pp. 249–278, 1993.

[32] F. Hill, A. Santoro, D. G. Barrett, A. S. Morcos, and T. Lillicrap, "Learning to make analogies by contrasting abstract relational structure," *arXiv preprint arXiv:1902.00120*, 2019.

[33] W. Ye, X. Lan, S. Chen, Y. Ming, X. Yu, H. Bao, Z. Cui, and G. Zhang, "Pvo: Panoptic visual odometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 9579–9589, 2023.

[34] G. Wang, J. Zhong, S. Zhao, W. Wu, Z. Liu, and H. Wang, "3d hierarchical refinement and augmentation for unsupervised learning of depth and pose from monocular video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1776–1786, 2022.

[35] S. Han, M. Li, H. Tang, Y. Song, and G. Tong, "Uvmo: Deep unsupervised visual reconstruction-based multimodal-assisted odometry," *Pattern Recognit.*, vol. 153, p. 110573, 2024.

[36] S. Chen, Z. Pu, X. Fan, and B. Zou, "Fixing defect of photometric loss for self-supervised monocular depth estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1328–1338, 2021.

[37] J. Li, S. Dong, Y. Gong, Y. He, and X. Wei, "Analogical learning-based few-shot class-incremental learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5493–5504, 2024.
[38] L. Liu, J. Zhang, R. He, Y. Liu, Y. Wang, Y. Tai, D. Luo, C. Wang, J. Li, and F. Huang, "Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6489–6498, 2020.
[39] K. Luo, C. Wang, S. Liu, H. Fan, J. Wang, and J. Sun, "Upflow: Upsampling pyramid for unsupervised optical flow learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1045–1054, 2021.
[40] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8934–8943, 2018.
[41] Y. Wang, L. Lipson, and J. Deng, "Sea-raft: Simple, efficient, accurate raft for optical flow," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 36–54, 2024.
[42] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, no. 07, pp. 11,908–11,915, 2020.
[43] H.-W. Chen, T.-H. Liao, H.-K. Yang, and C.-Y. Lee, "Pixel-wise prediction based visual odometry via uncertainty estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 2518–2528, 2023.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.
[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention (MICCAI)*, pp. 234–241, 2015.
[46] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016.
[47] C. Campos, R. Elvira, J. J. G. Rodríguez, J. 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.
[48] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, 2017.
[49] Z. Teed, L. Lipson, and J. Deng, "Deep patch visual odometry," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 39,033–39,051, 2023.
[50] X. Gao, R. Wang, N. Demmel, and D. Cremers, "Ldso: Direct sparse odometry with loop closure," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pp. 2198–2204, 2018.
[51] L. Lipson, Z. Teed, and J. Deng, "Deep patch visual slam," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 424–440, 2024.
[52] S. Wang, W. Li, Y. Wang, Z. Fan, Z. Huang, X. Cai, J. Zhao, and D. Li, "Mambavo: Deep visual odometry based on sequential matching refinement and training smoothing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1252–1262, 2025.
[53] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4104–4113, 2016.
[54] C. M. Parameshwara, G. Hari, C. Fermüller, N. J. Sanket, and Y. Aloimonos, "Diffposenet: Direct differentiable camera pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6845–6854, 2022.
[55] R. Murai, E. Dexheimer, and A. J. Davison, "Mast3r-slam: Real-time dense slam with 3d reconstruction priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16,695–16,705, 2025.

VI. Biography Section

Jituo Li received the PhD degree from Zhejiang University in 2006. He is an associate professor with the School of Mechanical Engineering, Zhejiang University. Before moving back to Zhejiang University in June 2010, he was with the Institute of Automation, Chinese Academy of Sciences. His research interests include intelligent robotics and CAD/CG.

Shunwang Sun received the B.S. degree in Robotics Engineering from Yanshan University in 2024. He is currently recommended to pursue the M.E. degree in Mechanical Engineering at Zhejiang University. His research interests include creating new generations of robots, navigation, localization, and control of mobile robots.

Tingxi Xue received the B.S. degree from Dalian Maritime University in 2023. He is currently working toward the master's degree with the College of Engineering, Zhejiang University.

Xinqi Liu received the B.S. degree from the Xi'an University of Technology in 2018 and the PhD degree from the College of Mechanical Engineering, Zhejiang University, China. He is a assistant professor with the School of Artificial Intelligence and Robotics, Hunan University (HNU). His research interests include 3D vision, AIGC and embodied intelligence.

Jialu Zhang received the B.S. degree from Shijiazhuang Tiedao University in 2021. He is currently working toward the PhD degree with the College of Mechanical Engineering, Zhejiang University. His research interests include computer vision and data-driven techniques.

Huixu Dong received the Ph.D. degree in robotics from the Robotics Research Centre, Nanyang Technological University in 2018. Since 2022, he has been a New Hundred-Talent Program Faculty, directing Grasp Laboratory, Zhejiang University, China. His current research interests include robotic perception and grasp in unstructured environments, and construction of robotic gripper.

He is an Associate Editor of IEEE Robotics and Automation Letters, IEEE Transactions on Automation Science and Engineering, ICRA 2023/2024, IROS 2022/2023/2024, and AIM 2022/2023/2024.

Guodong Lu received the B.S. master's and PhD degrees from Zhejiang University in 1983, 1993 and 2000, respectively. He is a professor with the College of Mechanical Engineering, Zhejiang University. He is the executive vice-director of Zhejiang University Robotics Institute.