SUM-VO: Semantic Uncertainty-based Monocular Visual Odometry in Outdoor Scenes

***Abstract*—Direct SLAM methods typically estimate the poses by minimizing the photometric error represented by intensity of grayscale images. However, due to the inherent non-convexity of images and sensitivity to illumination changes caused by the camera exposure in ambient outdoor scenes, existing methods often yield sub-optimal ego-pose estimates. In this work, we propose a novel approach, Semantic Uncertainty-based Monocular Visual Odometry (SUM-VO), which estimates poses by minimizing semantic uncertainty errors instead of photometric errors. Compared to grayscale, semantic uncertainty offers a more stable pixel-wise representation of the view especially under exposure changes, which encodes sparse yet essential semantic and geometric context for pose estimation. Leveraging semantic uncertainty as a natural byproduct of the segmentation network without extra training, we deeply integrate it with the predicted semantic labels to mitigate dynamic interference, an unavoidable challenge in outdoor scenes. To this end, we design the Semantic-Aware Pose Estimation (SAPE) module, which tightly couples label-based priors with uncertainty-aware motion coherence. This unified iterative optimization excludes large dynamic objects while concurrently refining the overall pose and geometry estimation. Experimental results on the KITTI dataset demonstrate that our method achieves outstanding performance in outdoor monocular odometry. We also evaluate our approach on representative sequences from the Complex Urban Dataset and a real-world Campus dataset collected by ourselves, verifying its promising generalizability and competitive performance across rural and urban scenes.**

***Index Terms*—Visual odometry, semantic information, autonomous vehicle**

***Note to Practitioners*—This work is motivated by the practical challenge of reliable ego-localization for autonomous vehicles and ground robots in dynamic and illumination-varying outdoor environments. Traditional vision-based navigation systems often experience severe performance degradation when camera exposure changes or when moving objects dominate the scene. To address these pain points, we propose replacing traditional light-sensitive tracking metrics with a novel semantic uncertainty representation. This framework offers a robust and ready-to-use solution for engineering practitioners. It utilizes pre-trained semantic segmentation networks without requiring any domain-specific fine-tuning or additional training for the uncertainty metric itself. By tightly coupling this semantic uncertainty with a pose estimation module, the proposed system inherently filters out dynamic interference and resists environmental lighting variations. To further elevate the system reliability under extreme conditions such as prolonged visual occlusion, future engineering efforts can integrate this pure vision framework with low-cost inertial sensors to form a resilient multi-sensor navigation suite.**

I. Introduction

SIMULTANEOUS localization and mapping (SLAM) is a fundamental component in robotics and various vision applications. After decades of development, visual SLAM systems can be generally classified as feature-based methods [1]–[3] and direct methods [4]–[6]. Feature-based methods

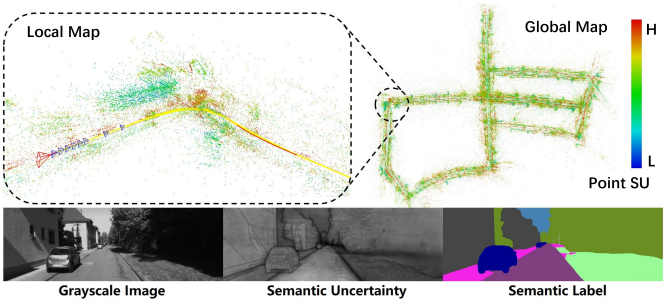


Fig. 1. Illustration of the estimated pose trajectory and reconstructed sparse semantic point cloud on the KITTI 05 sequence using SUM-VO. The color of points represents semantic uncertainty. Below are the grayscale image, semantic uncertainty map, and predicted semantics from left to right.

detect feature points or lines and use descriptors to establish correspondences, while direct methods estimate camera poses and reconstruct geometry by minimizing photometric errors, yielding a map similar to that shown in Fig.1. Although feature matching enables accurate pose estimation and mapping, its performance degrades in unstructured or textureless environments. In contrast, direct methods have demonstrated robust performance under such challenging conditions [7]. Beyond traditional approaches, learning-based methods [8], [9] integrate deep networks into classic geometry paradigm to enhance localization accuracy and robustness.

Despite their potential, existing direct methods depend on photometric error, defined as the difference in intensity between corresponding projected points across consecutive frames, which assumes a certain level of grayscale consistency, *i.e.*, stable or nearly stable pixel intensities over time. However, in outdoor scenes, particularly when vehicles traverse urban canyons or pass through dappled light in rural woodlands, camera exposure may undergo abrupt shifts, leading to global photometric shifts. As shown by three consecutive frames in Fig. 2 c), the entire scene brightens abruptly in the middle frame as the vehicle exits the overpass. These changes also cause spatially uneven illumination variations, as seen in the boxed region of Fig. 2 a), further violating the photometric consistency. Consequently, trajectory accuracy degrades, as illustrated on the left of Fig. 2 b), where the enlarged segment from Fig. 2 a) reveals a marked increase in relative pose error due to disturbed estimation.

Another significant challenge in outdoor environments stems from the interference of dynamic objects, particularly moving vehicles. While direct methods can inherently suppress the influence of small dynamic elements by iteratively rejecting outlier points during photometric optimization to reduce the risk of substantial drift, large or persistent dynamic obstacles still introduce considerable inconsistencies into photometric error items and pose estimation.

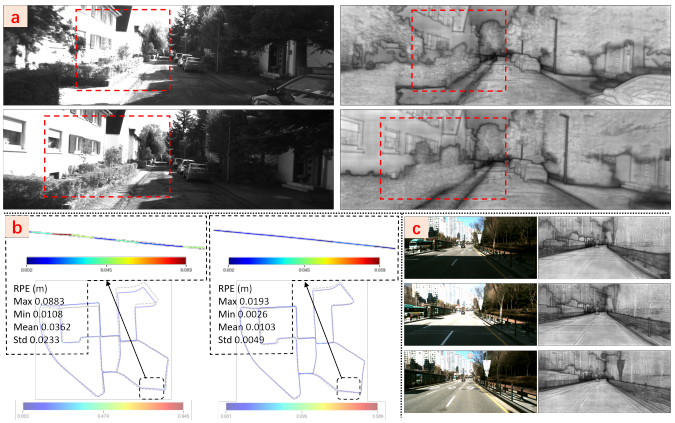


Fig. 2. Analysis of abrupt exposure changes in outdoor scenes. a) A segment from KITTI-00 with grayscale images and semantic uncertainty maps; top: overexposed, bottom: recovered. b) Trajectories with relative pose error comparison on KITTI-00 using photometric error (left) and semantic uncertainty error (right), with the segment from a) and its corresponding relative pose error highlighted. Semantic uncertainty maps are scaled to the 0–255 range for visualization. c) Another segment from KAIST-27 showing drastic global illumination shifts beneath an overpass.

Advances in semantic understanding have enabled the acquisition of pixel-wise label information, facilitating the approximate identification of potential dynamic objects and thereby keeping the odometry system vigilant. Moreover, an intuitive insight suggests that the semantic features exhibit greater stability against photometric variations induced by camera exposure than grayscale intensities. Building on these two aspects, we propose a novel semantic uncertainty-based monocular visual odometry (SUM-VO) in outdoor scenes. Inspired by prior studies that characterize network uncertainty through loss gradients, we quantify the pixel-wise semantic uncertainty as the norms of the gradients of the pixel-wise cross entropy loss. We reformulate the incremental ego-pose estimation problem (*i.e.*, tracking) by minimizing a novel semantic uncertainty error in place of conventional photometric error. As shown on the right side of Fig. 2 a) and c), semantic uncertainty is visually less sensitive to short-term varying illumination conditions caused by exposure fluctuations in outdoor scenes, while implicitly encoding stable semantic boundary context and preserving texture information vital for odometry. The trajectory on the right side of Fig. 2 b) demonstrates that using semantic uncertainty error reduces the impact of exposure variations in challenging segments, yielding lower mean and variance of relative pose error in the highlighted region. Notably, our heuristic formulation of semantic uncertainty is independent of neural network architectures and requires no additional modules or training.

To mitigate severe trajectory drift caused by large moving objects, we further exploit pixel-wise semantic labels to identify and exclude dynamic points. We seamlessly integrate the filter of dynamic points, grouped by semantic regions, into iterative pose optimization, forming the basis of our Semantic-Aware Pose Estimation (SAPE) module. Using semantics from the host frame, we make a prior assessment of object mobility at the topmost level of optimization and estimate poses by only points from static regions. If applying the estimated pose to a movable semantic region results in high semantic uncertainty error, the region is deemed dynamic and excluded from the optimization and point selection in the target frame.

Experiments on the KITTI dataset [10] demonstrate that our method achieves competitive performance compared to state-of-the-art monocular visual odometry approaches. Despite the learning-based semantic segmentation, experiments on urban scenes from the Complex Urban dataset [11] by KAIST and our self-collected Campus dataset confirm the promising generalizability without network fine-tuning. Fig. 1 depicts the estimated pose trajectory and reconstructed sparse semantic point cloud map generated by SUM-VO. In summary, our contributions are as follows:

- We design a novel monocular visual odometry approach in outdoor scenes that leverages semantic uncertainty to handle non-convexity and exposure-induced inconsistencies by optimizing semantic uncertainty error.
- We introduce a semantic-aware pose estimation module that integrates continuous semantic uncertainties with discrete semantic labels. This approach enables a unified optimization of pose and geometry estimation while mitigating the impact of dynamic objects significantly obstructing the field of view during odometry.
- Our odometry system achieves competitive performance on the KITTI dataset and demonstrates promising generalizability on the Complex Urban dataset as well as our self-collected dataset.

II. Related Work

A. Feature-based and Direct Visual SLAM

Over the past decades, visual SLAM has evolved into two primary paradigms: feature-based and direct methods. Early works predominantly focused on feature-based methods utilizing hand-crafted detectors and descriptors like Shi-Tomasi [12], FAST [13], KLT [14], and ORB [15] to extract and track feature points for pose estimation. Notably, ORB-SLAM [3] employs ORB features as a unified basis for tracking, mapping, relocalization, and loop closing, achieving high robustness in textured environments. In contrast, direct methods estimate pose by minimizing photometric error across grayscale frames, assuming photometric consistency in dense [4], semi-dense [5], or sparse [6] forms. They are intuitively and empirically more accurate in low-texture environments than feature-based methods. DSO [6] skips explicit feature matching through joint geometric and pose optimization directly, while LDSO [16] integrates loop closure detection using a feature-based bag-of-words (BoW) [17]. Moreover, learning-based approaches integrate deep networks into geometric frameworks to improve localization accuracy and robustness. DROID-SLAM [8] constructs an end-to-end differentiable recurrent architecture with dense bundle adjustment. DPV-SLAM [9] enhances the sparse patch-based optical flow updates and incorporates the camera proximity factor for loop closing, achieving substantially higher efficiency.

B. Semantic-enhanced Visual Odometry

Recently, the use of semantic information in visual odometry has gained attention, contributing to improvements in

feature matching, pose estimation, dynamic object removal, and semantic mapping. Most existing semantic SLAM systems are built upon feature-based frameworks. For example, SIVO [18] introduces neural network uncertainty by information theory into the feature selection process, using a Bayesian neural network to incorporate the classification entropy of features into the new features. VSO [19] proposes the semantic reprojection error, which measures the distance of points to the nearest region with the same semantic label and incorporates semantic error terms to optimize intermediate continuous tracking points. Object-oriented SLAM methods further expand the application of feature-based SLAM by integrating object segmentation [20]–[22]. For instance, QuadricSLAM [23] utilizes 2D object detection from multiple views to estimate 3D quadrics and localize camera positions. CubeSLAM [24] leverages 2D object detections and vanishing points to generate high-quality 3D bounding box proposals, enabling object-level data association even without depth sensors. DSP-SLAM [25] utilizes category-specific deep shape priors to reconstruct dense 3D models of foreground objects from sparse SLAM points while integrating them into a joint factor graph optimization alongside background landmarks and camera poses. Bowman *et al.* [26] estimates the correspondences between objects and observations via expectation maximization (EM). While these methods provide rich topological information, they often require more computational resources for object proposal generation and data association. Other approaches incorporate semantic cues to tackle specific challenges within SLAM. AVP-SLAM [27] obtains IPM (inverse perspective mapping) images and employs guide lines to assist visual localization, enabling map reconstruction in GPS-denied parking lot. SMORE-SLAM [28] utilizes semantic information from backgrounds, object-levels, and keypoints to facilitate reverse loop closure detection, thereby resolving the scale drift issues typically in outdoor scenes.

Semantic direct SLAM, despite its potential, remains less explored due to the difficulty of integrating discrete semantic labels into continuous photometric optimization. DeepFactors [29] presents a real-time probabilistic dense SLAM system that parametrizes scene geometry into learned compact latent codes and integrates them into a factor graph framework to jointly optimize photometric, reprojection, and sparse geometric errors. SDVO [30] attempts to jointly align category-wise semantic probability residuals and image intensity through gradient-based optimization, yet its applicability is constrained by a reliance on manually selected semantic categories that possess limited intra-class texture. Furthermore, a common limitation in these works is that semantics are often treated as an external constraint, rather than being deeply coupled with the geometric information of the tracking process. This gap motivates our work to incorporate semantic uncertainty directly into the pose and geometry optimization.

C. Semantic-assisted Visual SLAM in Dynamic Scenes

A fundamental assumption of traditional SLAM is the static world, which is frequently violated in real-world applications. Besides semantic-aided feature matching and tracking, dynamic object removal is crucial in SLAM, as dynamic objects can cause severe incorrect correspondences and tracking errors. Thanks to semantics, it becomes possible to identify potentially movable objects beforehand [31]. DS-SLAM [32] combines SegNet [33] outputs with epipolar constraints to detect motion inconsistencies. DynaSLAM [34] leverages Mask R-CNN [35] for more precise pixel-level segmentation and exploits historical observations for background filling to handle dynamic occlusions.

However, a movable object is not necessarily moving, and discarding such stationary landmarks can severely degrade localization accuracy, especially in cluttered urban environments. To tackle this, recent studies focus on motion consistency checks. Bao *et al.* [36] propose a point group consistency check to distinguish between static and dynamic instances of the same semantic class. TwistSLAM [37] introduces mechanical joint constraints between semantic clusters such as the road plane and observed objects to perform constrained estimation of both the poses and velocities of moving instances. SG-SLAM [38] integrates semantic segmentation with epipolar geometry to classify features as static or dynamic while adaptively adjusting the outlier rejection threshold based on the motion likelihood of identified objects. Panoptic-SLAM [39] employs panoptic segmentation to provide a more holistic view of the scene with short-term data association to filter out-of-distribution dynamic keypoints. SDS-SLAM [40] introduces semantic local ground manifolds to couple road planes and lane line information, providing a unified framework that leverages these geometric constraints to concurrently enhance localization and object tracking in driving scenarios. Additionally, certain methods perform tracking of moving objects while handling them for SLAM, organically unifying both tasks within a single framework [41]. SDPL-SLAM [42] incorporates line features into the dynamic SLAM framework to provide additional geometric constraints that improve the precision of multi object tracking and ego localization. Unlike existing methods, our approach exploits predicted semantic labels and derived semantic uncertainty by coupling them into the iterative optimization for poses and geometry to achieve seamless dynamic object removal and allow semantic uncertainty to inherently improve visual odometry robustness.

III. Methodology

In this work, we fully integrate both discrete and continuous semantic information into the odometry pipeline. Specifically, we design a semantic uncertainty error and a semantic-aware pose estimation module, both well-suited for gradient-based optimization in direct sparse odometry, aiming to improve odometry accuracy and robustness, especially in the presence of abrupt exposure changes and large moving objects common in outdoor scenes.

A. Pipeline

As depicted in Fig. 3, SUM-VO consists of three primary components: semantic information generation, semantic two-stage tracking, as well as loop closing and mapping.

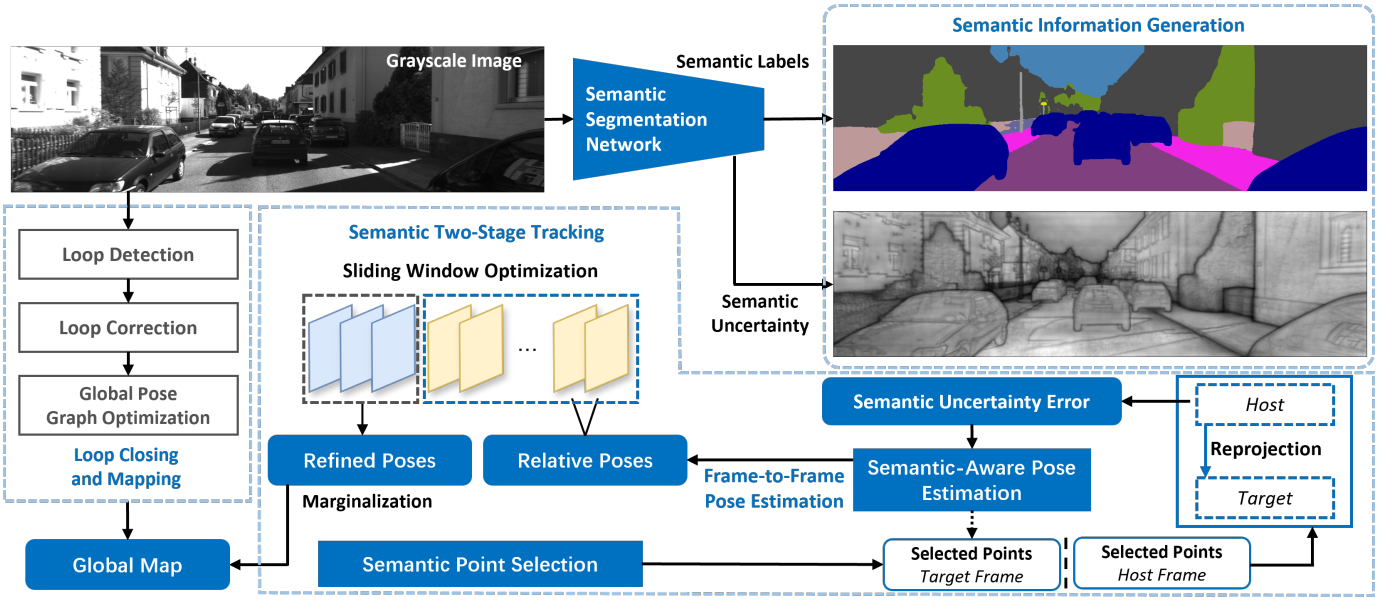


Fig. 3. System architecture of the proposed SUM-VO. The framework establishes a unified pipeline that couples semantic labels and semantic uncertainty, encompassing semantic perception, uncertainty-aware two-stage tracking, and globally consistent loop closing and mapping.

1) Semantic Information Generation: The system starts by generating semantic labels and uncertainty for each incoming monocular frame. We employ a pre-trained semantic segmentation network to obtain, for each pixel, discrete semantic labels and continuous semantic uncertainty. The choice of the semantic segmentation network is largely flexible, provided that its final layer consists of a convolution followed by softmax normalization. Notably, our formulation does not require any additional training or fine-tuning, as a closed-form solution for uncertainty estimation can be directly derived. This uncertainty measure effectively captures rich semantic boundary information while preserving intra-class texture details. Moreover, it inherits the robustness of semantic representations to illumination variations, making it particularly well-suited for visual tracking as well as pose and geometric estimation.

2) Semantic Two-Stage Tracking:

a) Front-end Frame-to-Frame Tracking: At the front end, we adopt a pyramid-based tracking strategy that integrates semantic labels and uncertainty. For a new frame, we perform semantic-aware pose estimation for frame-to-frame tracking with geometry reconstruction by minimizing the semantic uncertainty error. Specifically, static points from the previous frame are reprojected to formulate semantic uncertainty errors instead of photometric errors for optimization. A constant velocity model provides the initial pose, and each pyramid level refines the result for the next. Unlike conventional direct SLAM methods, which select points based on image gradients, we propose a semantic-aware point selection scheme that leverages semantic uncertainty to identify more suitable points for tracking.

b) Back-end Sliding Window Optimization: To maintain local consistency, we adopt a sliding window containing a fixed number of recent keyframes. We refine the camera poses and inverse depths of all active points through a joint optimization framework. The energy functional is constructed using the semantic uncertainty error across multiple views. We follow the marginalization strategy of DSO [6], where old keyframes and points that move out of the field of view are marginalized using the Schur complement to maintain a sparsified Hessian matrix, ensuring real-time performance while preserving historical geometric constraints.

3) Loop Closing and Mapping: The global mapping component ensures large-scale consistency by maintaining a sparse map of semantically labeled landmarks. Following the loop closing component of LDSO [16], we circumvent the inherent limitations of direct methods in place recognition by extracting and describing ORB features on keyframes solely for loop detection. These features remain independent of the semantic-based direct tracking process to maintain computational efficiency. When a potential loop is identified via a Bag-of-Words (BoW) dictionary, a $Sim(3)$ pose graph optimization is triggered to rectify accumulated drift in rotation, translation, and scale. The system finally yields a globally consistent trajectory and an uncertainty-aware sparse reconstruction.

B. Semantic Uncertainty Representation

Conventional direct methods primarily estimate camera poses by minimizing photometric error. However, this approach relies on the assumption of photometric consistency, which is often violated in outdoor environments due to frequent and abrupt camera exposure changes. To identify an optimization objective that is both robust to illumination variations and spatially continuous, we turn to the internal representations of semantic segmentation networks.

Although leveraging the distance fields of semantic label boundaries appears as a plausible alternative, our preliminary attempts indicate that such geometric priors lack sufficient texture inside the objects, inevitably resulting in unconstrained tracking shifts. As semantic segmentation is an upstream

task for semantic SLAM, we propose a generic and closed-form formulation of semantic uncertainty applicable to any network architecture. Inspired by prior work on quantifying network uncertainty via loss gradients (e.g., [43]–[45]), we avoid introducing any additional network. Based on the insight that diminishing gradients of the loss typically indicate better convergence, *i.e.*, lower uncertainty of network predictions, we directly quantify semantic uncertainty as the norm of the pixel-wise cross-entropy loss gradients from a pre-trained segmentation model, providing a continuous observation at each pixel for odometry.

Denote a semantic segmentation network with learnable parameters θ as $\mathcal{F}$, which takes images $x \in \mathbb{R}^{H\times W}$ as inputs and outputs softmax semantic probabilities $y_{uv} \in \mathbb{R}^C$ for each pixel (u, v), where C is the number of classes. In the absence of ground-truth semantic labels during inference, we generate one-hot pseudo-labels $\hat{y}$ and formulate the semantic uncertainty as the L_1-norm of gradients:

$$U = \|\nabla_\theta L(\mathcal{F}(x;\theta)|\hat{y})\|_1,\ \hat{y}_{uv} = \arg\max_c y^c_{uv}, \tag{1}$$

where y^c_{uv} is the value of $y_{uv} = \mathcal{F}(x;\theta)_{uv}$ at c-th dimension. Denote S as the softmax operation for C channels and upper marker c as c-th channel. Denote f as the predicted logits before softmax, then the cross entropy loss of pixel (u, v) is written as $L_{uv}(f_{uv}(x;\theta)|\hat{y})$. The gradients of this loss w.r.t. θ can be derived as:

$$\begin{aligned}
&\nabla_\theta L_{uv}(f_{uv}(x,\theta)|\hat{y})\\
&= \sum_{i=1}^{C} -\hat{y}^i_{uv} \frac{1}{S^i(f_{uv}(x;\theta))} \cdot \nabla_\theta S^i(f_{uv}(x;\theta))\\
&= \sum_{i=1}^{C}\sum_{j=1}^{C} \hat{y}^i_{uv} S^j(f_{uv}(x;\theta)) \cdot (1-\delta_{ij}) \cdot \nabla_\theta f^j_{uv}(x;\theta)\\
&= \sum_{j=1}^{C} S^j(f_{uv}(x;\theta)) \cdot (1-\hat{y}^j_{uv}) \cdot \nabla_\theta f^j_{uv}(x;\theta),
\end{aligned} \tag{2}$$

where δ is the Kronecker symbol.

For monocular odometry systems with stringent real-time requirements, performing backpropagation over all network parameters incurs prohibitive computational overhead. For simplification, we consider the last 1×1 convolutional layer, with learnable weights $\mathbf{W} \in \mathbb{R}^{C\times C'}$. Denote $g \in \mathbb{R}^{C'\times H\times W}$ as the feature map of the last hidden layer, then

$$f^c_{uv} = \sum_{k=1}^{C'} \mathbf{W}_{ck} \cdot g^k_{uv}. \tag{3}$$

According to Eq. 1-3, the semantic uncertainty at pixel (u, v) can be finally derived as follow:

$$U_{uv} = \sum_{c=1}^{C} \left[S^c(f_{uv}(x;\theta)) \cdot (1-\hat{y}^c_{uv}) \cdot \sum_{k=1}^{C'} |g^k_{uv}| \right]. \tag{4}$$

While this derivation can be extended backward through the network parameters via back-propagation, we limit our focus to the output layer for computational and memory efficiency. This restriction not only preserves but also enhances the generality of our formulation. It imposes no constraints on the network architecture, requiring only a 1×1 convolution or a point-wise linear layer as the final layer, a design widely used in convolutional semantic segmentation models [46]–[48], including those [49]–[51] based on Vision Transformers (ViT) [52]. Consequently, our formulation is readily plug-and-play across diverse architectures. Furthermore, the formulation in Eq. 4 integrates semantic probabilities, semantic labels, and texture-rich feature maps, thereby capturing rich semantic boundary, structural, and textural information, while remaining straightforward to derive, compute and generalize.

C. Semantic Uncertainty Error and Point Selection

Conventional direct odometry heavily relies on the assumption of photometric consistency, which is inherently brittle under the dynamic illumination common in outdoor environments. To transcend this limitation, we project the geometric tracking problem into a more robust semantic confidence space. Rather than minimizing the residuals of raw pixel intensities, we formulate tracking as the alignment of semantic uncertainty fields across frames. This continuous representation inherently exhibits distinct gradients along both geometric structures and semantic boundaries, providing ideal convergence basins for pose optimization. Furthermore, this semantic-aware field is seamlessly coupled with our novel point selection scheme to guarantee robust tracking and reliable loop closure detection.

Semantic uncertainty error formulation. Instead of comparing visual appearance, we constrain the system by enforcing that a physical 3D point should project consistently onto the semantic uncertainty distributions of sequential frames. For a point p_k selected from host frame i and observed in target frame j, the semantic uncertainty error $E_{i,j,k}$ over its local neighborhood pattern $\mathcal{N}_{p_k}$ is defined as:

$$E_{i,j,k} = \sum_{p\in\mathcal{N}_{p_k}} w_p \left\| (U_j[p'] - b_j) - \frac{t_j e^{a_j}}{t_i e^{a_i}} (U_i[p] - b_i) \right\|_\gamma, \tag{5}$$

where w_p is a heuristic weighting factor; a, b are coefficients to adapt for affine transformation across different domains; t_i, t_j are exposure time of frame i, j, respectively; $\|\cdot\|_\gamma$ is the Huber norm; and U denotes the semantic uncertainty normalized within the image range. The projected coordinates p' in the target frame are governed by the strict rigid body transformation:

$$p' = \Pi(R\Pi^{-1}(p, d_{p_k}) + t), \tag{6}$$

where Π and Π^{-1} are projection and back-projection functions, R and t are relative rotation and translation between frame i and j, and d_{p_k} is the inverse depth of point p_k.

While semantic uncertainty is generally robust to illumination changes, we retain the affine transformation coefficients a and b to serve vital function of online domain adaptation. Because the segmentation network is supervised on specific pre-training datasets, encountering out-of-distribution scenarios in real-world driving typically induces a global shift in the uncertainty distribution. When the scene significantly deviates, the overall semantic uncertainty tends to be higher, prompting

these coefficients to adjust to smaller values for accurate pose estimation. Co-optimizing these affine parameters during tracking enables the system to dynamically absorb this domain gap. By iteratively minimizing this semantic uncertainty error term, we accurately estimate both the camera poses and the point depths, as detailed in Section III-D.

Semantic point selection scheme. In direct sparse SLAM, point selection provides the observations used to construct the reprojection error for tracking and loop closure detection. Point selection consists of two types. The first is for tracking, where points are selected to construct and minimize the objective. The second selects a subset of points with already optimized geometry for loop closure queries. During tracking, points are selected based on the gradient of semantic uncertainty. At each level of the pyramid, *i.e.*, images at various resolutions, points with higher gradients are adaptively selected on a grid basis to ensure a uniform distribution across semantic and geometric boundaries as well as internal textured areas. As shown in Fig. 3, darker regions correspond to higher semantic uncertainty, typically found at semantic boundaries and geometric edges with discriminative structures or textures, intuitively facilitating effective tracking and loop closure detection.

Our loop closure detection builds upon the framework of LDSO [16] and follows the standard Bag-of-Words (BoW) procedure, but introduces a semantically augmented scoring mechanism. Instead of relying solely on local image intensity variations, we incorporate both geometric and semantic contexts to evaluate the existing sparse points with inverse depths already refined by the front-end sliding window. Specifically, the score of each candidate point is formulated as a weighted sum of its Shi-Tomasi corner response and its localized semantic uncertainty. The top-k points are then extracted to compute descriptors for BoW construction and global pose graph optimization. By fusing these two metrics, our system ensures that the selected representative features possess not only the local geometric repeatability required for reliable descriptor matching but also the long-term structural stability inherent to semantic boundaries.

D. Semantic-Aware Pose Estimation

Tracking points on dynamic objects can induce pose errors, especially when large moving objects occlude most previously tracked points. Although direct methods handle minor dynamics robustly, large dynamic objects may cause significant drift or failures. To tackle this, we propose the SAPE module to jointly estimate the camera pose and remove dynamic objects based on semantic labels and geometric consistency assessment. Thanks to the coarse-to-fine optimization on the image pyramid for bundle adjustment, our dynamic object filtering can be seamlessly integrated into this process by assessing the category-wise reprojection error. We introduce three specific criteria to detect and eliminate large dynamic objects, with a preliminary assessment at the top pyramid level and a definitive confirmation upon completing the optimization.

Denote the host and target frames as F_h and F_t, respectively. Let $\mathcal{C}_h$ be the semantic labels detected in F_h, where an individual element $s \in \mathcal{C}_h$ represents a specific semantic class.

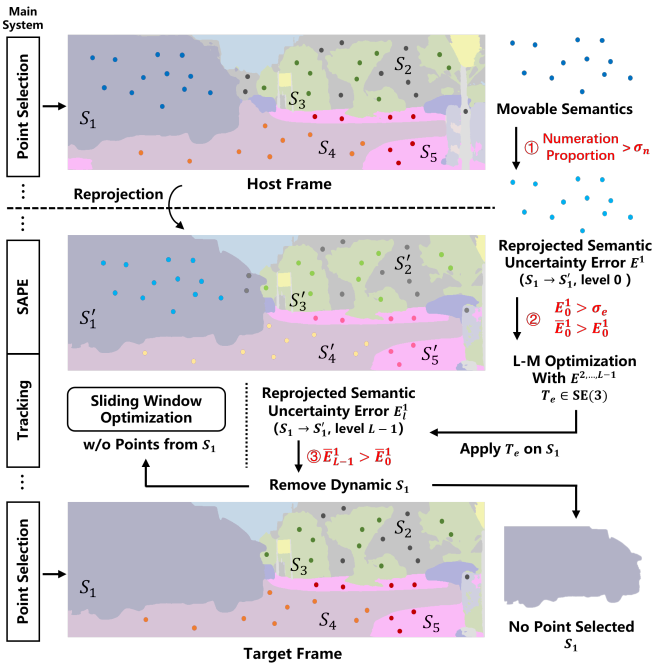


Fig. 4. Illustration of semantic-aware pose estimation during tracking.

Within this set, $\mathcal{M}_h \subset \mathcal{C}_h$ denotes the subset of potentially movable categories (*e.g.*, car, bus, or truck). To prevent repeated state assessments across consecutive frames and ensure temporal consistency, SAPE inherits the persistently dynamic categories from the previous tracking step, denoted as $\mathcal{D}_{prior}$. If a previously recognized dynamic category still occupies a substantial portion of the current frame, its points are directly excluded from the optimization process. For the remaining movable categories, we utilize further geometric criteria to ascertain whether they are moving. Empirically, direct methods remain robust to a small fraction of dynamic foreground due to the robust loss functions. Hence, our first criterion checks whether the points $\mathcal{P}_{s_i}$ belonging to a movable class $s_i \in \mathcal{C}_h$ occupy a significantly large proportion of the total selected points $\mathcal{P}$. If the ratio $|\mathcal{P}_{s_i}|/|\mathcal{P}|$ exceeds a designated threshold σ_n (set to 30%), this semantic category is flagged for deeper verification. We argue that only when a movable semantic area covers more than σ_n of the tracking points can it severely corrupt the energy function and substantially impact the accuracy of pose estimation.

Our second criterion relies on the semantic uncertainty error and its convergence behavior during the top-level (coarsest resolution) optimization. For an L-level image pyramid with ascending resolutions from level 0 to $L-1$, the average errors before and after optimization for level j of the points with semantic label s_i are denoted as E_j^i and $\bar{E}_j^i$, respectively. If the initial error E_0^i exceeds a threshold $\sigma_e = 20$, and meets the condition $\bar{E}_0^i > E_0^i$, it indicates that the optimization is struggling to fit these points under the rigid static scene assumption. Consequently, the semantic category s_i is temporarily marked as dynamic ($\mathcal{D}_{temp}$) to reflect its adverse effect on the tracking accuracy. To eliminate the influence of these suspicious dynamic regions, they are immediately

Algorithm 1: Semantic-Aware Pose Estimation

Input: Host frame F_h, Target frame F_t, Tracked points $\mathcal{P}$, All semantic labels $\mathcal{C}_h$, Movable labels $\mathcal{M}_h \subset \mathcal{C}_h$, Prior dynamic categories $\mathcal{D}_{prior}$, Pyramid $0 \dots L-1$, Thresholds $\sigma_n, \sigma_e, \sigma_o$

Output: Optimized relative pose T_e, Unoptimized temporary dynamic categories $\mathcal{D}_{temp}$, Updated dynamic categories $\mathcal{D}_{final}$

Initialize T_e via constant velocity model;
$\mathcal{S}_e \leftarrow \mathcal{C}_h$, $\mathcal{D}_{temp} \leftarrow \emptyset$, $\mathcal{D}_{final} \leftarrow \emptyset$;
// Inherit Prior Dynamic Status
for *each* $s_i \in \mathcal{D}_{prior}$ **do**
  if $|\mathcal{P}_{s_i}|/|\mathcal{P}| \geq \sigma_o$ **then**
    $\mathcal{D}_{final} \leftarrow \mathcal{D}_{final} \cup \{s_i\}$, $\mathcal{S}_e \leftarrow \mathcal{S}_e \setminus \{s_i\}$
  end
end
// Level 0 Optimization
$\mathcal{M}_{test} \leftarrow \mathcal{M}_h \setminus \mathcal{D}_{proir}$;
Calculate initial error E_0^i for each $s_i \in \mathcal{M}_{test}$ using T_e;
Optimize T_e at Level 0 using points in $\mathcal{P}_{\mathcal{S}_e}$;
Calculate post-optimization error $\bar{E}_0^i$ for each $s_i \in \mathcal{M}_{test}$;
for *each* $s_i \in \mathcal{M}_{test}$ **do**
  // Criterion 1: Point proportion
  if $|\mathcal{P}_{s_i}|/|\mathcal{P}| > \sigma_n$ **then**
    // Criterion 2: Optimization trend
    if $E_0^i > \sigma_e$ ***and*** $\bar{E}_0^i > E_0^i$ **then**
      $\mathcal{D}_{temp} \leftarrow \mathcal{D}_{temp} \cup \{s_i\}$, $\mathcal{S}_e \leftarrow \mathcal{S}_e \setminus \{s_i\}$
    end
  end
end
// Level 1 to $L-1$ Optimization
for *level* $j = 1$ **to** $L-1$ **do**
  Optimize T_e at level j using remaining points $\mathcal{P}_{\mathcal{S}_e}$;
end
// Final Motion Consistency Check
for *each* $s_i \in \mathcal{D}_{temp}$ **do**
  Project $\mathcal{P}_{s_i}$ using optimized T_e and compute $\bar{E}_{L-1}^i$;
  // Criterion 3: Consistency
  if $\bar{E}_{L-1}^i > \bar{E}_0^i$ **then**
    $\mathcal{D}_{final} \leftarrow \mathcal{D}_{final} \cup \{s_i\}$;
  end
end
return $T_e, \mathcal{D}_{temp}, \mathcal{D}_{final}$;

excluded from the optimization of subsequent levels, *i.e.*, we estimate a relative pose $T_e \in \mathrm{SE}(3)$ between F_h and F_t only with reliable points from the remaining semantic labels $\mathcal{S}_e := \mathcal{C}_h - s_i$ via Levenberg-Marquardt (L-M) algorithm. Similar to regular frame-to-frame bundle adjustment, and initialized by a constant velocity motion model, we minimize semantic uncertainty error E of points from $\mathcal{P}_{s_e} := \bigcup_{s \in \mathcal{S}_e} \mathcal{P}_s$ by iteratively solving the increments $\Delta x = [\Delta d_1, \cdots, \Delta d_n, \Delta \xi^T, \Delta a, \Delta b]^T$ in the normal equation:

$$\begin{aligned} (J^T W J + \lambda I)\Delta x &= -J^T W E, \\ J &= \left[\frac{\partial E}{\partial d_1}, \cdots, \frac{\partial E}{\partial d_n}, \frac{\partial E}{\partial \delta\xi}, \frac{\partial E}{\partial a}, \frac{\partial E}{\partial b} \right], \end{aligned} \tag{7}$$

where $d_1, \cdots, d_n$ refer to inverse depths of points $\mathcal{P}_{s_e}$, $\delta\xi \in \mathfrak{se}(3)$ is the Lie algebra left perturbation of pose T_e, a, b are semantic affine transformation coefficients, W is the weighting matrix, and λ is the L-M damping factor.

After optimizing T_e through the remaining pyramid levels (levels 1 to $L-1$), the third criterion performs a final motion consistency check. We apply the rotation and translation components R_e, t_e of the optimized pose T_e to project the temporarily discarded points $\mathcal{P}_{s_i}$ where $s_i \in \mathcal{D}_{temp}$:

$$p' = \Pi(R_e \Pi^{-1}(p, d_p) + t_e), \forall p \in \mathcal{P}_{s_i}, \tag{8}$$

and evaluate the fine-level average semantic uncertainty error $\bar{E}_{L-1}^i$ following Eq. 5. A semantic category s_i is definitively classified as persistently dynamic (denoted as $\mathcal{D}_{final}$) only if $\bar{E}_{L-1}^i > \bar{E}_0^i$. This strict condition implies that even with a highly accurate global pose estimated from the static background, the reprojection error for s_i continues to magnify at higher resolutions, firmly verifying its independent motion.

Subsequently, SAPE facilitates the downstream status update by explicitly guiding the global SLAM system. Points belonging to the temporarily dynamic categories $\mathcal{D}_{temp}$ are excluded from the backend sliding window optimization to preserve mapping accuracy, though they remain eligible for point selection for the next frame. Conversely, the definitively dynamic categories $\mathcal{D}_{final}$ serve as a semantic prior for the frontend to exclude them from subsequent point extraction. This set is then propagated to the next frame as $\mathcal{D}_{prior}$, ensuring the consistent exclusion of moving entities as long as their area proportion exceeds the threshold $\sigma_o = 15\%$. Notably, σ_n and σ_o are meticulously chosen to balance computational efficiency against the necessity of removing large moving objects, while σ_e is empirically set to prevent misclassification under the magnitude of semantic uncertainty. The pseudocode of SAPE is shown in Alg. 1, and its diagram in Fig. 4.

In summary, SAPE establishes a unified framework where camera pose estimation and dynamic object filtering are intrinsically coupled. This integration is realized through a systematic three-step process: evaluating the spatial distribution and initial residuals of movable semantic categories, estimating the camera pose using reliable static regions, and performing a final motion consistency check. Crucially, the same reprojection errors utilized to minimize the objective function are simultaneously repurposed to evaluate category-wise motion consistency, eliminating the need for separate detection steps. By leveraging semantic uncertainty and labels, SAPE jointly optimizes the pose and identifies dynamic points, effectively mitigating the influence of spatially large moving objects. This ensures that confirmed dynamic entities are excluded from both backend tracking and point selection, thereby maintaining system stability in challenging environments without incurring significant computational overhead.

E. Semantic-Informed Joint Optimization

Within the back-end joint optimization, the energy functional is constructed entirely using the proposed multi-view semantic uncertainty error. By inheriting the exact same error metric utilized in the front-end, this formulation ensures strict mathematical consistency across the entire pipeline, establishing a compatible objective for both tracking and mapping. This formulation allows the back-end to deeply integrate the robust semantic priors extracted during the tracking phase. Furthermore, to ensure real-time performance without discarding valuable historical geometric constraints, we follow the

TABLE I
APE (M) / RPE (M) OF VISUAL SLAM SYSTEMS ON KITTI

Seq.	ORB-SLAM2 *2016*	LDSO *2018*	DynaSLAM *2018*	ORB-SLAM3 *2020*	DROID-SLAM *2021*	SDVO *2022*	DPV-SLAM *2023*	Panoptic-SLAM *2024*	SUM-VO *Ours*
00	5.68 / 0.21	10.06 / 0.18	5.61 / 0.25	9.05 / 0.18	81.13 / 0.32	7.38 / -	103.25 / 0.89	6.28 / 0.18	**3.91 / 0.10**
02	21.84 / 0.18	38.68 / **0.12**	25.09 / 0.19	**20.58** / 0.14	111.39 / 0.50	24.73 / -	108.82 / 0.92	23.76 / 0.14	22.59 / **0.12**
03	1.14 / 0.04	2.57 / 0.06	0.92 / 0.05	0.66 / 0.04	1.89 / 0.05	1.93 / -	2.35 / 0.06	0.68 / **0.03**	**0.62** / 0.05
04	1.04 / 0.05	1.03 / 0.04	1.20 / 0.08	0.45 / 0.04	1.07 / 0.05	0.39 / -	1.10 / 0.07	0.74 / 0.04	**0.34 / 0.03**
05	5.42 / 0.09	4.02 / 0.38	5.36 / 0.09	5.78 / 0.09	56.02 / 0.34	3.66 / -	58.64 / 0.64	11.31 / 0.13	**3.02 / 0.05**
06	12.38 / 0.30	11.70 / 0.36	12.00 / 0.22	13.22 / 0.22	39.75 / 0.28	**5.40** / -	55.29 / 0.70	12.63 / **0.16**	10.95 / **0.16**
07	1.93 / 0.24	2.12 / 0.21	2.65 / 0.23	2.61 / 0.14	30.70 / 0.24	1.97 / -	17.33 / 0.28	2.41 / 0.15	**1.46 / 0.12**
08	47.60 / 0.44	51.97 / 0.53	31.14 / 0.49	46.07 / 0.44	74.56 / 0.44	53.56 / -	101.35 / 0.98	46.12 / 0.44	**30.38 / 0.42**
09	6.78 / 0.12	64.51 / 0.32	**6.14** / 0.09	8.33 / 0.10	81.13 / 0.32	30.73 / -	75.25 / 0.62	6.65 / 0.14	10.17 / **0.08**
10	7.93 / **0.08**	14.14 / 0.14	6.81 / 0.10	**6.07 / 0.08**	20.64 / 0.17	6.84 / -	13.93 / 0.22	6.86 / 0.15	6.47 / 0.09

marginalization strategy of DSO [6]. When the sliding window reaches its capacity, old keyframes and points moving out of the field of view are marginalized. By applying the Schur complement, we effectively encapsulate their prior knowledge into the system while maintaining a sparsified Hessian matrix, thereby strictly bounding the computational complexity.

IV. EXPERIMENTS

The experiments are organized as follows: We first evaluate our method on the KITTI benchmark [10], then test its generalizability on the Complex Urban dataset [11] with a large domain gap from KITTI. We conduct ablation studies to demonstrate the effectiveness of semantic uncertainty error, semantic point selection scheme and the semantic-aware pose estimation module. We validate the practicality of our approach on a self-collected Campus dataset and measure the runtime for various parts. Across these outdoor datasets, some frames exhibit short-term illumination shifts due to exposure changes as shown in Fig.2, highlighting the stability of semantic uncertainty for tracking. As our formulation of semantic uncertainty is agnostic to semantic segmentation network architectures, we employ DeepLabv3+ [46] without loss of generality. All methods are evaluated using Absolute Pose Error (APE) and Relative Pose Error (RPE) with the EVO toolkit [53]. Considering the inability of monocular odometry to capture the real-world scale, we use Umeyama's method [54] to align the scale for fair comparison.

A. The KITTI Dataset

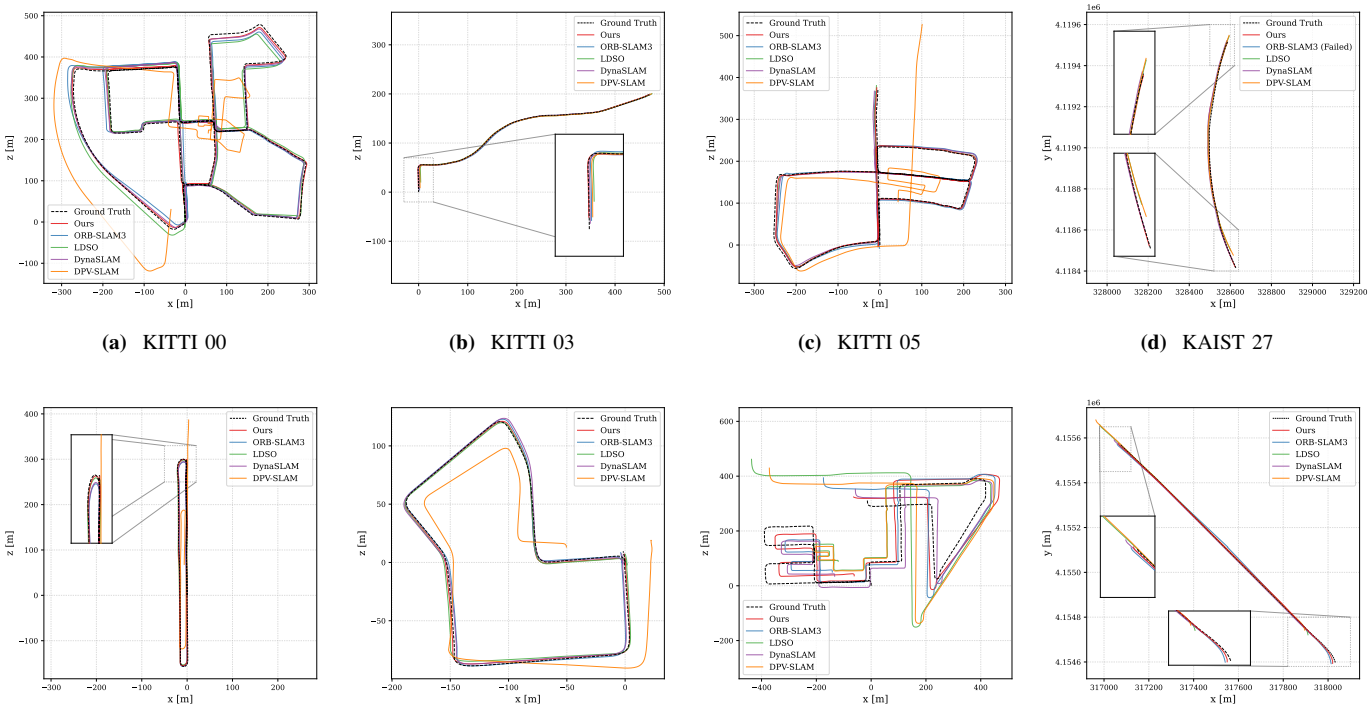
KITTI [10] is a widely used outdoor benchmark for evaluating localization accuracy. We employ a DeepLabv3+ model pre-trained exclusively on the Cityscapes [55] dataset to extract semantic information. It is worth noting that our semantic uncertainty formulation is completely training-free. By mathematically deriving the uncertainty without any fine-tuning, our method seamlessly generalizes to the diverse environments present in the KITTI dataset.

We benchmark SUM-VO with APE against several leading SLAM systems ranging from traditional feature-based [56], [57], direct [16] and learning-based [8], [9] methods, and semantic SLAM focusing on dynamic object removal [34], [39] or semantic probabilistic model-based optimization [30]. For sequence 01 collected in a highway, feature-based methods like ORB-SLAM2 [56] tend to fail in tracking and relocalization, while direct methods easily suffer scale drifts. As our method is also not designed for highway scenarios, we exclude all highway sequences from the evaluation on all benchmarks. We follow the formulation of RPE in [32], instead of [34], measuring the error between estimated and ground-truth relative poses in meters. As monocular SLAM systems are only able to recover the trajectory of the camera up to a similarity transform, this metric appropriately reflects the accuracy of frame-to-frame pose estimation.

The results of APE and RPE are presented in Table I. ORB-SLAM2 [56], LDSO [16], DynaSLAM [34], ORB-SLAM3 [57], DROID-SLAM [8], DPV-SLAM [9] are tested using their open-source implementations. For fair comparison, all methods are evaluated in their monocular versions. Since Panoptic-SLAM [39] claims monocular applicability by using only RGB images, we re-implement the monocular version according to their implementation of RGB-D camera for testing. Due to the absence of code for SDVO [30], we report the APE from its paper. As shown in Table I, SUM-VO delivers outstanding performance, securing the lowest APE on six sequences and the lowest RPE on eight sequences. The visualized trajectories are depicted in Fig. 5 (a)-(c) and (e)-(g). In monocular systems, APE is the primary indicator of global scale consistency. By minimizing the semantic uncertainty error rather than purely relying on photometric or geometric residuals, our joint optimization actively suppresses the scale drift caused by environmental dynamics and illumination inconsistencies, especially on the long sequence like 00. Alongside the highly accurate frame-to-frame tracking reflected by the RPE, these results confirm that SUM-VO provides a robust solution for outdoor monocular odometry.

B. The Complex Urban Dataset

To evaluate the generalizability of our method involving learning-based semantic information, we test on the Complex Urban dataset [11] with significant domain gap from KITTI. This dataset captures urban scenes via vehicle platforms from multiple cities, focusing on multi-lane intersections, diverse moving vehicles, dense traffic scenarios, and deep shadows cast by dense high-rise buildings, contrasting with the predominantly two-lane or rural roads in KITTI. Without additional training or fine-tuning of our model, following prior baselines,



(a) KITTI 00 **(b)** KITTI 03 **(c)** KITTI 05 **(d)** KAIST 27

(e) KITTI 06 **(f)** KITTI 07 **(g)** KITTI 08 **(h)** KAIST 32

Fig. 5. Qualitative comparison of the estimated trajectories and ground truth across eight sequences. To ensure a fair comparison and account for the scale ambiguity in monocular odometry, all trajectories are aligned to the ground truth using a similarity transformation (Umeyama's method) prior to plotting.

TABLE II
APE (M) / RPE (M) OF VISUAL SLAM SYSTEMS ON KAIST

Seq.	ORB-SLAM2 *2016*	LDSO *2018*	DynaSLAM *2018*	ORB-SLAM3 *2020*	DROID-SLAM *2021*	DPV-SLAM *2023*	Panoptic-SLAM *2024*	SUM-VO *Ours*
27 *Dongtan*	×	22.82 / 2.13	12.10 / 4.36	×	38.88 / 2.73	33.03 / 2.67	18.52 / 3.80	**6.23** / **1.97**
29 *Seongnam*	×	95.75 / 3.00	×	25.53 / 3.02	62.38 / 2.61	116.17 / 2.62	29.16 / **2.20**	**19.72** / 2.59
31 *Gangnam*	34.82 / 2.32	38.60 / 1.80	28.69 / 2.70	35.13 / 1.86	62.55 / 1.83	35.27 / 2.01	34.46 / 1.77	**27.55** / **1.74**
32 *Yeouido*	21.50 / 2.87	56.18 / 2.56	14.11 / 2.93	16.47 / 2.94	65.41 / 1.73	79.51 / 1.75	24.12 / 2.05	**10.52** / **1.66**

we conduct generalization tests on non-highway urban scenes. Without loss of generality, we select one sequence per region including Dongtan, Seongnam, Gangnam, and Yeouido. Due to the overlength of the sequences, we use the first 2000 frames from each for testing.

The quantitative evaluation results and qualitative trajectory visualizations are presented in Table II and Fig. 5 (d) and (h), respectively. Notably, both ORB-SLAM2 and ORB-SLAM3 fail in tracking and relocalization on Seq.27 specifically in the region shown in Fig. 2 c), where sudden overexposure leads to a sharp drop in feature points and reliable matches. The other two failures on Seq.29 are mainly due to prolonged initialization issues or severe drift caused by dense dynamic traffic. As shown in Table II, SUM-VO outperforms other methods in expansive urban main roads of the KAIST Urban dataset. This superiority demonstrates that our semantic uncertainty representation fundamentally overcomes the sensitivity to drastic illumination changes. Furthermore, unlike typical semantic or learning-based SLAM systems typically suffering from domain gaps, our semantic uncertainty-based odometry showcases strong generalizability to unseen scenarios during training, maintaining robustness across various domains without site-specific fine-tuning.

C. Ablation Studies

To systematically evaluate the individual contribution and the synergy of our proposed components, we conduct a series of ablation studies to test three core modules: the semantic uncertainty error metric (M1), the semantic point selection scheme for both tracking and loop closing (M2), and the Semantic-Aware Pose Estimation module especially for dynamic object exclusion (M3). While small dynamic objects typically do not cause significant pose estimation errors, SAPE targets large dynamic objects in the field of view. KITTI-07 is chosen as a representative case with a truck moving by and temporarily occupying a large proportion of the field of view (see Fig. 6). We also test on KITTI 00, 04, 05, and 06 sequences containing movable semantic regions like vehicles to assess robustness. We compare the performance and trajectories across four settings: 1) Baseline

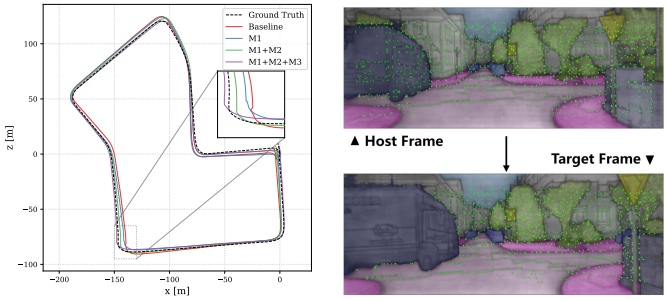


Fig. 6. Trajectories estimated by various methods when suffering severe interference from dynamic objects. The right panels visualize the point selection results when the truck enters the frame under SAPE.

TABLE III
ABLATION STUDIES (APE (M) / RPE (M)) ON KITTI SEQUENCES

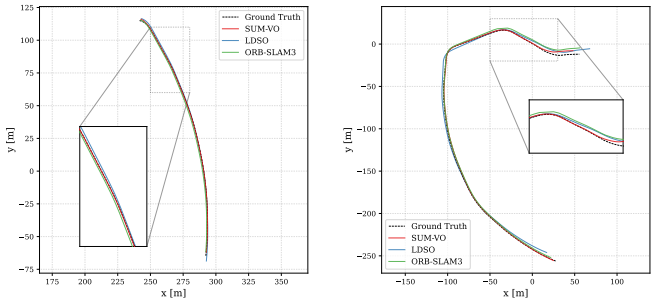
Seq.	Baseline LDSO	M1	M1 + M2	M1 + M2 + M3 SUM-VO
00	10.06 / 0.18	11.02 / 0.08	4.70 / 0.06	**3.91** / **0.05**
04	1.03 / 0.04	0.70 / 0.04	0.58 / 0.04	**0.50** / **0.03**
05	4.02 / 0.38	5.35 / 0.10	**2.99** / **0.05**	3.02 / **0.05**
06	11.70 / 0.36	14.92 / 0.17	12.16 / **0.15**	**11.92** / 0.16
07	2.12 / 0.21	1.90 / 0.18	1.84 / **0.12**	**1.46** / **0.12**

(LDSO): A standard direct formulation relying on photometric error and gradient-based point selection. 2) M1: A framework adopting semantic uncertainty error while retaining the point selection strategy according to grayscale images. 3) M1 + M2: A framework integrating the semantic uncertainty error optimization with the semantic point selection scheme. 4) M1 + M2 + M3 (SUM-VO): The complete system equipped with the full SAPE module.

Table III reports the ablation results. Compared to the baseline, M1 yields limited improvements in APE because the photometrically selected points used to construct the error are not only misaligned with the optimal locations for semantic uncertainty metric, but also remain vulnerable to photometric disturbances. Nevertheless, the consistent improvement in RPE highlights the robustness of semantic uncertainty metric for frame-to-frame tracking. When augmented with the semantic point selection scheme, M1 + M2 achieves a substantial performance leap, illustrating the synergy between the two modules. Explicitly leveraging semantic uncertainty for point selection ensures that the selected points are optimally tailored for the optimization. Finally, incorporating the SAPE module (*SUM-VO*) not only aggressively decreases the APE on highly dynamic sequences like KITTI-07, but also refines the average performance on scenes without challenging obstacles. Qualitatively, Fig. 6 demonstrates that SAPE prevents abrupt odometry shifts by suspending point selection only when a traversing truck exceeds the predefined threshold, effectively shielding the system from dynamic interference.

D. The Campus Dataset

To further assess the performance of our method in real-world scenes, still without fine-tuning, we collected data on the campus roads with an onboard monocular camera for



(a) Campus 00 **(b)** Campus 01

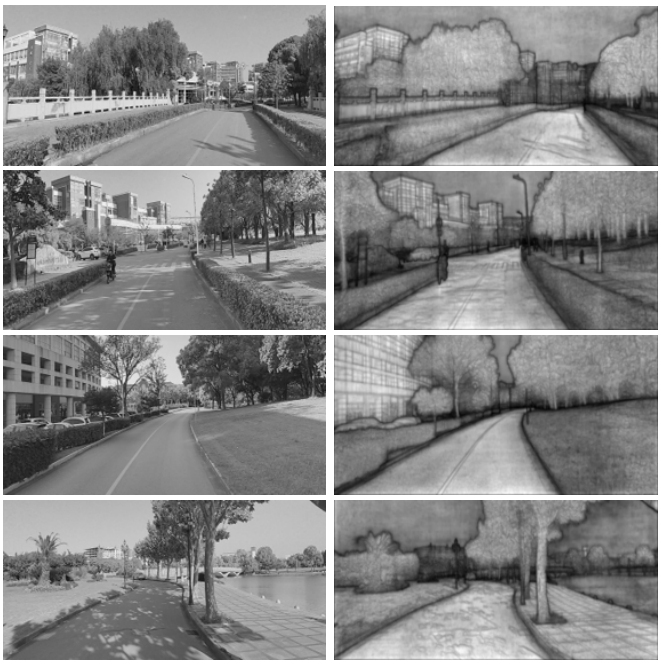


Fig. 7. Top: (a) and (b) show the estimated trajectories of the evaluated methods on the two sequences. Bottom: Grayscale images and the corresponding semantic uncertainty maps at selected locations within the campus dataset.

TABLE IV
APE (M) / RPE (M) OF VISUAL SLAM SYSTEMS ON CAMPUS

Seq.	ORB-SLAM3	LDSO	SUM-VO (L)	SUM-VO (S)
00	0.46 / 0.79	1.11 / 0.82	**0.38** / **0.40**	0.44 / 0.65
01	1.88 / 0.94	5.91 / 1.06	**1.21** / 0.85	1.76 / **0.82**

experiments. Specifically, we evaluate our system on two collected sequences consisting of 1300 and 2762 frames, respectively. Given the high-resolution nature of our raw data, we conduct evaluations at two downsampled scales, L (1920x640) and S (960x480), to systematically investigate the impact of image size on the odometry accuracy. These campus scenarios feature open environments and dense street trees, presenting significant challenges.

The results are detailed in Table IV and Fig. 7. Although ORB-SLAM3 [57] relies on feature matching to maintain scale consistency, the massive proportion of repetitive textures from campus trees severely corrupts its tracking and leads to drift. Similarly, LDSO encounters distinct challenges

TABLE V
RUNTIME COMPARISON OF KEY COMPONENTS (MS)

	Semantic Info	Front End			Back End	Loop
		PS	PE	Total		
LDSO	-	8.448	1.961	16.329	80.468	8.976
Ours	65.833 *KITTI* 83.284 *KAIST* 85.132 *Campus-L* 51.925 *Campus-S*	11.987	2.161	21.462	85.872	7.398

in these expansive environments. It experiences significant depth estimation errors during turns due to distant objects, which severely disrupts its overall scale. In contrast, SUM-VO successfully mitigates these issues through the proposed semantic uncertainty framework. These results show that our method performs well in real-world scenarios and shows strong practicality and generalization capabilities.

E. Runtime Analysis

We measure and analyze the runtime of each component in our system and compare it with that of LDSO [16], as both share a similar front and back end structure of direct methods. Loop closing is reported independently. We separately report the runtime of point selection (PS) and frame-to-frame pose estimation (PE), corresponding to our semantic point selection and SAPE modules. Tests are conducted on Ubuntu 22.04 with an Intel Core i7-13700H CPU, using the same GPU and network settings as earlier. KITTI-00 is selected for benchmarking due to its sufficient frames for stable averaging.

Compared to LDSO, our method introduces only millisecond-level increase of computational overhead, with negligible impact on system efficiency. The runtime of semantic information generation is clearly affected by the image resolution, with the highest being on Campus-L, which uses 1280×640 images. Lowering the resolution improves its speed to nearly 20 FPS while preserving odometry accuracy. Notably, semantic information generation runs as a separate thread, consistently above 12 FPS across different resolutions, meeting real-time requirements for many datasets like the KITTI and KAIST Urban dataset (10 FPS).

V. CONCLUSION

This paper presents a semantic uncertainty-based monocular visual odometry method in outdoor scenes, fully exploiting semantic information. We introduce a model-agnostic formulation of semantic uncertainty as a stable, continuous image representation and construct a novel semantic uncertainty error for both tracking and bundle adjustment. A semantic point selection scheme is developed by integrating geometric and semantic context. In addition, we propose a semantic-aware pose estimation module which seamlessly combines semantic-guided dynamic object removal with a hierarchical pose estimation approach without extra optimization costs. Our method achieves highly competitive performance on the KITTI dataset and prominent generalizability on the challenging Complex Urban Dataset without fine-tuning. Real-world experiments on our Campus dataset further validate its practicality.

REFERENCES

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[2] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *2007 6th IEEE and ACM international symposium on mixed and augmented reality*. IEEE, 2007, pp. 225–234.

[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[4] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2320–2327.

[5] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[6] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[7] 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 Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[8] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.

[9] L. Lipson, Z. Teed, and J. Deng, "Deep patch visual slam," in *European Conference on Computer Vision*. Springer, 2024, pp. 424–440.

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

[11] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban lidar data set," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 6344–6351.

[12] J. Shi *et al.*, "Good features to track," in *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE, 1994, pp. 593–600.

[13] D. G. Viswanathan, "Features from accelerated segment test (fast)," in *Proceedings of the 10th workshop on image analysis for multimedia interactive services, London, UK*, 2009, pp. 6–8.

[14] J. K. Suhr, "Kanade-lucas-tomasi (klt) feature tracker," *Computer Vision (EEE6503)*, pp. 9–18, 2009.

[15] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[16] X. Gao, R. Wang, N. Demmel, and D. Cremers, "Ldso: Direct sparse odometry with loop closure," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2198–2204.

[17] T. Botterill, S. Mills, and R. Green, "Bag-of-words-driven, single-camera simultaneous localization and mapping," *Journal of Field Robotics*, vol. 28, no. 2, pp. 204–226, 2011.

[18] P. Ganti and S. L. Waslander, "Network uncertainty informed semantic feature selection for visual slam," in *2019 16th Conference on Computer and Robot Vision (CRV)*. IEEE, 2019, pp. 121–128.

[19] K.-N. Lianos, J. L. Schonberger, M. Pollefeys, and T. Sattler, "Vso: Visual semantic odometry," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 234–250.

[20] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *2017 IEEE International Conference on Robotics and automation (ICRA)*. IEEE, 2017, pp. 4628–4635.

[21] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1352–1359.

[22] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an open-source library for real-time metric-semantic localization and mapping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1689–1696.

[23] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, 2018.

[24] S. Yang and S. Scherer, "Cubeslam: Monocular 3-d object slam," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.

[25] J. Wang, M. Rünz, and L. Agapito, "Dsp-slam: Object oriented slam with deep shape priors," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 1362–1371.

[26] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1722–1729.

[27] T. Qin, T. Chen, Y. Chen, and Q. Su, "Avp-slam: Semantic visual mapping and localization for autonomous vehicles in the parking lot," in *2020 IEEE/RSJ International Conference on intelligent robots and systems (IROS)*. IEEE, 2020, pp. 5939–5945.

[28] Y. Chen, F. Zhao, Y. Zhuge, J. Liu, J. Yan, and H. Luo, "Smore-slam: Semantic monocular slam with scale correction and reverse loop utilization in outdoor environments," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7870–7877.

[29] J. Czarnowski, T. Laidlow, R. Clark, and A. J. Davison, "Deepfactors: Real-time probabilistic dense monocular slam," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 721–728, 2020.

[30] Y. Bao, Z. Yang, Y. Pan, and R. Huan, "Semantic-direct visual odometry," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6718–6725, 2022.

[31] F. Zhong, S. Wang, Z. Zhang, and Y. Wang, "Detect-slam: Making object detection and slam mutually beneficial," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 1001–1010.

[32] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "Ds-slam: A semantic visual slam towards dynamic environments," in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2018, pp. 1168–1174.

[33] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

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

[35] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[36] Y. Bao, Y. Pan, Z. Yang, and R. Huan, "Utilization of semantic planes: Improved localization and dense semantic map for monocular slam in urban environment," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 6108–6115, 2021.

[37] M. Gonzalez, E. Marchand, A. Kacete, and J. Royan, "Twistslam: Constrained slam in dynamic environment," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6846–6853, 2022.

[38] S. Cheng, C. Sun, S. Zhang, and D. Zhang, "Sg-slam: A real-time rgb-d visual slam toward dynamic scenes with semantic and geometric information," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2022.

[39] G. F. Abati, J. C. V. Soares, V. S. Medeiros, M. A. Meggiolaro, and C. Semini, "Panoptic-slam: Visual slam in dynamic environments using panoptic segmentation," in *2024 21st International Conference on Ubiquitous Robots (UR)*. IEEE, 2024, pp. 01–08.

[40] Y. Liu, C. Guo, J. Zhan, and X. Wu, "Sds-slam: Vslam fusing static and dynamic semantic information for driving scenarios," *IEEE Transactions on Automation Science and Engineering*, 2025.

[41] S. Jiao, Y. Li, and Z. Shan, "Dfs-slam: A visual slam algorithm for deep fusion of semantic information," *IEEE Robotics and Automation Letters*, vol. 9, no. 12, pp. 11 794–11 801, 2024.

[42] A. Manetas, P. Mermigkas, and P. Maragos, "Sdpl-slam: Introducing lines in dynamic visual slam and multi-object tracking," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7893–7899.

[43] R. Huang, A. Geng, and Y. Li, "On the importance of gradients for detecting distributional shifts in the wild," *Advances in Neural Information Processing Systems*, vol. 34, pp. 677–689, 2021.

[44] T. Riedlinger, M. Rottmann, M. Schubert, and H. Gottschalk, "Gradient-based quantification of epistemic uncertainty for deep object detectors," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3921–3931.

[45] K. Maag. and T. Riedlinger., "Pixel-wise gradient uncertainty for convolutional neural networks applied to out-of-distribution segmentation," in *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2: VISAPP*, INSTICC. SciTePress, 2024, pp. 112–122.

[46] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[49] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[50] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[51] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[53] M. Grupp, "evo: Python package for the evaluation of odometry and slam." https://github.com/MichaelGrupp/evo, 2017.

[54] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 04, pp. 376–380, 1991.

[55] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

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

[57] 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 transactions on robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.