Online Indoor Visual Odometry with Semantic Assistance under Implicit Epipolar Constraints

Yang Chen[a], Lin Zhang [a,*], Shengjie Zhao [a], Yicong Zhou [b]

[a]*School of Software Engineering, Tongji University, Shanghai 201804, China*
[b]*Department of Computer and Information Science, University of Macau, Macau 999078, China*

Abstract

Among solutions to the tasks of indoor localization and reconstruction, compared with traditional SLAM (Simultaneous Localization And Mapping), learning-based VO (Visual Odometry) has gained more and more popularity due to its robustness and low cost. However, the performance of existing indoor deep VOs is still limited in comparison with their outdoor counterparts mainly owing to large areas of textureless regions and complex indoor motions containing much more rotations. In this paper, the above two challenges are carefully tackled with the proposed SEOVO (Semantic Epipolar-constrained Online VO). On the one hand, as far as we know, SEOVO is the first semantic-aided VO under an online adaptive framework, which adaptively reconstructs low-texture planes without any supervision. On the other hand, we introduce the epipolar geometric constraint in an implicit way for

*This work was supported in part by the National Natural Science Foundation of China under Grant 62272343 and Grant 61936014; in part by the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission under Grant 21SG23; and in part by the Fundamental Research Funds for the Central Universities.

*Corresponding author.
Email address: cslinzhang@tongji.edu.cn (Lin Zhang)

improving the accuracy of pose estimation without destroying the global scale consistency. The efficiency and efficacy of SEOVO have been corroborated by extensive experiments conducted on both public datasets and our collected video sequences.
Keywords: Indoor visual odometry, self-supervised learning, unsupervised semantic segmentation, epipolar geometric constraint, online learning.

---

1. Introduction

With the development of various 3D visual perceptual tasks such as smart homes, 3D indoor navigation and Augmented Reality (AR), accurate localization [1] and high-quality 3D indoor reconstruction [2] have been turning into two fundamental tasks of great concern. In recent years, visual SLAM has been developed rapidly as an excellent solution to both tasks, and is successfully applied to various mobile reconstruction platforms, such as the mobile robot, the mobile backpack and the micro UAV (Unmanned Aerial Vehicle).

As an indispensable part of visual SLAM, VO (Visual Odometry) is responsible for calculating the camera pose between adjacent frames and recovering the local map. According to the optimization frameworks adopted, existing VO schemes mainly fall into two categories, the traditional ones and the learning-based ones. Next, we will analyze both schemes and summarize their limitations.

The traditional VO schemes suffer from low-level manual features, which are fragile to illumination variations and texture distributions. Nearly all of the classic SLAM systems are built on the golden rule of feature corre-

spondence [3, 4] and often perform reliably in ideal environments with rich textures. However, the robustness of these point-based SLAM systems in low-texture regions needs to be improved. Later, line-based SLAMs [5, 6, 7] and plane-based SLAMs [8, 9] came into existence. However, the former ones rely on the line descriptors which are time-consuming in extraction while the latter ones usually assume that the camera poses are known. Also, their utilized line (plane) features are still low-level and handmade, leading to the extractions of lines and planes being vulnerable to occlusion and the damage of edge patterns.

Existing learning-based VO schemes usually have limitations in *generalization abilities* whether they are supervised or not. Compared with traditional VOs, they extract high-level features through convolutional networks and jointly learn the prediction of monocular depth and ego-motion, showing excellent performance in locating and mapping. However, the majority of them are based on offline learning. Accordingly, they optimize network models on the training data and inference on the testing data with model parameters fixed. Thus, the performance of these offline VOs will significantly decrease when working in new scenes different from the training set.

In addition, few learning-based VO methods pay attention to the *challenging indoor environments* which bring difficulty to the regression of depths and poses due to the low-texture regions and complex indoor motions. Thus, the indoor performance of learning-based VOs is still limited. Besides, although there exist several unsupervised methods leveraging semantic information [10, 11, 12, 13] for monocular depth estimation, they either rely on pretrained semantic segmentation networks with fixed weights or need man-

ually labelled class labels for multi-task training. Last but not least, most existing self-supervised VOs guide their optimization mainly based on the traditional photometric loss but ignore the geometric relationships in the scene, which greatly limits their performance for pose estimation.

On account of the aforementioned limitations, we aim to deal with the indoor challenges by introducing both the semantic and the geometric information of indoor scenes and improve the generalization ability of the learning framework through online optimization. Specifically, we propose an online adaptive VO with semantic assistance under an implicit epipolar constraint, namely SEOVO (Semantic Epipolar-constrained Online VO). To our best knowledge, SEOVO is the first online monocular deep VO with semantic assistance. In summary, our contributions are mainly threefold:

1. The first semantics-aided VO following an online adaptive framework, SEOVO, is proposed. Different from offline training schemes which inference on the testing data with the model parameters fixed, SEOVO is optimized under an online meta-learning framework [14, 15], adapting itself to every new frame.
2. A novel epipolar constrained photometric loss is designed to implicitly introduce the epipolar geometric constraint free of the scale uncertainty problem to guide the pose regression. Considering that the epipolar pose solved based on epipolar geometry suffers from scale ambiguity, SEOVO determines its scale by aligning it with the predicted pose to construct this newly designed photometric loss for pose estimation instead of explicitly taking it as pose supervision.
3. A "multi-grad" map fusing the gradients embedded with the photo-

metric, the semantic and the geometric information is designed. Compared with the single photometric gradient map, this fused gradient map can better distinguish different instances, especially in textureless regions. Additionally, the "multi-grad" map can be of great benefit in ensuring a sharp depth map along the object edges. To make our results reproducible, our codes and data are available at `https://cslinzhang.github.io/SEOVO/SEOVO.html`.

The remainder of this paper is organized as follows. Sec. 2 introduces related studies. Sec. 3 provides some preliminary knowledge for better understanding this article. Details of the proposed SEOVO are presented in Sec. 4. Experimental results are reported in Sec. 5. Finally, Sec. 6 concludes the paper.

2. Related Work

2.1. Self-supervised Depth and Pose Estimation

Early works of depth estimation [16, 17] are mostly supervised and achieve excellent performance, but it is expensive to capture ground-truth data in many real-world scenes. Compared with the above schemes, jointly training the depth and pose network from unlabelled monocular videos [18, 19] shows its simplicity and effectiveness, which attracts a lot of researchers' interests and inspires a series of works including ours.

The motions involved in outdoor SLAM datasets are dominated by translations, which are easier for the pose network to regress [20]. Lately, researchers got to find that current approaches could not achieve comparative performance on indoor datasets as outdoor ones. To fill in this gap to some

extent, a few solutions focusing on the characteristics of indoor environments and indoor motions emerged. MonoIndoor++ [21, 22] took a different strategy of progressively estimating the rotations via a residual pose module instead of directly removing them. MovingIndoor [23] proposed an optical-flow based training paradigm which reduced the difficulty of unsupervised learning by utilizing the results of pretrained optical-flow network as the supervision. Zhao *et al.* [24] employed a differentiable two-view triangulation layer to generate a sparse depth map as the self-supervision on depth. However, their triangulated depth map is sensitive to mismatch. CEGVO [25] proposed an end-to-end global-context-aware visual odometry an augmented-attention-enhanced block in its model to learn the long-range dependencies and internal correlation. Work [26] gave a concise derivation of the pure pose function and designed a novel two-view imaging loss function for self-supervised learning. Notably, few of the aforementioned works explicitly deal with the challenging weakly textured scenes, which are commonly encountered in indoor environments and have not been well coped with yet.

2.2. Solutions to Low-Texture Environments

Most traditional visual SLAM systems are based on point features [3, 4, 7]. However, in some low-texture environments, these features are extremely sparse or distributed unevenly, leading to the degradation or even failure of the system. Therefore, the line-based [5, 6] and the plane-based [8, 9] SLAM systems were developed. Yang *et al.* [5] extracted sufficient line features on the premise of ensuring real-time performance, so as to obtain a visual SLAM with higher accuracy and robustness. Planes were also used to reconstruct the low-texture areas. However, these plane-based methods

assume that the poses have been provided by traditional SLAM and only focus on depth estimation. Pop-up SLAM [9] popped the plane based on the room layout and fused point-based and plane-based SLAM together to enhance depth estimation. The aforementioned methods are all based on low-level features and tend to fail when the target regions are occluded or the boundaries are broken. Instead, Zhou *et al.* [23] resorted to deep learning to capture high-level features in the scene. In their work, a well-designed SF-Net actively propagated the sparse initial flows at low-textured regions to the entire image. Although it improves the performance in low-textured regions in most cases, it requires a certain amount of sparse seeds which cannot be satisfied when low-textured regions are relatively large. On the other side, several approaches resorted to semantic understanding for better indoor reconstruction. For example, Concha *et al.* [27] suggested using the layout of the room to generate a prior depth map for dense mapping. But, they didn't track and update the layout, so theoretically their solution can only work in small space.

3. Preliminaries

In this section, we provide some preliminaries for a better understanding of our work. Specifically, we give a brief introduction of the common loss terms widely utilized in self-supervised VOs, which were mainly proposed in [18, 19]. These losses also motivate us to design our optimization objectives.

Photometric loss. The photometric loss penalizes the difference between the target frame and the warped one synthesized by the predicted pose and depth. Specifically, with the target image $\boldsymbol{I}_t$ and the source image $\boldsymbol{I}_s$ as

inputs, the network can output the local transformation matrix from the target view to the source one denoted by $\boldsymbol{T}_{t\to s}^{L}$ and their individual depth maps denoted by $\boldsymbol{D}_t$ and $\boldsymbol{D}_s$, respectively. The warping transformation can be formulated as:

$$\boldsymbol{p}_t'(\boldsymbol{\theta}_D, \boldsymbol{\theta}_P) = \frac{1}{Z_{\boldsymbol{p}_t'}} \boldsymbol{K} \boldsymbol{T}_{t\to s}^{L}(\boldsymbol{\theta}_P) \boldsymbol{D}_t(\boldsymbol{p}_t; \boldsymbol{\theta}_D) \boldsymbol{K}^{-1} \boldsymbol{p}_t \tag{1}$$

where $\boldsymbol{K}$ denotes the camera intrinsic matrix, $Z_{\boldsymbol{p}_t'}$ is the depth of the warped 3D point in the adjacent camera's coordinate system, and $\boldsymbol{p}_t$ and $\boldsymbol{p}_t'$ are the pixel coordinates before and after warping. $\boldsymbol{\theta}_D$ and $\boldsymbol{\theta}_P$ denote the network parameters to be optimized from DepthNet and PoseNet, respectively. Based on the mapping relationship in Eq. 1, we can sample in $\boldsymbol{I}_s$ to obtain the warped image $\boldsymbol{I}_t'$ by

$$\boldsymbol{I}_t'(\boldsymbol{p}_t; \boldsymbol{\theta}_D, \boldsymbol{\theta}_P) = \boldsymbol{I}_s(\boldsymbol{p}_t'(\boldsymbol{\theta}_D, \boldsymbol{\theta}_P)) \tag{2}$$

The photometric loss calculates the distance between $\boldsymbol{I}_t$ and $\boldsymbol{I}_t'$ as [19]:

$$\begin{aligned} L_{pho} = \frac{1}{\|\boldsymbol{\mathcal{V}}\|} \sum_{\boldsymbol{p}_t \in \boldsymbol{\mathcal{V}}} [\lambda \left\| \boldsymbol{I}_t(\boldsymbol{p}_t) - \boldsymbol{I}_t'(\boldsymbol{p}_t; \boldsymbol{\theta}_D, \boldsymbol{\theta}_P) \right\|_1 \\ + (1-\lambda) \frac{1 - SSIM_{tt'}(\boldsymbol{p}_t; \boldsymbol{\theta}_D, \boldsymbol{\theta}_P)}{2}] \end{aligned} \tag{3}$$

where $\boldsymbol{\mathcal{V}}$ stands for the set of valid points without boundary or occluded points. $SSIM_{tt'}(\cdot)$ calculates element-wise similarity between $\boldsymbol{I}_t$ and $\boldsymbol{I}_t'$ by the Structured Similarity (SSIM) function [28], and λ can be empirically set to 0.15 as suggested by Bian *et al.* [19].

Depth smoothness loss. The depth smoothness loss L_{sm} is employed to regularize the depth estimation, which indeed ensures the smoothing of the depth map to be guided by the image gradient. Formally,

$$L_{sm} = \frac{1}{\|\boldsymbol{\mathcal{V}}\|} \sum_{\boldsymbol{p}_t \in \boldsymbol{\mathcal{V}}} (exp(-\boldsymbol{G}_{pho}(\boldsymbol{p}_t)) \cdot \nabla \boldsymbol{D}_t(\boldsymbol{p}_t; \boldsymbol{\theta}_D))^2 \tag{4}$$

where ∇ is the gradient operator and $\boldsymbol{G}_{pho}$ is the gradient map of the target image calculated by $\boldsymbol{G}_{pho} = \nabla \boldsymbol{I}_t$.

Scale consistency loss. The scale consistency loss L_{sc} constrains the frame-to-frame depth scale to maintain a globally consistent map. Given the source depth map $\boldsymbol{D}_s$ and the synthesized one $\boldsymbol{D}'_s$ generated by the target depth map $\boldsymbol{D}'_t$ and the pose $\boldsymbol{T}^L_{t\to s}$, the scale consistency loss is defined as:

$$L_{sc} = \frac{1}{\|\boldsymbol{\mathcal{V}}\|} \sum_{\boldsymbol{p}_t \in \boldsymbol{\mathcal{V}}} \frac{\|\boldsymbol{D}_s(\boldsymbol{p}_t; \boldsymbol{\theta}_D) - \boldsymbol{D}'_s(\boldsymbol{p}_t; \boldsymbol{\theta}_D, \boldsymbol{\theta}_P)\|_2}{\boldsymbol{D}_s(\boldsymbol{p}_t; \boldsymbol{\theta}_D) + \boldsymbol{D}'_s(\boldsymbol{p}_t; \boldsymbol{\theta}_D, \boldsymbol{\theta}_P}). \tag{5}$$

4. Methodology

The overall pipeline of our SEOVO is shown in Fig. 1 and the whole optimization process is conducted under the meta-learning framework [14, 15], which online fine-tunes the network models for each coming frame. It takes consecutive frames from the monocular video as inputs and utilizes three main networks, the DepthNet, the PoseNet, and the FlowNet to predict the depth of the target image, the relative pose between adjacent frames and the corresponding optical flows, respectively. Lying at the core of SEOVO are two innovative modules, namely the pose alignment module and the semantic extraction module. In the following, we will give detailed explanations to the key components of SEOVO, the pose alignment module, the semantic extraction module, the multi-grad fusion and the loss function, respectively.

4.1. Pose Alignment Module

The pose alignment module is responsible for generating the scaled pose $\boldsymbol{T}^S$(the subscript $t \to s$ is omitted for brevity) to introduce geometric con-

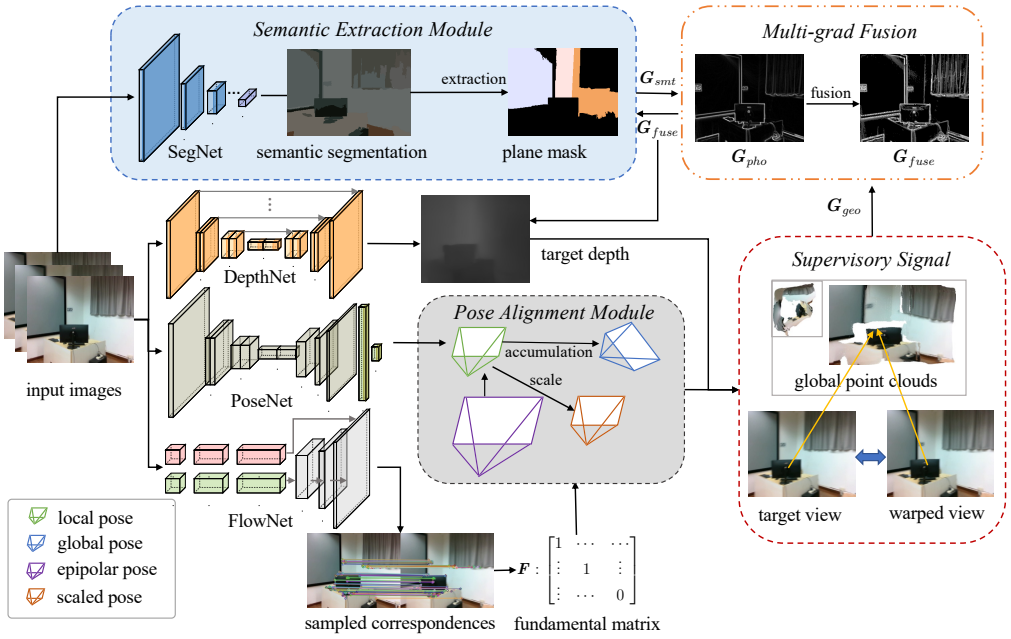

Figure 1: The system pipeline of our SEOVO. The "Pose Alignment Module" introduces the geometric constraints to pose regression through the alignments of the predicted pose while the "Semantic Extraction Module" utilizes the semantic prior to extract the low-texture planes. Also, we generate a multi-grad map $\boldsymbol{G}_{fuse}$ through "Multi-grad Fusion" to fuse the photometric, the semantic and the geometric gradients denoted by $\boldsymbol{G}_{pho}, \boldsymbol{G}_{smt}$ and $\boldsymbol{G}_{geo}$, respectively.

straints into SEOVO. Denoting the local predicted pose from PoseNet by $\boldsymbol{T}^L$, this module conducts *scale alignment* based on $\boldsymbol{T}^L$ to obtain $\boldsymbol{T}^S$.

As we know, compared with outdoor sequences, indoor ones contain much more rotational motions which bring difficulty to pose regression. We try to solve this problem with the aid of the epipolar geometric constraint. However, owing to the scale uncertainty of this 2D-2D constraint, the translation vector of the epipolar pose $\boldsymbol{T}^E$ is usually normalized, which is free of any scale information. In order to make $\boldsymbol{T}^E$ have proper scale, we align $\boldsymbol{T}^E$ with the

predicted one $\boldsymbol{T}^L$ to produce the scaled pose $\boldsymbol{T}^S$ as:

$$\boldsymbol{T}^S = \begin{bmatrix} \boldsymbol{T}^E_{(R)} & s \cdot \boldsymbol{T}^E_{(t)} \\ \boldsymbol{0}^T & 1 \end{bmatrix}, s = \frac{\boldsymbol{T}^L_{(t)}[i]}{\boldsymbol{T}^E_{(t)}[i]}, i = \arg\max_{0 \le i \le 2} |\boldsymbol{T}^L_{(t)}[i]| \tag{6}$$

where $\boldsymbol{T}^E_{(R)}, \boldsymbol{T}^E_{(t)}$ are the rotation matrix and the translation vector of $\boldsymbol{T}^E$ respectively, and $\boldsymbol{T}^L_{(t)}$ is the translation vector in $\boldsymbol{T}^L$. $[i]$ returns the $i\text{-}th$ element in the translation vector $\boldsymbol{T}_{(t)}$. Through Eq. 6, $\boldsymbol{T}^S$ keeps the reliable rotation matrix and translational direction calculated based on epipolar constraints, while preserving the global scale at the same time.

So far, we have obtained the target $\boldsymbol{T}^S$. Then the corresponding loss term L_{epho} can be generated, which is given in Eq. 10. By this loss term, we can ensure the multi-view geometric consistency in pose regression.

4.2. Semantic Extraction Module

Considering that low-texture environment is a challenging problem in indoor VOs, we design the semantic extraction module to extract textureless areas so that explicit constraints can be imposed on these regions to guarantee the reconstruction quality. Technically, we propose a coarse-to-fine extraction strategy containing two stages, namely the coarse extraction and the fine extraction.

Coarse extraction. First, we design a screening principle based on the "photometric overfitting" phenomenon to form a coarse set of texturelss pixels denoted by $\boldsymbol{\mathcal{C}}$. Specifically, we have an interesting insight that the photometric errors of the pixel correspondences tend to be extremely small in textureless areas. The underlying reason is that pixels in these areas usually have similar colors, and small photometric errors can be kept even if the

correspondences are mismatched, which is called the phenomenon of "photometric overfitting". In order to make use of this phenomenon to help extract textureless regions, we form a photometric error map $\boldsymbol{M}_{pho}$ where $\boldsymbol{M}_{pho}(\boldsymbol{p}_t)$ records the photometric error of an arbitrary pixel $\boldsymbol{p}_t$ in the target image $\boldsymbol{I}_t$. Then 2D points with small photometric errors are extracted by Eq. 7 to form a coarse textureless pixel set $\boldsymbol{\mathcal{C}}$.

$$\boldsymbol{\mathcal{C}} = \{\boldsymbol{p}_i | \boldsymbol{M}_{pho}(\boldsymbol{p}_i) < \mu_{pho} + \sigma_{pho}, \boldsymbol{p}_i \in \boldsymbol{I}_t\} \tag{7}$$

where μ_{pho} and σ_{pho} are the mean and the variance of $\boldsymbol{M}_{pho}$, respectively. Second, having obtained the sparse textureless points, we gather the corresponding class labels of all $\boldsymbol{p}_i$ in set $\boldsymbol{\mathcal{L}}$ based on the segmentation map $\boldsymbol{M}_{seg}$ predicted by SegNet. Based on the assumption that textureless regions are mostly planar segments [27], we collect pixels whose semantic labels are in $\boldsymbol{\mathcal{L}}$ and divide them into different planes according to their classes to form a coarse set of textureless planes denoted by $\boldsymbol{\mathcal{P}}^C$.

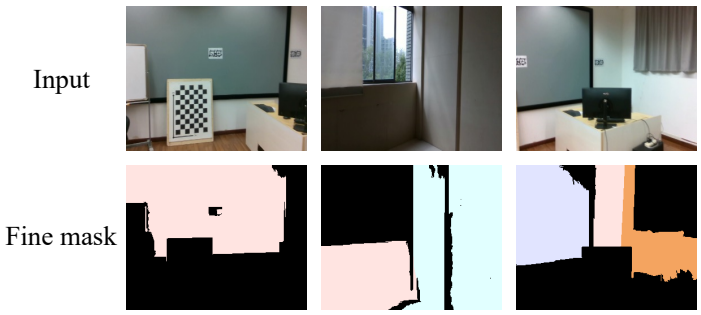

Figure 2: The visualization results of the final extracted textureless planes in different scenes. Different planes are filled with different colors.

Fine extraction. Although the extraction strategy in the coarse stage takes the semantic information into consideration, it is still unable to separate

textureless planes with the same semantic labels. A typical example in indoor scenes is walls in the corner as shown in Fig. 3 (b). To deal with this problem, we further refine the coarse extraction results in this stage. First, for each coarse region $P^C \in \boldsymbol{\mathcal{P}}^C$, only its largest connected region is kept to get rid of small planar regions in consideration of robustness and efficiency. Then we borrow advantage of our designed multi-grad map $\boldsymbol{G}_{fuse}$, which will be introduced in Sec. 4.3, to segment different instances with the same semantic labels. Thanks to the rich edge information provided by $\boldsymbol{G}_{fuse}$, the separating line (such as the wall joint) of different instances can be easily detected although they belongs to the same class so as to achieve a finer planar segmentation. The final set of extracted planes is denoted by $\boldsymbol{\mathcal{P}}^H$ and some visualized results are illustrated in Fig. 2, in which textureless planes are nicely extracted and segmented so that corresponding plane constraints can be applied.

4.3. Multi-grad Fusion

In order to further improve the performance of depth estimation in low-texture areas, inspired by the depth smoothness loss in [19], we design a fused smoothness loss which utilizes a multi-grad map to guide the optimization of depth with smoothness priors. Apart from the commonly used photometric gradient map $\boldsymbol{G}_{pho}$, our proposed multi-grad map also integrates a semantic gradient map and a geometric gradient map denoted by $\boldsymbol{G}_{smt}$ and $\boldsymbol{G}_{geo}$ respectively to help capture more edges consistent with perception. In detail, $\boldsymbol{G}_{smt}$ represents the gradient map of the segmentation results while $\boldsymbol{G}_{geo}$ stands for the gradient map of a pseudo-colored position map denoted by $\boldsymbol{I}_{geo}$. In $\boldsymbol{I}_{geo}$, the 3D coordinates of all pixels are stored in the corresponding

grids. Then the geometric gradient $\boldsymbol{G}_{geo}$ is acquired by $\boldsymbol{G}_{geo} = \nabla^2 \boldsymbol{I}_{geo}$.

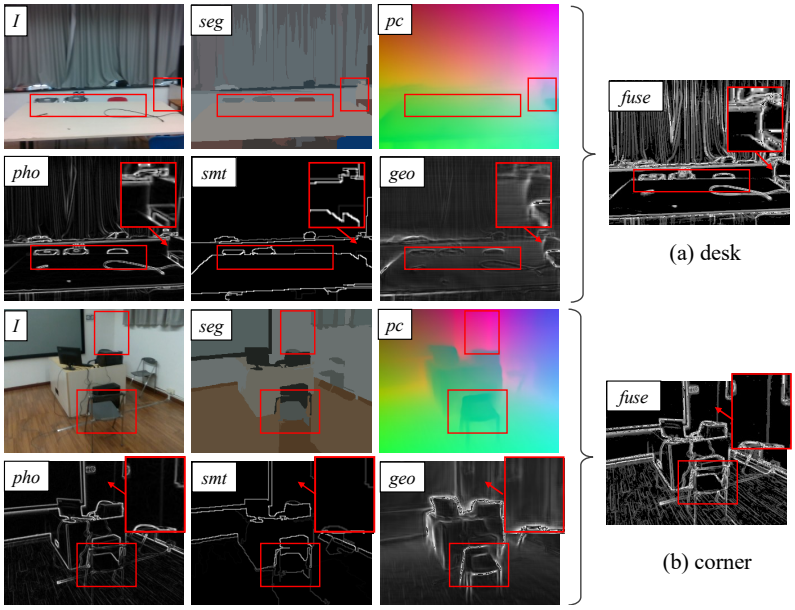

Figure 3: Visualization of "multi-grad" maps in two typical scenes. The label "I", "seg" and "pc" denote the original RGB image, the pseudo-color maps of the segmentation result and that of the point clouds, respectively. Under them are the corresponding gradient maps, where "pho", "smt", and "geo" represent the visualization results of $\boldsymbol{G}_{pho}, \boldsymbol{G}_{smt}$ and $\boldsymbol{G}_{geo}$, respectively. The final fused "multi-grad" map is marked by $fuse$. The red boxes represent difficult regions for edge detecting where our fused gradient map outperforms other gradient maps

Having obtained the multi-source gradients $\boldsymbol{G}_{pho}, \boldsymbol{G}_{smt}$ and $\boldsymbol{G}_{geo}$, we can formulate the final multi-grad gradient map $\boldsymbol{G}_{fuse}$ as:

$$\boldsymbol{G}_{fuse} = (exp(\boldsymbol{G}_{geo}) + \frac{\mu_{geo}}{\mu_{smt}} \boldsymbol{G}_{smt}) * \boldsymbol{G}_{pho} \tag{8}$$

where μ_{geo} and μ_{smt} are the mean of $\boldsymbol{G}_{geo}$ and $\boldsymbol{G}_{smt}$, respectively, and $*$ is the elementwise multiplication. It can be seen from Fig. 3 that more reasonable edges in the indoor scenes can be observed from $\boldsymbol{G}_{fuse}$ instead of

$\boldsymbol{G}_{pho}$ thanks to the assistance of the semantic and the geometric information. In addition, the edges detected in $\boldsymbol{G}_{fuse}$ are also used in the fine stage of the semantic extraction module discussed in Sec. 4.2 to obtain more accurate segmentation results of textureless planes.

During training, the fused gradient map $\boldsymbol{G}_{fuse}$ is of great utility in two aspects. On the one hand, $\boldsymbol{G}_{fuse}$ can be utilized in the semantic extraction module to select more accurate textureless planes, based on which the semantic loss L_{smt} (Eq. (11)) can be generated. On the other hand, $\boldsymbol{G}_{fuse}$ is the input of the smoothness loss L_{fs} (Eq. (12)), making L_{fs} better guide the optimization of the depth map to ensure sharper edges in it.

4.4. Loss Function

The overall loss function of our SEOVO is given as:

$$\begin{aligned} L_{total}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_P) =& L_{epho}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_P) + w_1 L_{smt}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_P) \\ &+ w_2 L_{sc}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_P) + w_3 L_{fs}(\boldsymbol{\theta}_D) \end{aligned} \tag{9}$$

where L_{epho} is the epipolar constrained photometric loss, L_{smt} is the semantic loss, L_{fs} stands for the fused smoothness loss, and L_{sc} refers to the scale consistency loss, which was proposed in [19]. The hyper-parameters are set to $w_1 = 0.5$, $w_2 = 0.1$ and $w_3 = 0.5$. $\boldsymbol{\theta}_D$ and $\boldsymbol{\theta}_P$ denote the network parameters to be optimized from DepthNet and PoseNet, respectively. Among all loss terms, L_{sc} has been given in Eq. 5. Next, we will introduce the other three loss terms in detail.

Epipolar constrained photometric loss term. L_{epho} forces PoseNet to regress pose under the epipolar geometric constraint. However, directly penalizing the distance between $\boldsymbol{T}^S$ and $\boldsymbol{T}^L$, which is considered as an explicit

strategy to impose the epipolar geometric constraint on PoseNet, shows poor performance in regressing the translation motion. The underlying reason is that the depth range of indoor sequences varies a lot across different frames, and thus the scale of $\boldsymbol{T}^S$ will also change abruptly. Instead, we propose L_{epho} to avoid unstable scales, and to implicit impose the epipolar constraint on pose regression. Technically, similar to Eq. 3, L_{epho} penalizes the photometric inconsistency between $\boldsymbol{I}_t$ and the synthesized target image $\hat{\boldsymbol{I}}_t$ which is warped from $\boldsymbol{I}_t$ based on $\boldsymbol{T}^S$ and output depth $\boldsymbol{D}_t$. Formally,

$$\begin{aligned} L_{epho}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_P) = \frac{1}{\|\boldsymbol{\mathcal{V}}\|} \sum_{\boldsymbol{p}_t \in \boldsymbol{\mathcal{V}}} [\lambda \left\| \boldsymbol{I}_t(\boldsymbol{p}_t) - \hat{\boldsymbol{I}}_t(\boldsymbol{p}_t; \boldsymbol{\theta}_D, \boldsymbol{\theta}_P) \right\|_1 \\ + (1-\lambda) \frac{1 - SSIM_{t\hat{t}}(\boldsymbol{p}_t; \boldsymbol{\theta}_D, \boldsymbol{\theta}_P)}{2}] \\ \hat{\boldsymbol{I}}_t(\boldsymbol{p}_t; \boldsymbol{\theta}_D, \boldsymbol{\theta}_P) = \boldsymbol{I}_s(\frac{1}{Z_{\hat{\boldsymbol{p}}_t}} \boldsymbol{K} \boldsymbol{T}^S(\boldsymbol{\theta}_P) \boldsymbol{D}_t(\boldsymbol{p}_t; \boldsymbol{\theta}_D) \boldsymbol{K}^{-1} \boldsymbol{p}_t) \end{aligned} \tag{10}$$

where $Z_{\hat{\boldsymbol{p}}_t}$ is the depth of the 3D point warped by $\boldsymbol{T}^S$ in adjacent camera's coordinate system, $SSIM_{t\hat{t}}(\cdot)$ calculates element-wise similarity between $\boldsymbol{I}_t$ and $\hat{\boldsymbol{I}}_t$, and λ is set to 0.15.

Semantic loss. Weakly textured places are difficult to be well reconstructed due to the lack of photometric features. To cope with this problem, we apply L_{smt} to enforce the surface points in a planar region to share the same normal direction:

$$\begin{aligned} L_{smt}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_P) = \sum_{P^H \in \boldsymbol{\mathcal{P}}^H} \left\| var(\boldsymbol{\mathcal{N}}(P^H(\boldsymbol{\theta}_D, \boldsymbol{\theta}_P))) \right\|_2^2 \\ + \frac{1}{N^H} \sum_{P^H \in \boldsymbol{\mathcal{P}}^H} \sum_{\boldsymbol{p}^H \in P^H} \left\| \nabla \boldsymbol{D}_t(\boldsymbol{p}^H; \boldsymbol{\theta}_D) \right\|_2^2 \end{aligned} \tag{11}$$

where $\boldsymbol{\mathcal{N}}(P^H)$ returns the normal map of P^H, $var(\cdot)$ calculates the channel-wise variances of the given normal map, and N^H is the number of all pixels recorded in $\boldsymbol{\mathcal{P}}^H$.

Fused smoothness loss. Different with the commonly used smoothness loss term L_{sm} introduced in Eq. 4, our fused smoothness loss L_{fs} better guides the optimization of depths, relying on our multi-grad map $\boldsymbol{G}_{fuse}$ instead of $\boldsymbol{G}_{pho}$, which introduces richer edge information consistent with perception. Specifically, L_{fs} is given as:

$$L_{fs}(\boldsymbol{\theta}_D) = \frac{1}{\|\boldsymbol{\mathcal{V}}\|} \sum_{\boldsymbol{p}_t \in \boldsymbol{\mathcal{V}}} (exp(-\boldsymbol{G}_{fuse}(\boldsymbol{p}_t)) \cdot \nabla \boldsymbol{D}_t(\boldsymbol{p}_t; \boldsymbol{\theta}_D))^2. \tag{12}$$

5. Experimental Results

5.1. Experimental Setup

Implementation details. All experiments in this paper were conducted on the same desktop computer equipped with a GPU of NVIDIA GeForce RTX 3070. The training and evaluation codes of SEOVO were implemented using PyTorch. Our SEOVO can process the input data set at about 10FPS. As for network architectures, we borrowed FlowNet from [24] and used the same PoseNet and DepthNet as [19]. Also, the pre-training of FlowNet follows the same training process as in [24] while the pre-training of PoseNet and DepthNet follows [19]. In terms of SegNet, we employed the unsupervised image segmentation network in [29] and this segmentation network strictly follows the online learning framework via back propagation. It can quickly complete the training process online without pre-training and generate semantic segmentation results for each input image. This online self-supervised network does not rely on semantic labels as input and exhibits better generalization capability, enabling us to handle textureless regions effectively. All images were resized to 320×256 and then fed to SEOVO in time order. The weights

sequence	resolution	RGB	depth
Lab	640 × 480	856	856
Corridor	848 × 480	2021	2021

Table 1: Specification of our collected dataset

of network models were initialized by the pretrained weights on the NYUv2 dataset [30]. The initialized weights of FlowNet and SegNet are frozen while those of DepthNet and PoseNet are tuned during training. During online training under the meta-learning framework proposed in [14, 15], the Adam optimizer was used with 10^{-4} as the initial learning rate.

Datasets. First, we evaluated our proposed SEOVO on two public indoor datasets, RGB-D 7-scenes [31] and ScanNet [32]. Then to better confirm the benefits of our SEOVO in textureless environments, we also collected two video sequences, namely "Lab" and "Corridor" respectively, by a Realsense d453i camera for online testing. These sequences include large areas of texturelss regions which are difficult to be handled by existing deep VO schemes. The content and scale of this dataset are given in Tabel 1. All the utilized datasets provide RGB sequences and the associated depth maps. In addition, public datasets also provide frame-to-frame poses for evaluating the accuracy of pose regression. Considering that we follow an online learning framework, we took the official training splits of NYUv2 dataset [30] as the *training set*. As for the *testing set*, we evaluated all online methods on the typical sequences of public datasets [31, 32] as well as all of our collected data with online fine-tuning.

Evaluation metrics. The evaluation of SEOVO mainly focused on two

aspects, the performance on depth estimation and pose estimation. For the evaluation of depth estimates, we followed [19] to use the mean absolute relative error (AbsRel), the root mean square error (RMS), and the accuracy under threshold ($\sigma_i \leq 1.25^i, i = 1, 2, 3$) as metrics.

To measure the accuracy of the output poses, we selected two widely used metrics in measuring the SLAM trajectory, the Absolute Pose Error (APE) and the Relative Pose Error (RPE). Specifically, we applied the above two pose metrics on both the translation ($/m$) and the rotation components ($/^\circ$), represented by $APE_t \downarrow$, $APE_r \downarrow$, $RPE_t \downarrow$, and $RPE_r \downarrow$. The subscript "t" corresponds to the translation component while "r" indicates the rotation one.

Table 2: Characteristics of related self-supervised VO methods

Methods	Optimization	Cues		
		Photometry	Geometry	Semantics
SC-SfMLearner [19]	offline	√	×	×
MonoIndoor++ [22]	offline	√	×	×
TrianFlow [24]	offline	√	√	×
CEGVO [25]	offline	√	√	×
GeoConst [26]	offline	√	√	×
SC-online18	online	√	×	×
Trian-online	online	√	√	×
OnlineVO[15]	online	√	√	×
SEOVO (ours)	online	√	√	√

Compared methods. We compared our SEOVO with some typical self-

Table 3: Quantitative comparison on depth estimation with related methods on the 7-scenes dataset

Methods/Scenes		Chess		Fire		Heads		Office		Pumpkin		RedKitchen		Stairs	
		AbsRel ↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑
Offline	SC-SfMLearner[19]	0.103	0.880	0.089	0.916	0.124	0.862	0.096	0.912	0.083	**0.946**	0.101	0.896	0.106	0.855
	MonoIndoor++[22]	0.097	0.888	0.077	0.939	0.106	0.889	0.083	0.934	**0.078**	0.945	0.112	0.893	0.139	0.821
	TrianFlow[24]	0.114	0.817	0.107	0.874	0.173	0.755	0.126	0.848	0.112	0.893	0.139	0.821	0.167	0.746
	CEGVO [25]	0.104	0.866	0.082	0.925	0.114	0.906	0.091	0.918	0.089	0.911	0.927	0.900	0.108	0.859
	GEOConst [26]	0.096	0.843	0.080	0.931	0.100	0.901	0.093	0.922	0.112	0.899	0.099	0.909	0.103	0.855
Online	SC-online18	0.098	0.882	0.080	0.919	0.099	0.912	0.086	0.934	0.124	0.868	0.100	0.899	0.106	0.871
	Trian-online	0.107	0.867	0.091	0.907	0.102	0.908	0.096	0.913	0.101	0.904	0.117	0.871	0.129	0.831
	OnlineVO[15]	0.101	0.878	0.079	0.912	0.132	0.884	0.100	0.900	0.123	0.893	0.098	0.903	0.129	0.829
	SEOVO (ours)	**0.091**	**0.892**	**0.075**	**0.941**	**0.097**	**0.925**	**0.080**	**0.936**	0.103	0.901	**0.088**	**0.917**	**0.097**	**0.876**

supervised VO methods, including SC-SfMLearner [19], MonoIndoor++ [22], TrianFlow [24], CEGVO [25], and GeoConst [26]. Considering that there are few public online VO methods, we extended offline works [19, 24] which are closely related to our SEOVO to the online versions for comparison, denoted by SC-online18 and Trian-online, respectively. Technically, these extended online schemes were optimized based on the same network architectures and loss functions as their original works, but under an online framework. In general, we give the characteristics of the above methods in Table 2. It can be seen from Table 2 that only our SEOVO combines the photometric, geometric and semantic cues mined from the input images under an online adaptive optimization framework. This fusion technique accounts for our outstanding performance on estimation accuracy.

5.2. Quantitative Experiment

Depth estimation on public datasets. We first present the quantitative depth estimation results of our SEOVO and several offline or online competitors on 7-scenes and ScanNet in Table 3 and Table 4. From the results presented in the above tables, we make the following observations.

Table 4: Quantitative comparison on depth estimation with related methods on the ScanNet dataset

Scenes	0000_00		0059_00		0101_04		0106_00		0169_00		0181_00		0241_00	
	AbsRel ↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑
CEGVO [25]	0.068	0.949	0.095	0.904	0.103	0.891	0.166	0.761	0.104	0.883	0.223	0.669	0.109	0.877
GeoConst [26]	0.074	0.940	0.099	0.906	0.105	0.884	0.173	0.754	0.102	0.890	0.187	0.718	0.103	0.901
SC-online18	0.066	0.954	0.100	0.895	0.096	0.909	0.163	0.769	0.101	0.889	0.189	0.722	0.103	0.896
Trian-online	0.065	**0.956**	0.109	0.882	0.126	0.882	0.181	0.747	0.109	0.865	0.186	0.722	0.160	0.878
OnlineVO[15]	0.066	0.952	0.122	0.849	0.102	0.903	0.159	0.770	0.113	0.862	0.186	0.715	0.117	0.888
SEOVO (ours)	**0.061**	**0.956**	**0.094**	**0.912**	**0.093**	**0.911**	**0.150**	**0.775**	**0.096**	**0.905**	**0.153**	**0.771**	**0.089**	**0.926**

First, both the offline SC-SfMLearner and TrianFlow underperform their online variants SC-online18 and Trian-online, respectively, which implies the superiority of online learning in generalization. Second, SEOVO shows an overwhelming performance over other competitors in nearly all scenes owing to the fact that it integrates the geometric structure and the semantic layout of the scene under an online optimization framework.

Depth estimation on our dataset. To demonstrate the performance of SEOVO in tackling the challenging environment with weak textures, we also conducted experiments on our collected testing data containing large areas of textureless regions. To directly evaluate on our testing sequences, we compared our SEOVO with online schemes free of finetune training and give their evaluation results in Table 5. It can be seen from Table 5 that our scheme exhibits clear performance advantages over all counterparts on this challenging dataset owing to the semantic loss which directly constrains the depth regression in the textureless regions.

Pose estimation on public datasets. To evaluate the accuracy of our visual odometry for pose estimation, we first aligned the output trajectories predicted by our SEOVO and typical competitors to the ground truth trajectories respectively. And then we measured the aligned results based on

Table 5: Quantitative comparison with online methods on our collected dataset

Scene	Method	AbsRel↓	RMS↓	σ_1↑	σ_2↑	σ_3↑
Lab	SC-online18	0.118	0.720	0.872	0.991	0.995
	Trian-online	0.155	0.955	0.786	0.948	0.986
	OnlineVO[15]	0.119	0.765	0.867	0.992	0.997
	SEOVO (ours)	**0.113**	**0.708**	**0.893**	**0.996**	**0.999**
Corridor	SC-online18	0.139	0.776	0.837	0.970	0.989
	Trian-online	0.176	0.958	0.808	0.937	0.969
	OnlineVO[15]	0.146	0.815	0.827	0.967	0.988
	SEOVO (ours)	**0.137**	**0.752**	**0.844**	**0.972**	**0.990**

Table 6: Quantitative comparison on pose estimation with related methods on the 7-scenes dataset

Methods/Scenes		Chess		Fire		Heads		Office		Pumpkin		RedKitchen		Stairs	
		APE_t	APE_r	APE_t	APE_r	APE_t	APE_r	APE_t	APE_r	APE_t	APE_r	APE_t	APE_r	APE_t	APE_r
Offline	SC-SfMLearner[19]	0.416	11.774	0.496	29.478	0.176	26.431	0.311	33.455	0.379	23.744	0.229	18.878	0.533	159.473
	MonoIndoor++[22]	0.288	8.658	0.477	25.469	0.164	26.334	0.254	29.647	0.374	21.156	0.265	16.024	0.439	50.612
	TrianFlow[24]	0.457	12.035	0.302	60.459	0.182	43.651	0.425	41.623	0.377	25.491	0.301	24.389	0.513	89.435
	CEGVO [25]	0.388	9.654	0.404	36.431	0.149	33.854	0.229	30.016	0.256	14.764	0.187	15.213	0.648	49.342
	GEOConst [26]	0.325	5.492	0.331	44.226	0.134	23.412	0.195	27.853	0.278	19.855	0.205	18.247	0.355	37.561
Online	SC-online18	0.318	10.457	0.397	24.039	0.148	23.482	0.189	25.717	0.244	7.840	1.074	13.025	0.410	101.35
	Trian-online	0.443	15.781	0.460	60.016	0.166	17.384	0.435	31.052	0.495	38.235	0.263	36.75	0.756	29.731
	OnlineVO[15]	0.278	4.678	0.338	25.191	0.144	16.944	0.190	26.255	0.204	10.260	0.178	12.431	0.394	57.758
	SEOVO (ours)	**0.246**	**4.385**	**0.284**	**23.111**	**0.125**	**13.567**	**0.151**	**21.483**	**0.185**	**7.437**	**0.169**	**10.758**	**0.303**	**22.852**

the four metrics mentioned in Sec. 5.1. The results on 7-scenes, ScanNet are reported in Table 6 and Table 7, respectively. Whether from the absolute metrics or the relative ones, our SEOVO gives the most accurate pose estimations. The underlying reason is that our epipolar constrained photometric loss guides SEOVO to optimize the estimated pose under the epipolar geometric constraint and to maintain the scale consistency at the same time.

Table 7: Quantitative comparison on pose estimation with related methods on the ScanNet dataset

Scenes	0000_00		0059_00		0101_04		0106_00		0169_00		0181_00		0241_00	
	RPE_t	RPE_r	RPE_t	RPE_r	RPE_t	RPE_r	RPE_t	RPE_r	RPE_t	RPE_r	RPE_t	RPE_r	RPE_t	RPE_r
CEGVO[25]	0.145	3.897	0.175	9.475	0.209	10.337	0.205	9.730	0.290	3.967	0.225	10.305	0.135	6.895
GeoConst[26]	0.142	3.792	0.176	9.396	0.214	10.169	0.204	9.949	0.267	3.337	0.244	14.149	0.137	6.244
SC-online18	0.144	3.783	0.176	9.482	0.208	10.107	0.204	10.068	0.284	3.742	0.245	13.252	0.133	6.723
Train-online	0.137	3.417	0.183	9.432	0.216	10.240	0.193	8.892	**0.246**	4.274	0.245	13.194	0.135	5.979
OnlineVO[15]	0.143	3.878	0.180	9.502	0.213	9.082	0.206	9.693	0.273	3.441	0.230	13.753	0.136	6.228
SEOVO (ours)	**0.131**	**3.076**	**0.173**	**9.225**	**0.202**	**9.057**	**0.086**	**7.877**	0.272	**3.402**	**0.206**	**6.572**	**0.132**	**5.719**

5.3. Qualitative Experiment

Visualized depth maps. In order to qualitatively examine our performance in depth regression, depth estimation results on 7-scenes of our SEOVO and other competitors are visualized in Fig. 4. Besides, the ground truth depths denoted by "GT" are also provided as reference. It can be clearly seen from Fig. 4 that our depth maps are closer to the ground truth maps, which is in line with our quantitative results in Table 3. Besides, our depth maps have sharper edges thanks to our multi-grad maps.

Reconstruction results. To show the performance of our SEOVO in both depth estimation and pose regression qualitatively, some typical results of the point clouds synthesized by poses and depths output from SC-online18, OnlineVO and our SEOVO are offered in Fig. 5. In detail, for the completeness of maps, we transformed the synthesized point clouds from several consecutive frames (usually 3 to 5 keyframes) to a unified global coordinate system which takes the first frame as reference. It can be seen from Fig. 5 that the results of other two competitors have obvious misalignments among frames while our SEOVO best keeps the consistency of global maps, corrob-

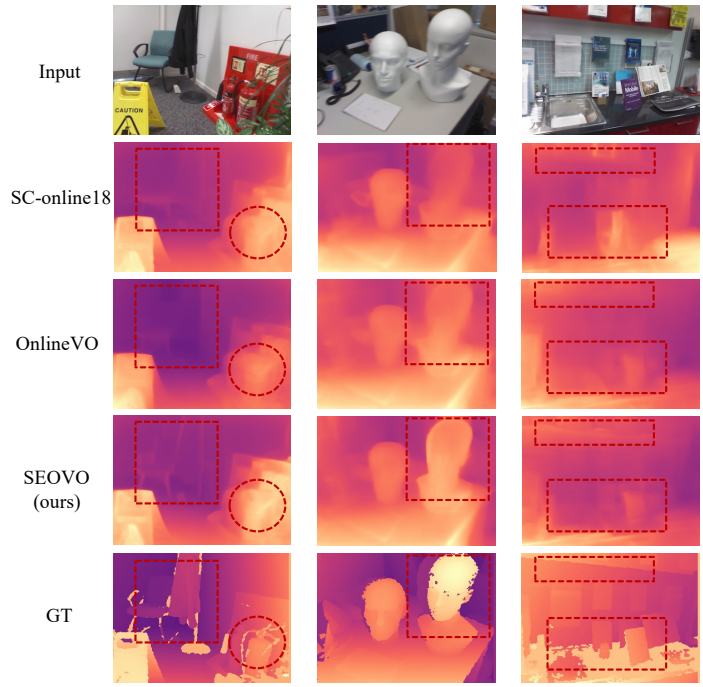

Figure 4: Depth estimation results on the 7-scenes dataset. Thanks to our two innovated modules introducing the semantic and geometric information, SEOVO performs favorably compared to other competitors in better prediction results and preserves sharper edges.

orating the claim that our scheme produces the most accurate depths and poses.

5.4. Ablation Study of Loss Terms

We demonstrate how the two important loss terms, the epipolar constrained photometric loss and the semantic loss in our loss function affect the results by comparing SEOVO with its two variants, EOVO and SOVO. Specifically, compared with SEOVO, EOVO is optimized without the semantic loss L_{smt} while SOVO is without the epipolar constrained photometric loss L_{epho}. Furthermore, to verify the effectiveness of the implicit strategy

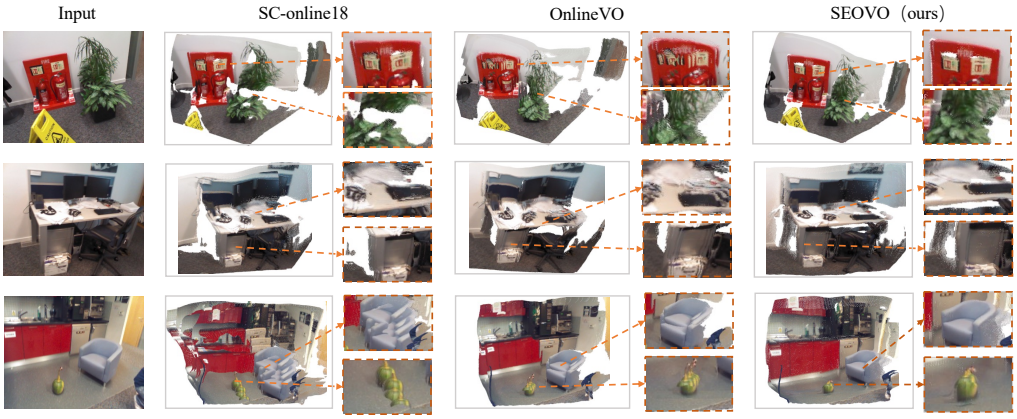

Figure 5: Illustration of the point clouds synthesized by the poses and depths generated by the online baseline SC-online18, OnlineVO and our SEOVO on the 7-scenes dataset. To show the global consistency of the estimation, we warp the maps of several consecutive keyframes to a unified global coordinate system for a more complete model.

adopted by L_{epho}, we constructed another competitor using explicit epipolar constraints named SEEOVO which replaces L_{epho} with the direct pose supervision $L_{ee} = \left\|\boldsymbol{T}^L - \boldsymbol{T}^S\right\|_2^2$ where $\boldsymbol{T}^L$ is the predicted pose and $\boldsymbol{T}^S$ is the scaled pose aligned with $\boldsymbol{T}^L$ in scale.

Table 8 and Table 9 tabulate the quantitative results of SEOVO and its rivals for depth estimation and pose estimation, respectively. In Table 8, we adopted the first-best (black bold) and second-best (blue bold) highlighting in Table IV These two tables demonstrate that our SEOVO outperforms other variants for both depth and pose regression except in the "Stairs" scene. The reason of our unsatisfactory performance in this scene lies in the strong reflection of sunshine on the stairs, which coincides with our failure cases analyzed in Sec. 5.5. Technically, EOVO imposes the geometric constraints on pose regression while SOVO guarantees the performance in

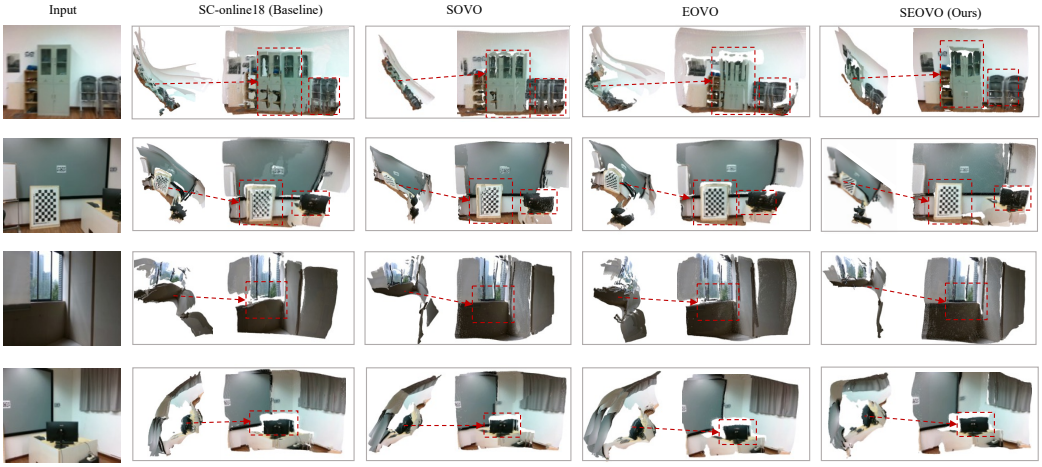

Figure 6: Comparison of the maps reconstructed by the online baseline SC-online18 and our two variants on our collected dataset. To clearly demonstrate the reconstruction results in textureless regions, we provide both the top views and the font views in each block.

textureless regions with semantic assistance. Compared with them, SEOVO combines both geometric and semantic cues, undoubtedly achieving the best performance. Besides, there's an interesting finding that SEEOVO performs worst in most scenes, which validates our analysis in Sec. 4.4 that explicitly introducing epipolar constraints will lead to unsatisfactory predictions.

In addition to the quantitative comparison, Fig. 6 qualitatively shows the reconstruction results of SEOVO and its variants, SOVO and EOVO, on our collected dataset. The synthesized point clouds of SC-online18, which utilizes neither geometric constraints nor semantic priors, are also given in Fig. 6 as the baseline. For clearer observations, we provide both the top and the front views of these maps. It can be clearly seen from the top view in Fig. 6 that SOVO achieves remarkable results in reconstructing the geometric structures

Table 8: Performance of networks trained with various combinations of loss terms on the 7-scenes and our collected dataset

Scenes	EOVO		SOVO		SEEOVO		SEOVO	
	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑	AbsRel↓	σ_1↑
Chess	**0.096**	**0.883**	0.097	0.882	0.104	0.876	**0.091**	**0.892**
Fire	0.087	0.911	**0.086**	**0.917**	0.090	0.903	**0.075**	**0.925**
Heads	**0.100**	0.913	0.101	**0.916**	0.103	0.901	**0.097**	**0.925**
Office	**0.087**	**0.930**	0.092	0.926	0.091	0.920	**0.080**	**0.936**
Pumpkin	**0.127**	**0.874**	0.128	0.873	0.132	0.831	**0.103**	**0.901**
RedKitchen	**0.095**	**0.910**	0.108	0.885	0.118	0.872	**0.088**	**0.911**
Stairs	**0.102**	**0.888**	0.115	0.857	0.134	0.833	**0.097**	**0.876**
Lab	**0.120**	0.861	0.121	0.860	0.122	**0.870**	**0.113**	**0.893**
Corridor	**0.142**	**0.832**	0.158	0.802	0.147	0.821	**0.137**	**0.844**

of low-texture regions only inferior to our SEOVO, demonstrating the effectiveness of the semantic loss which enforces the extracted surface points to be on the same plane. However, it fails to output poses with high accuracy, thus generating obvious misalignments among frames as clearly shown in the front view. On the other hand, although EOVO guarantees the pose accuracy by the epipolar geometric constraint introduced in L_{epho}, it still highly relies on the photometric consistency to guide the depth optimization, which is unreliable in low-texture regions and thus causes the poor reconstruction quality of these places. Fortunately, our SEOVO integrates both the geometric cues and the semantic ones in the scenes, generating the best reconstruction results observed from both the front view and the top view. These results lead us to express the belief that the loss function of SEOVO is well designed and both L_{epho} and L_{smt} play essential roles in it.

Table 9: Quantitative comparison of different variants on the 7-scenes dataset for pose estimation

Scenes	EOVO		SOVO		SEEOVO		SEOVO	
	RPE_t	RPE_r	RPE_t	RPE_r	RPE_t	RPE_r	RPE_t	RPE_r
Chess	0.047	1.437	0.501	1.525	0.068	2.020	**0.040**	**1.211**
Fire	0.074	2.627	0.080	2.549	0.072	2.485	**0.054**	**2.087**
Heads	0.026	1.699	0.026	1.542	0.028	1.618	**0.024**	**1.370**
Office	0.033	1.586	0.037	1.734	0.041	1.856	**0.033**	**1.582**
Pumpkin	0.074	1.152	0.072	1.137	0.092	1.398	**0.067**	**1.085**
RedKitchen	0.040	1.095	0.043	1.083	0.041	1.167	**0.038**	**1.074**
Stairs	0.239	**1.793**	0.195	1.868	**0.109**	1.990	0.172	1.945

5.5. Failure Case Analysis

By observing and analyzing the experimental results, we found that the illumination and reflection of the scene have an obvious influence on SEOVO's performance. For example, when the surface of an object surface is made of strongly reflective materials such as glass or mirror, the information contained in the corresponding area of the image is relatively confusing. Actually, with complex reflections, the basic assumption of the photometric consistency is violated and the impact of noises will be more notable. This problem poses a challenge to SEOVO in extracting high-quality features and the correspondences generated by FlowNet tend to be mismatched, resulting in poor performance as shown in Fig. 7. It can be seen from Fig. 7 that the glass window in front of the artwork is hard to be reconstructed while the artifacts in the mirror cause the pseudo depths.

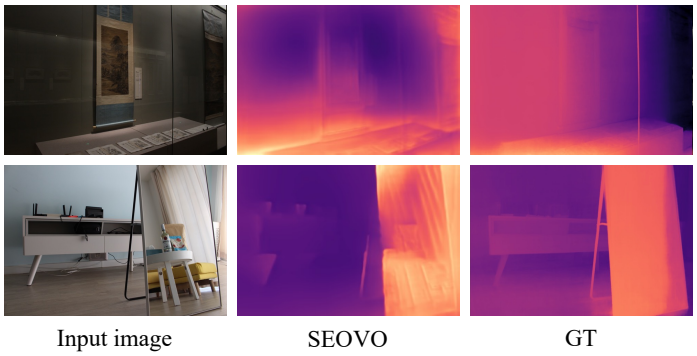

Figure 7: Depth estimation of failure case with the complex reflection owing to the glass and the mirror. In detail, the glass part is hard to distinguish and recover the true depth while the artifacts in the mirror is mistaken for the objects with real depth.

6. Conclusion

In this paper, we studied a practical problem, the indoor deep visual odometry which is of great significance in visual perception tasks such as AR, indoor navigation and smart homes. Considering the limitations of existing methods in accuracy and generalization ability, we proposed a fully self-supervised solution under an online adaptive framework namely SEOVO. To the best of our knowledge, we are the first to introduce semantic assistance under an online framework to make full use of the semantics of each new frame. For the indoor localization task, we introduce the epipolar geometric constraint in an implicit manner to keep the global scale consistency, which accounts for our superiority in trajectory prediction compared with other VO schemes. In terms of the reconstruction task, our newly designed multi-grad map is of great importance in capturing the complex edge information even in difficult weakly-textured areas, which enables our SEOVO's overwhelming performance in indoor 3D reconstruction. In conclusion, the

success of our SEOVO demonstrates that multiple cues, such as photometric information, geometric structures and semantics can compensate each other to handle challenging indoor scenes, and online learning has great potential to narrow the performance gap of a network between the training and the testing phrases.

Application field. Online VO/SLAM is crucial in various fields, enhancing both navigation and environmental understanding. One prominent application is in robotics, where autonomous robots and drones utilize SLAM for real-time mapping and navigation in dynamic environments, allowing them to operate effectively in unstructured spaces. In the realm of augmented reality (AR) and virtual reality (VR), our produced pose and depths play a vital role by enabling devices to accurately track their position relative to the real world, ensuring virtual objects are seamlessly integrated into the user's environment. Similarly, in smart homes, robotic vacuum cleaners employ online SLAM to create efficient cleaning paths and adapt to changing room layouts.

Limitations and future work. Though SEOVO can work well in most cases, its performance is still not satisfactory when working in environments with complex reflections as analyzed in Sec. 5.5. In our future work, we will continue to devote our efforts to this area. For instance, considering that the assumption of multi-view appearance consistency does not hold under this circumstance, the real Euclidean space can be decomposed into multiple virtual subspaces, in which the multi-view consistency can be satisfied. Based on this hypothesis, we will replace the single output depth map of DepthNet with multi-space outputs. Based on these subspace depths and the predicted

pose, we can synthesize the multi-space photometric maps and perform a weighted sum of them to get the final image. Fortunately, several reflection datasets are publicly available to support the training.

References

[1] X. Song, H. Li, L. Liang, W. Shi, G. Xie, X. Lu, X. Hei, TransBoNet: Learning camera localization with Transformer Bottleneck and Attention, Pattern Recognition 146 (2024), pp. 109975:1-11.

[2] S. Song, K. G. Truong, D. Kim, S. Jo, Prior depth-based multi-view stereo network for online 3D model reconstruction, Pattern Recognition, 136 (2023), pp. 109198:1-12.

[3] G. Klein, D. Murray, Parallel tracking and mapping for small AR workspaces, in: Proc. IEEE Int. Symp. Mixed Augmented Reality, 2007, pp. 225-234.

[4] R. Mur-Artal, J. M. M. Montiel, J. D. Tardós, ORB-SLAM: A versatile and accurate monocular slam system, IEEE Trans. Robot. 31 (5) (2015) 1147-1163.

[5] G. Yang, Q. Wang, P. Liu, H. Zhang, An improved monocular PL-SlAM method with point-line feature fusion under low-texture environment, in: Proc. 4th Int. Conf. Contr. Comput. Vis., 2021, pp. 119-125.

[6] A. Pumarola, A. Vakhitov, A. Agudo, A. Sanfeliu, F. Moreno-Noguer, PL-SLAM: Real-time monocular visual SLAM with points and lines, in: Proc. IEEE Int. Conf. Robot. Automat., 2017, pp. 4503-4508.

[7] J. Zhang, J. Yang, F. Fu, J. Ma, PlaneAC: Line-guided planar 3D reconstruction based on self-attention and convolution hybrid model, Pattern Recognition, in press, 2024.

[8] Y. Furukawa, J. Ponce, Accurate, dense, and robust multiview stereopsis, IEEE Trans. Pattern Anal. Mach. Intell. 32 (8) (2010) 1362-1376.

[9] S. Yang, Y. Song, M. Kaess, S. Scherer, Pop-up SLAM: Semantic monocular plane SLAM for low-texture environments, in: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., 2016, pp. 1222-1229.

[10] V. Casser, S. Pirk, R. Mahjourian, A. Angelova, Unsupervised monocular depth and ego-motion learning with structure and semantics, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2019, pp. 381-388.

[11] V. Guizilini, R. Hou, J. Li, A. Gaidon, Semantically-guided representation learning for self-supervised monocular depth, in: Proc. Int. Conf. Learn. Represent., 2020, pp. 1-14.

[12] L. Huynh, P. Nguyen-Ha, J. Matas, J. Matas, E. Rahtu, J. Heikkila, Guiding monocular depth estimation using depth-attention volume, in: Proc. Eur. Conf. Comput. Vis., 2020, pp. 581-597.

[13] X. Xu, Z. Chen, F. Yin, Multi-scale spatial attention-guided monocular depth estimation with semantic enhancement, IEEE Trans. Image Process. 30 (2021) 8811-8822.

[14] S. Li, X. Wang, Y. Cao, F. Xue, Z. Yan, H. Zha, Self-supervised deep

visual odometry with online adaptation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6338-6347.

[15] S. Li, X. Wu, Y. Cao, H. Zha, Generalizing to the open world: Deep visual odometry with online adaptation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 13179-13188.

[16] A. Saxena, S. H. Chung, A. Y. Ng, Learning depth from single monocular images, in: Proc. Adv. Neural Inf. Process. Syst., 2006, pp. 1161-1168.

[17] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, A. Yuille, Towards unified depth and semantic prediction from a single image, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 2800-2809.

[18] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M. M. Cheng, I. Reid, Unsupervised scale-consistent depth and ego-motion learning from monocular video, in: Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 35-45.

[19] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M. M. Cheng, I. Reid, Unsupervised scale-consistent depth learning from video, Int. J. Comput. Vis. 129 (9) (2021) 2548-2564.

[20] Y. Cao, X. Zhang, F. Luo, P. Peng, C. Lin, K. Yang, Y. Li, Learning generalized visual odometry using position-aware optical flow and geometric bundle adjustment, Pattern Recognition 136 (2023), pp. 109262:1-11.

[21] P. Ji, R. Li, B. Bhanu, Y. Xu, MonoIndoor: Towards good practice of self-supervised monocular depth estimation for indoor environments, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 12767-12776.

[22] R. Li, P. Ji, Y. Xu, B. Bhanu, MonoIndoor++: Towards better practice of self-supervised monocular depth estimation for indoor environments, IEEE Trans. Circuits Syst. Video Technol. 33 (2) (2023) 830-846.

[23] J. Zhou, Y. Wang, K. Qin, W. Zen, Moving Indoor: Unsupervised video depth learning in challenging environments, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 8617-8626.

[24] W. Zhao, S. Liu, Y. Shu, Y. J. Liu, Towards better generalization: Joint depth-pose learning without PoseNet, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 9148-9158.

[25] Z. Ji, K. Nan, Y. Xu, H. Wang, X. Li, J. Bai, Global-context-aware visual odometry system with epipolar-geometry-constrained loss function, IEEE Trans. Instrum. Meas. 73 (2024) 1-11.

[26] M. Xiong, Z. Zhang, J. Liu, T. Zhang, H. Xiong, Monocular depth estimation using self-supervised learning with more effective geometric constraints, Eng. Appl. Artif. Intell. 128 (2024).

[27] A. Concha, W. Hussain, L. Montano, J. Cibvera, Incorporating scene priors to dense monocular mapping, Auton. Robots 39 (3) (2015) 279-292.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600-612.

[29] A. Kanezaki, Unsupervised image segmentation by backpropagation, in:

Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2018, pp. 1543-1547.

[30] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, in: Proc. Eur. Conf. Comput. Vis., 2012, pp. 746-760.

[31] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Fitzgibbon, Scene coordinate regression forests for camera relocalization in RGB-D images, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp. 2930-2937.

[32] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Niessner, Scannet: Richly-annotated 3d reconstructions of indoor scenes, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5828-5839.