RGB-Based Visual-Inertial Odometry via Knowledge Distillation from Self-Supervised Depth Estimation with Foundation Models

Anonymous ICCV submission

Paper ID *****

Abstract

Autonomous driving represents a transformative advancement with the potential to significantly impact daily mobility, including enabling independent vehicle operation for individuals with visual disabilities. The commercialization of autonomous driving requires guaranteed safety and accuracy, underscoring the need for robust localization and environmental perception algorithms. In cost-sensitive platforms such as delivery robots and electric vehicles, cameras are increasingly favored for their ability to provide rich visual information at low cost. However, estimating 3D positional changes using only 2D image sequences remains a fundamental challenge, primarily due to inherent scale ambiguity and the presence of dynamic scene elements. In this paper, we present a visual-inertial odometry framework incorporating a depth estimation model trained without ground-truth depth supervision. Our approach leverages a self-supervised learning pipeline enhanced with knowledge distillation via foundation models, including both self-distillation and geometry-aware distillation. The proposed method improves depth estimation performance and consequently enhances odometry estimation, without modifying the network architecture or increasing the number of parameters. The effectiveness of the proposed method is demonstrated through comparative evaluations on both the public KITTI dataset and a custom campus driving dataset, showing performance improvements over existing approaches.

1. Introduction

Recent advances in deep learning, driven by improvements in both hardware and software, have enabled its widespread application across various industries. In the field of computer vision, significant progress has been made not only in well-established areas such as object detection [29] and semantic segmentation [14], but also in depth estimation [5]. Depth estimation is the task of inferring the real-world distance from the camera to the scene surface at each pixel. This technique has become a key for applications in robotics, including autonomous navigation of mobile robots and drones [2, 3], as well as for providing auxiliary information in minimally invasive surgery [21]. Due to its broad applicability across various domains, depth estimation continues to receive increasing attention. Although depth estimation has seen significant progress, it still faces fundamental challenges. Monocular methods are limited by scale ambiguity, making absolute depth recovery ill-posed without additional cues. In contrast, stereo approaches must trade off between long-range accuracy and occlusion handling, which depends on the baseline length. In this paper, we propose a novel monocular depth estimation method based on a foundation model, aimed at improving the performance of simultaneous localization and mapping (SLAM).

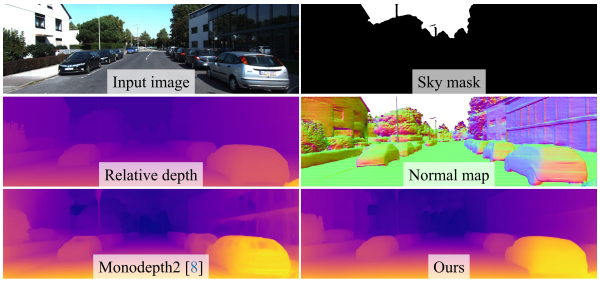


Figure 1. Geometric cues used for knowledge distillation and their effects. Sky mask, relative depth, and surface normal are derived using foundation model-based methods [11, 12, 28]. These geometric cues guide our knowledge distillation framework, which outperforms [8], especially in handling transparent objects.

As a data-driven approach, the performance of deep neural networks is largely determined by the quality and quantity of the training data. Constructing high-quality datasets typically requires ensuring domain diversity, filtering out noisy or erroneous samples, and generating accurate ground truth annotations. In recent years, increasing awareness of ethical and legal concerns—such as privacy and copyright issues—has made data collection more cautious. To address these challenges, recent research has increasingly

focused on the use of foundation models trained on large-scale datasets. In this work, we propose a method that fine-tunes a depth foundation model pretrained on a large-scale dataset comprising synthetic data generated from carefully designed virtual environments, using a target dataset.

A straightforward and widely used approach for training depth estimation models is the supervised learning pipeline. This method minimizes a loss function defined based on the error between the estimated depth map and the ground-truth depth map. It typically yields high performance, as the model is trained directly using explicit supervisory signals. However, generating accurate ground-truth depth data requires expensive light detection and ranging (LiDAR) sensor and precise calibration between sensors. Moreover, such ground-truth data is often domain-specific, leading to overfitting to the data acquisition environment and resulting in limited generalization performance. As an alternative, self-supervised learning approaches based on image reprojection have been proposed to mitigate these limitations. These methods estimate a depth map from an input image and leverage adjacent frames to reproject their content into the viewpoint of the input image, using the photometric error between the synthesized and original views as the supervision signal. Due to their complex and indirect supervision pipelines, self-supervised methods typically exhibit lower accuracy than supervised approaches. To address this limitation, we introduce a self-supervised framework that improves depth estimation by incorporating auxiliary depth cues via knowledge distillation, as shown in Figure 1, while preserving image reprojection-based supervision.

SLAM is a technique that enables a robot to simultaneously construct a map of its environment and estimate its position within it in real time. The performance of SLAM algorithms is primarily influenced by the sensor modality. Recent advancements in SLAM research have given rise to two primary system categories: LiDAR-based approaches and camera-based approaches. LiDAR-based methods are known for their high accuracy, primarily due to their capability to directly capture detailed and reliable 3D points. However, they face several challenges, including high cost and reduced reliability in environments such as highways, which are characterized by repetitive or low-texture geometric features. Additionally, they require extra processing to mitigate issues caused by light reflection and material transparency. Camera-based SLAM systems, on the other hand, are generally more robust in such scenarios and offer a more cost-effective alternative. Nevertheless, due to the absence of direct 3D measurements, camera-based SLAM systems generally underperform compared to LiDAR-based approaches. In this work, we demonstrate that the proposed camera-based SLAM system with integrated depth estimation outperforms existing methods on a real-world outdoor dataset.

2. Related work

Monodepth2 [8] serves as a commonly used baseline in recent self-supervised depth estimation research. Godard et al. introduced a method to ensure that only pixels satisfying the photometric consistency assumption in reprojection-based training are used for supervision. To address occlusion issues, they introduced a loss function that warps multiple source images into a single target image and computes the reprojection error, selecting only the minimum error across the sources. In addition, they proposed an automatic masking strategy to exclude pixels that violate the parallax assumption, such as those belonging to static background regions. In some cases, this strategy may also filter out moving objects whose motion is consistent with the camera, as these can otherwise degrade the quality of the supervision signal. These contributions enhanced the reliability of the underlying assumptions in the training pipeline and led to significant performance improvements. In contrast to previous methods that primarily suppress unreliable training signals, we introduce a knowledge distillation framework designed to provide more direct and semantically enriched supervision for self-supervised depth estimation.

Knowledge distillation refers to the transfer of learned representations from a teacher model to a student model. It has been extensively studied in deep learning across various tasks including depth estimation. In the field of depth estimation, knowledge distillation has been applied both to reduce model complexity and to transfer informative geometric cues that contribute to improved depth prediction performance. For example, Wang et al. [26] focused on enabling real-time depth estimation on edge devices by reducing model complexity. Rather than relying on the final output of the teacher, they proposed a framework in which the intermediate features extracted from the decoders of both teacher and student models are aligned using a pairwise loss. Pilzer et al. [16] proposed a training framework that enforces cycle consistency over repeated reprojections of a single image, allowing the student network to learn more stable depth representations in a self-supervised setting. In contrast, Poggi et al. [17] proposed a self-distillation framework where the student shares the same structure as the teacher, and can even outperform it. Their method incorporates pixel-level uncertainty into the loss function via a negative log-likelihood formulation, allowing the student model to account for uncertainty during training. Song et al. [23] designed a modified model architecture and loss function tailored for effective knowledge distillation of foundation depth model. Their method demonstrated state-of-the-art performance on the KITTI online benchmark, providing empirical evidence of the strong generalization capability of foundation models when used as sources of transferable knowledge. Foundation models often require networks with a large number of parameters

to fully leverage their representational capacity. However, practical applications such as SLAM benefit from models with significantly fewer parameters due to computational and resource constraints. To bridge this gap, we propose a knowledge distillation framework in which lightweight student model is guided by both segmentation and depth foundation models through geometry-aware supervision.

Foundation models refer to large-scale models pretrained using extensive computational resources and massive datasets. These models were initially developed for natural language processing tasks and have subsequently exhibited substantial impact in computer vision. Oquab et al. introduced DINOv2 [15], a robust vision transformer capable of extracting generalizable visual features from unseen images, enabling its use across various downstream tasks such as classification, segmentation, and depth estimation. Liu et al. proposed GroundingDINO [12], a vision-language model designed for object detection tasks involving unseen classes, which leverages diverse forms of text prompts to achieve strong performance. Kirillov et al. introduced the Segment Anything Model (SAM) [11], a prompt-driven segmentation framework capable of processing various input types—such as points, bounding boxes, masks, and text—alongside image data. In this work, we utilize GroundingDINO and SAM to explicitly distinguish sky regions, which are unsuitable for quantitative evaluation and potentially detrimental to depth model training. Ranftl et al. proposed the Dense Prediction Transformer (DPT) [19], which is trained on a large meta-dataset constructed by aggregating multiple existing depth datasets. Yang et al. introduced Depth Anything v2 [28], a model composed of a DINOv2-based encoder and a DPT-based decoder, trained using both labeled synthetic images and unlabeled real-world images to enhance generalization. The foundation depth model can reliably estimate relative depth even for unseen images; however, fine-tuning is required to achieve accurate absolute depth estimation. To address this limitation, we do not directly incorporate the absolute outputs of the foundation model into the knowledge distillation process. Instead, surface normal map is derived from relative depth prediction and incorporated as auxiliary geometric supervision in the training framework.

SLAM has increasingly integrated Inertial Measurement Unit (IMU) sensors, which provide measurements of linear acceleration and angular velocity. While IMUs provide valuable measurements, they are prone to cumulative drift due to inherent sensor noise characteristics. Therefore, integrating complementary sensor modalities is often necessary to improve the robustness of state estimation. Representative examples include LiDAR–Inertial Odometry (LIO), which integrates LiDAR and IMU data, and Visual–Inertial Odometry (VIO), which combines camera and IMU measurements. Bai et al. proposed a Faster-LIO [1], introducing efficient data structures for handling point cloud representations. Their method offers computational efficiency while maintaining reliable and accurate state estimation. In this study, it is utilized to generate ground-truth odometry for our custom dataset. This enables quantitative evaluation of odometry accuracy in VIO. Qin et al. introduced VINS-Mono [18], a widely adopted baseline VIO framework that fuses monocular images with IMU data. Building upon the original VINS-Mono framework, Shan et al. introduced VINS-RGBD [20], which incorporates depth data from RGB-D camera. Leveraging prior frameworks, we propose an RGB-based VIO pipeline guided by foundation models, designed to enhance the robustness of vision-based autonomous navigation systems.

3. Method

A detailed explanation of each component in the proposed VIO pipeline is presented in this section, as shown in Figure 2. Section 3.1 and 3.2 describe the image reprojection-based supervision strategy and the self-distillation scheme using a transformer-based foundation depth model. Section 3.3 describes the process of generating sky segmentation masks from foundation models for use in geometry-aware distillation. Section 3.4 presents a geometry-aware distillation strategy that transfers boundary and geometric cues from a pretrained foundation model to enhance student prediction accuracy. Finally, Section 3.5 presents the integration of the student depth model into a VIO pipeline.

3.1. Image reprojection based training

Our proposed depth estimation training pipeline incorporates an image reprojection-based self-supervised learning strategy, following prior successful work in the field [8]. Given a target image I_t, a corresponding depth map D_t^s is predicted by the student depth model. The relative camera transformation from the target view I_t to a reference view $I_{t'}$ is denoted as $T_{t\to t'}$ to facilitate reprojection. In stereo settings, this transformation is derived from known extrinsic calibration parameters. In monocular settings, it is estimated by a ResNet [9]-based pose network that takes the image pair $(I_t, I_{t'})$ as input. Given the camera intrinsic matrix K, the reprojected view $I_{t'\to t}$ is synthesized from the reference image $I_{t'}$.

$$I_{t'\to t} = I_{t'}\langle \text{proj}(D_t^s, T_{t\to t'}, K)\rangle. \tag{1}$$

Here, $\text{proj}(\cdot)$ denotes the projection operation into the target frame, and $\langle\cdot\rangle$ indicates bilinear sampling. A photometric loss L_p is employed, combining the Structural Similarity Index Measure (SSIM) [27] and the L1-norm with a weighting factor α:

$$L_p = \min_{t'} pe(I_t, I_{t'\to t}), \tag{2}$$

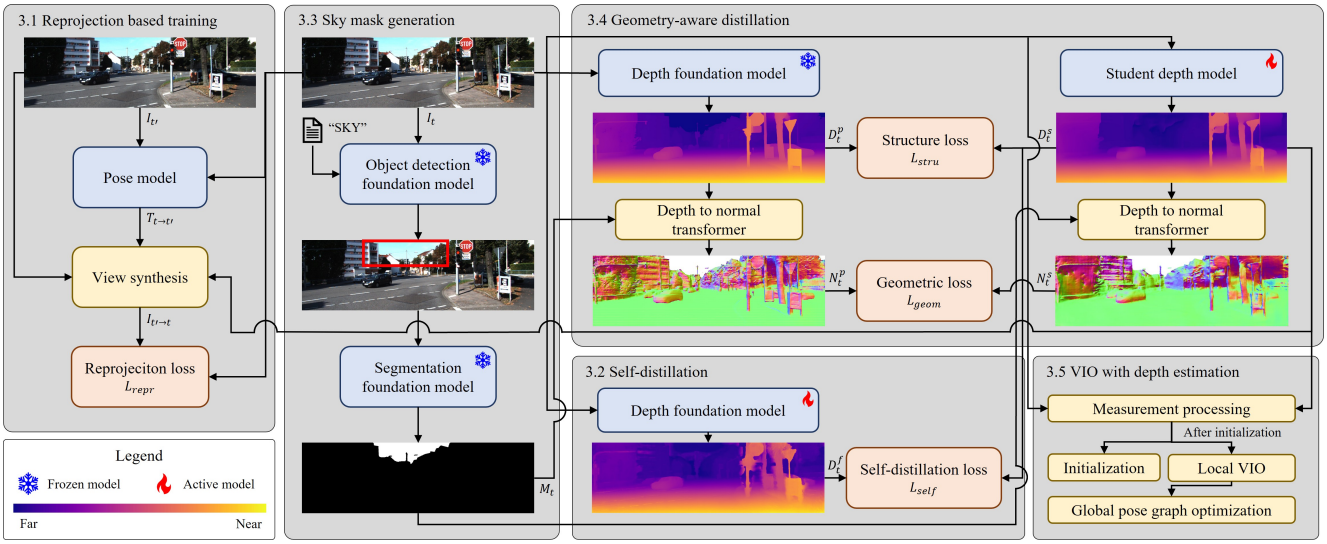


Figure 2. Overview of the proposed visual–inertial odometry pipeline. Yellow boxes represent non-learnable geometric modules, blue boxes indicate deep neural networks, and red boxes denote loss functions used for training the student network.

$$pe(I_t, I_{t'\to t}) = \frac{\alpha}{2}(1 - \text{SSIM}(I_t, I_{t'\to t})) + (1-\alpha)\|I_t - I_{t'\to t}\|_1. \quad (3)$$

To account for occlusions, the minimum photometric error is computed across multiple reference frames, ensuring supervision is derived from the most photometrically consistent view. Since this approach assumes a static scene with a moving camera, a binary mask μ is introduced to restrict the loss computation to valid regions. It is computed as follows:

$$\mu = \left[\min_{t'} pe(I_t, I_{t'\to t}) < \min_{t'} pe(I_t, I_{t'})\right], \quad (4)$$

where $[\cdot]$ denotes the Iverson bracket. And we adopted an auxiliary loss term known as the edge-aware smoothness loss L_s, which encourages depth smoothness in textureless regions while preserving object boundaries:

$$L_s = |\partial_x D_t^s| e^{-|\partial_x I_t|} + |\partial_y D_t^s| e^{-|\partial_y I_t|}. \quad (5)$$

Here, ∂_x and ∂_y denote the partial derivatives with respect to the x-axis and y-axis, respectively. The final reprojection loss is defined as a weighted combination of the masked photometric loss and the smoothness loss:

$$L_{\text{repr}} = \mu L_p + \beta L_s. \quad (6)$$

In all experiments, the hyperparameter β is empirically set to 0.001. The reprojection-based approach is supported by established theoretical principles and demonstrates reliable performance in coarse depth estimation. However, it tends to struggle with fine-grained details such as object boundaries and transparent surfaces. To address these limitations, two knowledge distillation techniques are additionally employed, as detailed in the following sections.

3.2. Self-distillation of dense prediction transformer

Recent approaches [17, 23] have explored self-distillation techniques in which a model fine-tuned on the target dataset is used to generate pseudo ground truth depth map D_t^f. The pseudo label is then used as supervision signal to train the student model by minimizing a depth loss function. Although the generated pseudo depth may be imperfect, the availability of dense supervision provides explicit guidance during training, often resulting in improved performance. Prior studies have shown that, even when the teacher and student models share identical architectures, the student model can achieve superior performance through self-distillation. In contrast, our work aims to bridge the performance gap between a high-capacity teacher model and a lightweight student model with significantly fewer parameters. The depth models adopt the architecture of a depth foundation model based on DPT [19], utilizing a ViT-based encoder [4] to extract rich visual representations. The decoder consists of a reassemble block and a fusion block, followed by a prediction head that reconstructs the depth map in the image space. To better leverage dense depth cues, the student model is augmented with parameter-shared multi-prediction heads. Deep supervision is applied to all intermediate predictions to encourage the extraction of semantically meaningful features. For the self-distillation loss, we adopt the scale-invariant error, a widely used metric in supervised depth estimation [5], which compares the predicted depth D_t^s with the pseudo ground truth D_t^f in logarithmic space:

$$L_{\text{self}} = \frac{1}{n}\sum_{i=1}^{n} {\epsilon_i}^2 - \frac{\lambda}{n^2}\left(\sum_{i=1}^{n}\epsilon_i\right)^2, \quad (7)$$

$$\epsilon_i = \log d_i^s - \log d_i^f. \tag{8}$$

Here, n denotes the number of valid pixels, while d_i^s and d_i^f represent the predicted and pseudo depth values at pixel i.

3.3. Sky mask generation via foundation model

In depth estimation, the sky region is inherently unsuitable for accurate prediction due to its near-infinite depth, which fundamentally differs from other regions that exhibit observable geometric structure. Recent approaches, such as Depth Anything v2 [28], address this issue by training models to predict values that are inversely proportional to depth, thereby encouraging near-zero outputs for sky pixels. However, when applying similar strategies to relatively limited real-world custom dataset, we observed a substantial degradation in both quantitative accuracy and generalization performance. Therefore, a sky segmentation mask M_t is generated for input image I_t to explicitly exclude these regions from the distillation process. The mask M_t is generated in two stages, beginning with the input of the RGB image I_t and the textual prompt "sky" into GroundingDINO [12], a state-of-the-art open-vocabulary object detection model recognized for its strong performance in zero-shot scenarios. This model produces bounding boxes that roughly localize sky regions. These bounding boxes then serve as strong prompts for SAM [11], which takes the RGB image I_t and outputs high-quality segmentation mask corresponding to the detected sky area.

3.4. Geometry-aware knowledge distillation

Depth foundation models pre-trained on large-scale datasets are capable of producing sharp and geometrically consistent relative depth maps D_t^p, even for previously unseen images. Leveraging this capability, we propose two geometry-aware distillation losses designed to transfer the boundary-aware and surface-consistent knowledge from the pretrained model to the student model. We introduce a structure consistency loss L_{stru}, which promotes edge preservation by evaluating local structural patterns instead of penalizing absolute depth differences. Inspired by the structural similarity term in the SSIM metric, the loss measures local geometric coherence between the foundation model prediction D_t^p and the student output D_t^s.

$$L_{\text{stru}} = \frac{1}{n}\sum_{i=1}^{n}(1 - \frac{\sigma_i^{ps} + k}{\sigma_i^p \sigma_i^s + k}), \tag{9}$$

where σ_i^{ps} denotes the local covariance between D_t^p and D_t^s at pixel i, while σ_i^p and σ_i^s represent the local standard deviations of D_t^p and D_t^s, respectively. These statistics are computed over a 3×3 window centered at each pixel. The constant k, which is empirically set to 30 in accordance with the SSIM formulation, stabilizes the computation by preventing division by near-zero values.

To enhance the model's understanding of object shapes through geometric context, we incorporate surface normal supervision, which provides richer structural information than depth values alone. To generate surface normal supervision, we employ Depth-to-Normal Transformer (D2NT) [6], a recently proposed method that achieves both high accuracy and computational efficiency. As normal estimation directly impacts the training speed of our pipeline, fast and reliable normal computation is a critical requirement. Both D_t^p and D_t^s, along with the corresponding sky mask M_t, are passed through D2NT to produce surface normal maps N_t^p and N_t^s, respectively. We then define a geometric consistency loss L_{geom} based on cosine similarity between the predicted and reference normal vectors:

$$L_{\text{geom}} = \frac{1}{m}\sum_{i=1}^{m}(1 - n_i^p \cdot n_i^s), \tag{10}$$

where m denotes the number of pixels outside the masked region defined by M_t, and n_i^p and n_i^s represent the normal vectors at pixel i in N_t^p and N_t^s, respectively.

By incorporating both boundary-sensitive and surface-aware supervision, the proposed method facilitates enhanced structural understanding within the student network, thereby promoting more stable and accurate depth estimation. Consequently, the total loss used for training the student depth model is formulated as a weighted combination of the following components:

$$L_{\text{total}} = L_{\text{repr}} + w_1 L_{\text{self}} + w_2 L_{\text{stru}} + w_3 L_{\text{geom}} \tag{11}$$

In all experiments, we set the loss weights to $w_1 = 0.1$, $w_2 = 1$, and $w_3 = 0.1$ based on empirical tuning.

3.5. VIO with depth estimation

We construct a VIO pipeline based on VINS-RGBD [20], integrating RGB images, the predicted depth maps from the student model, and inertial measurements from an IMU sensor. Given the distinct characteristics of visual and inertial modalities, we apply modality-specific preprocessing steps. The IMU operates at a significantly higher sampling rate than the camera and is subject to considerable sensor noise; thus, we employ pre-integration techniques to fuse the high-rate inertial data effectively. In the visual processing pipeline, feature points are identified in each RGB frame using the Shi–Tomasi corner detection algorithm [22], and their inter-frame correspondences are established via the Kanade–Lucas–Tomasi (KLT) sparse optical flow method [13]. In contrast to conventional RGB-based VIO systems that estimate depth through the perspective-n-point (PnP) algorithm, the proposed method directly incorporates depth values aligned with tracked features, as

obtained from the predicted depth map D_t^s. During system initialization, visual–inertial initialization [18] is conducted to jointly estimate the metric scale, gravity direction, and initial pose. Upon successful completion of the initialization phase, subsequent preprocessing outputs are propagated into a sliding-window-based local VIO optimization framework. When a previously visited location is recognized using a Bag-of-Words-based visual retrieval method, the accumulated drift in relative pose estimates is corrected through pose graph optimization. Our experiments demonstrate that the proposed method enhances the robustness of localization in challenging outdoor environments.

4. Experiments

4.1. Experimental setup

All experiments were conducted on a workstation equipped with an AMD EPYC 7313P 16-core processor and two NVIDIA RTX 4090 GPUs. Pretrained weights from the foundation model trained on the meta-dataset [19] were used to initialize both the teacher and student depth networks. The models were optimized using the Adam optimizer [10] with a weight decay of 10^{-2}, where the initial learning rate was set to 10^{-4} for the baseline [8] and 10^{-5} for the foundation-based models. To mitigate overfitting, standard data augmentation techniques were employed. These included horizontal flipping and color jittering, with brightness, contrast, and saturation factors randomly sampled from the range 0.8 to 1.2, and hue from -0.1 to 0.1. Each augmentation was applied with a probability of 50%. Although all sequences were captured at a frame rate of 10 Hz, different temporal triplet configurations were adopted to account for variations in motion dynamics: the triplet [-1, 0, 1] was used for the KITTI dataset, while [-3, 0, 3] was applied to the custom dataset.

Depth estimation was evaluated using three accuracy metrics (δ_1, δ_2, δ_3) and six error metrics (RMSE, RMSEi, AbsRel, SqRel, log10, SIlog). Accuracy metrics δ_j denote the percentage of pixels satisfying $\max(\hat{d}/d, d/\hat{d}) < 1.25^j$, where $\hat{d}$ and d denote the predicted and ground-truth depth values, respectively. Detailed definitions and formulations of the error metrics are available in previous work [5]. In the tables presenting quantitative results for depth estimation, error metrics are indicated with a red background, while accuracy metrics are marked in blue. For odometry evaluation, ground-truth trajectories were generated using an existing LiDAR-based SLAM algorithm [1]. As evaluation metrics, we used the relative pose error (RPE) in both translation and rotation, and the RMSE of the absolute trajectory error (ATE), as defined in the RGB-D SLAM benchmark [24]. Bold and underlined values indicate the best and second-best performance, respectively, for each task.

4.2. KITTI dataset

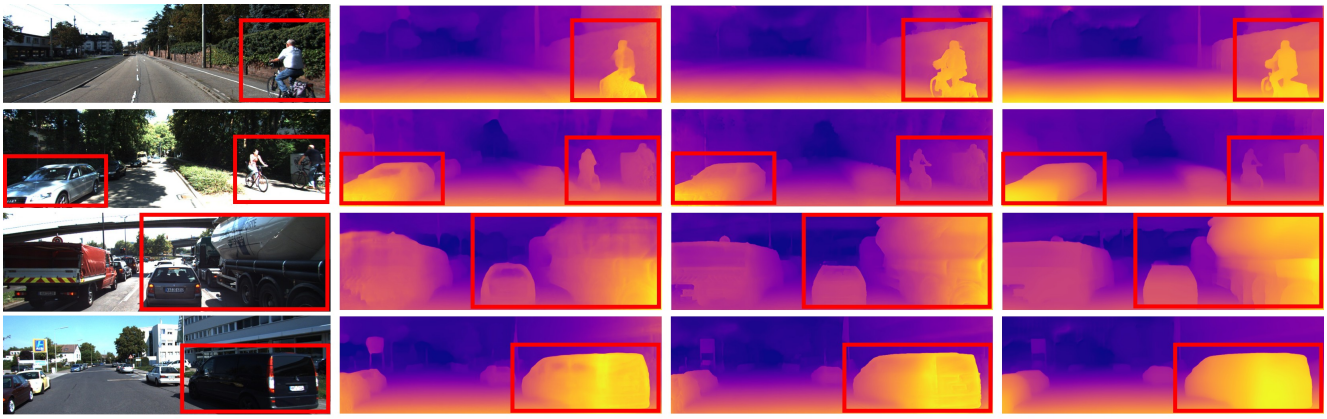
To empirically validate the effectiveness of the proposed depth estimation framework, we conduct experiments on the KITTI public dataset [7]. The KITTI dataset was constructed to facilitate a broad spectrum of computer vision tasks, including depth estimation, stereo matching, optical flow, and 3D object detection and tracking. It was collected using a vehicle equipped with stereo cameras, a 3D LiDAR scanner, and GPS/IMU sensors, driving through real-world urban, rural, and highway environments in Karlsruhe, Germany. We employ the Eigen split [5], a standardized data partitioning widely adopted for evaluating depth estimation methods. This split comprises 39,810 monocular triplets for training, 4,424 for validation, and 697 for evaluation.

Quantitative and qualitative evaluations on the KITTI dataset are presented in Table 1 and Figure 3, respectively. We compare the proposed method against Monodepth2 [8] and a finetuning strategy that uses pretrained parameters from Depth Anything V2 [28] under a reprojection-based training pipeline. All methods were evaluated using monocular inputs at a fixed resolution of 1024×320 pixels, and median scaling was applied following standard practice. The number of parameters in Depth Anything V2 varies depending on the Vision Transformer [4] backbone: ViT-Large (vitl) yields 335.3M parameters, while ViT-Small (vits) results in 24.7M. While the vitl-based model achieves strong performance due to its high representational capacity, its substantially higher computational cost may hinder deployment in resource-constrained settings. In contrast, Monodepth2, with only 14.8M parameters, is highly efficient in terms of computational resources but requires improvement in estimation accuracy for real-world deployment. The vits-based Depth Anything V2 model achieves better performance than Monodepth2 when fine-tuned, but still falls short of the vitl model overall. Although our proposed method shares the same backbone architecture as the vits-based model, it integrates additional loss functions for training, which lead to the best performance on four error metrics and two accuracy metrics. A slight decrease was observed in AbsRel and δ_1, indicating a minor reduction in absolute distance estimation accuracy. However, improvements in SIlog and δ_3 suggest that the model has better learned the relative geometric structure within scenes. Qualitative results in Figure 3 further support these findings. In columns 1 and 2, the proposed method better captures fine structures such as pedestrians and bicycles. Furthermore, as shown in columns 3 and 4, the method that relies solely on reprojection loss in the third row exhibits texture-induced artifacts in the predicted depth, whereas the proposed method produces smoother and more geometrically consistent depth maps.

Table 2 presents an ablation analysis that investigates the individual contributions of each loss component in addition

Table 1. Quantitative evaluation of depth estimation performance on the Eigen validation split of the KITTI dataset. All methods use monocular input with a resolution of 1024×320. DAv2 denotes Depth Anything V2 [28].

Method	Params	SIlog	AbsRel	SqRel	RMSE	RMSEi	log10	δ_1	δ_2	δ_3
Monodepth2 [8]	14.8 M	18.293	0.109	0.832	4.648	0.186	0.048	0.888	0.963	0.982
DAv2(vits) [28]	24.7 M	16.994	**0.099**	0.740	4.314	0.173	**0.043**	**0.908**	**0.969**	0.985
Ours	24.7 M	**16.856**	0.101	**0.687**	**4.280**	**0.172**	0.044	0.899	**0.969**	**0.986**
DAv2(vitl) [28]	335.3 M	16.406	0.090	0.639	4.040	0.166	0.040	0.924	0.971	0.985



RGB image | Monodepth2 [8] | DAv2 [29] | Ours

Figure 3. Qualitative comparisons of depth estimation results on the Eigen validation split of the KITTI dataset.

Table 2. Ablation study of loss function configurations on the Eigen validation split of the KITTI dataset. All methods use stereo input with a resolution of 640×192.

Loss function	SIlog	AbsRel	RMSE	δ_1
L_{repr}	19.688	**0.118**	5.230	0.853
L_{repr}, L_{self}	18.926	0.120	5.055	0.852
L_{repr}, L_{stru}	18.872	0.121	5.067	0.854
L_{repr}, L_{geom}	18.703	0.115	5.024	**0.861**
L_{total}	**18.521**	0.119	**5.017**	**0.861**

to the baseline reprojection loss L_{repr}, which is consistently employed across all model variants. To ensure a consistent and equitable evaluation, all experiments were carried out using stereo image inputs with a fixed resolution of 640×192. When incorporating L_{self} or L_{stru} during training, we observed moderate performance gains in SIlog, RMSE, and δ_1, indicating improvements in both relative depth accuracy and structural consistency. The adoption of L_{geom}, which supervises surface normals, led to performance improvements across four metrics, including AbsRel. The full model trained with all loss terms L_{total} achieved superior performance on most metrics, with only a marginal decline in AbsRel, indicating improved overall depth quality.

4.3. Campus driving dataset

We further evaluate the performance of the proposed algorithm on depth and odometry estimation using a custom-collected real-world dataset, and compare it against existing methods. As illustrated in Figure 4, the dataset was acquired using a compact electric vehicle equipped with an RGB stereo camera and a 3D LiDAR sensor. The vehicle was driven at speeds ranging from 6 to 12 km/h within a university campus. Although the campus is an outdoor environment, it features rich textures and geometric structures in both 3D point clouds and camera imagery, making it suitable for generating reliable ground truth using LiDAR-based SLAM and for validating the application of vision-based SLAM systems. The RGB stereo and LiDAR data were recorded at 10 Hz, while the IMU embedded in the LiDAR system was recorded at 100 Hz. The dataset consists of 12 driving sequences, each lasting between 100 and 400 seconds. For both tasks, the dataset is partitioned into 8 sequences for training, 1 for validation, and 3 for testing, corresponding to 15,025 stereo pairs for training, 1,205 for validation, and 6,416 for testing. Reliable ground truth was established through offline extrinsic calibration [25] using optimization-based methods between the camera–IMU and camera–LiDAR sensor pairs.

Table 3 presents depth estimation results on the custom campus driving dataset, which reveal patterns consistent with those observed on the public dataset. Foundation model-based approaches significantly outperformed Monodepth2 across all evaluation metrics, demonstrating the effectiveness of pretraining on large-scale data. Compared to the baseline finetuning strategy using only reprojection loss, the proposed method achieved further improvements,

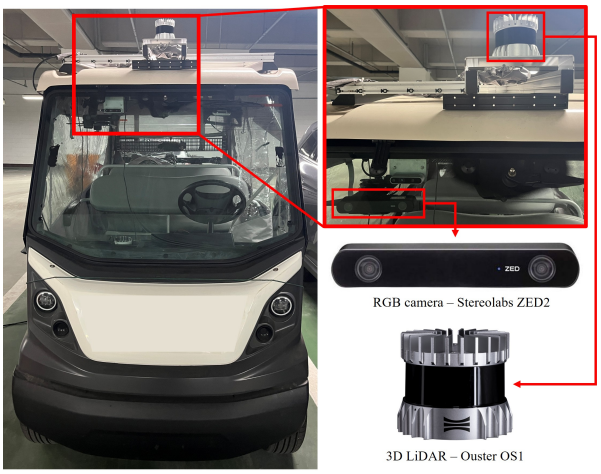


Figure 4. Sensor setup for the custom dataset collection.

Table 3. Quantitative evaluation of depth estimation performance on the custom dataset. DAv2 denotes Depth Anything V2 [28].

Method	SIlog	AbsRel	RMSE	δ_1
Monodepth2 [8]	19.448	0.121	3.106	0.886
DAv2(vits) [28]	17.883	**0.096**	2.997	**0.918**
Ours	**17.631**	0.097	**2.857**	0.903

particularly in SIlog and RMSE, indicating enhanced relative and overall depth accuracy. As illustrated in Figure 5, the qualitative comparisons highlight distinct improvements in visual prediction quality. Models trained solely with reprojection loss tend to overfit to image textures, resulting in depth artifacts around regions such as tree foliage and ground shadows. In contrast, the proposed method effectively suppresses such artifacts while providing sharper delineation of structural elements like building pillars and bollard edges, along with globally smoother and more coherent depth predictions.

As an RGB-based method, the proposed VIO system leverages depth predictions from the student network and is evaluated against both RGB-only [18] and RGB-D [20] VIO baselines, as shown in Table 4. In terms of the RMSE of ATE, which serves as a principal metric for evaluating odometry accuracy, the proposed method consistently outperformed the RGB-only baseline across all test cases, achieving significantly lower error values. In case 1, the proposed method demonstrated slightly superior performance even compared to the RGB-D baseline. Although the translation error of the RPE varied across individual cases, the proposed method achieved a higher average performance. In terms of the rotation error of the RPE, the RGB-based method exhibited slightly better performance than the RGB-D based approach. However, as all VIO methods demonstrated consistently low rotation errors, the differences were not statistically significant.

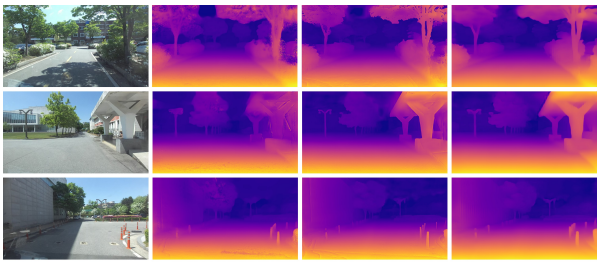


Figure 5. Qualitative comparisons of depth estimation results on the custom dataset. DAv2 denotes Depth Anything V2 [28].

Table 4. Quantitative evaluation of odometry estimation performance on the campus driving dataset.

Scenario	Case 1	Case 2	Case 3	Average
Distance [m]	318.12	394.75	502.55	
Method	RMSE of ATE [m]			
VINS-Mono [18]	4.9564	8.2617	7.8689	7.0290
Ours	**4.1027**	**5.2186**	**7.0312**	**5.4508**
VINS-RGBD [20]	4.3050	4.7788	6.1443	5.0760
Method	Translation error of RPE [m]			
VINS-Mono [18]	**0.3712**	0.6332	**0.3648**	0.4564
Ours	0.3993	**0.4627**	0.3921	**0.4180**
VINS-RGBD [20]	0.3957	0.4092	0.3399	0.3816
Method	Rotation error of RPE [deg]			
VINS-Mono [18]	**0.3116**	**0.2635**	**0.3291**	**0.3014**
Ours	0.3319	0.3666	0.3733	0.3573
VINS-RGBD [20]	0.2975	0.2959	0.3440	0.3125

5. Conclusion

In this study, we fine-tune a pretrained foundation model through a self-supervised learning framework and integrate it into a VIO system, resulting in performance that surpasses existing RGB-based VIO. Our self-supervised training strategy effectively distills knowledge from the pretrained foundation models into a lightweight student depth network, enabling it to inherit the structural understanding learned from large-scale data. In the depth estimation, the proposed method demonstrates notable improvements in qualitative performance. However, metrics related to absolute depth accuracy exhibit slight degradation. Nevertheless, the proposed network exhibits robust performance in capturing object boundaries and recognizing transparent surfaces, achieving results comparable to those of the pretrained foundation model. These capabilities cannot be reliably obtained using reprojection-based supervision alone, highlighting the necessity of our knowledge distillation approach. Furthermore, as transparent objects pose inherent challenges not only in depth prediction but also in reliable ground-truth acquisition, future research should investigate both improved estimation techniques and more appropriate evaluation strategies for such regions.

References

[1] Chunge Bai, Tao Xiao, Yajie Chen, Haoqian Wang, Fang Zhang, and Xiang Gao. Faster-lio: Lightweight tightly coupled lidar-inertial odometry using parallel sparse incremental voxels. *IEEE Robotics and Automation Letters*, 7(2):4861–4868, 2022. 3, 6

[2] Yingxiu Chang, Yongqiang Cheng, Umar Manzoor, and John Murray. A review of uav autonomous navigation in gps-denied environments. *Robotics and Autonomous Systems*, 170:104533, 2023. 1

[3] Jiyu Cheng, Hu Cheng, Max Q-H Meng, and Hong Zhang. Autonomous navigation by mobile robots in human environments: A survey. In *2018 IEEE international conference on robotics and biomimetics (ROBIO)*, pages 1981–1986. IEEE, 2018. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 6

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 1, 4, 6

[6] Yi Feng, Bohuan Xue, Ming Liu, Qijun Chen, and Rui Fan. D2nt: A high-performing depth-to-normal translator. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 12360–12366. IEEE, 2023. 5

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6

[8] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 1, 2, 3, 6, 7, 8

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1, 3, 5

[12] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 1, 3, 5

[13] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 5

[14] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021. 1

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[16] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019. 2

[17] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3227–3237, 2020. 2, 4

[18] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE transactions on robotics*, 34(4):1004–1020, 2018. 3, 6, 8

[19] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3, 4, 6

[20] Zeyong Shan, Ruijian Li, and Sören Schwertfeger. Rgbd-inertial trajectory estimation and mapping for ground robots. *Sensors*, 19(10):2251, 2019. 3, 5, 8

[21] Shuwei Shao, Zhongcai Pei, Weihai Chen, Baochang Zhang, Xingming Wu, Dianmin Sun, and David Doermann. Self-supervised learning for monocular depth estimation on minimally invasive surgery scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7159–7165. IEEE, 2021. 1

[22] Jianbo Shi et al. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pages 593–600. IEEE, 1994. 5

[23] Jimin Song and Sang Jun Lee. Knowledge distillation of multi-scale dense prediction transformer for self-supervised depth estimation. *Scientific Reports*, 13(1):18939, 2023. 2, 4

[24] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 573–580. IEEE, 2012. 6

[25] Darren Tsai, Stewart Worrall, Mao Shan, Anton Lohr, and Eduardo Nebot. Optimising the selection of samples for robust lidar camera calibration. In *Proceedings of the IEEE international intelligent transportation systems conference (ITSC)*, pages 2631–2638. IEEE, 2021. 7

[26] Yiran Wang, Xingyi Li, Min Shi, Ke Xian, and Zhiguo Cao. Knowledge distillation for fast and accurate monocular depth estimation on mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2457–2465, 2021. 2

[27] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[28] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1, 3, 5, 6, 7, 8

[29] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 1