Incorporating Learned Depth Perception Into Monocular Visual Odometry to Improve Scale Recovery

Hamza Mailka, Mohamed Abouzahir, Mustapha Ramzi
High School of Technology of Salé Laboratory of Systems Analysis-Information Processing
and Industrial Management (LASTIMI),
Mohammed V University in Rabat, Morocco

***Abstract*—A growing interest in autonomous driving has led to a comprehensive study of visual odometry (VO). It has been well studied how VO can estimate the pose of moving objects by examining the images taken from onboard cameras. In the last decade, it has been proposed that deep learning under supervision can be employed to estimate depth maps and visual odometry (VO). In this paper, we propose a DPT (Dense Prediction Transformer)-based monocular visual odometry method for scale estimation. Scale-drift problems are common in traditional monocular systems and in recent deep learning studies. In order to recover the scale, it is imperative that depth estimation to be accurate. A framework for dense prediction challenges that bases its computation on vision transformers instead of convolutional networks is characterized as an accurate model that is utilized to estimate depth maps. Scale recovery and depth refinement are accomplished iteratively. This allows our approach to simultaneously increase the depth estimate while eradicating scale drift. The depth map estimated using the DPT model is accurate enough for the purpose of achieving the best efficiency possible on a VO benchmark, eliminating the scaling drift issue.**

***Keywords*—Visual odometry; scale recovery; depth estimation; DPT model**

I. Introduction

The greatest anticipated technological advancement in the near future is autonomous ground vehicles (AGV), which operate entirely automatically. A vehicle operating autonomously requires precise and reliable information regarding its position for a successful navigation [1], [2], [3], [4]. There are currently many popular methods of providing comparatively reliable positioning information, such as the Global Navigation Satellite System (GNSS), Visual Odometry, and the Inertial Navigation System (INS) [5], [6], [7], [8], [9], [10], [11].

A growing interest in autonomous vehicles has led to well-developed novel approaches based on VO. Different approaches have been well studied since they estimate the position and orientation of moving objects based on the analysis of image sequences [12], [13], [14], [15]. A precise VO system is one of the most crucial techniques in the area of mobile robots. [16], [17], [18], [19]. The way conventional monocular VO systems operate is by assuming that the scale is one or by using ground truth for an approximation of the scale. Due to significant drift, monocular VO systems cannot operate on image sequences without ground truth or estimate the pose with significant drift [20], [21], [22], [23]. Despite the fact that several traditional monocular VO systems have been developed, they have still performed poorly or are unable to work in some conditions, like monotonous scenes that lack visible texture information or large-scale camera movements. When it comes to learning-based VO systems, they are developed by training neural networks in supervised or unsupervised manners through end-to-end pose estimation [24], [25], [26], [27]. Moreover, the efficiency of networks completely determines how accurate pose estimation is. Even with numerous training datasets and a network structure optimized, it is unavoidable for a network to encounter issues such as insufficient accuracy when estimating rotational pose.

Over the years, a lot of work has gone into developing a reliable and precise VO system. In terms of traditional VO algorithms, two main types exist: feature-based [13], [28] and direct methods [29], [30]. Calibration of the camera, identifying and matching features, rejecting outliers (using RANSAC), estimating motion, and estimating scale are typical components of feature-based techniques (e.g., Bundle Adjustment). However, finding the right features to reconstruct certain motions is still difficult. The motion of the pixel is tracked, and pose predictions are obtained by minimizing photometric error, so it is highly sensitive to light variations. Additionally, the classic monocular VO's absolute scale estimation requires the use of additional data or knowledge (such as the camera's height). In monocular systems, obtaining scale information is complicated and typically depends on an earlier, predefined absolute reference. A reference scale can be provided by integrating with some other sensors, like an inertial measurement unit. Scale drift [31] is often addressed by local optimization techniques like bundle adjustment and loop closure detection. In addition, researchers employ the depth estimation [32] from images to approximate the scale and adjust the calculated translation in addition to other approaches, like a ground plane estimation using the camera height, which is assumed to remain stable during motion.

The vast amount of training data (ground truth) required by supervised deep learning methods is usually collected with RGB-D cameras indoors and 3D laser scanners. Nevertheless, since ground truth is required, the supervised technique has a number of drawbacks. At first, the sensors' own inaccuracy and noise may have an impact on the network. Second, these sensors cannot record high-resolution information as well as images since their measurements are often sparser. Lastly, those sensors may not be able to obtain ground truth in some

locations. Because of this, researchers have begun to pay greater attention to unsupervised approaches that only need training data. In this context, the goal of our contribution is to propose a reliable localization that just requires images from a monocular camera in order to obtain an estimation of motion. The suggested strategy deals with the problem of scale estimation by using a dense prediction transformer model to estimate the depth map of the environment. As shown in this work, according to KITTI odometry benchmarks, the system's scale estimation performs in a manner comparable to that of the state-of-the-art.

The remainder of this paper is organized as follows: Section II represents a brief summary of related work. The DPT model utilized in our paper is summarized in Section III. In Section IV, we provide details on our end-to-end approach. Section V shows experimental results on the KITTI. In Section VI, the paper's contributions are summarized, as well as it concludes with some recommendations for future work.

II. Related Work

Numerous studies have investigated how depth can be estimated from images employing stereo and monocular images, or multi-view images. By using conventional and traditional techniques in a single-view image, it is difficult to recognize the structure of the scene. Luckily, since the innovative research of [33], deep learning has progressed greatly in the computer vision field. Most CNN-based depth-map prediction approaches are supervised. To learn parameters, these techniques require more than one labeled dataset. We will look at how to solve the scale estimation problem in the following parts: In this section, we provide a brief summary of the most closely related work that is needed to assess the scene depth estimation and camera motion prediction.

A. Recovering Camera Poses and Depth using Conventional Techniques

Researchers in computer vision have long been interested in recovering depth-maps and camera poses. A learning depth approach based on 2D to 3D image conversion using examples is proposed by Konrad et al. [34]. They create an efficient and simplified version of the current 2D to 3D frame conversion algorithms. A feasible depth generation method from sequences of images was described in [35] employing auxiliary data from non-parametric depth sampling. The performance of this method was superior to all current standard depth methods. Camera posture research another important topic of study in the discipline of computer vision, has a great success using conventional methods. The most well-known conventional approach that is used to estimate camera pose using images is called ORB-SLAM [36]. It uses the feature matching approach for mapping and localization combined with a single monocular image. The process of this method has four stages: loop closure, tracking, mapping, and relocalization. But each stage needs to be carefully planned. Gao et al. [37] expanded this knowledge to build reconstructions of 3D objects from 2D images based on motion techniques. A method for predicting 3D structures and camera projections was put out in [38]. The strategy is in the area of estimating 3D symmetric objects from 2D symmetric perspectives and forms using numerous intra-class objects as an input model. It was suggested by Ma et al. [39] that in remote sensing imagery, rigid and non-rigid structures can be matched using a locally linear transformation model. To determine scene depth, all of the approaches mentioned above either rebuild 3D geometry or establish pixel-by-pixel correspondences between input views. Nevertheless, the input data for these methods is multi-view images.

B. A Supervised Learning Approaches using Monocular Images

Since we can't obtain the structural properties from a single view image, determining a depth-map using a monocular camera is a difficult issue. Depth estimation has recently been viewed by some academics as a supervised learning approach. A network with two factors was proposed by Eigen et al. [40], the first of which assesses the scene's overall structure and the second of which refines it using local information. As one of the few papers using deep CNN to estimate scene depth using monocular images. Three separate computer vision issues were handled simultaneously via a framework that Eigen et al. [41] created (prediction of surface normals, depth estimates, and semantic identification) based on prior research. A completely convolutional architecture was suggested by Laina et al. [42] to describe the uncertain mapping between depth maps and monocular pictures. Li et al. [43] introduced a multi-streamed CNN architecture for depth estimation that is quick to train. Yan et al. [44] used a reference as the surface normal to aid in the monocular depth estimation problem. Up to this point, some studies on monocular depth prediction have combined CNNs and Random Forests. In certain studies on monocular depth prediction, CNNs and random forests were merged. Regression based on deep CNN characteristics was used by Li et al. [45] to overcome this issue, together with conditional random fields for post-processing refinement. Using only one image as a source of depth prediction, Roy et al. [46] introduced a deep regression forest approach that blends CNNs with random forests. As a result of depth data being continuous, Liu's formulation of depth prediction as the "random field learning with continuous conditions" problem [47] was developed. Although the aforementioned techniques have shown precise monocular-depth prediction, the ground truth is used as a basis for training, which limits the model's capacity for generalization.

C. An Unsupervised Learning Approaches using Monocular Images

Several unsupervised learning techniques that address the monocular depth estimation issue have recently been introduced to get over the ground-truth issue. Garg et al. [48] built a CNN to approximate the complex non-linear transformation that turns stereo images into depth maps using input camera motions. The proposed method of [49], [50] constructed a model based on Garg's work to incorporate a spatial smoothness loss into the unsupervised optical flow total loss function. Their efforts and outcomes are comparable. When training, Godard et al. [51] approached the difficulty of predicting depth as an issue with image reconstruction using epipolar geometry constraints. To determine the relationship between the rectified stereo images, a loss function is developed. In a semi-supervised manner, Kuznietsov et al. [52] employed

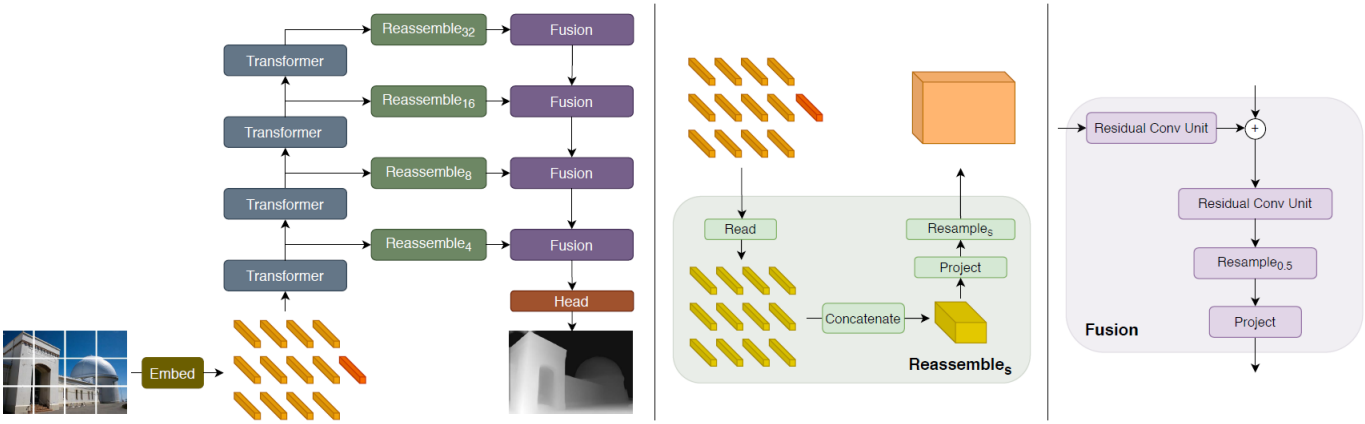

Fig. 1. The architecture of dense prediction transformer model that we used in our approach to estimate disparity maps from monocular cameras. This enables the recovery of precise metric estimates [61].

projected inverse and sparse ground truth depth data. Their models depend on a 3D laser sensor and camera that are precisely calibrated externally. A learning approach called CCRFN (Convolutional Conditional Random Field Network) was proposed by Yan Hua and Hu Tian [53] For estimating depth and identifying features by using the learning approach. It offers two benefits: first, it doesn't require manually created features, and second, it uses the relationship between individual features to estimate depth. Based on the research of Zhou[54], Yin et al. [55] introduced the GeoNet architecture for unsupervised learning, which jointly predicts optical flow, monocular depth, and dynamic object detection. Luo et al. [56] from SenseTime Research suggested stereo matching as the first sub-issue to address after the monocular depth estimation method. The above unsupervised learning approaches were trained on mono-image sequences or using stereo images with precise calibration. Temporal information cannot be fully utilized by stereo pictures. Due to depth ambiguity, which can occur in monocular images, different depths may correlate to objects that appear to be similar in the image. Although these unsupervised models succeeded in their objective of estimating scene depth without the need for ground truth, they have received little attention for their joint use of stereo and monocular sequences for depth prediction.

III. DPT Model

Due to the quick advancement of computer technology and digital imaging sensors. The camera sensor is progressively becoming more advantageous, and as a result, navigation using visual assistance and its related combined system have emerged as a significant component of the integrated navigation system. Since Transformer was so successful in natural language processing (NLP) [57], the computer vision community has given it a lot of attention. It has recently demonstrated exceptional performance on a variety of computer vision tasks, including semantic segmentation, object identification, imaging classification, and depth estimation. The standard architecture for dense prediction is fully-convolutional networks [58], [59]. Although many variations of this fundamental pattern have been presented throughout time, all current architectures use convolution and subsampling as their core components to learn multi-scale models that can make use of a sufficiently wide context. When trained on enormous datasets and deployed as high-capacity architectures, transformer models have proven particularly effective. Attention processes have been adapted to image analysis in a number of publications. In particular, it has recently been shown that a direct application of effective token-based transformer designs in NLP may produce competitive performance on image categorization [60]. This work's most important finding was that, similar to transformer models in NLP, visual transformers require a substantial quantity of training data to reach their full potential.

In contrast to the state-of-the-art CNN-based method, Ranftl et al. [61] reported improved relative performance using the dense prediction transformer (DPT) model for monocular depth estimation. This is why we decided to estimate the depth map using the DPT network since precise depth estimation improves scale estimation [20]. A ViT serves as the basis of the DPT model. The frame is divided into regions, which are subsequently incorporated as flattening depictions of ResNet-50 network-derived features [61]. The CNN feature extractor's embedding step turns the model into a hybrid one (DPT-Hybrid Architecture [61]). Following the original terminology of the transformer architecture, we shall refer to the embedded patches as tokens and the image patches as "words" in NLP tasks. The tokens are transformed using layers composed of multi head self-attention blocks. A reassemble operation is used to reassemble the output of the transformer layers, and then fusion blocks are used to gradually fuse the characteristics. In the embedding stage, the DPT-Hybrid architecture makes use of features that were taken from layers of the ResNet 50, and Fig. 1 depicts the entire dense prediction transformer model.

IV. Pipeline of Monocular Visual Odometry

The pipeline of monocular visual odometry is based basically on the DPT model to calculate depth estimation. Feature detection using a fast feature detection approach, matching features with optical flow, depth estimation by DPT, scale, and motion estimation blocks are the primary block components of the pipeline shown schematically in Fig. 2, which is also illustrated in Algorithm 1.

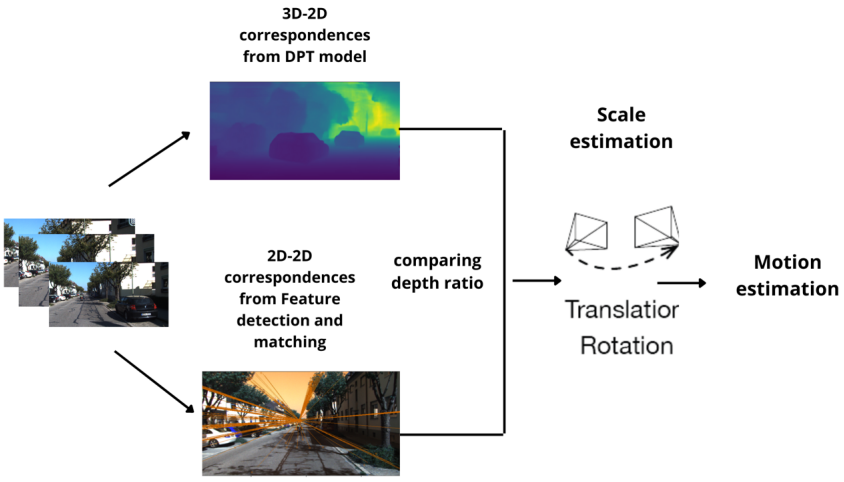

Fig. 2. The proposed approach's architecture estimates the scale from a deep-learning model to estimate the disparity map from a monocular camera. This enables the recovery of precise metric estimates.

Algorithm 1 Proposed Visual Odometry algorithm

Require: $Model$: $dpt\_hybrid\_kitti\text{-}cb926ef4.pt$
$Frames$: $[\boldsymbol{F}_1, ..., \boldsymbol{F}_k]$
Ensure: Vehicle poses: $[\boldsymbol{T_1}, \boldsymbol{T_2}, ..., \boldsymbol{T_k}]$
Initialization: n=2, N=number_Of_Frame
Prev_Feature=**FastFeatureDetection(F1)**
Last_Frame=F_1
while n $\leq$ N **do**
  Prev_Feature,Cur_Feature=**featureTracking(Last_Frame,** F_n**,Prev_Feature)**
  Compute **E** using Prev_Feature and Cur_Feature
  compute **[R, t]** using Essential matrix E
  Get Depth frame prediction D_n
  Get D'_n from Triangulation between Prev_Feature and Cur_Feature
  α : Scale estimation from comparison between D'_n,D_n
  if $|\alpha - absoluteScale| < \xi$ **then**
    $T_n = [R, \alpha\hat{\mathbf{t}}]$
  else
    3D-2D correspondences using D_n, Prev_Feature and Cur_Feature
    Compute $[R, t]$ from PnP
  end if
  n++
  Last_Frame=F_n
  Prev_Feature=Cur_Feature
end while

A. 2D-2D Correspondences

Monocular VO uses a single camera to combine images in an effort to progressively estimate an agent's motion. Epipolar geometry is one of the fundamental approaches that can be used to compute the pose from frame sequences using monocular or stereo cameras. Epipolar geometry is based on many steps, from 2D-2D correspondence to solving the essential matrix and the fundamental matrix **(E,F)**. From the intense optical flow, the 2D-2D correspondences are recovered. Given a pair of frames, $(F_k; F_{k+1})$, optical flow can be used to characterize the feature variation of time and provide correspondences for all the features that were derived from F_i and their correspondences in F_j. For the purpose of solving the fundamental matrix F and the essential matrix E, epipolar constraint is used based on the intrinsic calibration matrix **K** also indicates that the projection characteristics of the camera, where $F_k = K^{-T}E_kK^{-1}$ and the motion of the vehicle can be estimated using the following equation:

$$T = \begin{bmatrix} R_k & t_k \\ 0 & 1 \end{bmatrix} \tag{1}$$

with $R_k \in SO(3)$ and $t_k \in \mathbb{R}^{3\times1}$ are the rotation matrix and translation vector, respectively, that illustrate how the camera rotated and translated from instant $k-1$ to k The following are the manners in which the camera motion is associated with the essential matrix:

$$\boldsymbol{E} = [t]_\times \boldsymbol{R} \tag{2}$$

1) Fast feature detection: In the feature detection stage, interesting features in each frame, such as corners, are found. These locations are known as keypoints or features, and the next frames should be able to clearly identify them so that feature matching may be used.

Rosten and Drummond [62] developed the FAST feature detector (Features from Accelerated Segment Test). Fast criteria for interest point identification have grown in popularity as cutting-edge techniques with strict real-time limitations In Fig. 3, a feature is shown at pixel p if the intensities of at least nine surrounding pixels in a 16 pixel circle are all either lower than or higher than $I(p)$ by a threshold score. Training a decision tree further accelerated the algorithm, which examines candidate pixels into corners and is not based on as few pixels as possible. The procedure was accelerated even more by instructing a decision tree to look at as few pixels as possible to identify whether a candidate pixel is a corner or not. The segment test characteristics cannot be directly suppressed using

a non-maximal approach because no corner response function has been constructed. Therefore, a scoring function must be calculated for each detected corner in order to delete any corners that have an adjacent corner with a higher $C.C$ is provided by:

$$C = \max\left(\sum_{i\in S_{\text{bright}}} |I_{p\to i} - I_p| - t, \sum_{i\in S_{\text{dark}}} |I_p - I_{p\to i}| - t\right) \quad (3)$$

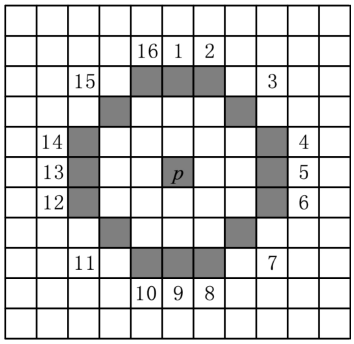

Fig. 3. Fast Feature Detection.

An analysis of similarity is performed at the end of the feature extraction process to compare each keypoint in a frame to all other keypoints in the following frame. In order to match detected features, the Lucas-Kanade method is used to calculate optical flow between frames iteratively.

2) Optical flow: Compared to other computer vision issues, ego-motion estimation has a fundamentally different basis, which places a greater focus on geometric motion inside individual video frames. The camera's output frame varies over time and could be seen as a function of time, and the assumption of photometric constancy is the basis for the optical flow computation. Alternatively, every frame has the same spatial location and a predetermined pixel intensity value. The following characteristics apply in the case of a pixel shifting to $(x + \Delta x, y + \Delta y)$ at a time of $t + \Delta t$:

$$\boldsymbol{M}(x + \Delta x, y + \Delta y, t + \Delta t) = \mathbf{M}(x, y, t) \quad (4)$$

On the left side of Eq. (4), we may carry out the first-order Taylor expansion:

$$\boldsymbol{M}(x+\Delta x, y+\Delta y, t+\Delta t) \approx \boldsymbol{M}(x, y, t)+D_x\Delta x+D_y\Delta y+D_t\Delta t \quad (5)$$

where:

$\frac{\partial \boldsymbol{M}}{\partial x}$, $\frac{\partial \boldsymbol{M}}{\partial y}$, and $\frac{\partial \boldsymbol{M}}{\partial t}$ are the frame's gradients Dx, Dy, and Dz in the x, y, and t axes, respectively. u=$\frac{dx}{dt}$ and v=$\frac{dy}{dt}$ are the pixels' rates of movement on the x and y axes, respectively. and The future grayscale equals the prior one based on photometric consistency, so:

$$uDx + vDy = -Dt \quad (6)$$

Eq. (6) can be expressed as a matrix:

$$[\boldsymbol{D}_x, \boldsymbol{D}_y]\begin{bmatrix} u \\ v \end{bmatrix} = -\boldsymbol{D}_{\text{t}} \quad (7)$$

The conventional approach is to use the Lucas-Kanade (LK) method to introduce the least squares solution to establish the u, v pixel motion. By doing this, we can determine how quickly pixels change between frames. Fig. 4 shows an example of feature tracking using optical flow.

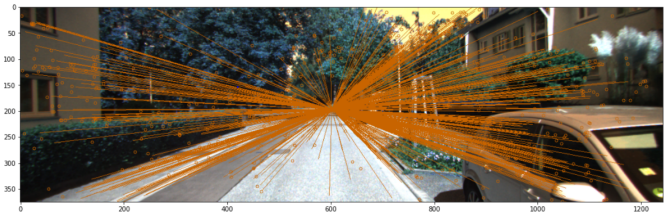

Fig. 4. Feature tracking using optical flow.

B. 3D-2D Correspondences

It is possible to construct 2D-2D and 3D-2D correspondences given a depth prediction from the DPT model and the features extracted using fast feature detection. Either PnP (3D-2D) or the essential matrix can be used to solve the relative camera pose.

1) Depth estimation: The problem of dense regression is frequently used to model a monocular depth estimate. Massive datasets can be formed from sources of data that already exist if certain considerations are made in how many depth representations are combined into a single representation and common ambiguities (like scale ambiguity) are addressed properly in the training loss. Since it is well known that transformers only perform to their greatest potential when a wealth of training data is provided, Our research primarily relied on monocular depth estimation using the DPT model.

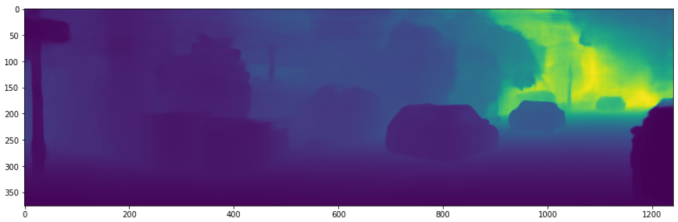

Fig. 5. The DPT model was used to create a depth view of the KITTI dataset image.

Fig. 5 shows an example of using the DPT model to create a depth view based on the Kitti dataset. A convolutional decoder is used by DPT to gradually merge tokens from various stages of the vision transformer into full-resolution predictions. The translation vector can be corrected and the relative scale in MVO estimated in a variety of ways. The scale can be estimated using the depth information and also using the earlier knowledge of camera height. The scale recovery approach used by Zhan et al [63]. is based on CNN depth estimation for lining

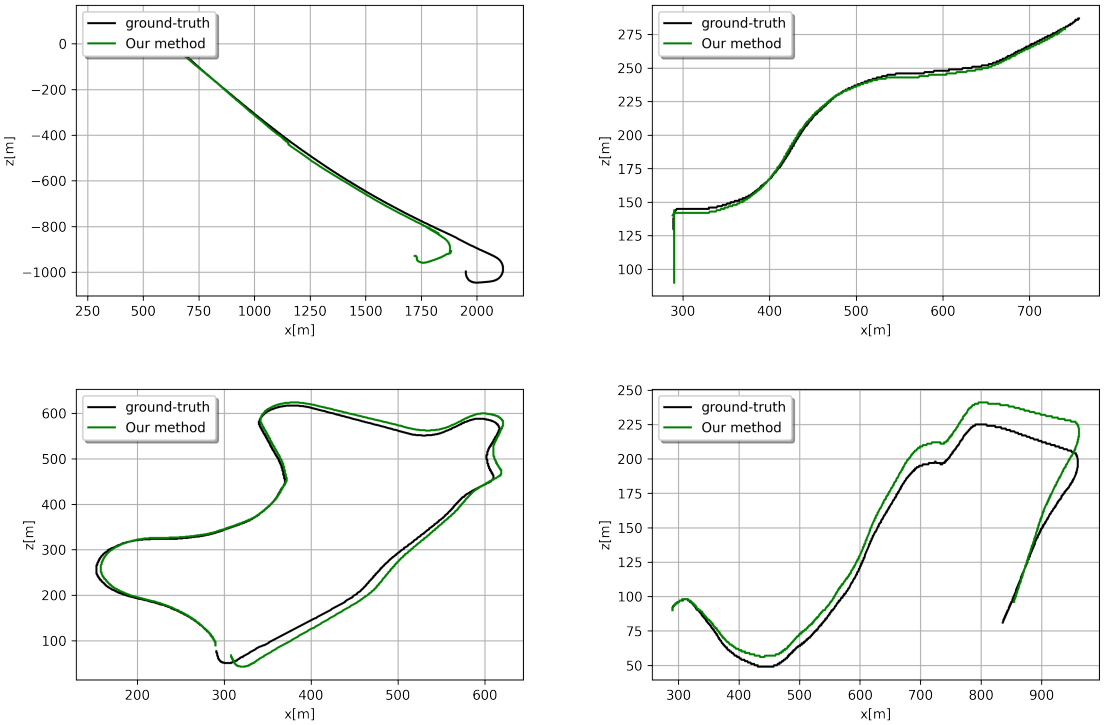

Fig. 6. Localization results of depth visual odometry compared to ground-Truth of KITTI dataset.

up depth based on a triangulated approach with the estimated depth map from deep learning. The realistic 3D structures are presumed to be known in this scenario since the depth sensor is supposed to be the deep learning model. Once the triangulation procedure has removed any earlier outliers, let M represent the number of keypoints that are still matching. A vector representing depth ratios is utilized to establish the scale, as is a RANSAC regressor.

$$\mathbf{D} = \left[\frac{\hat{D_0}}{D_0}, \frac{\hat{D_1}}{D_1}, ..., \frac{\hat{D_M}}{D_M} \right]^T \tag{8}$$

2) PnP(Perspective from n-point)-based motion estimation: To address this pose estimation problem using 3D-to-2D correspondences, either a nonlinear strategy (perspective from n points, PnP approach) or a linear technique (DLT-based estimation) can be used.The keypoints' depth is considered to be known given the estimated depth map produced by the deep learning model. Therefore, using a 2D projection of the corresponding 3D point, the PnP predicts the camera motion $\mathbf{T}_k$ as follows:

$$\arg\min_{\mathbf{T}_k} \sum_i \left\| \mathbf{K} \left(\mathbf{R} \mathbf{P}_{k-1}^i + \mathbf{t} \right) - \mathbf{x}_k^i \right\|_2 \tag{9}$$

V. Results and Discussion

As a primary method for evaluating our trajectories, we employ the KITTI relative error metric. According to the error metrics, we compute the average RMSE error for the rotational r_{error} and translational t_{error} errors using different sequences of KITTI Dataset. The relative pose error, which is particularly helpful for the evaluation of visual odometry approaches since it correlates to the drift of the trajectory, assesses the local accuracy of the trajectory over a set time period $Delta$. at time step i let's define the relative pose error matrix as follows:

$$Ei := \left(\hat{P_i}^{-1} \hat{P_j} \right)^{-1} \left(P_i^{-1} P_j \right) \tag{10}$$

The relative pose error matrix m is obtained from a sequence of N camera poses where $M = N - \Delta$ where the estimated and the real camera poses are $\hat{P} \in SE(3)$ and $P \in SE(3)$, respectively. Typically, the translation and rotation parts of the RPE are separated.

$$\text{trans}_{\text{error}}^i = \left(\frac{1}{M} \sum_1^M \left\| \text{trans} \left(E_i \right) \right\|^2 \right)^{\frac{1}{2}} \tag{11}$$

As for the rotation component, we use the mean error approach:

$$\text{rot}_{error}^i = \frac{1}{M} \sum_1^M \angle \left(\text{rot} \left(E_i^{\Delta} \right) \right) \tag{12}$$

Averaging both the translation and rotation components of SLAM systems is a sensible approach for evaluating these systems.

TABLE I. COMPARING THE RMSE OF THE SEQUENCES 01 AND 03 USING DIFFERENT METHODS

Method \Sequences	Sequence 01				Sequence 03			
	$tran_{error}$	rot_{error}	$RPE(m)$	RPE(°)	$tran_{error}$	rot_{error}	$RPE(m)$	RPE(°)
ORB-SLAM2(without LC)	107.57	**0.89**	2.970	**0.098**	**0.97**	**0.19**	0.031	0.055
DF-VO(Mono-SC Train)	66.98	17.04	1.281	0.725	2.67	0.50	**0.030**	**0.038**
VISO2	61.36	7.68	1.413	0.432	30.21	2.21	0.226	0.157
SfM-Learner	22.41	2.79	0.660	0.133	12.56	4.52	0.077	0.158
Depth-VO-Feat	23.78	1.75	0.547	0.133	15.76	10.62	0.168	0.308
Our Method	**21.53**	1.64	**0.471**	0.125	6.81	2.23	0.201	0.1689

TABLE II. COMPARING THE RMSE OF THE SEQUENCES 09 AND 10 USING DIFFERENT METHODS

Method \Sequences	Sequence 09				Sequence 10			
	$tran_{error}$	rot_{error}	$RPE(m)$	RPE(°)	$tran_{error}$	rot_{error}	$RPE(m)$	RPE(°)
ORB-SLAM2(Without LC)	9.30	**0.26**	0.128	0.061	2.57	0.32	**0.045**	0.065
DF-VO(Mono)	**2.47**	0.30	**0.055**	**0.037**	**1.96**	**0.31**	0.047	**0.042**
VISO2	18.06	1.25	0.284	0.125	26.10	3.26	0.442	0.154
SfM-Learner	11.32	4.07	0.103	0.159	15.25	4.06	0.118	0.171
Depth-VO-Feat	11.89	3.60	0.164	0.233	12.82	3.41	0.159	0.246
Our Method	3.41	1.42	0.094	0.147	16.25	7.82	0.148	0.187

$$\angle S := \arccos\left(\frac{\mathrm{tr}(S)-1}{2}\right) \tag{13}$$

We performed a qualitative experiment comparing this approach to visual odometry with numerous cutting-edge VO techniques to evaluate its applicability for estimating scale, including traditional monocular,stereo approaches, and learning approaches for monocular odometry such as ORB-SLAM2[64], DF-VO[63], VISO2[65], SfM-Learner[54], and Depth-VO-Feat[66]. Conventional monocular VO techniques necessitate posture with a prior knowledge of ground truth and are unable to recover the absolute scale. The global loop-closure detection of the ORB-SLAM2 has been deactivated in order to create an equivalent comparison. The keyframe trajectories of the ORB-SLAM2 are matched to ground truth via similarity transformation because it cannot retrieve the absolute scale.

On the KITTI odometry dataset's sequences 01, 03, 09, and 10, respectively, Fig. 6 compares the trajectories obtained by our approach to the ground truth trajectories. How closely the trajectory produced by our technique and the ground truth in the numbers 01, 03, 09, and 10 correspond. The metrics were computed using the KITTI evaluation toolkit for the KITTI sequences with ground truth. The quantitative results are shown in Tables I and II, and the best metric values are denoted by bolded values. Our strategy produced good results and was comparable with the other approaches, but the DF-VO remains the accurate framework not only for the four trajectories but for all trajectories of KITTI-Dataset. Additionally, in the various four sequences, the $terr$ and $rerr$ both produced good results that were greater to those of some other methods in the four trajectory. For rotation and translation RPEs, our model performed significantly enough.

This shows that to reduce scale drift problems, it is essential to have a depth map computed by a high accuracy and precise model in scenarios where depth-map estimation is used to compute the scale. We showed that a transformer-based method can produce results that are comparable to or even better than CNN-based techniques when used as a monocular visual odometry system component.

VI. CONCLUSION

In this paper, we present a method for estimating scale in visual odometry using a dense prediction transformer model. Due to our model's high performance in estimating depth maps from a monocular camera, scale drifts were reduced in multiple visual odometry sequences of the KITTI dataset. As a result of our experimental results on the KITTI odometry benchmark, we are confident that our proposed method is not only accurate enough but also shows a similar result to state-of-the-art approaches.

While data fusion-based visual approaches provide the highest accuracy of localization, they have some limitations, such as being computationally expensive. Real-time operation on resource-constrained systems is still possible if implemented efficiently. As a part of our future work, we will concentrate on developing a real-time integration of GNSS, IMU, and Lidar data using one of the extensions of the Kalman filter. In order to optimize the time consumption of low-cost embedded system implementation without losing accuracy.

ACKNOWLEDGMENT

We would like to express our gratitude to the Moroccan National Center for Scientific and Technical Research (CNRST) for its encouragement (grant number: 37 UM5R2022) during the period June 2022 to April 2023.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] PHAM, Minh et XIONG, Kaiqi. A survey on security attacks and defense techniques for connected and autonomous vehicles. Computers & Security, 2021, vol. 109, p. 102269.

[2] MANUEL, Melvin P., FAIED, Mariam, KRISHNAN, Mohan, et al. Robot Platooning Strategy for Search and Rescue Operations. Intelligent Service Robotics, 2022, vol. 15, no 1, p. 57-68.

[3] KUMAR, Amit, OJHA, Aparajita, YADAV, Sonal, et al. Real-time interception performance evaluation of certain proportional navigation based guidance laws in aerial ground engagement. Intelligent Service Robotics, 2022, vol. 15, no 1, p. 95-114.

[4] TIAN, Ying, YAO, Qiangqiang, WANG, Chengqiang, et al. Switched model predictive controller for path tracking of autonomous vehicle considering rollover stability. Vehicle System Dynamics, 2022, vol. 60, no 12, p. 4166-4185.

[5] LI, Qingqing, QUERALTA, Jorge Peña, GIA, Tuan Nguyen, et al. Multi-sensor fusion for navigation and mapping in autonomous vehicles: Accurate localization in urban environments. Unmanned Systems, 2020, vol. 8, no 03, p. 229-237.

[6] SABIHA, Ahmed D., KAMEL, Mohamed A., SAID, Ehab, et al. Real-time path planning for autonomous vehicle based on teaching–learning-based optimization. Intelligent Service Robotics, 2022, vol. 15, no 3, p. 381-398.

[7] MENG, Lingbo, YE, Chao, et LIN, Weiyang. A tightly coupled monocular visual lidar odometry with loop closure. Intelligent Service Robotics, 2022, vol. 15, no 1, p. 129-141.

[8] YEONG, De Jong, VELASCO-HERNANDEZ, Gustavo, BARRY, John, et al. Sensor and sensor fusion technology in autonomous vehicles: A review. Sensors, 2021, vol. 21, no 6, p. 2140.

[9] DU, Hao, WANG, Wei, XU, Chaowen, et al. Real-time onboard 3D state estimation of an unmanned aerial vehicle in multi-environments using multi-sensor data fusion. Sensors, 2020, vol. 20, no 3, p. 919.

[10] XU, Xiaobin, ZHANG, Lei, YANG, Jian, et al. A review of multi-sensor fusion slam systems based on 3D LIDAR. Remote Sensing, 2022, vol. 14, no 12, p. 2835.

[11] MARKOVIĆ, Lovro, KOVAČ, Marin, MILIJAS, Robert, et al. Error state extended Kalman filter multi-sensor fusion for unmanned aerial vehicle localization in GPS and magnetometer denied indoor environments. In : 2022 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE, 2022. p. 184-190.

[12] ZHOU, Yi, GALLEGO, Guillermo, et SHEN, Shaojie. Event-based stereo visual odometry. IEEE Transactions on Robotics, 2021, vol. 37, no 5, p. 1433-1450.

[13] ABOUZAHIR, Mohamed, ELOUARDI, Abdelhafid, LATIF, Rachid, et al. Embedding SLAM algorithms: Has it come of age?. Robotics and Autonomous Systems, 2018, vol. 100, p. 14-26.

[14] FERRERA, Maxime, EUDES, Alexandre, MORAS, Julien, et al. OV ^{2} SLAM: A Fully Online and Versatile Visual SLAM for Real-Time Applications. IEEE Robotics and Automation Letters, 2021, vol. 6, no 2, p. 1399-1406.

[15] CHENG, Jun, ZHANG, Liyan, CHEN, Qihong, et al. A review of visual SLAM methods for autonomous driving vehicles. Engineering Applications of Artificial Intelligence, 2022, vol. 114, p. 104992.

[16] CAMPOS, Carlos, ELVIRA, Richard, RODRÍGUEZ, Juan J. Gómez, et al. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics, 2021, vol. 37, no 6, p. 1874-1890.

[17] YU, Zhelin, ZHU, Lidong, et LU, Guoyu. Tightly-coupled Fusion of VINS and Motion Constraint for Autonomous Vehicle. IEEE Transactions on Vehicular Technology, 2022.

[18] QIN, Tong, LI, Peiliang, et SHEN, Shaojie. Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics, 2018, vol. 34, no 4, p. 1004-1020.

[19] SATTLER, Torsten, TORII, Akihiko, SIVIC, Josef, et al. Are large-scale 3d models really necessary for accurate visual localization?. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 1637-1646.

[20] YIN, Xiaochuan, WANG, Xiangwei, DU, Xiaoguo, et al. Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields. In : Proceedings of the IEEE international conference on computer vision. 2017. p. 5870-5878.

[21] ZHANG, Hui, WANG, Xiangwei, YIN, Xiaochuan, et al. Geometry-Constrained Scale Estimation for Monocular Visual Odometry. IEEE Transactions on Multimedia, 2021.

[22] ÖLMEZ, Burhan et TUNCER, Temel Engin. Metric scale and angle estimation in monocular visual odometry with multiple distance sensors. Digital Signal Processing, 2021, vol. 117, p. 103148.

[23] TIAN, Rui, ZHANG, Yunzhou, ZHU, Delong, et al. Accurate and robust scale recovery for monocular visual odometry based on plane geometry. In : 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021. p. 5296-5302.

[24] LI, Ruihao, WANG, Sen, et GU, Dongbing. Ongoing evolution of visual slam from geometry to deep learning: Challenges and opportunities. Cognitive Computation, 2018, vol. 10, no 6, p. 875-889.

[25] ARSHAD, Saba et KIM, Gon-Woo. Role of deep learning in loop closure detection for visual and lidar slam: A survey. Sensors, 2021, vol. 21, no 4, p. 1243.

[26] CHAPLOT, Devendra Singh, GANDHI, Dhiraj, GUPTA, Saurabh, et al. Learning to explore using active neural slam. arXiv preprint arXiv:2004.05155, 2020.

[27] DUAN, Chao, JUNGINGER, Steffen, HUANG, Jiahao, et al. Deep learning for visual SLAM in transportation robotics: a review. Transportation Safety and Environment, 2019, vol. 1, no 3, p. 177-184.

[28] PEDRAZA, Luis, RODRIGUEZ-LOSADA, Diego, MATIA, Fernando, et al. Extending the limits of feature-based SLAM with B-splines. IEEE Transactions on Robotics, 2009, vol. 25, no 2, p. 353-366.

[29] ENGEL, Jakob, STÜCKLER, Jörg, et CREMERS, Daniel. Large-scale direct SLAM with stereo cameras. In : 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2015. p. 1935-1942.

[30] SILVEIRA, Geraldo, MALIS, Ezio, et RIVES, Patrick. An efficient direct approach to visual SLAM. IEEE transactions on robotics, 2008, vol. 24, no 5, p. 969-979.

[31] STRASDAT, Hauke, MONTIEL, J., et DAVISON, Andrew J. Scale drift-aware large scale monocular SLAM. Robotics: Science and Systems VI, 2010, vol. 2, no 3, p. 7.

[32] STRASDAT, Hauke, MONTIEL, J., et DAVISON, Andrew J. Scale drift-aware large scale monocular SLAM. Robotics: Science and Systems VI, 2010, vol. 2, no 3, p. 7.

[33] KRIZHEVSKY, Alex, SUTSKEVER, Ilya, et HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, vol. 60, no 6, p. 84-90.

[34] KONRAD, Janusz, WANG, Meng, et ISHWAR, Prakash. 2d-to-3d image conversion by learning depth from examples. In : 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012. p. 16-22.

[35] KARSCH, Kevin, LIU, Ce, et KANG, Sing Bing. Depth extraction from video using non-parametric sampling. In : European conference on computer vision. Springer, Berlin, Heidelberg, 2012. p. 775-788.

[36] MUR-ARTAL, Raul, MONTIEL, Jose Maria Martinez, et TARDOS, Juan D. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE transactions on robotics, 2015, vol. 31, no 5, p. 1147-1163.

[37] GAO, Yuan et YUILLE, Alan L. Symmetric non-rigid structure from motion for category-specific object structure estimation. In : European Conference on Computer Vision. Springer, Cham, 2016. p. 408-424.

[38] GAO, Yuan et YUILLE, Alan L. Exploiting symmetry and/or manhattan properties for 3d object structure estimation from single and multiple images. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 7408-7417.

[39] MA, Jiayi, ZHOU, Huabing, ZHAO, Ji, et al. Robust feature matching for remote sensing image registration via locally linear transforming. IEEE Transactions on Geoscience and Remote Sensing, 2015, vol. 53, no 12, p. 6469-6481.

[40] EIGEN, David, PUHRSCH, Christian, et FERGUS, Rob. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems, 2014, vol. 27.

[41] EIGEN, David et FERGUS, Rob. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In : Proceedings of the IEEE international conference on computer vision. 2015. p. 2650-2658.

[42] LAINA, Iro, RUPPRECHT, Christian, BELAGIANNIS, Vasileios, et al. Deeper depth prediction with fully convolutional residual networks. In : 2016 Fourth international conference on 3D vision (3DV). IEEE, 2016. p. 239-248.

[43] LI, Jun, KLEIN, Reinhard, et YAO, Angela. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In : Proceedings of the IEEE International Conference on Computer Vision. 2017. p. 3372-3380.

[44] YAN, Han, ZHANG, Shunli, ZHANG, Yu, et al. Monocular depth estimation with guidance of surface normal map. Neurocomputing, 2018, vol. 280, p. 86-100.

[45] LI, Bo, SHEN, Chunhua, DAI, Yuchao, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 1119-1127.

[46] ROY, Anirban et TODOROVIC, Sinisa. Monocular depth estimation using neural regression forest. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 5506-5514.

[47] LIU, Fayao, SHEN, Chunhua, et LIN, Guosheng. Deep convolutional neural fields for depth estimation from a single image. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 5162-5170.

[48] GARG, Ravi, BG, Vijay Kumar, CARNEIRO, Gustavo, et al. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In : European conference on computer vision. Springer, Cham, 2016. p. 740-756.

[49] ILG, Eddy, MAYER, Nikolaus, SAIKIA, Tonmoy, et al. Flownet 2.0: Evolution of optical flow estimation with deep networks. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 2462-2470.

[50] YU, Jason J., HARLEY, Adam W., et DERPANIS, Konstantinos G. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In : European Conference on Computer Vision. Springer, Cham, 2016. p. 3-10.

[51] GODARD, Clément, MAC AODHA, Oisin, et BROSTOW, Gabriel J. Unsupervised monocular depth estimation with left-right consistency. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 270-279.

[52] KUZNIETSOV, Yevhen, STUCKLER, Jorg, et LEIBE, Bastian. Semi-supervised deep learning for monocular depth map prediction. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 6647-6655.

[53] HUA, Yan et TIAN, Hu. Depth estimation with convolutional conditional random field network. Neurocomputing, 2016, vol. 214, p. 546-554.

[54] ZHOU, Tinghui, BROWN, Matthew, SNAVELY, Noah, et al. Unsupervised learning of depth and ego-motion from video. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 1851-1858.

[55] YIN, Zhichao et SHI, Jianping. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 1983-1992.

[56] LUO, Yue, REN, Jimmy, LIN, Mude, et al. Single view stereo matching. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 155-163.

[57] VASWANI, Ashish, SHAZEER, Noam, PARMAR, Niki, et al. Attention is all you need. Advances in neural information processing systems, 2017, vol. 30.

[58] DEVLIN, Jacob, CHANG, Ming-Wei, LEE, Kenton, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[59] RADFORD, Alec, NARASIMHAN, Karthik, SALIMANS, Tim, et al. Improving language understanding by generative pre-training. 2018.

[60] DOSOVITSKIY, Alexey, BEYER, Lucas, KOLESNIKOV, Alexander, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[61] RANFTL, René, BOCHKOVSKIY, Alexey, et KOLTUN, Vladlen. Vision transformers for dense prediction. In : Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. p. 12179-12188.

[62] ROSTEN, Edward, PORTER, Reid, et DRUMMOND, Tom. Faster and better: A machine learning approach to corner detection. IEEE transactions on pattern analysis and machine intelligence, 2008, vol. 32, no 1, p. 105-119.

[63] ZHAN, Huangying, WEERASEKERA, Chamara Saroj, BIAN, Jia-Wang, et al. DF-VO: What Should Be Learnt for Visual Odometry?. arXiv preprint arXiv:2103.00933, 2021.

[64] MUR-ARTAL, Raul et TARDÓS, Juan D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE transactions on robotics, 2017, vol. 33, no 5, p. 1255-1262.

[65] GEIGER, Andreas, ZIEGLER, Julius, et STILLER, Christoph. Stereoscan: Dense 3d reconstruction in real-time. In : 2011 IEEE intelligent vehicles symposium (IV). Ieee, 2011. p. 963-968.

[66] ZHAN, Huangying, GARG, Ravi, WEERASEKERA, Chamara Saroj, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 340-349.