A New Dense Hybrid Stereo Visual Odometry Approach

Ziming Liu, Ezio Malis, Philippe Martinet

A New Dense Hybrid Stereo Visual Odometry Approach

Ziming Liu [1] and Ezio Malis [1] and Philippe Martinet [1]

***Abstract*— Visual odometry is an important part of the perception module of autonomous robots. Recent advances in deep learning approaches have given rise to hybrid visual odometry approaches that combine both deep networks and traditional pose estimation methods. One limitation of deep learning approaches is the availability of ground truth data needed to train the neural networks. For example, it is extremely difficult, if not impossible, to obtain a ground truth dense depth map of the environment to be used for stereo visual odometry. Even if unsupervised training of networks has been investigated, supervised training remains more reliable and robust. In this paper, we propose a new hybrid dense stereo visual odometry approach in which a dense depth map is obtained with a network that is supervised using ground truth poses that can be more easily obtained than ground truth depths maps. The depth map obtained from the neural network is used to warp the current image into the reference frame and the optimal pose is obtained by minimizing a cost function that encodes the similarity between the warped image and the reference image. The experimental results show that the proposed approach, not only improves state-of-the-art depth maps estimation networks on some of the standard benchmark datasets, but also outperforms the state-of-the-art visual odometry methods.**

I. INTRODUCTION

Visual odometry is an important task for the localization of autonomous robots and several approaches have been proposed in the literature [1]. In this paper, we consider visual odometry approaches that use stereo images since the depth of the observed scene can be correctly estimated. These approaches do not suffer from the scale factor estimation problem of monocular visual odometry that can only estimate the translation up to a scale factor [1].

Traditional model-based visual odometry approaches are generally divided into two steps. Firstly, the depths of the observed scene are estimated from the disparity obtained by matching left and right images. Then, the depths are used to obtain the camera pose. The depths can be computed for selected features like in sparse methods [2], [3] or for all possible pixels in the image like in dense direct methods [4]. It has been shown that dense direction methods are more robust than sparse-based visual odometry methods because they use all the possible information and they avoid feature extraction that is prone to errors.

Recently, more and more end-to-end deep learning visual odometry approaches have been proposed, including supervised [5] and unsupervised [6] models. However, it has been shown that hybrid visual odometry approaches can achieve better results [7], [8], [9]. Hybrid visual odometry models predict depths with a deep neural network, and estimate the camera poses with model-based methods. However, previous works focused on combining deep neural network with sparse model-based pose estimation. In this paper we will propose a new hybrid approach that combines a deep neural network with direct dense model-based pose estimation.

Many deep learning-based approaches have focused on using monocular images to estimate depth maps [10], [6], [11], [9]. But monocular depth estimation is an ill-posed problem and the scale factor cannot be correctly estimated. In this work, we consider stereo depth estimation DNN since they are more reliable and stable [12]. Moreover, recent works have shown that deep learning-based approaches [12] can perform much better than the traditional stereo matching approaches [13], [14] and provide more accurate depth estimation. Some deep learning-based methods follow the traditional stereo matching pipeline, making this pipeline to be differential and trainable with deep learning [12]. Other works use a simple encoder-decoder architecture to predict the disparity and depth [15].

Deep neural network training can be unsupervised or supervised. Unsupervised approaches can use the stereo geometric constraint like in [10], [11] or use videos temporal reconstruction of appearance images or depth maps with the camera pose predicted by another pose CNN [6], [9], [11]. Supervised stereo depth estimation networks can be trained with the ground truth depth maps [12]. However, ground truth depth maps are extremely difficult to be obtained for all pixels in real and variable environments. We can consider that this is almost impossible for dense depth maps.

The main contributions of this paper are (i) to propose a paradigm shift in DNN supervised approaches by using a pose-supervised depth-estimation network (PDENet) instead of a depth-supervised depth-estimation network and (ii) to propose a new dense hybrid stereo visual odometry approach, which is composed of a pose-supervised depth-estimation network and a direct dense stereo model-based pose estimation (DPE) module. The proposed PDENet pose-supervised paradigm outperforms state-of-the-art results without using ground truth depth maps for training. Combined with the DPE module, the hybrid stereo visual odometry approach achieve state-of-the-art results on standard benchmarks.

This paper is organized as follows. In Section II we describe the proposed method and main contributions. The experimental results are shown in Section III. Finally, Section IV concludes the paper and presents ideas for future work.

[1] All authors are with ACENTAURI team at INRIA (Sophia Antipolis, France) and 3IA Côte d'Azur Institute, Université Côte d'Azur (Nice, France). {first name}.{name}@inria.fr

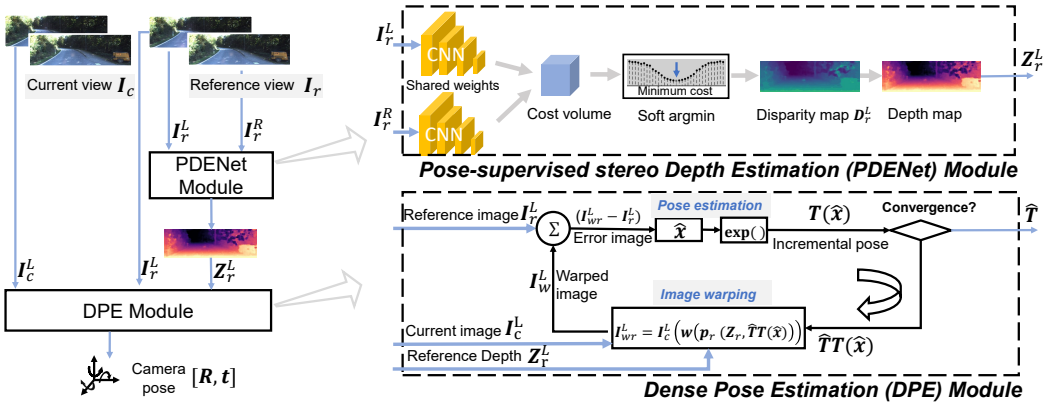


Fig. 1. The architecture of the proposed Dense Hybrid Stereo Visual Odometry Method (on the left) and details on the Pose-supervised stereo Depth Estimation (PDEnet) and the Dense Pose Estimation (DPE) modules (on the right).

II. PROPOSED METHOD

In this paper, we propose a new dense hybrid stereo visual odometry method that is composed of two main parts: a pose-supervised depth estimation network, which will be named the PDENet module, and a dense model-based pose estimation module, that will be named the DPE module. The architecture of the proposed method is illustrated in Figure 1.

The main idea of the hybrid method is to combine a new neural network for accurate dense depth estimation with a dense visual odometry approach inspired by [4]. In this section, we firstly introduce some notations and mathematical backgrounds. Then we will explain the architecture of the depth estimation network (PDENet module), and how to train this DNN with a pose-supervised function. Finally, we will describe in detail the dense model-based pose estimation module (DPE module). It must be emphasized that the main contribution of this paper is in the global architecture more than the design of the specific modules.

A. Notations

Let $\mathbf{I}(\mathbf{p}) = [I_R(\mathbf{p}), I_G(\mathbf{p}), I_B(\mathbf{p})]$ be the $(rows \times cols \times 3)$ tensor containing the RGB image values, where $\mathbf{p} = [u; v]$ is the 2×1 vector containing the pixel coordinates. $\mathbf{I}_r(\mathbf{p}_r)$ will represent the image acquired in a reference frame, $\mathbf{p}_r = [u_r; v_r]$ being the 2×1 vector containing the pixel coordinates in that reference frame. $\mathbf{I}_c(\mathbf{p}_c)$ will represent the image acquired in a current frame, $\mathbf{p}_c = [u_c; v_c]$ being 2×1 vector containing the pixel coordinates in that current frame. We will assume that there is a warping function transforming the current image into the reference image:

$$\mathbf{I}_r(\mathbf{p}_r) = \mathbf{I}_c(\mathbf{w}(\mathbf{K}, {}^c\mathbf{T}_r, \mathbf{Z}_r, \mathbf{p}_r)) \tag{1}$$

where $\mathbf{K}$ is the 3×3 matrix containing the camera intrinsic parameters (we suppose that the reference and current images are taken by the same camera), ${}^c\mathbf{T}_r \in SE(3)$ is the 4×4 matrix containing the pose of the reference frame relative to the current frame and $\mathbf{Z}_r$ is a $(rows \times cols)$ matrix containing the Z coordinates of the 3D points of the scene in the reference frame.

$\mathscr{D}_r = \mathbf{p}_{r1}, \mathbf{p}_{r2}, ..., \mathbf{p}_{rn}$ represents the set of all pixels coordinates defining a domain in the reference image.

$\mathbf{S} = (\mathbf{I}^L, \mathbf{I}^R)$ represents the stereo images, where $\mathbf{I}^L$ is the left image and $\mathbf{I}^R$ the right image. $\mathbf{S}_r = (\mathbf{I}_r^L, \mathbf{I}_r^R)$ represents the stereo pair acquired in a reference frame and $\mathbf{S}_c = (\mathbf{I}_c^L, \mathbf{I}_c^R)$ represents the stereo pair acquired in a current frame.

B. The PDENet Module

As already mentioned, in order to obtain a depth map with correct scale we use a stereo based depth estimation network to extract features from stereo RGB images, and output predicted depth maps. In this paper, we modify a disparity-supervised network PSMNet [12] into a pose-supervised stereo based depth estimation network (that we will name PDENet). Different from PSMNet, which is trained with ground truth disparity maps in a supervised way, PDENet does not require ground truth depth or disparity map for training but is trained of camera pose annotations. The architecture of the PDENet network is illustrated on the top right of Figure 1. The images of the left and right cameras are fed into a CNN sub-network to extract features. These features form a cost volume, which is processed with 3D convolution layers including up-sampling [12]. Finally, we use disparity regression [16] to obtain the disparity map with the correct scale. With the predicted disparity map, we can obtain the depth map of the left camera. The predicted depth map will be used as the input of the warping functions of the warping losses, and the input of the DPE module.

We train the PDENet depth estimation network with proposed Pose-supervised Stereo Reconstruction (PSR) loss. PSR loss uses the ground truth camera poses, instead of depth (or disparity), as the supervision signal to train the PDENet. The PSR loss is based on two parts, the stereo geometric-based image warp (L_{stereo}) and the temporal pose-based image warp (L_{pose}).

$$L_{PSR} = (1-\alpha)(L_{stereo,R\to L} + L_{pose,c\to r} + L_{pose,r\to c}) + \alpha L_{SSIM} \tag{2}$$

The loss ratio α is set as 0.85.

$$L_{stereo,R\to L} = \frac{1}{n}\sum_{\mathbf{p}^L\in\mathscr{D}_r} \|(\mathbf{I}^L(\mathbf{p}^L) - \mathbf{I}^L_w(\mathbf{p}^L))\|, \tag{3}$$

$$L_{pose,c\to r} = \frac{1}{n}\sum_{\mathbf{p}_r\in\mathscr{D}_r} \|(\mathbf{I}_r(\mathbf{p}_r)) - \mathbf{I}_{wr}(\mathbf{p}_r))\|, \tag{4}$$

$$L_{pose,r\to c} = \frac{1}{n}\sum_{\mathbf{p}_c\in\mathscr{D}_c} \|(\mathbf{I}_c(\mathbf{p}_c)) - \mathbf{I}_{wc}(\mathbf{p}_c))\| \tag{5}$$

$$\begin{aligned} \mathbf{I}^L_w(\mathbf{p}^L) &= \mathbf{I}^R(\mathbf{w}(\mathbf{D}^L, \mathbf{p}^L)), \\ \mathbf{I}_{wr}(\mathbf{p}_r)) &= \mathbf{I}_c(\mathbf{w}(\mathbf{Z}_r, \mathbf{p}_r, cTr)), \\ \mathbf{I}_{wc}(\mathbf{p}_c)) &= \mathbf{I}_r(\mathbf{w}(\mathbf{Z}_c, \mathbf{p}_c, rTc)) \end{aligned} \tag{6}$$

$$L_{SSIM} = \sum_{i=\{r,c\},j=\{L\}} \frac{1}{2}(1 - SSIM(\mathbf{I}^j_i, \mathbf{I}^j_{wi})) \tag{7}$$

For the above equations, $\mathbf{I}, \mathbf{D}, \mathbf{Z}$ represent the image, disparity and depth maps, $*^L, *^R, *_r, *_c$ represent left view, right view, reference view and current view of stereo image sequences. $\mathbf{w}$ refers to the image warping operation.

Stereo-based loss $L_{stereo,R\to L}$: For the well-calibrated stereo images, we can warp the right image to the left image if we have the disparity map of the left view. L1 loss is computed based on the warped image and source image. The Eq. 3 shows how to compute the $L_{stereo,R\to L}$. Specifically, the PDENet predicts a disparity map $\mathbf{D}^L$ of the left image view, which is used for warping the right image to the left image view. Then we compute the L1 loss metrics on the left image $\mathbf{I}^L$ and warped left image $\mathbf{I}^L_w$. In practice, we compute the stereo-based loss from right image to left image on both the reference view and current view.

Pose-based loss $L_{pose,c\to r}$ and $L_{pose,r\to c}$: With the given temporal camera poses, we can warp photometric images from the current image to the reference image or from the reference image to the current image. Then we construct the $L_{pose,c\to r}$ and $L_{pose,r\to c}$ on the warped images and the source images. The pose estimation module depends on the predicted depth, training the PDENet with ground-truth poses can help to keep the temporal consistency.

To keep the consistency between the reference view and current view, we compute L_{pose} on both the current view and reference view, i.e. warp photometric images from the reference view to the current view and from the current view to the reference view with camera pose rTc, cTr.

For the example of $L_{pose,c\to r}$, given the ground truth relative camera pose ${}^c\mathbf{T}_r$ from the reference view to the current view, we can warp the current image $\mathbf{I}$ into to the reference view with the help of reference view depth map $\mathbf{Z_r}$, obtaining the warped reference image $\mathbf{I}_{wr}$, as shown in eq. 6. After obtaining the warped reference image $\mathbf{I}_{wr}$, we compute the L1 loss based on $\mathbf{I}_r, \mathbf{I}_{wr}$, as shown in Eq. 4.

SSIM loss:

The SSIM is a metric to measure the similarity between two images [17]. We also use structural similarity (SSIM) on all warped images and ground truth images. SSIM loss is a common method for depth estimation task [10], [11].

In addition, we also add a disparity smoothness loss on the predicted disparity map, like the previous work [10]. These two losses are small tricks to keep good disparity quality.

C. The DPE module

Composed with the depth results of the PDENet network, the dense pose estimation module (DPE module) outputs the 6Dof camera poses. The DPE module is a traditional appearance-based visual odometry algorithm inspired by [4]. In this part, we introduce the DPE module and its way to communicate with the depth estimation network.

Firstly, we formulate the pose estimation problem which is directly computed on the photometric image intensities by a non-linear model. This model accounts for dense 3D geometric configuration.

As the goal of DPE module is to estimate the camera pose $\widehat{\mathbf{T}} \in SE(3)$, we can transform the absolute camera pose estimation problem to a problem of estimating the incremental pose $\mathbf{T}(\mathbf{x})$, where $\mathbf{x}$ is a minimal parametrization of the Lie algebra $se(3)$. Therefore $\exists\, \widehat{\mathbf{x}} : \widehat{\mathbf{T}}\mathbf{T}(\widehat{\mathbf{x}}) = \overline{\mathbf{T}}$, where $\overline{\mathbf{T}}$ is the optimal pose. The estimated pose $\widehat{\mathbf{T}}$ will be updated iteratively with the incremental homogeneous transformation $\widehat{\mathbf{T}}_{k+1} = \widehat{\mathbf{T}}_k\mathbf{T}(\mathbf{x})$.

The parameters $\mathbf{x} \in \mathbb{R}^6$ are defined with Lie algebra, $\mathbf{x} = (\omega\Delta t, v\Delta t)$ which refers to the integral of the constant velocity twist. v is the linear velocity, ω is the angular velocity.

The relation between the incremental pose $\mathbf{T}(\mathbf{x})$ and the twist is as follows.

$$\mathbf{T}(\mathbf{x}) = \exp\left(\begin{bmatrix} [\omega\Delta t]_\times & v\Delta t \\ 0 & 0 \end{bmatrix}\right) \tag{8}$$

where the $[\cdot]_\times$ is a skew symmetric matrix operator.

Finally, the pose estimation problem becomes an optimization problem of minimizing a non-linear least squared cost function, shown as follows.

$$C(\mathbf{x}) = \sum_{\mathbf{p}_r\in\mathscr{D}_r} \|\mathbf{I}_r(\mathbf{p}_r) - \mathbf{I}(\mathbf{w}(\mathbf{Z}_r, \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}), \mathbf{p}_r)\|^2 \tag{9}$$

In our implementation, we use Gauss-Newton method to minimize this cost function $C(\mathbf{x})$, with an HUBER M-Estimator [18].

III. EXPERIMENTS

In this section, we firstly introduce the experimental setup. Since our hybrid dense visual odometry approach is composed of two modules we will first present results to assess the accuracy of the deep stereo depth network (PDENet) and then we will present the results on the complete hybrid visual odometry approach (PDENet+DPE).

A. Experimental setup

1) Datasets: We used the following well known datasets.

KITTI Depth (Eigen split [19]): To compare the performance of the PDENet with other state-of-the-art works, we train our PDENet deep network on KITTI Depth dataset

with the Eigen train/test split [19]. There are 22,600 training stereo images, 888 validation stereo images, and 697 benchmark testing stereo images. To be fair, we use the same code provided by [10] to generate the ground truth depth maps, which re-projects 3D points viewed from the velodyne laser to the left RGB camera. We use the same cropping operation in [19], and test the depth results with original image resolution.

KITTI Odometry: KITTI Odometry dataset contains 11 sequences (seq $00 \rightarrow 10$) with camera pose annotations. Following most previous works, we use seq $00 \rightarrow 08$ for training, seq $09 \rightarrow 10$ for testing.

Virtual KITTI2: Virtual KITTI2 is a well-annotated simulation dataset for various tasks of auto driving. There are 6 Scenes with different weather conditions (clone, fog, morning, overcast, rain, sunset) and camera degrees (15-deg-left, 15-deg-right, 30-deg-left, 30-deg-right). We conduct ablation studies on this simulation dataset, with 6 sequences of 6 scenes of 15-deg-left.

2) Metrics: We consider the following metrics that are generally used by most of the papers in the literature for the KITTI datasets:

$t_{err}(\%)$*:* Evaluate the average translation error of sub-sequences of length (100, 200, ..., 800) meters.

$r_{err}(^\circ/100m)$*:* Evaluate the Average rotational errors of sub-sequences of length (100, 200, ..., 800) meters.

ATE (Absolute Trajectory Error): It directly measures the difference between estimated trajectory points and ground truth trajectory points at each frame.

We first break down the sequence into 5-frame segments, ATE is computed on 5-frame segments and averaged over the whole sequence. On each 5-frame segment, We first align the predicted absolute poses $a = [x;y;z]$ and ground truth absolute poses $a^{'} = [x^{'};y^{'};z^{'}]$ and optimize the scale factor γ [2], [6]. Assume segment length $L = 5$ and step-size $S = 1$, N is the length of the sequence. Relative mean square error $rmse_i = \frac{\sqrt{\sum_{k=i}^{i+L-1} ||(\gamma \cdot a_k^{'} - a_k)^2||_1}}{L}$, scale $\gamma = \frac{\sum_{k=i}^{i+L-1} ||(a_k^{'} * a_k)||}{\sum_{k=i}^{i+L-1} ||a_k^2||}$, $RMSE = \{rmse_i\}, i \in [1,N]$, i indexes the frame id of sequences, $ATE_{mean} = \frac{\sum_{i=1}^{N} rmse_i}{N}$, $ATE_{std} = \sqrt{\frac{\sum_{i=1}^{N} (rmse_i - ATE_{mean})^2}{N}}$.

Depth Error Metrics: abs rel, rel sqr, rmse, rmse log refer to absolute relative error, relative square error, root-mean-square error and log root-mean-square error. Each of them is defined as $abs\ rel = \frac{1}{N}\sum_{i=1}^{N}\frac{|z_i - z_i^*|}{z_i}$, $rel\ sqr = \frac{1}{N}\sum_{i=1}^{N}\frac{|z_i - z_i^*|^2}{z_i}$, $rmse = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|z_i - z_i^*|}$, $rmse\ log = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\log z_i - \log z_i^*|}$. z_i^*, z_i are the ground truth depth and predicted depth values.

Depth Accuracy Metric: Accuracy with threshold $\tau < thr$, percent of z_i such that $max(\frac{z_i}{z_i^*}, \frac{z_i^*}{z_i}) < thr$, where $thr = 1.25, 1.25^2, 1.25^3$.

B. Experiments for the PDENet module

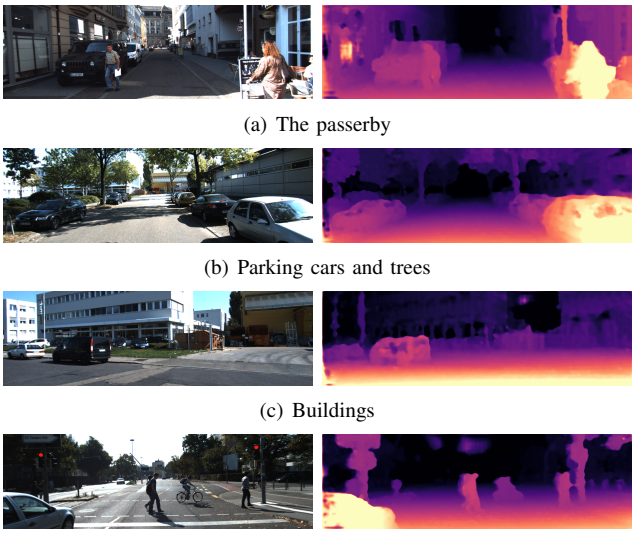
To show the advantage of the proposed stereo depth estimation network PDENet, we firstly compare the results of PDENet with other state-of-the-art depth estimation works on KITTI Depth dataset [26] with Eigen split [19]. Since the ground truth poses are not provided in KITTI Depth, we match the image samples between KITTI Depth dataset and KITTI Odometry dataset. Finally, there are 13217 KITTI Depth images which can be corresponded to KITTI Odometry dataset. It's about $58.5\%, (13217/22600)$ of the original KITTI Depth dataset. In this paper, PDENet achieves state-of-the-art result on KITTI Depth with only 58.5% samples.

1) Results on KITTI Depth Dataset: The comparison of the proposed PDENet with recent years' unsupervised depth estimation state-of-the-art methods is provided in Table I. We achieve better results compared with those monocular-based networks trained with ground truth depth maps [19], [7] or introducing an additional pose CNN [11], [8] which estimates a pose to be used as a pseudo ground-truth. This shows the advantage of stereo matching network architecture compared with monocular-based methods. As shown in Table I, PDENet achieves the best results on almost all metrics. The improvement of the metric *abs rel,* $\tau < 1.25$ *accuracy* is extremely obvious. PDENet makes the accuracy of $\tau < 1.25$ to a new level ($> 90\%$) for the first time. Note that to obtain the best accuracy $\tau < 1.25^3$ result, DVSO [7] obtains

year	model	Error Metrics				Accuracy Metric			supervision		
		abs rel	rel sqr	rmse	rmse log	$\tau < 1.25$	$< 1.25^2$	$< 1.25^3$	depth	pose	stereo
2014	Multi-scale-depth [19]	0.203	1.548	6.307	0.282	0.702	0.890	0.958	✓		
2015	Mono-depthDCN [20]	0.202	1.614	6.523	0.275	0.678	0.895	0.965	✓		
2017	SfmLearner [6]	0.208	1.768	6.856	0.283	0.678	0.885	0.957		✓	
2017	Monodepth [10]	0.148	1.344	5.927	0.247	0.803	0.922	0.964			✓
2018	DVSO[7]	0.097	**0.734**	4.442	0.187	0.888	0.958	**0.980**	✓		✓
2018	GeoNet [21]	0.149	1.060	5.567	0.226	0.796	0.935	0.975		✓	
2019	Comp collaboration [22]	0.148	1.149	5.464	0.226	0.815	0.935	0.973		✓	
2019	EPC++ [23]	0.128	0.935	5.011	0.209	0.831	0.945	0.979		✓	✓
2019	Monodepth2 [11]	0.106	0.806	4.630	0.193	0.876	0.958	**0.980**		✓	✓
2020	D3VO [8]	0.099	0.763	4.485	**0.185**	0.885	0.958	0.979		✓	✓
2021	Attention multi-warp [24]	0.121	0.971	5.206	0.214	0.843	0.944	0.975			✓
2021	Recursive stereo distillation [25]	0.105	0.842	4.810	0.196	0.861	0.947	0.978			✓
	PDENet(ours)	**0.080**	0.795	**4.146**	**0.185**	**0.922**	**0.959**	0.976		✓	✓

TABLE I

COMPARE WITH STATE-OF-THE-ART DEPTH ESTIMATION MODELS. THE PDENET IS TRAINED WITH 1024×320 IMAGE'S RESOLUTION.



(a) The passerby

(b) Parking cars and trees

(c) Buildings

(d) The traffic crossing

Fig. 2. Examples of depth maps from KITTI Depth test set (Eigen split).

0.980 by training with ground truth disparity annotations, monodepth2 [11] obtains same accuracy using and additional pose CNN and more data. PDENet achieves a quite close $\tau < 1.25^3$ accuracy with only 58.5% samples. However, they are all lower than ours in the most difficult accuracy metric $\tau < 1.25$. These results suggest that the ground truth depth and additional pose CNN are not essential for obtaining state-of-the-art depth estimation results.

Figure 2 shows some examples of estimated depth with the proposed PDENet network.

2) Results on Virtual KITTI2 Dataset: We use the simulated dataset Virtual KITTI2 [27] to evaluate the performance of the baseline model [28], [12] and PDENet. Indeed, this dataset provides accurate depth annotations. We used sequences: 01, 02, 18 for training and sequence 06 for testing.

As shown in Tables II, PDENet with pose-supervised PSR loss considerably improve the results compared with unsupervised stereo geometric-based approach proposed in [28]. With the help of ground truth camera poses, we can considerably improve the depth map quality. Note again, that ground truth camera poses can be much more easily obtained than ground truth depths.

Loss	Depth Error Metrics			
	rel	rel sqr	rmse	rmse log
baseline [28], [12]	0.0884	1.4229	6.7763	0.2236
PDENet	**0.0565**	**1.3299**	**5.9259**	**0.1880**
Loss	Depth Accuracy Metrics			
	$\tau < 1.25$	$< 1.25^2$	$< 1.25^3$	
baseline [28], [12]	0.9101	0.9584	0.9751	
PDENet	**0.9495**	**0.9717**	**0.9819**	

TABLE II

COMPARE DEPTH ERROR AND ACCURACY WITH OTHER METHOD.

C. Experiments for the Hybrid Visual Odometry Approach

In this section, we show the comparison with the state-of-the-art visual odometry approaches using the KITTI Odometry dataset [26]. The recent state-of-the-art deep learning based visual odometry approaches can be classified into three groups, including model-based, end-to-end deep learning models which are the mainstream and hybrid approaches. Similar to most papers, we make the state-of-the-art comparison on KITTI Odometry benchmark training the PDENet on sequences from $00 \rightarrow 08$, and testing on sequences $09, 10$.

year	model	seq.9		seq.10		input	
		t_{err}	r_{err}	t_{err}	r_{err}	monocular	stereo
	model-based						
2015	ORB [2]	15.30	0.26	3.68	0.48	✓	
	end-to-end						
2017	SfmLearner [6]	17.84	6.78	37.91	17.80	✓	
2018	Geonet [21]	43.76	16.00	35.60	13.8	✓	
2019	Wang [29]	9.30	3.50	7.21	3.90	✓	
2019	Li [30]	8.10	2.81	12.90	3.17	✓	
2021	TAPE [31]	6.72	2.60	8.66	3.13	✓	
2021	F2FPE [31]	2.36	1.06	3.00	1.28	✓	
	hybrid						
2018	DVSO [7]	0.83	0.21	0.74	0.21	✓	
2019	UnOS [32]	5.21	1.80	5.20	2.18		✓
2020	DFVO [9]	2.07	0.23	2.06	0.36	✓	
	Ours	0.87	0.28	0.83	0.54		✓

TABLE III

t_{err}, r_{err}. KITTI ODOMETRY, TEST ON SEQUENCE 09 AND 10.

year	model	seq.9	seq.10	input	
		ATE	*ATE*	mono	stereo
	model-based				
2015	ORB full [2]	0.0140±0.0080	0.0120±0.0110	✓	
2015	ORB short [2]	0.0640±0.1410	0.0640±0.1300	✓	
	end-to-end				
2017	SfmLearner [6]	0.0160±0.0090	0.0130±0.0090	✓	
2018	Geonet [21]	0.0120±0.0070	0.0120±0.0090	✓	
2018	Vid2depth [33]	0.0130±0.0100	0.0120±0.0110	✓	
2019	Com Col [22]	0.0120±0.0070	0.0120±0.0080	✓	
	hybrid				
2019	UnOS [32]	0.0120±0.0060	0.0130±0.0080		✓
	Ours	**0.0109±0.0068**	**0.0105±0.0088**		✓

TABLE IV

ATE. KITTI ODOMETRY, TEST ON SEQUENCE 09, 10.

Table III show the results for the metric t_{err} and r_{err} that are used to evaluate the camera pose results on KITTI Odometry. The proposed dense hybrid stereo visual odometry approach achieves competitive results compared with all the state-of-the-art methods. Similar to other hybrid methods, dense hybrid stereo visual odometry achieves better results than state-of-the-art end-to-end deep learning methods and model-based ORB SLAM methods. Compared to the hybrid method, the proposed approach achieves the best result on sequence 10 and the next-best result on sequence 09.

Although DVSO [7] achieves better results on some metrics, it uses a ground truth disparity map for training that is not always available. In contrast, our PDENet can achieve similar or better results without the need of a ground truth disparity map. When compared to the newest hybrid visual odometry method DFVO [9], which is based on sparse pose estimation solving a perspective-n-point problem, the proposed method achieve better results since we use a more accurate direct approach exploiting all the possible information in the image. In addition, the DFVO approach is much more complex since it uses two networks to generate both depth map and optical flow for pose estimation.

Table IV shows the absolute trajectory error ATE on the

KITTI Odometry dataset. Note that not all works in the literature provide results using these metrics, therefore the list of methods is different from the one in Table III. Again, even using this different metric, the proposed approach compares favorably relative to existing methods in the literature.

Figure 3 shows the estimated trajectory of our method quite close to the ground truth.

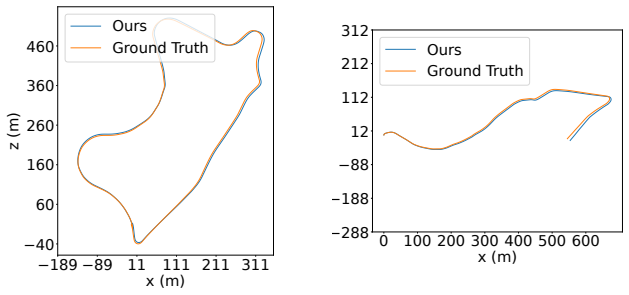


Fig. 3. The estimated trajectories for sequences 09 and 10 on the KITTI Odometry dataset.

IV. CONCLUSION

In this paper, we proposed a new approach for hybrid visual odometry combining a supervised artificial neural network for depth estimation with dense model-based stereo visual odometry. The main advantage of the proposed approach is that dense ground truth depths are not needed to train the neural network. We use instead ground truth poses that are much easier to obtain. Encouraging experimental results have been obtained on benchmark datasets. The proposed approach can outperform most of the previous methods or can obtain similar results while being much easier to set up. Future work will be dedicated to investigate improvements in each module and to assess the effectiveness of the approach on more benchmarks and on a robotic system.

ACKNOWLEDGMENTS

This work was supported by 3IA institute at Sophia Antipolis, France. The results have been obtained using OPAL computing cluster of INRIA and Université Côte d'Azur.

REFERENCES

[1] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE robotics & automation magazine*, vol. 18, no. 4, pp. 80–92, 2011.
[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *TRO*, vol. 31, no. 5, pp. 1147–1163, 2015.
[3] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.
[4] A. I. Comport, E. Malis, and P. Rives, "Real-time quadrifocal visual odometry," *IJRR*, vol. 29, no. 2-3, pp. 245–266, 2010.
[5] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *ICRA*. IEEE, 2017, pp. 2043–2050.
[6] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*. IEEE, 2017, pp. 1851–1858.
[7] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *ECCV*. Springer, 2018, pp. 817–833.
[8] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *CVPR*. IEEE, 2020, pp. 1281–1292.
[9] H. Zhan, C. S. Weerasekera, J.-W. Bian, and I. Reid, "Visual odometry revisited: What should be learnt?" in *ICRA*. IEEE, 2020, pp. 4203–4210.
[10] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*. IEEE, 2017, pp. 270–279.
[11] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*. IEEE, 2019, pp. 3828–3838.
[12] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *CVPR*. IEEE, 2018, pp. 5410–5418.
[13] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *CVPR*, vol. 2. IEEE, 2005, pp. 807–814.
[14] K. Yamaguchi, D. McAllester, and R. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *ECCV*. Springer, 2014, pp. 756–771.
[15] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*. IEEE, 2016, pp. 4040–4048.
[16] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *ICCV*. IEEE, 2017, pp. 66–75.
[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[18] P. J. Huber, *Robust statistics*. John Wiley & Sons, 2004, vol. 523.
[19] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *NeurIPS*, vol. 27, pp. 2366–2374, 2014.
[20] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *TPAMI*, vol. 38, no. 10, pp. 2024–2039, 2015.
[21] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *CVPR*. IEEE, 2018, pp. 1983–1992.
[22] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *CVPR*. IEEE, 2019, pp. 12 240–12 249.
[23] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, "Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2624–2641, 2019.
[24] C. Ling, X. Zhang, and H. Chen, "Unsupervised monocular depth estimation using attention and multi-warp reconstruction," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.
[25] X. Ye, X. Fan, M. Zhang, R. Xu, and W. Zhong, "Unsupervised monocular depth estimation via recursive stereo distillation," *IEEE Transactions on Image Processing*, vol. 30, pp. 4492–4504, 2021.
[26] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*. IEEE, 2012, pp. 3354–3361.
[27] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," *arXiv preprint arXiv:2001.10773*, 2020.
[28] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," *arXiv preprint arXiv:1709.00930*, 2017.
[29] R. Wang, S. M. Pizer, and J.-M. Frahm, "Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth," in *CVPR*. IEEE, 2019, pp. 5555–5564.
[30] Y. Li, Y. Ushiku, and T. Harada, "Pose graph optimization for unsupervised monocular visual odometry," in *ICRA*. IEEE, 2019, pp. 5439–5445.
[31] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Transformer guided geometry model for flow-based unsupervised visual odometry," *Neural Computing and Applications*, pp. 1–12, 2021.
[32] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu, "Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos," in *CVPR*. IEEE, 2019, pp. 8071–8081.
[33] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *CVPR*. IEEE, 2018, pp. 5667–5675.