Kuwait J.Sci., Vol.49, No.(3),July.2022,pp(1-23)

Monocular weakly supervised depth and pose estimation method based on multi-information fusion

Zhimin Zhang, Jianzhong Qiao*, Shukuan Lin
Dept. of Computer Science and Engineering, Northeastern University, China

**Corresponding author: qiaojz_neu@163.com*

Abstract

The depth and pose information are the basic issues in the field of robotics, autonomous driving, and virtual reality, and are also the focus and difficult issues of computer vision research. The supervised monocular depth and pose estimation learning are not feasible in environments where labeled data is not abundant. Self-supervised monocular video methods can learn effectively only by applying photometric constraints without expensive ground true depth label constraints, which results in an inefficient training process and suboptimal estimation accuracy. To solve these problems, a monocular weakly supervised depth and pose estimation method based on multi-information fusion is proposed in this paper. First, we design a high-precision stereo vision method to generate a depth and pose data as the "Ground Truth" labels to solve the problem that the ground truth labels are difficult to obtain. Then, we construct a multi-information fusion network model based on the "Ground truth" labels, video sequence, and IMU information to improve the estimation accuracy. Finally, we design the loss function of supervised cues based on "Ground Truth" labels cues and self-supervised cues to optimize our model. In the testing phase, the network model can separately output high-precision depth and pose data from a monocular video sequence. The resulting model outperforms mainstream monocular depth and poses estimation methods as well as the partial stereo matching method in the challenging KITTI dataset by only using a small number of real training data(200 pairs)

Keywords: Depth-pose estimation; "Ground Truth" labels; inertial measurement unit; multi-information fusion; weakly supervised learning

1. Introduction

In recent decades, science researchers have gradually shifted from the cognition of the computer's two-dimensional plane image to the computer's processing of the real three-dimensional world in the objective scene. How to reconstruct the three-dimensional information of the scene from single or multiple images, that is, image depth estimation is a very important basic topic in the current computer vision research. With the in-depth research in the field of computer vision, three-dimensional information such as the depth of images has gradually been applied to the fields of intelligent robots, intelligent medical care, unmanned driving, target detection, and tracking, face recognition, 3D video production, and object detection which has great social value and economic value (Albarqouni *et al.*, 2016; Chen *et al.*, 2017; Cunha *et al.*, 2011; Cui *et al.*, 2018; Fang *et al.*, 2002; Feng *et al.*, 2017; Kao *et al.*, 2016; Shotton *et al.*, 2011; Sielhorst *et al.*, 2006; Zhou & Koltun, 2014; Xie *et al.*, 2016; Zhou *et al.*, 2021). Generally, the acquisition methods of depth information are mainly tackled with two types of technical methodologies namely active depth acquisition and passive depth acquisition. Active methods mainly include laser scanning (Biber *et al.*, 2004), TOF (Time of Flight) camera (Foix *et al.*, 2011; Zhu *et al.*, 2010), and structured light camera (Han *et al.*, 2013; Scharstein C Szeliski, 2003). However, these depth sensor devices often have certain limitations. For example, lidar is very expensive and the depth information collected is sparse and uneven (Scharstein & Szeliski, n.d.), while structured light and TOF cameras are subject to light (indoor only) and distance limitations (below 5m). While camera sensor is widely concerned because of its low cost, simple hardware setting, and long shooting distance.

The current image-based depth estimation methods are mainly tackled with two types of technical methodologies: stereo matching (Ladicky *et al.*, 2014; Wu *et al.*, 2011; Luo *et al.*, 2016) and monocular vision(Saxena *et al.*, 2008; Chen *et al.*, 2016).

Since the stereo matching methods have high accuracy and a few assumptions about imaging equipment, it is currently the most widely used depth estimation algorithm. Recent advances in learning based methods (Kendall *et al.*, 2017; Chang & Chen, 2018; Song *et al.*, 2020; Zhang, Chen, Bai, Yu, Yu, Li & Yang, 2020; Shen *et al.*, 2021) show that the estimation accuracy can be significantly improved by deep models trained with pre-trained data and finetuned on another dataset with a limited amount of ground truth data. However, the binocular methods need more expensive special stereo camera equipment and are prone to camera calibration error and synchronization problems. It is easy to be limited by the camera volume during the deployment process. In practical application, the monocular camera is more popular, but the accuracy of traditional monocular depth estimation is limited due to its ill-posed and geometrically ambiguous problem. It is a very challenging problem to estimate the depth of information of the scene in only one image.

In recent years, deep learning technology has shown remarkable advantages in various fields, including natural language processing(Khan *et al.*, 2016), image processing(Minaee *et al.*, 2021), and computer vision(Al-Hmouz, 2020; Gao *et al.*, 2018), etc. Researchers have begun to use learning-based methods to solve the ill-posed monocular depth estimation problem(Liu *et al.*, 2015, 2016). The current mainstream monocular depth estimation methods based on supervised and self-supervised learning generally have some problems. First of all, the supervised method to solve this depth prediction problem almost entirely relies on the semantic information of a single image, which direct matches it with the ground truth depth. However, it is difficult and impractical to obtain a large amount of high-quality ground truth depth data corresponding to the input image. Self-supervised methods (Zhan *et al.*, 2018; Yang *et al.*, 2017) usually train a deep network model to find the dense correspondence disparity field and then warps the source view into the target view to form the image alignment constraints by using the image wrap technique "spatial transformer" (Jaderberg *et al.*, 2015). However, these learning methods (Garg *et al.*, 2016; Godard *et al.*, 2017) rely on a large amount of high-quality data and effective learning that are often ill-posed and geometrically ambiguous without ground truth labels in theory, so the result is usually suboptimal. Currently, the SFMlearner models (Zhou *et al.*, 2017; Godard *et al.*, 2019) are based on the traditional structure of the geometric principle of motion (SFM). This method not only follows the matching principle of stereo vision but also a single-view estimation model. Although the self-supervised SFMlearner model can achieve better results, it is only from photometric and temporal consistency between consecutive frames in monocular videos, which are prone to overly smoothed depth map estimations. We need more auxiliary constraint information to achieve higher precision depth and pose estimation.

To solve this problem, we propose a novel weakly supervised monocular depth and pose estimation model based on a multi-information fusion, which is illustrated as shown in Figure 1. Current learning-based stereo matching uses the insights of decades of multi-view geometry research(Hernandez *et al.*, 2008) to guide modeling, rather than constructing a black box model(Cheng *et al.*, 2018; Zhang, Chen, Bai, Yu, Yu, Li & Yang, 2020). This allows the network to learn the entire model end to end while leveraging our geometric knowledge of the stereo problem. What's more, existing learning-based stereo matching models(Zhang *et al.*, 2019; Zhang, Qi, Yang, Prisacariu, Wah & Torr, 2020) can effectively perform migration training by being pre-trained in the synthetic dataset(Mayer *et al.*, 2016) and finetuning in the application dataset with limited ground truth depth data. Inspired by these ideas, we design a "Ground Truth" labels module based on a stereo matching model and pose method to generate a depth and pose data offline as weakly supervised labels. At the same time, based on the existing self-supervised Depth and pose estimation model(Godard *et al.*, 2019), we have improved the network model and added an inertial measurement unit data pose estimation module and introduced a multi-sensor information fusion module to synchronize motion information from visual perception and inertial data. Unlike SFM-Learner (Zhou *et al.*, 2017), which only uses the temporal image loss of continuous monocular images, in the training phase, we also use the left and right stereo pairs of consecutive

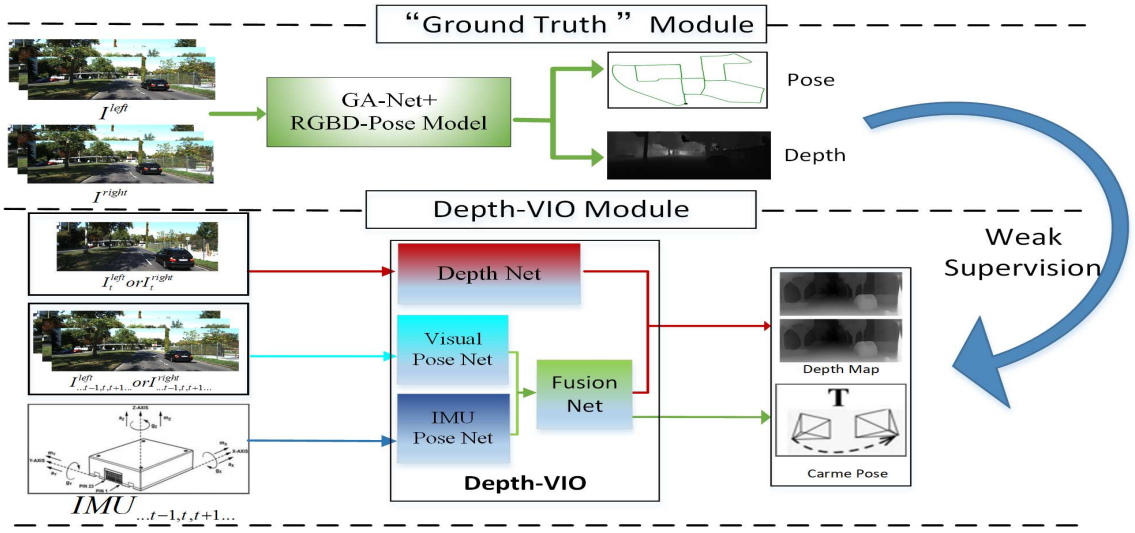


Fig. 1. An overall overview of the proposed method. The overview is divided into two parts, namely the "Ground Truth" module, Depth VIO module. First, the "Ground Truth" module can generate "Ground Truth" labels as the weak supervision constraint of and the Depth-VIO network. Then, we input unlabeled left or right continuous multi-frame view data and IMU data into the proposed the Depth_VIO model to estimate depth information and pose information. In the testing phase, we can estimate the depth map or the pose data by inputting a single frame of data into the deep network or continuous frame data into the pose network.

multiple frames to construct the spatial image loss of the stereo image pairs and weakly supervised constraint to optimize the model. In the testing phase, we only use continuous monocular images as input data to estimate the scene depth and camera pose. We train and verify the correctness and advancement of the proposed method on the challenging autonomous driving data set KITTI(Geiger, 2012). Through the data fusion training between the end-to-end visual pose and the pose of the inertial measurement unit, it is possible to eliminate the frequent manual synchronization and calibration of the time stamp between the camera and the IMU when processing the KITTI data set in the traditional method. Moreover, the "Ground Truth" labels are used to pre-train the visual-inertial odometry(VIO) network, which effectively solves the over-fitting problem in the training process of the CNN-LSTM hybrid network. In summary, our contributions are as follows:

(1) This paper proposes a weakly supervised monocular depth and poses an estimation method based on multi-information fusion for improving estimation accuracy.

(2) We designed a label generation model based on small sample data, which can generate "Ground Truth" labels to solve the problem of difficult access to ground truth data.

(3) The paper designs a multi-network fusion model including a depth estimation network, visual pose estimation network, IMU pre-integration network, and fusion network.

(4) This paper combines the cost of weak supervision and self-supervision to construct a new weak supervision joint optimization loss function and adopts a stepwise optimization method to solve the problem that the CNN-LSTM network is difficult to train end-to-end.

The rest of the paper is organized as follows. In Sec. 2, the related work on depth estimation. In Sec. 3, the problem sets, network model, and loss function of the proposed method are introduced. In Sec. 4, the algorithm flow, experimental process, and test results of the proposed method are displayed. In Sec.5, the conclusion.

2. Related work

2.1 Stereo matching

Depth estimation from stereo matching obeys geometric principles and has little assumption, so it is the most widely applicable technique in practical applications. Traditional stereo matching methods are usually divided into four steps by Scharstein and Szeliski (Scharstein & Szeliski, 2002): matching cost, cost aggregation, disparity calculation or optimization, and disparity refinement. In these four steps, cost aggregation is vital for eliminating ambiguity and mismatching, so many cost aggregation methods (Yang, 2012; Hosni *et al.*, 2012; Mei *et al.*, 2013) have been modified to refine the cost volume and achieve better estimates. In recent years, stereo matching models based on deep neural networks, especially the current end-to-end deep learning, have become very popular. In (Pang *et al.*, 2017), authors proposed a different cascade of residual convolutional neural network architecture composed of two stages to tackle the problem that affects the estimation accuracy in ill-posed regions. In (Kendall *et al.*, 2017), the authors adopted a novel regressing stereo disparity model (GC-net) that incorporated contextual information by 3D convolutions over the cost volume, which was formed based on the problem's geometry, and then used a differentiable argmin function to regress disparity values. In (Chang & Chen, 2018), the paper used a pyramid stereo matching network (PSM-Net) that included a spatial pyramid pooling module and 3D CNN modules to resolve the correspondence in ill-posed regions. In (Guo *et al.*, 2019), the authors proposed a group correlation cost volume construction method to improve the estimation accuracy. In (Cheng *et al.*, 2020), authors proposed the end-to-end hierarchical framework by incorporating the gold standard pipeline for deep stereo matching into the neural architecture search framework. In (Zhang, Chen, Bai, Yu, Yu, Li & Yang, 2020), the paper was based on the PSM-Net model to add constraints by filtering the unimodal distribution peak of the cost volume at the true disparities, thereby improving performance. To reduce the consumption of computational resources by 3D con-volution, researchers have gradually used some modules to reduce or replace 3D convolution. GA-net (Zhang *et al.*, 2019) adds two new network layers, introducing semi-global cost aggregation and local cost aggregation into deep learning, making deep learning follow the traditional cost filtering strategy to refine the network structure, and can replace the 3D convolutional layer with high computing cost and memory consumption. In (Zhang, Qi, Yang, Prisacariu, Wah & Torr, 2020), the model was developed based on GA-Net by proposing a novel domain normalization approach and a trainable non-local graph-based filter for further performance improvement. In (Wang *et al.*, 2019), the authors proposed a multi-resolution parallax real-time stereo matching method based on a U-Net network. In (Tankovich *et al.*, 2021; Yee and Chakrabarti, 2020), papers followed this design idea to improve the DCNN inference speed. In (Wang *et al.*, 2021), the models employ a recurrent unit to iteratively update disparity estimations at high resolution for achieving a trade-off between accuracy and efficiency. To reduce the dependence on ground truth disparity labels, few kinds of literature have proposed unsupervised stereo matching methods. In (Zhong *et al.*, 2017), authors designed a self-supervised deep network based on GC-net (Kendall *et al.*, 2017) with image warping error as the loss function to compute dense disparity maps directly. Reference(Chao *et al.*, 2017) presented a deep unsupervised stereo matching framework to learn cost-volume with iteratively updating network parameters and guide the training by the left-right check. Then appropriate matches were selected as training data in future iterations. In (Zhang *et al.*, 2018), the authors proposed the self-supervised active stereo matching method based on deep learning for the first time, which combines the cues of active lighting and passive light to improve the accuracy of depth prediction further. In (Liu *et al.*, 2020), the model leverage the geometric constraints about stereo video sequences to perform disparity and optical flow estimation.

2.2 Monocular depth estimation

Due to volume, calibration errors, and synchronization problems that may exist in the deployment and the setting of binocular cameras, monocular cameras are still much more preferred in most scenarios. Therefore, a large number of researchers have currently devoted themselves to monocular depth estimation, and some research results have been achieved in the literature.

Traditional geometry-based methods mainly rely on correspondences search(Bian *et al.*, 2017), model

fitting(Bian *et al.*, 2019), and multiview triangulation(Hernandez *et al.*, 2008). Therefore, the model needs at least two views from different perspectives in the same scene as input to calculate depth. Before the emergence of learning-based methods, estimating depth from a single perspective was an inherent geometric ambiguity and an ill-posed problem. Generally, learning-based methods can be divided into two categories according to the presence or absence of ground truth labels: supervised depth estimation, self-supervised depth estimation. The supervised depth estimation address this issue by utilizing the relationship between the input image and the corresponding ground truth depth data to fit the predictive model. References(Eigen *et al.*, 2014; Eigen & Fergus, 2015) first used two convolutional neural networks(CNN) to integrate global and local information. In (Laina *et al.*, 2016), authors proposed an end-to-end deeper encoder-decoder hourglass network architecture based fully convolutional network(FCN)(Long *et al.*, 2015) and used the reverse Huber loss to optimize the network model. Subsequently, a lot of related works(Laina *et al.*, 2016; Ummenhofer *et al.*, 2017; Fu *et al.*, 2018; Tang & Tan, 2018; Yin *et al.*, 2019; Garg *et al.*, 2019; Huynh *et al.*, 2020; de Queiroz Mendes *et al.*, 2021) began to appear. Supervised monocular depth estimation needs vast and expensive ground truth depth data, and the measure of depth maps is much sparser.

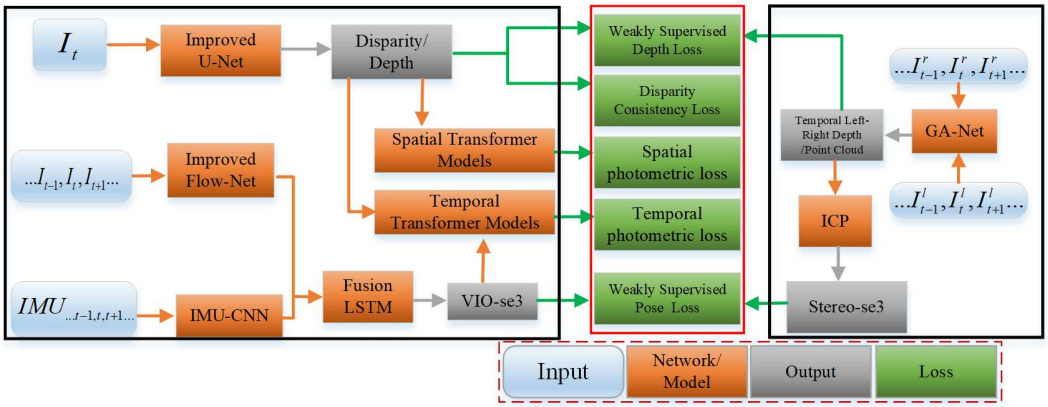


Fig. 2. Illustration of our model architecture. The model architecture mainly includes three parts: the left black box is the proposed Depth_VIO model; the right black box is the "Ground Truth" label generation model, and the middle red box is the loss function constructed by the left and right black boxes.

An alternative to the supervised depth estimation method is a self-supervised method that poses depth prediction as a view synthesis problem. In these methods, the loss of image reconstruction is taken as the primary model constraint to optimize the network. In (Xie *et al.*, 2016), model trained a deep convolutional network to obtain a probability disparity-like map from a single left view for reconstructing the right view by using the image-based rendering algorithm. Reference(Garg *et al.*, 2016) end-to-end trained the deep network model to predict the disparity map for synthesizing the right view. Besides, the loss function is linearized by the Taylor expansion method to make it fully differentiable. In (Godard *et al.*, 2017), authors used the same depth estimation idea but introduced a more sophisticated image reconstruction loss and left-right disparity consistency loss. To achieve a more accurate dense correspondence depth estimation, reference(Kuznietsov *et al.*, 2017) proposed a semi-supervised depth estimation method, which takes advantage of self-supervised dense matching and supervised unambiguous depth estimation. To make the training process respect the geometric principles, reference(Luo *et al.*, 2018) decomposed the monocular depth estimation into a self-supervised view synthesis process and the supervised stereo matching process. Based on the traditional theory(Geiger *et al.*, 2011; Leutenegger *et al.*, 2015; Mur-Artal & Tardós, 2017) of struct from motion or multi-view method, the literatures (Zhou *et al.*, 2017; Vijayanarasimhan *et al.*, 2017; Mahjourian *et al.*, 2018; Yang *et al.*, 2018; Zhou *et al.*, 2019; Li *et al.*, 2020; Poggi *et al.*, 2020; Guizilini *et al.*, 2020) adopt two CNN networks to end-to-end learning depth and pose separately by inputting image sequence at the same time. The pose network provides

the relative camera transformation for the image warp technique to synthesis adjacent images sequence. Following this idea, researchers (Zhan *et al.*, 2018; Li *et al.*, 2018) combined spatial image losses of a stereo image pair and temporal image losses of consecutive monocular Images for further performance improvement. Since moving objects violate the assumption of a static world in-depth estimation, reference(Zhou *et al.*, 2017; Godard *et al.*, 2019) used a deep network to learn a mask to mask out the moving objects. Many subsequent methods either add optical flow networks(Zhichao & Jianping, 2018; Chen *et al.*, 2019) or leverage semantic information(Casser *et al.*, 2019; Huynh *et al.*, 2020; Lee *et al.*, 2021) to detect moving objects. These methods require additional complex network training and related labels. CNN-SVO(Luo *et al.*, 2019) and D3VO (Yang *et al.*, 2020) also train depth and pose networks on the calibrated stereo videos by improving network and use geometric loss optimization to further improve performance. In (Bian *et al.*, 2021), authors proposed a geometry consistency loss to penalizes the in-consistency of predicted depth and proposed a self-discovered mask to automatically localize moving objects. Most of the pose networks of the above methods adopt the PoseNet model(Kendall *et al.*, 2015). Recent studies(Wang *et al.*, 2017; Almalioglu *et al.*, 2019; Zhan *et al.*, 2020) on visual mileage show that recursive convolutional neural networks (RCNNs) are more accurate in estimating camera pose. In addition, performing data fusion(Clark *et al.*, 2017; Shamwell *et al.*, 2018; Li & Waslander, 2020) of visual odometry and inertial measurement unit data at the intermediate feature representation level can further improve the estimation accuracy.

The above-mentioned models(Almalioglu *et al.*, 2019; Zhan *et al.*, 2020; Li & Waslander, 2020) based on monocular visual odometry or fused inertial measurement unit data usually require ground truth pose data as a supervision label. The self-supervised visual odometry method combined with depth estimation only optimizes the pose and depth network through the loss of photometric consistency, which often makes the result sub-optimal. In this paper, a high-precision stereo vision odometry method is used to generate "Ground Truth" labels. At the same time, a multi-information fusion model based on vision and inertial measurement unit data is constructed to improve the estimation accuracy.

3. Method

Here we describe our weakly supervised monocular depth estimation method for multi-information fusion. We first i ntroduce t he i mplementation p rocess o f t his m ethod, t hen d escribe t he " Ground Truth" labels model, and finally, d escribe t he n etwork a nd l oss f unction o f t he w eakly s upervised monocular depth and pose estimation model of multi-information fusion.

3.1 Problem Setup

The proposed multi-information fusion weakly supervised monocular depth and pose estimation model architecture is shown in Figure 2. We denote the left and right training image sequences as $[I_1^l, ..., I_k^l]$ and $[I_1^r, ..., I_k^r]$ respectively, where the subsequent representation I_t is a target left or right image, and I_t^s is a target stereo pair. The upper part is that we build a "Ground Truth" label generation module based on the Guided Aggregation network(GA-Net)(Zhang *et al.*, 2019) and the well-known Iterative Closest Point (ICP)(Yang *et al.*, 2015) algorithm. Input the left and right continuous video frames I_t^s to the model to generate depth map$[D_t^l, D_t^r]$ and 6-DOF Lie algebra pose data $[\rho_t, \phi_t]$ as "Ground Truth" labels. The lower part is our multi-information fusion monocular weakly supervised depth and poses estimation model(depth-VIO). From the Figure 2, we can see that the model consists of depth estimation network, visual pose network, Imu pose network, and fusion network. We can input left or right view I_t to depth network and input left or right continuous video frames $[..., I_{t-1}, I_t, I_{t+1}, ...]$ to visual pose network. At the same time, the inertial measurement unit(IMU) data($[..., IMU_{t-1}, IMU_t, IMU_{t+1}, ...]$) is input into the IMU pose network. Finally, the depth network outputs left and right disparity maps $[\hat{D}_t^l, \hat{D}_t^r]$. The fusion LSTM is used to integrate the visual pose feature and IMU pose feature to produce the final 6-DoF relative pose(VIO-se3) $[..., \hat{\varepsilon}_{t-1}, \hat{\varepsilon}_t, \hat{\varepsilon}_{t+1}), ...$ In the training phase, we can construct three types of losses to optimize the model: weak supervision losses of the "Ground Truth" label, spatial image losses of a stereo image pair, and temporal image losses of consecutive monocular images. The loss function will be discussed in detail in section 2.4. In the testing phase, we can estimate the depth and camera pose separately by inputting a monocular video sequence.

3.2 "Ground Truth" Labels Model

Traditional semi-global matching (SGM)(Hirschmuller, 2007) and cost filtering(Hosni *et al.*, 2012) are all robust and efficient c ost a ggregation m ethods t hat h ave b een w idely u sed i n m any industrial products. GA-Net(Zhang *et al.*, 2019) introduces a semi-global guided aggregation layer (SGA) which implements a differentiable approximation of semi-global matching (SGM) and aggregates the matching cost in different directions over the whole image. This makes the learning-based stereo matching method more follow the traditional geometric process, improve the accuracy of disparity estimation. The bad pixels average percentage of 200 test images in the KITTI2015 dataset is shown in Table 1. According to the benchmark test(Menze and Geiger, 2015), we believe that the disparity error is less than 3 pixels, then the estimation of this pixel is correct. We have compared the performance on four evaluation indexes of "D1-bg", "D1-fg", "D1-all", and "Runtimes". Therefore, we adopt the GA-Net(Zhang *et al.*, 2019) network model to estimate the left and right disparity from left-right stereo pairs, thereby generating depth maps and 3D point cloud data. Then, we calculate the relative pose(Stereo-se3) of 6-DoF from the point cloud sequence by the well-known Iterative Closest Point (ICP) method (Yang *et al.*, 2015). Here we will not introduce the network model of GA-Net, but mainly introduce the generation of the subsequent depth map, point cloud and pose data.

Table 1. The bad pixels average percentage. Among them, "D1" refers to the percentage of outliers in the first frame of stereo disparity, "bg" refers to the average percentage of outliers only in the background area, "fg" refers to the average percentage of outliers only in the foreground area and "all" refers to the average percentage of outliers only in the ground truth pixels.

Error	D1-bg(%)	D1-fg(%)	D1-all(%)
All / All	1.48	3.46	1.81
All / Est	1.48	3.46	1.81
Noc / All	1.34	3.11	1.63
Noc / Est	1.34	3.11	1.63

Depth Labels: GA-Net output the left disparity map D_t^l and right disparity map D_t^r of the target view from stereo pairs. According to the epipolar geometry theorem, we can calculate the depth Z_t of the scene, as shown in Eq. 1.

$$Z_t = \frac{f * b}{D_t^l} = \frac{f * b}{x_l - x_r} \tag{1}$$

where f denotes the focal length of the binocular camera, and B denotes the baseline of the binocular camera. Ideally, we can also get the same depth map Z_t from the estimated right disparity D_t^r.

At the same time, we also show the qualitative results on Figure3 about "Ground Truth" labels.

Pose Labels: We use the traditional ICP(Iterative Closest Point) algorithm to obtain pose labels. First, we need to calculate the 3D point cloud data c based on the obtained depth map D_t^l or D_t^r. To reduce the depth estimation error from the GA-Net network, we set the depth value range to 0-80m. If the estimated depth value is lower than the lowest value or higher than the highest value, we set the value to the nearest or farthest depth value. The 3D point cloud c_t of the view at time t is calculated as Eq.2.

$$c_t(x_c, y_c, z_c) = \mathbf{K}^{-1} Z_t [x_u, y_v]^T \tag{2}$$

where K is the intrinsic parameter of the left camera.

Then we can use the ICP algorithm to obtain the transformation matrix $SE(3)$ from $t-1$ frame to t frame based on the calculated 3D point cloud c_t. $SE(3)$ contains the rotation matrix $\mathbf{R} \in SO(3)$ and the transformation vector $\mathbf{t} \in \mathbb{R}^3$.

$$\begin{aligned} SE(3) &= ICP(c_{t-1}, c_t) \\ &= \left\{ T = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{4\times4} \right\} \end{aligned} \tag{3}$$

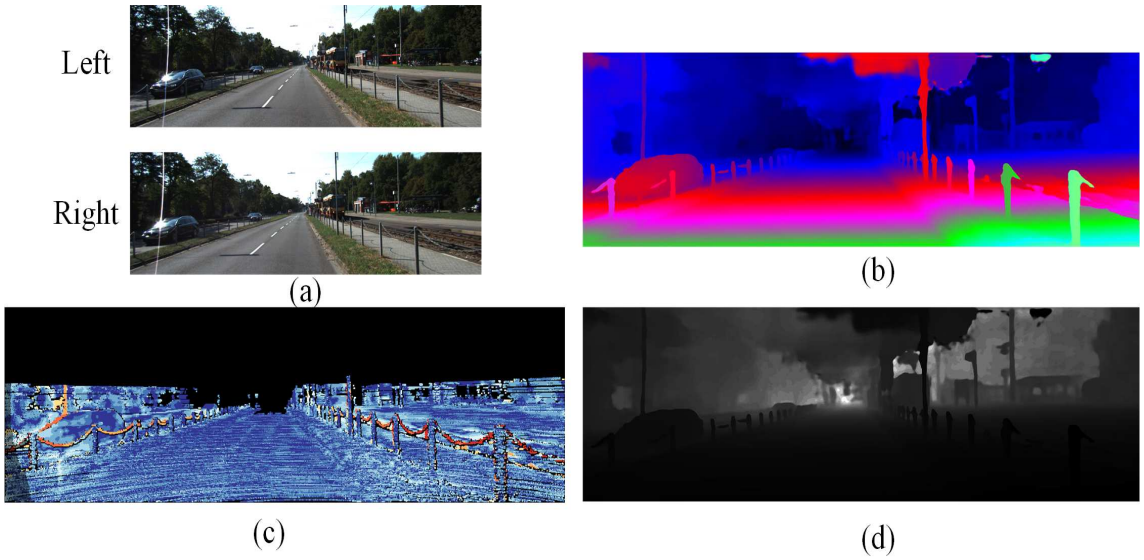


Fig. 3. The qualitative results of "Ground Truth" depth. Above, (a) is the left or right input images, (b) is the estimated result, (c) is the error map, and (d) is the depth map.

$$SO(3) = \{\mathbf{R} \in \mathbb{R}^{3\times3} | \mathbf{R}\mathbf{R}^T = \mathbf{I}, det(\mathbf{R} = 1)\} \tag{4}$$

However, for a 3-DOF rotation, the expression of $SO(3)$ in nine quantities is too redundant and too restrictive, so we convert the transformation matrix T into Lie Algebra $se(3)$, which includes a three-dimensional rotation vector ϕ and a three-dimensional translation vector ρ .

$$\begin{aligned} se(3) &= log(T) \\ &= \left\{ \varepsilon = \begin{bmatrix} \rho \\ \phi \end{bmatrix} \in \mathbb{R}^6, \rho \in \mathbb{R}^3, \phi \in \mathbb{R}^3 \right\} \end{aligned} \tag{5}$$

3.3 Depth_VIO Network

Our Depth_VIO network builds upon MonoDepth2(Godard *et al.*, 2019) and extends it by improving depth estimation network, replacing visual pose model using CNN-RNN network and introducing inertial measurement unit (IMU) model. This section mainly introduces the Depth-VIO network architecture. The model mainly includes two networks, namely depth estimation network, visual inertial odometry network(VIO) pose network.

3.3.1 Depth Estimation Network

The depth estimation network is shown in Figure 5 that estimates a depth map from a single target RGB view. With reference to the performance of the encoding-decoding network structure(Garg *et al.*, 2016; Godard *et al.*, 2017, 2019) used in the current monocular depth estimation, our depth estimation network is built upon by improving the U-Net network(Ronneberger *et al.*, 2015). We adopt Resnet(He *et al.*, 2015) as our encoder and the decoding network uses bilinear interpolation for upsampling. In the Resnet network, we use three small convolutional layers of 3×3 to replace the sizeable convolutional layer of 7×7 and use the convolutional layer of stride 2 to replace the pooling layer of stride 2. We take advantage of skip connections from the encoder's activation block to the decoder blocks with the same resolution to resolve higher resolution details. Inspired by the refinement module in(Dosovitskiy *et al.*, 2015), we reconnect the multi-scale disparity output after upsampling to the high-resolution features to improve the refinement accuracy.

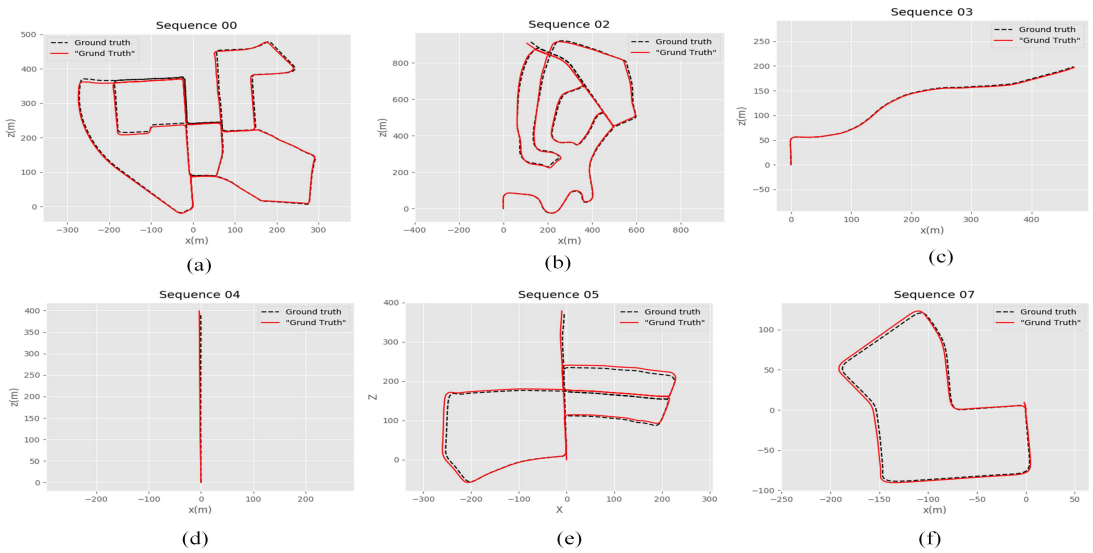


Fig. 4. Trajectories of "Ground Truth" pose on video sequences. Where, (a) is sequence 00,(b) is sequence 02, (c) is sequence 03, (c) is sequence 04, (c) is sequence 05 and (d) is sequence 07.

3.3.2 Visual Inertial Odometry Network:

Visual Odometry Network: Taking inspiration from (Dosovitskiy *et al.*, 2015), the FlowNet network without the refinement module is used as the feature extraction part of the visual odometry(VO) network of our VIO model. The configuration of the network is shown in Table 2. The RGB image frame is normalized by subtracting the mean training set and dividing by the variance, resizing to a new size as 640 × 192. A monocular image tensor sequence formed by stacking multiple consecutive sets of front and rear frame images (feature map is 6) is used as input. The conv6_1 and conv6_2 output translation and rotation feature side by side, respectively. This output feature will be used as part of the input data of the fusion network.

IMU Pre-integration Network: Usually, the sampling frequency of the inertial measurement unit is several times higher than the sampling frequency of the camera. To enable the network to learn how to implicitly estimate the time offset between the camera and IMU data, we choose the raw IMU measurement as the input sample of the network. IMU data include raw linear acceleration $\alpha \in \mathbb{R}^3$ and

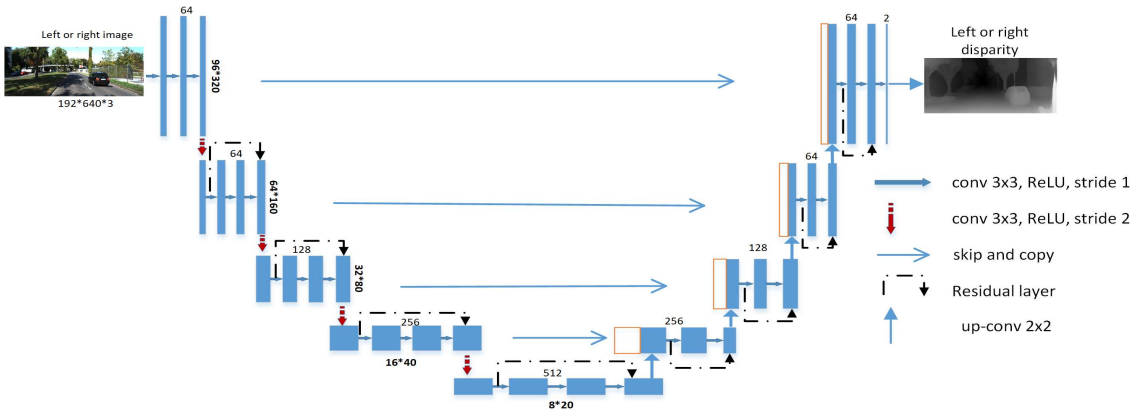


Fig. 5. The depth network architecture of our model. This model is improved on the basis of U-Net network architecture.

Table 2. The configuration of the Visual Odometry Network.

Layer	Receptive File	Stride	Feature Map	Output	Input
Input			6	640*192	
Conv1	7x7	2	64	320*96	Input
Conv2	5x5	2	128	160*48	Conv1
Conv3	5x5	2	256	80*24	Conv2
Conv3_1	3x3	1	256	80*24	Conv3
Conv4	3x3	2	512	40*12	Conv3_1
Conv4_1	3x3	1	512	40*12	Conv4
Conv5	3x3	2	512	20*6	Conv4_1
Conv5_1	3x3	1	512	20*6	Conv5
Conv6_1	3x3	2	1024	10*3	Conv5_1
Conv6_2	3x3	2	1024	10*3	Conv5_1

raw angular velocity $\omega \in \mathbb{R}^3$ from $t-1$ to $t+1$. We can obtain raw measurements $M \in \mathbb{R}^{n\times 6}$, where n is the number of IMU samples. Following the VIOlearner(Shamwell *et al.*, 2019), the IMU pre-integration network uses two parallel branches of 7 convolutional layers to extract pose information from IMU data. Each branch begins with 2 convolutional layers each of 64 single-stride with kernel size 3x5 followed by one convolutional layer of 128 filters each of stride 2 with kernel size 3x5 and one convolutional layer of 256 filters each of stride 2 with kernel size 3x5. Next, one convolutional layer of 512 filters is applied with strides of 2, 1, and 1, and kernels of size 3x5, 3x3, and 3x1. The final networks use three filters of kernel size 1 and stride 1 in the angular velocity and linear acceleration pathways. Finally, each output is 1×3 tensors and concatenated together into a tensor.

Fusion Network: The visual pose feature and IMU pose are concatenated into a tensor to feed into the fusion network, which is a two layers LSTM network. The LSTM network is followed by a fully connected layer that regresses the fused pose, which maps the features to a 6-DoF pose vector. The final outputs are a $batch \times (n-1) \times 6$(n is the length of the image sequence) tensor for translation and rotation parameters, representing the n motion of the camera between a time window $t - n \times \delta t$ to $t + n \times \delta t$, δt represents the time difference between adjacent video frames.

3.4 Loss Function

We formulate a total loss function L_θ that is composed of the weak supervision losses L_{wl}, the spatial image losses L_{sl}, and the temporal image losses L_{tl}. Each constraint term adds a trade-off parameter to limit the percentage of the cost in the loss function.

$$L_\theta = \alpha L_{wl} + \beta L_{sl} + \gamma L_{tl} \tag{6}$$

Where α, β, γ are the trade-off parameters of each loss term.

3.4.1 Weak Supervision Loss

We use "Ground Truth" labels to supervise our model, thereby constructing a weakly supervised constraint function.

$$L_{wd} = L_1(D_t^l, \hat{D}_t^l) + L_1(D_t^r, \hat{D}_t^r) \tag{7}$$

$$L_{wp} = L_1(\rho_t, \hat{\rho}^t) + \kappa L_1(\phi_t, \hat{\phi}^t) \tag{8}$$

where L_1 is the $L1$ norm operation. L_{wd} refers to the left and right weakly supervised depth loss, and L_{wp} refers to the left or right weakly supervised pose loss. We can construct weakly supervised constraint functions $L_{wl} = \omega_1 L_{wd} + \omega_2 L_{wp}$.

3.4.2 Self-supervised Spatial Image Losses

Appearance Matching Loss: The view reconstruction loss is an image alignment error between original stereo pairs and synthesized stereo pairs. Taking inspiration from the loss function(Zhao *et al.*, 2015) , we use a combination loss of an $L1$ loss L_1 and a structural similarity($SSIM$) (Wang *et al.*, 2004) loss L_{ssim} as the view reconstructed loss L_{ia}.

$$\begin{aligned} L_{ia}^{l} = L_{ssim}(I_t^l, \hat{I}_t^l) + (1-\rho)L_1(I_t^l, \hat{I}_t^l) \\ +(I_t^r, \hat{I}_t^r) + (1-\rho)L_1(I_t^r, \hat{I}_t^r) \end{aligned} \tag{9}$$

Among them, ρ is a proportional coefficient, and the value is 0.85.

b. Left-Right Disparity Consistency Loss

The left-right disparity maps can transform each other according to the translation of geometric relations. We define the consistency loss by taking advantage of this characteristic of the disparity map to improve the prediction accuracy. Inspired by the consistency loss (Godard *et al.*, 2017), we define the loss formula L_{dc} using L_1 penalty:

$$\begin{aligned} L_{dc} = L_1(\hat{D}_t^l(x), \hat{D}_t^r(x + \hat{D}_t^l(x))) \\ +L_1(\hat{D}_t^r(x), \hat{D}_t^l(x + \hat{D}_t^r(x))) \end{aligned} \tag{10}$$

Where x denotes the pixel position of left or right disparity maps.

We can construct the self-supervised spatial image loss $L_{sl} = s_1 L_{ia} + s_2 L_{wd}$.

3.4.3 Self-supervised Temporal Image Losses

We can project each pixel coordinate p_t of the target view I_l^t onto the $\hat{p}_s$ in the source view I_s using Eq. 12. Then we use the differentiable bilinear sampling mechanism proposed in the spatial transformer network(Jaderberg *et al.*, 2015) to calculate the value of point $\hat{p}_s$ by using the 4-pixel neighborhood of $\hat{p}_s$ ((0,0),(0,1),(1,0),(1,1)). Then the calculated value is the warped image $\bar{I}^t$ pixel at location p_t. The weighting scale for bilinear sampling is $\sum_{i,j} \omega_{ij} = 1$

$$\hat{p}_s \cong K \hat{T}_{t \to s} \hat{D}_t(p_t) K^{-1} p_t \tag{11}$$

$$\bar{I}^t(p_t) = I^s(\hat{p}_s) = \sum_{i,j} \omega_{ij} I^s(i,j) \tag{12}$$

where $i \in (0,1)$, $j \in (0,1)$, $\hat{p}_s$ is the homogeneous coordinates of a pixel in the source view, K is the camera intrinsics matrix, $\hat{T}_{t \to s}$ is the relative pose from target view to source view, p_t is the pixel coordinates of the mapped target view.

$$\begin{aligned} L_{tl} = L_{ssim}(I_t^l, \bar{I}_t^l) + (1-\rho)L_1(I_t^l, \bar{I}_t^l) \\ +L_{ssim}(I_t^l, \bar{I}_t^l) + (1-\rho)L_1(I_t^l, \bar{I}_t^l) \end{aligned} \tag{13}$$

4. Experiments

4.1 Implementation Details

We train our model on the rectified KITTI odometry dataset without using any ground truth depth and use data split as proposed by (Eigen *et al.*, 2014) to test our model for depth estimation. We use sequences 00-08 as the training sample and sequences 09-10 as the test sample for pose estimation. We divide the training samples into 17871 pairs (left and right views) of training images and 2466 pairs of verification images. In the training process, the model randomly selects a pair of views as the target images and uses the continuous view pairs with the target images as the center as the source images. Corresponding 100 Hz IMU data are collected from the KITTI raw datasets and for each target image,

Algorithm 1: "Ground Truth" module

Result: GA-Net:θ_s;ICP
Input: Stereo pairs:$[I_1,...,I_k]$
Output: Disparity pairs:$[D_1,...,D_k]$; Camera poses:$[\varepsilon_1,...,\varepsilon_k]$
Initialization: Load θ_s to GA-Net;$t=1$;
while $t \le k$ **do**
  Get GA-Net left-right disparity:$D_i = \{D_t^l, D_t^r\}$;
  Compute depth Z_t and cloud c_t from D_t;
  Get camera poses $[\rho_t, \phi_t]$ using ICP;
  $\varepsilon_t = [\rho_t, \phi_t]$
end

Algorithm 2: Monocular Weakly Supervised Depth and Pose Estimation Method Based on Multi-information Fusion

Result: Depth-CNN:θ_d;Pose-RCN:θ_p;IMU-CNN:θ_i;Fusion-Net:θ_o
Input: Left view:$[I_1^l,...,I_k^l]$
Output: Disparity pairs:$[\hat{D}_1,...,\hat{D}_k]$; Camera Poses:$[\hat{\varepsilon}_1,...,\hat{\varepsilon}_k]$
Initialization: Load pretrained model θ_p and initialize weights for VIO(First train VIO);
Initialize hyperparameters(η,α,β,γ,and so on);
for *each* $i \in \{0,...epoch\}$ **do**
  for *each* $j \in \{0,...nbatch\}$ **do**
    Get VIO-Net predictions:$\hat{\varepsilon}_j$;
    Compute forward-backward loss:L_{pl};
    Using SGD to optimize θ_{VIO};
    $\theta^*_{VIO} = \underset{argmin}{\theta_{VIO}}$;
  end
  Save $\theta_{VIO} = [\theta_p, \theta_i, \theta_o]$
end
Load pretrained model θ_p;θ_i;θ_o;
for *each* $i \in \{0,...epoch\}$ **do**
  for *each* $j \in \{0,...nbatch\}$ **do**
    Get Depth_VIO predictions:$\hat{D}_j$;$\hat{\varepsilon}_j$;
    Compute forward-backward loss:L_θ Using SGD to optimize θ;
    $\theta^* = \underset{argmin}{\theta}$
  end
  Save $\theta = [\theta_d, \theta_p, \theta_i, \theta_o]$
end

Table 3. Evaluation metrics of our model and the current mainstream depth estimation model on the KITTI dataset. Supervision refers to the way of supervision, in which D refers to using ground truth depth data as supervision, stereo refers to the supervision with stereo temporal pairs, and M refers to the common training of monocular temporal sequence

Method	Supervision	RMSE	RMSE(log)	ARD	SRD	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
		–lower is better–				–higher is better–		
Eigen *et al.*(2014) Coarse	D	6.215	0.271	0.204	1.598	0.695	0.897	0.960
Eigen *et al.*(2014) Fine	D	6.138	0.265	0.195	1.531	0.734	0.904	0.966
Liu *et al.*(2015)	D	6.471	0.273	0.201	1.584	0.680	0.898	0.967
Zhou *et al.*(2017)	M	6.709	0.270	0.183	1.595	0.734	0.902	0.959
Vid2Depth(Mahjourian *et al.*, 2018)	M	6.220	0.250	0.163	1.240	0.762	0.916	0.968
Zhou *et al.*(2019)	M	4.945	0.197	0.121	0.8370	0.853	0.955	0.982
MonoDepth2 (Godard *et al.*, 2019)	M	5.180	0.205	0.129	1.112	0.851	0.952	0.978
Li *et al.*(2020)	M	5.138	0.209	0.130	0.950	0.843	0.948	0.978
PackNet-SFM (Guizilini *et al.*, 2020)	M	**4.538**	0.186	0.107	0.802	0.889	**0.962**	0.981
Bian *et al.*(2021)	M	4.706	0.191	0.114	0.813	0.873	0.960	**0.982**
Ours-no "Ground Truth"	M	5.168	0.190	0.115	0.882	0.864	0.951	0.978
Ours	M	4.601	**0.182**	**0.105**	**0.751**	**0.890**	0.960	**0.982**
Garg *et al.*(2016)	S	5.104	0.273	0.169	1.080	0.740	0.904	0.962
MonoDepth (Godard *et al.*, 2017)	S	5.927	0.247	0.148	1.344	0.803	0.922	0.964
Zhan *et al.*(2018)	S	5.869	0.241	0.144	1.391	0.803	0.928	0.969
UndeepVO(Li *et al.*, 2018)	S+M	6.570	0.268	0.183	1.730	-	-	-
DVSO(Yang *et al.*, 2018)	S	4.442	0.187	0.097	**0.734**	**0.888**	0.958	0.980
MonoDepth2 (Godard *et al.*, 2019)	S+M	5.029	0.203	0.114	0.991	0.864	0.951	0.978
D3VO(Yang *et al.*, 2020)	S+M	4.485	0.185	0.099	0.763	0.885	0.958	0.979
Ours-no "Ground Truth"	S	5.145	0.196	0.112	0.908	0.859	0.950	0.976
Ours-no pretrain VIO	S	5.141	0.182	0.103	0.854	0.881	0.958	0.980
Ours-Res18	S	4.716	0.174	0.099	0.742	0.882	0.960	0.981
Ours-Res50	S	**4.405**	**0.171**	**0.094**	0.740	0.886	**0.961**	**0.982**

the preceding 100 ms and the following 100 ms of IMU data are combined yielding a tensor of size 20 x 6 (100ms between the source images and target).

We download the pre-trained GA-Net model from this link at `https://drive.google.com/open?id=19hVQXpcXwp7SrHgJ5Tlu7_iCYNi4Oj9u`. The "Ground Truth" labels are obtained by inputting stereo pairs into the stereo matching network model method based on transfer learning.

Since our depth_VIO network is a combination of CNN and RNN networks, it can be seen from the discussion in Literature (Wang *et al.*, 2017) that the CNN-RNN network training process is prone to overfitting. Therefore, we use a two-step training method in the training process. First, we use the weakly supervised pose loss L_{wp} to train the VIO network and then load the trained VIO model to the Depth_VIO model for full network training. The model is trained on the experiment platform that is both e5-2698v4 processors, 503 GB memory, and eight 32 GB Tesla V100 graphics cards. The 1242 × 375 resolution stereo pairs are resized into 640 × 192 resolution views for training and test data of our model. In the first stage, the Flow-CNN network can be initialized by the pre-trained weight of FlowNet (Download link: `https://drive.google.com/drive/folders/0B5EC7HMbyk3CbjFPb0RuODI3NmM`. Other network weights are initialized by the gaussian distribution with a standard deviation of 0.01. Stochastic Gradient Decent (SGD) with an RMSProp adaptive learning rate is used to update the weights of the networks. The epoch is set at 250 with 8 batch sizes and the sequence length of each batch is 7. In the second stage, the depth encoder network is initialized by the pre-trained model obtained from the ImageNet classification task (He *et al.*, 2015). We use Resnet18 or Resnet50 as an encoder for training. Stochastic gradient descent also is used to update the weights with a batch size of 8 for 50 epochs. In the optimization process, we set the default learning rate as 10^{-5} and keep the default learning rate unchanged in the first 30 steps, and then reduce it by a factor of 2 every ten steps until the end to avoid shock. The predicted disparity cap is constrained to 0.3 × the output disparity map width by using sig-

Fig. 6. Qualitative depth estimation results for different methods on the Kitti dataset. Where, (a) is the ground truth disparity map. (b) is SFMlearner(Zhou *et al.*, 2017), (c) is UndeepVO (Li *et al.*, 2018), (d) is Monodepth (Godard *et al.*, 2017), (e) is Monodepth2(Godard *et al.*, 2019), (f) is our model without the ”Ground Truth” label, (g) is our model

moid non-linearity. To adjust the effect of each loss function on the model, we set the parameter value of each loss term:$\kappa = 100, \omega_1 = 1, \omega_2 = 0.1, s_1 = 1, s_2 = 0.8$. The training mode is determined by setting the value of α, β, γ. When the supervision mode is a stereo pair, the value is $\alpha = 1, \beta = 1, \gamma = 1$, and when the supervision mode is monocular, the value is $\alpha = 1, \beta = 0, \gamma = 1$.

4.2 Algorithm

This part describes the algorithm flow of the proposed method, which is mainly divided into two parts, namely Algorithm 1 and Algorithm 2. Algorithm 1 is the process of generating "Ground Truth" depth labels. Algorithm 2 is the realization process of the proposed monocular weakly supervised depth and pose estimation method based on multi-information fusion.

4.3 Results

4.3.1 Prediction Results Analysis

We test our model on KITTI, a challenging autonomous driving dataset. Table 3 shows the estimation metrics of our method and other methods so that we can quantitatively analyze the performance of our model and other models. It can be seen that our model is better than other self-supervised monocular methods in most estimation indexes, whether it is monocular supervision or stereo pair supervision. We also tested the estimation results of our model without a weak supervision label (i.e., no "Ground Truth"). It can be seen from the results in the table that the estimation accuracy has been greatly improved after the addition of weak supervision labels. We also can see from the table that the accuracy of the pre-trained model is greatly improved compared with the model that is not pre-trained through the first step of weakly supervised pose loss. The encoder of Resnet50 has improved estimation accuracy than Resnet18.

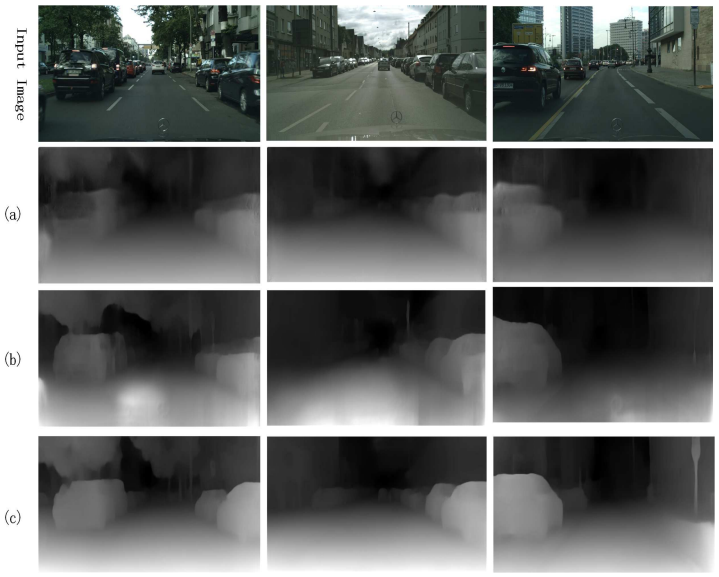


Fig. 7. Qualitative comparison of generalization for different models on Cityscapes. (a) is Zhou *et al.*(2017) , (b) is Godard *et al.*(2019) , (c) is our model trained on the KITTI dataset and tested on the Cityscape dataset.

In short, it can be seen from the table that the accuracy of the method proposed in this paper has been greatly improved, which proves the effectiveness of the method.

Figure 6 shows the visual disparity map about our model and the current mainstream monocular self-supervised methods. Perhaps unsurprisingly, the ground truth disparities obtained by the 3D scanner can provide better visual effects, but the sampled data points are sparse and sampling equipment is expensive. As we can see from the figure, although the current method (Godard *et al.*, 2017, 2019; Garg *et al.*, 2016) can obtain a better depth estimate from a single view in the scene, the proposed model can describe more clearly the details of object edge and depth information. Figure 6 also qualitatively illustrates the visual disparity map of our proposed model under self-supervised(no "Ground Truth" labels) and weakly supervised ("Ground Truth" labels) mode. The weakly supervised mode has a significant improvement in effect.

To show that our method can be generalized to other data sets, we test the comparison results of several models on the Cityscapes (Cordts *et al.*, 2016) dataset and verify that model trained on the KITTI dataset can be generalized well to the Cityscapes dataset. Figure 7 qualitatively compared the results of different models on the Cityscapes dataset. We can see from the comparison results that our model trained on the KITTI dataset shows good generalization performance on the Cityscapes dataset for accurate disparity estimation. We can use this model to predict the depth of similar scenes.

3.4 Pose results

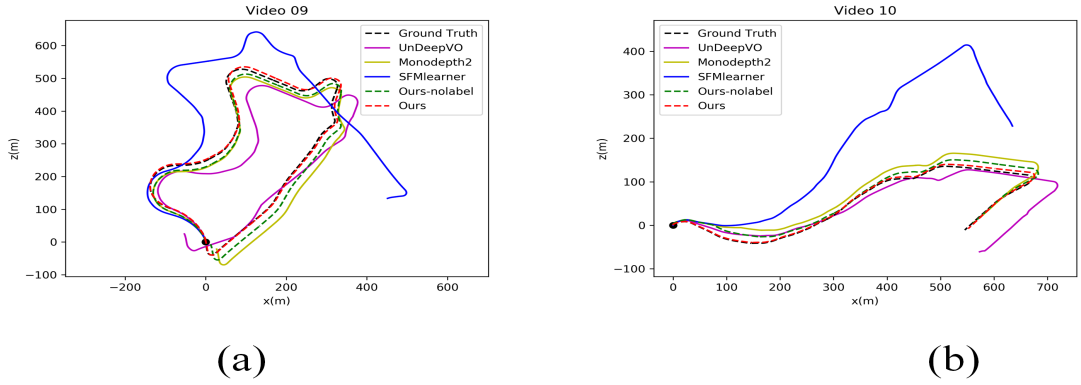
In this section, we comparatively discuss the pose estimation performance of our method in terms of both no-labels and with-labels estimation modes.

We tested the pose estimation model on sequences 09 and 10 that were not used in the training. These results are shown in Table 4. We can see from the table that compared with other self-supervised monocular depth and pose estimation methods, our model can have a good performance on the most estimated metric. After pre-training the VIO network, the accuracy of pose estimation is further improved. The metrics in Table 4 prove the effectiveness of the proposed method.

We also visualize the performance of the pose estimation model in Figure 8. For the KITTI data set, the camera in the scene basically only moves in a straight line, and the angle changes a little, so the displacement estimation is very sensitive to this scene. In the process of displacement change, the

Table 4. VO results with our proposed method and other mainstream models on Kitti sequence 09 and 10. t_{rel} refers to average translational RMSE drift (%) on length of 100m-800m. r_{rel} refer to average rotational RMSE drift (°/100m) on length of 100m-800m.

Method	Seq.09			Seq.10		
	ATE(m)	t_{err}(%)	r_{err}(°/100m)	ATE(m)	t_{err}(%)	r_{err}(°/100m)
VISO2-M(Geiger *et al.*, 2011)	52.62	18.06	1.25	57.25	26.10	3.26
OKVIS(Leutenegger *et al.*, 2015)	-	9.77	2.97	-	17.30	2.82
ORB-SLAM2-M(Mur-Artal and Tardós, 2017)	38.77	9.30	0.26	5.42	2.57	0.32
SfMLearner(Zhou *et al.*, 2017)	77.79	19.15	6.82	67.34	40.40	17.69
DeepVO-Feat(Zhan *et al.*, 2018)	52.12	11.89	3.60	24.70	12.82	3.41
UndeepVO(Li *et al.*, 2018)	-	7.01	3.60	-	10.63	4.60
MonoDepth2(Godard *et al.*, 2019)	45.22	12.17	3.85	18.35	8.68	5.31
Zhan *et al.*(2020)	10.88	2.61	**0.29**	3.72	2.29	**0.37**
Bian *et al.*(2021)	13.40	5.08	1.05	7.99	4.32	2.34
Our-no "Ground Truth"	21.40	4.11	1.94	18.99	3.96	2.06
Our-no pretrained VIO	12.52	2.03	1.82	8.56	1.98	1.62
Our	**10.45**	**1.63**	1.52	**6.35**	**1.74**	1.21



(a) (b)

Fig. 8. Sample trajectories comparison for different models on Sequence 09 and 10 of Kitti dataset.

movement in the Y direction is very small, so in our visual view, only the absolute displacement changes in the X and Z directions are displayed. From the figure, we can see that compared to other unsupervised monocular VO models, our weakly supervised multi-information fusion method can obtain more accurate pose estimation results.

5. Conclusion

In this paper, we propose a new monocular weakly supervised depth and pose estimation method based on multi-information fusion. Our model is improved on the current research and methods and is trained and tested on KITTI data. Our method uses the migration learning model of stereo matching to obtain the "Ground Truth" label with very few ground truth samples, but still obtains excellent results, which indicates that this method has certain research value. However, due to the some error between the "Ground Truth" labels and the ground truth labels, there are still some gaps between our algorithm and the fully supervised method. In the future, we will combine traditional algorithm and deep learning to make the "Ground Truth" label more obeying geometric principles and having higher precision. In addition, we will also propose more suitable models for monocular depth and pose estimation methods.

References

Al-Hmouz, R. (2020), 'Deep learning autoencoder approach: Automatic recognition of artistic arabic calligraphy types', *Kuwait Journal of Science* **47**(3).

Albarqouni, S., Konrad, U., Wang, L., Navab, N. & Demirci, S. (2016), 'Single-view x-ray depth recovery: toward a novel concept for image-guided interventions', *International journal of computer assisted radiology and surgery* **11**(6), 873–880.

Almalioglu, Y., Turan, M., Sari, A. E., Saputra, M. R. U., de Gusmão, P. P., Markham, A. & Trigoni, N. (2019) , 'Selfvio: Self-supervised deep monocular visual-inertial odometry and depth estimation', *arXiv preprint arXiv:1911.09968* .

Bian, J., Lin, W.-Y., Matsushita, Y., Yeung, S.-K., Nguyen, T.-D. & Cheng, M.-M. (2017) , Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 4181–4190.

Bian, J.-W., Wu, Y.-H., Zhao, J., Liu, Y., Zhang, L., Cheng, M.-M. & Reid, I. (2019) , 'An evaluation of feature matchers for fundamental matrix estimation', *arXiv preprint arXiv:1908.09474* .

Bian, J.-W., Zhan, H., Wang, N., Li, Z., Zhang, L., Shen, C., Cheng, M.-M. & Reid, I. (2021), 'Unsupervised scale-consistent depth learning from video', *International Journal of Computer Vision* pp. 1–17.

Biber, P., Andreasson, H., Duckett, T. and Schilling, A. (2004) , 3d modeling of indoor environments by a mobile robot with a laser scanner and panoramic camera, *in* '2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)', Vol. 4, IEEE, pp. 3430–3435.

Casser, V., Pirk, S., Mahjourian, R. & Angelova, A. (2019) , Unsupervised monocular depth and ego-motion learning with structure and semantics, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops', pp. 0–0.

Chang, J.-R. & Chen, Y.-S. (2018), Pyramid stereo matching network, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5410–5418.

Chao, Z., Hong, Z., Shen, X. & Jia, J. (2017), Unsupervised learning of stereo matching, *in* 'IEEE International Conference on Computer Vision'.

Chen, L., Tang, W., John, N. W., Wan, T. R. & Zhang, J. J. (2017), 'Augmented reality for depth cues in monocular minimally invasive surgery', *arXiv preprint arXiv:1703.01243* .

Chen, W., Fu, Z., Yang, D. & Deng, J. (2016) , Single-image depth perception in the wild, *in* 'Advances in neural information processing systems', pp. 730–738.

Chen, Y., Schmid, C. & Sminchisescu, C. (2019) , Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 7063–7072.

Cheng, X., Wang, P. & Yang, R. (2018), 'Learning depth with convolutional spatial propagation network', *arXiv preprint arXiv:1810.02695* .

Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Drummond, T., Li, H. & Ge, Z. (2020), 'Hierarchical neural architecture search for deep stereo matching', *arXiv preprint arXiv:2010.13501* .

Clark, R., Wang, S., Wen, H., Markham, A. & Trigoni, N. (2017), Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 31.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. & Schiele, B. (2016) , The cityscapes dataset for semantic urban scene understanding, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3213–3223.

Cui, J., Zhang, H., Han, H., Shan, S. & Chen, X. (2018), Improving 2d face recognition via discriminative face depth estimation, *in* '2018 International Conference on Biometrics (ICB)', IEEE, pp. 140–147.

Cunha, J., Pedrosa, E., Cruz, C., Neves, A. J. & Lau, N. (2011), 'Using a depth camera for indoor robot localization and navigation', *DETI/IEETA-University of Aveiro, Portugal* p. 6.

de Queiroz Mendes, R., Ribeiro, E. G., dos Santos Rosa, N. & Grassi Jr, V. (2021) , 'On deep learning techniques to boost monocular depth estimation for autonomous navigation', *Robotics and Autonomous Systems* **136**, 103701.

Dosovitskiy, A., Fischery, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Smagt, P. V. D., Cre-mers, D. & Brox, T. (2015) , Flownet: Learning optical flow with convolutional networks, *in* 'IEEE International Conference on Computer Vision', pp. 2758–2766.

Eigen, D. & Fergus, R. (2015) , Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 2650–2658.

Eigen, D., Puhrsch, C. & Fergus, R. (2014), Depth map prediction from a single image using a multi-scale deep network, *in* 'International Conference on Neural Information Processing Systems', pp. 2366–2374.

Fang, Y., Masaki, I. & Horn, B. (2002) , 'Depth-based target segmentation for intelligent vehicles: Fusion of radar and binocular stereo', *IEEE transactions on intelligent transportation systems* **3**(3), 196–202.

Feng, L., Yang, X. & Xiao, S. (2017) , Magictoon: A 2d-to-3d creative cartoon modeling system with mobile ar, *in* '2017 IEEE Virtual Reality (VR)', IEEE, pp. 195–204.

Foix, S., Alenya, G. & Torras, C. (2011) , 'Lock-in time-of-flight (tof) cameras: A survey', *IEEE Sensors Journal* **11**(9), 1917–1926.

Fu, H., Gong, M., Wang, C., Batmanghelich, K. & Tao, D. (2018) , Deep ordinal regression network for monocular depth estimation, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2002–2011.

Gao, J., He, Q., Gao, H., Zhan, Z. & Wu, Z. (2018) , 'Design of an efficient multi-objective recognition approach for 8-ball billiards vision system', *Kuwait Journal of Science* **45**(1).

Garg, R., Bg, V. K., Carneiro, G. & Reid, I. (2016) , Unsupervised cnn for single view depth estimation: Geometry to the rescue, *in* 'European conference on computer vision', Springer, pp. 740–756.

Garg, R., Wadhwa, N., Ansari, S. & Barron, J. T. (2019) , Learning single camera depth estimation using dual-pixels, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 7628–7637.

Geiger, A. (2012), Are we ready for autonomous driving? the kitti vision benchmark suite, *in* 'Computer Vision and Pattern Recognition', pp. 3354–3361.

Geiger, A., Ziegler, J. & Stiller, C. (2011) , Stereoscan: Dense 3d reconstruction in real-time, *in* '2011 IEEE intelligent vehicles symposium (IV)', Ieee, pp. 963–968.

Godard, C., Aodha, O. M. & Brostow, G. J. (2017) , Unsupervised monocular depth estimation with left-right consistency, *in* 'Computer Vision and Pattern Recognition', pp. 6602–6611.

Godard, C., Mac Aodha, O., Firman, M. & Brostow, G. J. (2019) , Digging into self-supervised monocular depth estimation, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 3828–3838.

Guizilini, V., Ambrus, R., Pillai, S., Raventos, A. & Gaidon, A. (2020) , 3d packing for self-supervised monocular depth estimation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 2485–2494.

Guo, X., Yang, K., Yang, W., Wang, X. & Li, H. (2019) , Group-wise correlation stereo network, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 3273–3282.

Han, J., Shao, L., Xu, D. & Shotton, J. (2013) , 'Enhanced computer vision with microsoft kinect sensor: A review', *IEEE transactions on cybernetics* **43**(5), 1318–1334.

He, K., Zhang, X., Ren, S. & Sun, J. (2015) , 'Deep residual learning for image recognition', pp. 770–778.

Hernandez, C., Vogiatzis, G. & Cipolla, R. (2008) , 'Multiview photometric stereo', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(3), 548–554.

Hirschmuller, H. (2007), 'Stereo processing by semiglobal matching and mutual information', *IEEE Transactions on pattern analysis and machine intelligence* **30**(2), 328–341.

Hosni, A., Rhemann, C., Bleyer, M., Rother, C. & Gelautz, M. (2012), 'Fast cost-volume filtering for visual correspondence and beyond', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(2), 504–511.

Huynh, L., Nguyen-Ha, P., Matas, J., Rahtu, E. & Heikkilä, J. (2020) , Guiding monocular depth estimation using depth-attention volume, *in* 'European Conference on Computer Vision', Springer, pp. 581–597.

Jaderberg, M., Simonyan, K., Zisserman, A. *et al.* (2015) , Spatial transformer networks, *in* 'Advances in neural information processing systems', pp. 2017–2025.

Kao, J.-Y., Tian, D., Mansour, H., Vetro, A. & Ortega, A. (2016) , Moving object segmentation using depth and optical flow in car driving sequences, *in* '2016 IEEE International Conference on Image Processing (ICIP)', IEEE, pp. 11–15.

Kendall, A., Grimes, M. & Cipolla, R. (2015) , Posenet: A convolutional network for real-time 6-dof camera relocalization, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 2938–2946.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A. & Bry, A. (2017) , End-to-end learning of geometry and context for deep stereo regression, *in* 'IEEE International Conference on Computer Vision', pp. 66–75.

Khan, W., Daud, A., Nasir, J. A. & Amjad, T. (2016) , 'A survey on the state-of-the-art machine learning models in the context of nlp', *Kuwait journal of Science* **43**(4).

Kuznietsov, Y., Stuckler, J. & Leibe, B. (2017) , Semi-supervised deep learning for monocular depth map prediction, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 6647–6655.

Ladicky, L., Shi, J. & Pollefeys, M. (2014) , Pulling things out of perspective, *in* 'IEEE Conference on Computer Vision and Pattern Recognition', pp. 89–96.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F. & Navab, N. (2016) , Deeper depth prediction with fully convolutional residual networks, *in* '2016 Fourth international conference on 3D vision (3DV)', IEEE, pp. 239–248.

Lee, S., Im, S., Lin, S. & Kweon, I. S. (2021) , 'Learning monocular depth in dynamic scenes via instance-aware projection consistency', *arXiv preprint arXiv:2102.02629* .

Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R. & Furgale, P. (2015) , 'Keyframe-based visual–inertial odometry using nonlinear optimization', *The International Journal of Robotics Research* **34**(3), 314–334.

Li, C. & Waslander, S. L. (2020) , Towards end-to-end learning of visual inertial odometry with an ekf, *in* '2020 17th Conference on Computer and Robot Vision (CRV)', IEEE, pp. 190–197.

Li, H., Gordon, A., Zhao, H., Casser, V. and Angelova, A. (2020) , 'Unsupervised monocular depth learning in dynamic scenes', *arXiv e-prints* pp. arXiv–2010.

Li, R., Wang, S., Long, Z. & Gu, D. (2018) , Undeepvo: Monocular visual odometry through unsupervised deep learning, *in* '2018 IEEE international conference on robotics and automation (ICRA)', IEEE, pp. 7286–7291.

Liu, F., Shen, C. & Lin, G. (2015) , Deep convolutional neural fields for depth estimation from a single image, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 5162–5170.

Liu, F., Shen, C., Lin, G. & Reid, I. (2016), 'Learning depth from single monocular images using deep convolutional neural fields', *IEEE Transactions on Pattern Analysis & Machine Intelligence* **38**(10), 2024–2039.

Liu, P., King, I., Lyu, M. R. & Xu, J. (2020) , Flow2stereo: Effective self-supervised learning of optical flow and stereo matching, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 6648–6657.

Long, J., Shelhamer, E. & Darrell, T. (2015) , Fully convolutional networks for semantic segmentation, *in* 'IEEE Conference on Computer Vision and Pattern Recognition', pp. 3431–3440.

Luo, C., Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R. & Yuille, A. (2019) , 'Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding', *IEEE transactions on pattern analysis and machine intelligence* **42**(10), 2624–2641.

Luo, W., Schwing, A. G. & Urtasun, R. (2016) , Efficient deep learning for stereo matching, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5695–5703.

Luo, Y., Ren, J., Lin, M., Pang, J., Sun, W., Li, H. & Lin, L. (2018) , Single view stereo matching, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 155–163.

Mahjourian, R., Wicke, M. & Angelova, A. (2018) , Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5667–5675.

Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A. & Brox, T. (2016) , A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, *in* 'Computer Vision and Pattern Recognition', pp. 4040–4048.

Mei, X., Sun, X., Dong, W., Wang, H. & Zhang, X. (2013) , Segment-tree based cost aggregation for stereo matching, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 313–320.

Menze, M. & Geiger, A. (2015) , Object scene flow for autonomous vehicles, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3061–3070.

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N. & Terzopoulos, D. (2021) , 'Image segmentation using deep learning: A survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Mur-Artal, R. & Tardós, J. D. (2017) , 'Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras', *IEEE transactions on robotics* **33**(5), 1255–1262.

Pang, J., Sun, W., Ren, J. S., Yang, C. & Yan, Q. (2017) , Cascade residual learning: A two-stage convolutional neural network for stereo matching, *in* 'Proceedings of the IEEE International Conference on Computer Vision Workshops', pp. 887–895.

Poggi, M.and Aleotti, F., Tosi, F. & Mattoccia, S. (2020) , 'On the uncertainty of self-supervised monocular depth estimation'.

Ronneberger, O., Fischer, P. & Brox, T. (2015) , U-net: Convolutional networks for biomedical image segmentation, *in* 'International Conference on Medical image computing and computer-assisted intervention', Springer, pp. 234–241.

Saxena, A., Chung, S. H. & Ng, A. Y. (2008) , '3-d depth reconstruction from a single still image', *International Journal of Computer Vision* **76**(1), 53–69.

Scharstein, D. & Szeliski, R. (2002) , 'A taxonomy and evaluation of dense two-frame stereo correspondence algorithms', *International journal of computer vision* **47**(1-3), 7–42.

Scharstein, D. and Szeliski, R. (2003) , High-accuracy stereo depth maps using structured light, *in* '2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.', Vol. 1, IEEE, pp. I–I.

Scharstein, D. & Szeliski, R. (n.d.), 'A taxonomy and evaluation of dense two-frame stereo correspondence algorithms', *International Journal of Computer Vision* **47**(1-3), 7–42.

Shamwell, E. J., Leung, S. & Nothwang, W. D. (2018) , Vision-aided absolute trajectory estimation using an unsupervised deep network with online error correction, *in* '2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)', IEEE, pp. 2524–2531.

Shamwell, E. J., Lindgren, K., Leung, S. & Nothwang, W. D. (2019), 'Unsupervised deep visual-inertial odometry with online error correction for rgb-d imagery', *IEEE transactions on pattern analysis and machine intelligence* **42**(10), 2478–2493.

Shen, Z., Dai, Y. & Rao, Z. (2021) , Cfnet: Cascade and fused cost volume for robust stereo matching, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 13906–13915.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. & Blake, A. (2011), Real-time human pose recognition in parts from single depth images, *in* 'CVPR 2011', Ieee, pp. 1297–1304.

Sielhorst, T., Bichlmeier, C., Heining, S. M. & Navab, N. (2006) , Depth perception–a major issue in medical ar: evaluation study by twenty surgeons, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 364–372.

Song, X., Zhao, X., Fang, L., Hu, H. & Yu, Y. (2020) , 'Edgestereo: An effective multi-task learning network for stereo matching and edge detection', *International Journal of Computer Vision* **128**(4), 910–930.

Tang, C. & Tan, P. (2018), 'Ba-net: Dense bundle adjustment network', *arXiv preprint arXiv:1806.04807* .

Tankovich, V., Hane, C., Zhang, Y., Kowdle, A., Fanello, S. & Bouaziz, S. (2021) , Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 14362–14372.

Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A. & Brox, T. (2017) , Demon: Depth and motion network for learning monocular stereo, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5038–5047.

Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R. & Fragkiadaki, K. (2017) , 'Sfm-net: Learning of structure and motion from video', *arXiv preprint arXiv:1704.07804* .

Wang, H., Fan, R., Cai, P. & Liu, M. (2021) , 'Pvstereo: Pyramid voting module for end-to-end self-supervised stereo matching', *IEEE Robotics and Automation Letters* **6**(3), 4353–4360.

Wang, S., Clark, R., Wen, H. & Trigoni, N. (2017) , Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks, *in* '2017 IEEE International Conference on Robotics and Automation (ICRA)', IEEE, pp. 2043–2050.

Wang, Y., Lai, Z., Huang, G., Wang, B. H., Van Der Maaten, L., Campbell, M. & Weinberger, K. Q. (2019) , Anytime stereo image depth estimation on mobile devices, *in* '2019 International Conference on Robotics and Automation (ICRA)', IEEE, pp. 5893–5900.

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004) , 'Image quality assessment: from error visibility to structural similarity', *IEEE Trans Image Process* **13**(4), 600–612.

Wu, C., Agarwal, S., Curless, B. & Seitz, S. M. (2011) , Multicore bundle adjustment, *in* 'Computer Vision & Pattern Recognition'.

Xie, J., Girshick, R. & Farhadi, A. (2016) , Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks, *in* 'European Conference on Computer Vision', Springer, pp. 842–857.

Yang, J., Li, H., Campbell, D. & Jia, Y. (2015), 'Go-icp: A globally optimal solution to 3d icp point-set registration', *IEEE transactions on pattern analysis and machine intelligence* **38**(11), 2241–2254.

Yang, N., Stumberg, L. v., Wang, R. & Cremers, D. (2020) , D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 1281–1292.

Yang, N., Wang, R., Stuckler, J. & Cremers, D. (2018) , Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry, *in* 'Proceedings of the European Conference on Computer Vision (ECCV)', pp. 817–833.

Yang, Q. (2012) , A non-local cost aggregation method for stereo matching, *in* '2012 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1402–1409.

Yang, Z., Wang, P., Xu, W., Zhao, L. & Nevatia, R. (2017) , 'Unsupervised learning of geometry with edge-aware depth-normal consistency', *arXiv preprint arXiv:1711.03665* .

Yee, K. & Chakrabarti, A. (2020) , Fast deep stereo with 2d convolutional processing of cost signatures, in 'Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision', pp. 183–191.

Yin, W., Liu, Y., Shen, C. & Yan, Y. (2019) , Enforcing geometric constraints of virtual normal for depth prediction, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 5684–5693.

Zhan, H., Garg, R., Saroj Weerasekera, C., Li, K., Agarwal, H. & Reid, I. (2018) , 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', pp. 340–349.

Zhan, H., Weerasekera, C. S., Bian, J.-W. & Reid, I. (2020) , Visual odometry revisited: What should be learnt?, *in* '2020 IEEE International Conference on Robotics and Automation (ICRA)', IEEE, pp. 4203–4210.

Zhang, F., Prisacariu, V., Yang, R. & Torr, P. H. (2019) , Ga-net: Guided aggregation net for end-to-end stereo matching, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 185–194.

Zhang, F., Qi, X., Yang, R., Prisacariu, V., Wah, B. & Torr, P. (2020) , Domain-invariant stereo matching networks, *in* 'European Conference on Computer Vision', Springer, pp. 420–439.

Zhang, Y., Chen, Y., Bai, X., Yu, S., Yu, K., Li, Z. & Yang, K. (2020) , Adaptive unimodal cost volume filtering for deep stereo matching, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 34, pp. 12926–12934.

Zhang, Y., Khamis, S., Rhemann, C., Valentin, J., Kowdle, A., Tankovich, V., Schoenberg, M., Izadi, S., Funkhouser, T. & Fanello, S. (2018) , Activestereonet: End-to-end self-supervised learning for active stereo systems, *in* 'Proceedings of the European Conference on Computer Vision (ECCV)', pp. 784–801.

Zhao, H., Gallo, O., Frosio, I. & Kautz, J. (2015) , 'Is l2 a good loss function for neural networks for image processing? arxiv preprint', *arXiv preprint arXiv:1511.08861* .

Zhichao, Y. & Jianping, S. (2018) , 'Geonet: Unsupervised learning of dense depth, optical flow and camera pose', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* .

Zhong, Y., Dai, Y. & Li, H. (2017) , 'Self-supervised learning for stereo matching with self-improving ability', *arXiv preprint arXiv:1709.00930* .

Zhou, J., Wang, Y., Qin, K. & Zeng, W. (2019) , Moving indoor: Unsupervised video depth learning in challenging environments, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 8618–8627.

Zhou, Q.-Y. & Koltun, V. (2014) , 'Color map optimization for 3d reconstruction with consumer depth cameras', *ACM Transactions on Graphics (TOG)* **33**(4), 1–10.

Zhou, T., Brown, M., Snavely, N. & Lowe, D. G. (2017) , Unsupervised learning of depth and ego-motion from video, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1851–1858.

Zhou, T., Fan, D.-P., Cheng, M.-M., Shen, J. & Shao, L. (2021) , 'Rgb-d salient object detection: A survey', *Computational Visual Media* pp. 1–33.

Zhu, J., Wang, L., Yang, R., Davis, J. E. *et al.* (2010) , 'Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps', *IEEE transactions on pattern analysis and machine intelligence* **33**(7), 1400–1414.

Submitted: 10/03/2021
Revised: 16/07/2021
Accepted: 23/08/2021
DOI: 10.48129/kjs.12929