Adversarial Learning for Joint Optimization of Depth and Ego-Motion

Anjie Wang, Zhijun Fang, *Senior Member, IEEE*, Yongbin Gao, Songchao Tan, Shanshe Wang, Siwei Ma, *Member, IEEE*, and Jenq-Neng Hwang, *Fellow, IEEE*

***Abstract*— In recent years, supervised deep learning methods have shown a great promise in dense depth estimation. However, massive high-quality training data are expensive and impractical to acquire. Alternatively, self-supervised learning-based depth estimators can learn the latent transformation from monocular or binocular video sequences by minimizing the photometric warp error between consecutive frames, but they suffer from the scale ambiguity problem or have difficulty in estimating precise pose changes between frames. In this paper, we propose a joint self-supervised deep learning pipeline for depth and ego-motion estimation by employing the advantages of adversarial learning and joint optimization with spatial-temporal geometrical constraints. The stereo reconstruction error provides the spatial geometric constraint to estimate the absolute scale depth. Meanwhile, the depth map with an absolute scale and a pre-trained pose network serves as a good starting point for direct visual odometry (DVO). DVO optimization based on spatial geometric constraints can result in a fine-grained ego-motion estimation with the additional backpropagation signals provided to the depth estimation network. Finally, the spatial and temporal domain-based reconstructed views are concatenated, and the iterative coupling optimization process is implemented in combination with the adversarial learning for accurate depth and precise ego-motion estimation. The experimental results show superior performance compared with state-of-the-art methods for monocular depth and ego-motion estimation on the KITTI dataset and a great generalization ability of the proposed approach.**

***Index Terms*— Depth estimation, ego-motion, self-supervised, adversarial learning, direct visual odometry.**

I. Introduction

THE computer vision society has regarded 3D localization and reconstruction as a pure geometric problem in the past. The epipolar geometry constraint between two views is widely used for visual odometry (VO), which estimates the relative pose between two frames. In recent years, deep learning has become prevalent for a variety of computer vision tasks. Numerous researchers attempt to design deep learning methods to solve the depth and ego-motion problems.

Most depth estimation and VO models are trained in a supervised manner, in which enormous ground truth depth and pose datasets are required. However, it is difficult and expensive to acquire the massive datasets, not to mention the issues that arise when the quality of the datasets varies with the use of different sensors and the capture process under distinct conditions. When less ground truth depth maps are used for learning, the depth estimation performance degrades severely. In contrast, unsupervised learning [1] provides an alternative way to force a deep network to learn the transformation relations between frames. Zhou *et al.* [2] propose to use the photometric warp error between adjacent frames as the supervised signal to jointly train two deep networks for monocular depth and ego-motion estimation, respectively.

However, the current monocular framework, which relies on the visual difference between consecutive frames, suffers

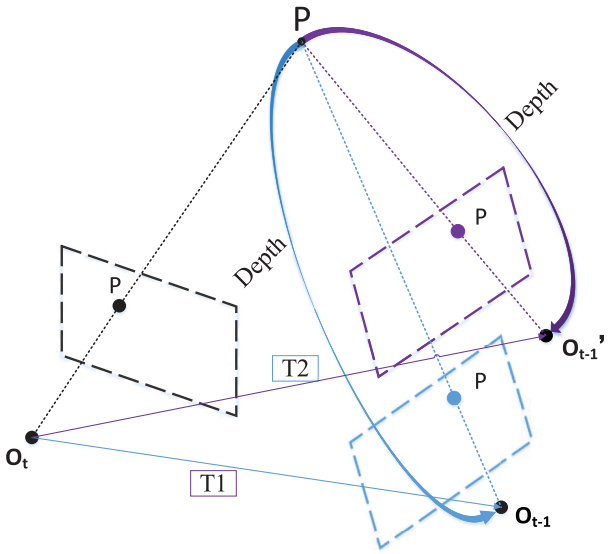

Fig. 1. Scale ambiguity in monocular vision, in which the lack of the absolute scale of the scene results in different estimated depths of P with varying displacement d. Thus, the projected pixels in the image plane might have equal coordinates when using different pose scales based on the structure from motion methods for monocular cameras. Different d values result in distinct depth estimations of the 3D point P.

Manuscript received May 14, 2019; revised November 26, 2019 and January 11, 2020; accepted January 12, 2020. Date of publication January 28, 2020; date of current version February 6, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61772328, Grant 61802253, and Grant 61831018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Weiyao Lin. *(Corresponding authors: Zhijun Fang; Yongbin Gao.)*

Anjie Wang, Zhijun Fang, and Yongbin Gao are with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China (e-mail: anjie@sues.edu.cn; zjfang@sues.edu.cn; gaoyongbin@sues.edu.cn).

Songchao Tan, Shanshe Wang, and Siwei Ma are with the Institute of Digital Media, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: songchaotan@pku.edu.cn; sswang@pku.edu.cn; swma@pku.edu.cn).

Jenq-Neng Hwang is with the Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: hwang@uw.edu).

Digital Object Identifier 10.1109/TIP.2020.2968751

from the scale ambiguity problem as shown in Fig. 1. More specifically, the lack of the absolute scale results in various pairs of poses T (including rotation R and translation t) and depths D with different scales. Thus, the projected pixels in the image plane might have equal coordinates when using different pose scales based on the structure from motion(SfM) methods for monocular cameras. Although the use of binocular video sequences can solve scale ambiguity problems, the learned deep networks will be restricted for use with the same stereo camera setups. Simultaneously, depth estimation based on deep learning is due to its semantic learning capabilities rather than direct geometric constraints.

Steinbrucker *et al.* [3] propose a direct VO (DVO) method to model the relationship between input dense depth maps and output pose predictions. The optimization objective is to minimize the pixel-level photometric warp error for sequential RGB-D images. DVO methods show better accuracy when the motion between frames is small, since they use the dense direct geometry information.

In the past few years, generative models achieve impressive success in modeling complex high-dimensional data distributions, of which Generative Adversarial Networks (GANs) [4] have become a powerful framework for generative and potential spatial learning. However, depth estimation, as a more difficult generative task, has yet to be fully explored.

In this paper, we proposed a joint self-supervised learning based pipeline for depth and ego-motion estimation by integrating adversarial learning with spatial-temporal geometric constraints. Specifically, a generator is trained to combine both a depth-pose net and DVO to generate a reconstructed view conditioned on a warping formula; and the original view is then fed into a discriminator that is adversarially trained to distinguish whether the reconstructed view is plausible or not. The rationale behind our idea is that a generator producing accurate depth maps and poses will also lead to better reconstructed images, which are harder to be distinguished from original unwarped images by the discriminator, thereby pushing the generator to build more realistic warped images and thus more accurate depth and pose predictions. The part of this work was presented in IEEE ICME conference [63], the contributions of our proposed architecture are summarized as follows:

1. We incorporate deep learning-based depth estimation with the spatial geometrical constraint of stereo reconstruction and the temporal geometrical constraint of DVO. This framework enables making full use of the semantic learning ability of deep learning for accurate dense depth estimation, while solving the scale ambiguity problem.

2. Ego-motion estimation based on deep learning learns only the latent transformation between frames, while it is not accurate for small pose changes. The DVO used to refine the estimated ego-motion by the pose-net, which provides a fine-grained pose estimation for adjacent frames, and gives an effective back-propagation gradient to the depth estimation network.

3. The proposed joint training pipeline of adversarial learning and the traditional geometry method enables an iterative coupling optimization process for accurate depth and fine-grained ego-motion estimation. This pipeline also exploits conditional GAN during training to refine the reconstructed view from a generator, taking advantage of the latent distribution rather than a binary variable from the discriminator to train both the discriminator and generator.

Extensive experiments demonstrate that our framework outperforms most advanced technologies; the depth estimation results are not only quantitatively more accurate but also qualitatively more detailed and VO estimation is also closer to the ground truth.

II. Related Work

A. Depth Estimation from images, as one of the crucial techniques in 3D scene reconstruction, has been playing an important role for many computer vision applications. However, the estimated depth map by traditional methods has defects, for example, depth inhomogeneity, local depth deletion, high computational complexity, and the requirement of a postprocessing module. All these problems make the traditional depth estimators encounter difficulty when applied to real-world scenarios.

With the development of neural network, Eigen *et al.* [5], [6] proposed a series of works using deep learning methods to estimate depth in a supervised manner, thus using ground-truth depth at training time. More specifically, a multiscale deep network with scale-invariant losses is proposed to learn representations directly from image pixels instead of extracting features from image batches. Similar works have also been reported [7]–[13]. However, supervised deep learning methods need a vast amount of training images with high quality ground truth depth data, which are expensive and intractable to capture.

To alleviate this laborious ground truth collection problem, many unsupervised learning frameworks are proposed for depth estimation. Garg *et al.* [14] propose an unsupervised deep learning method by using stereo image pairs that are captured by two cameras whose pose translation relationship is known. Godard *et al.* [15] improve the estimation performance by adding the left-right consistency to the loss function of the network. In addition, some works [16]–[19] use stereo matching in combination with other advantageous strategies to estimate the depth and achieve the desired effect.

Zhou *et al.* [2] train a depth network and a pose network from monocular sequences. To obtain promising performance in depth and ego-motion estimation by using the photometric error in videos as the supervisory information in training. The works in [20], [21] improve the method in [2], but all of them do not solve the scale ambiguity problem.

There are other strategies for solving monocular depth estimation problems, such as specially designed network structures [22]–[24] combined with information such as optical flow [25]–[27]. Some works [28], [29] use edge information for network optimization and had achieved ideal results.

Similar to the works of [14], [15], [19] we train the depth estimator that uses the baseline constraint between stereo images to obtain the absolute scale. In addition, the stereo spatial reconstruction constraint and temporal constraint from monocular videos are jointly imposed in the deep learning

framework to produce more accurate ego-motion estimation, resulting in a more reliable depth map.

B. Visual Odometry techniques compute the location and orientation for moving robots in an unknown environment from a series of images captured by the mounted cameras. In the past decades, model-based and geometry-based VO approaches have been studied widely. Steinbruecker *et al.* [3] proposed a DVO method to model the relationship between input dense depth maps and output pose predictions. The target was to minimize the pixel-level photometric warp error for sequential RGB-D images. Compared with VO approaches using sparse features, DVO approaches show better accuracy when the motion between frames is small since they use all the dense information.

Recently, data driven-based or deep learning-based VO methods have shown promising results due to their powerful learning ability and good robustness in various challenging environments. Kendall *et al.* [30] presented a convolutional neural network (CNN)-based approach, of which RGB images are the input and the six-degree pose is the regression output. Wang *et al.* [31] proposed an end-to-end DVO pipeline adopting deep recurrent CNNs. Moreover, ego-motion estimation based on deep learning methods are further studied in the works of [32].

It is worth noting that, the abovementioned approaches require either manually labeled camera poses or ground truth of depth data. In contrast, unsupervised learning methods, which are trained with only unlabeled raw data, are much easier to deploy and have more practical potential. Zhou *et al.* [2] trained an end-to-end unsupervised learning network containing a depth network and a multiview pose network using monocular sequences. Unfortunately, it suffers from the scale ambiguity problem since only the orientation of camera movement is known while the absolute metric is unknown.

Furthermore, the authors in [33], [34] adopt stereo image pairs to recover the absolute scale and solve the scale ambiguity issue. We are thus motivated to integrate the absolute scale recovery module and DVO to realize a self-supervised pose estimation deep learning framework, which is able to not only avoid the problems of scale ambiguity but also improve the ego-motion estimation accuracy.

C. Generative adversarial networks (GANs) [4] generates high-quality samples leveraging the idea of a unique zero-sum game and confrontational training and have more powerful feature learning and expression capabilities than traditional machine learning algorithms. The representations learned by GANs can be used in a variety of applications, including image and the text generation [35], [36], style transfer [37], image super-resolution [38], latent space learning [39], [40], etc.

The GANs have attracted substantial attention for their advantage in data generation problems. The backpropagation algorithm (BP) is used to train two networks. Specifically, the generator and discriminator networks can work together to provide a powerful framework for creating unsupervised learning models. However, some problems exist in the vanilla GAN, such as the vanishing gradients, the difficult training procedures, the inability of the loss function of the generator and the discriminator to guide the training process, the lack oof diversity of sample generation, the susceptibility of training to overfitting, etc.

Arjovsky *et al.* [41] propose a Wasserstein GAN (WGAN) that uses the Wasserstein distance (also known as earth-mover's distance, EMD), instead of using the Jensen¨CShannon (JS) or Kullback-Leibler (KL) divergence to measure the distance between real and generated samples. Subsequently, Gulrajani *et al.* [42] propose an improved WGAN structure, named WGAN-GP denoting a WGAN with a gradient penalty, to implement the Lipschitz constraint method instead of the weight clipping in WGAN. The WGAN-GP method is shown to produce a generated sample with a higher quality than that of WGAN with stable training.

Yang *et al.* [43] reconstruct the complete three-dimensional (3D) structure of a given object from a single arbitrary depth view. Using the adversarial learning combination of the autoencoder generation capability and the conditional GAN (CGAN) framework, the precise and refined three-dimensional (3D) structure of the objects in the high-dimensional voxel space is inferred.

Pilzer *et al.* [44] propose a reconstruction of disparity maps by employing two generator subnetworks through joint training of adversarial learning, organized in a circular form to impose constraints during the supervision. Different from [44], although the adversarial training is used, [44] is based on the loss of cyclic consistency, and only the constraint information of the binocular stereo pair is used; however, we are conducting the adversarial training under the joint spatial and temporal constraint. The generation function proposed by Kumar *et al.* [45] learns the adjacent image depth map and the relative target pose, the discriminant function of which learns the distribution of monocular images and correctly classifies the authenticity of the composite image. Similar strategies are also applied to stereo pairs [46]. Kim *et al.* [47] use adversarial loss to fit the distribution of depth predictions with the ground-truth depth, producing perceptually more convincing solutions. Our pipeline also exploits the GAN during training to refine the reconstructed view from the generator, taking advantage of the latent distribution rather than a binary variable from the discriminator to train both the discriminator and generator.

III. The Proposed Approach

The proposed joint deep learning pipeline is shown in Figure 2, where the absolute scale of the dense depth map is estimated by minimizing the photometric warp error between left-right images with a stereo reconstruction error being used as a spatial geometrical constraint. Meanwhile, the deep learning pose network is used as an initial point for the geometrical optimization and the DVO estimates the fine-grained ego-motion even for a small pose change by using the inverse compositional algorithm with the temporal constraint between adjacent frames. The joint training pipeline benefits from adversarial learning and traditional geometrical optimization.

A. Spatial Geometric Constraint for Depth Estimation

To calculate the absolute scale of the depth map, stereo images are used to train the depth-network as shown

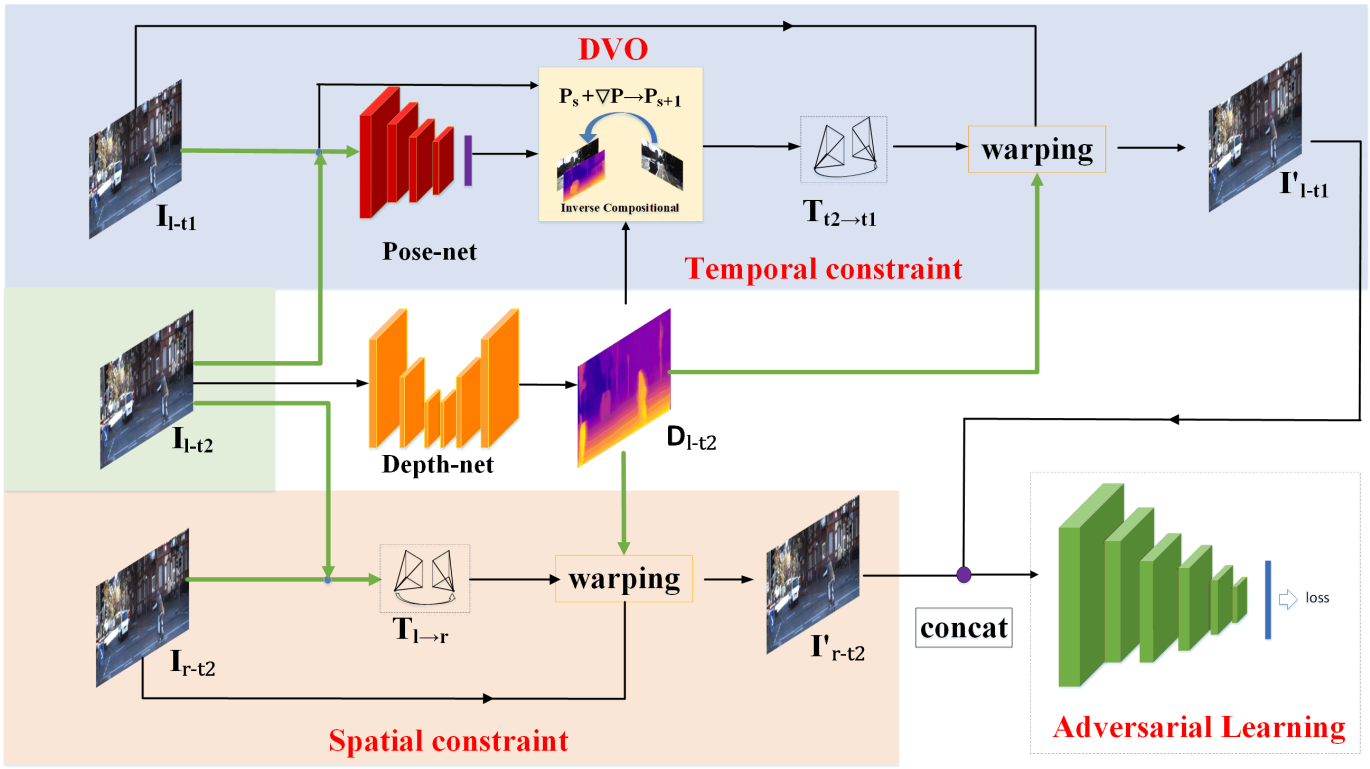

Fig. 2. The framework of the proposed method. Given a pair of spatial stereo images I_{l-t2}, I_{r-t2} and an extra temporal image I_{l-t1}, the spatial constraint aims to minimize the photometric warp error between I_{l-t2} and its reconstruction I'_{l-t2} from I_{r-t2}, while the temporal constraint that establishes on the DVO attempts to minimize the photometric warp error between I_{l-t2} and its reconstruction I'_{l-t1} from I_{l-t1}. Finally, the reconstructed view and the target view are further fed into the discriminator, aiming to distinguish the distributions of latent representations of real and reconstructed views, while the generator is trained to force these two distributions to resembled each other.

in Figure 2. Given the baseline of a stereo system, which is provide clear geometric constraints in the spatial domain, the depth estimation network is adopted to obtain depth maps. Consequently, we can use the baseline distance and focal length information to convert the depth data into disparity data, and then use it to reconstruct the left/right views. The photometric consistency between the reconstructed view and the target view is used as supervisory information for training the depth estimation network.

Specifically, for those overlapping regions of a stereo image pair, the corresponding pixels of a left-right image-pair can be matched by their disparity d, which is calculated as:

$$d = bf \times D_{inv}, \tag{1}$$

where b is the baseline distance of the stereo camera; f is the focal length; and D_{inv} is the inverse depth value of the corresponding pixel.

Stereo matching is based on the work of Godard *et al.* [15], such that given a stereo image-pair I_{l-t2} and I_{r-t2}, where I_{l-t2} is the target view and I_{r-t2} is the reference view, we can synthesize the reference view I_{l-t2} as I'_{r-t2} from the source view I_{r-t2} by using the horizontal distance d. Similarly, we can reconstruct the right image from the given left image. A hybrid loss function $\mathcal{L}_{s-l1}$, which combines the structural similarity (SSIM) index [48] and L1 norm, is used as the left and right photometric consistency loss of the synthesized view and the target view:

$$\mathcal{L}_l = \sum\nolimits_{x_i} \{\mathcal{L}_{s-l1}[I_{l-t2}(x_i), I'_{r-t2}(x_i)]\}, \tag{2}$$

$$\mathcal{L}_r = \sum\nolimits_{x_i} \{\mathcal{L}_{s-l1}[I_{r-t2}(x_i), I'_{l-t2}(x_i)]\}, \tag{3}$$

$$\mathcal{L}_s = \mathcal{L}_l + \mathcal{L}_r. \tag{4}$$

The summation of the left and right photometric consistency loss is the spatial geometry constraint. Since the pose between the left and right images is fixed in a stereo system, there is no pose estimation involved. Thus, the direct supervision of the depth-net by the left and right consistency loss is added in the network, and absolute scale of the monocular video sequence can be calculated, resulting in an accurate depth map.

B. Temporal Geometric Constraint for Ego-Motion Estimation

In addition, the temporal geometric constraint can be exploited for ego-motion estimation. Given a temporal image-pair I_{l-t1} and I_{l-t2}, during the training of pose-net, the adjacent frames are reconstructed by the image warping process. For each pixel point p_{l-t2} in the target view I_{l-t2}, we first project it to the adjacent reference view I_{l-t1} according to the predicted depth and the predicted pose of the camera. Then, the value of the pixel of the target view I_{l-t2} is reconstructed at position x_i using bilinear interpolation.

Algorithm 1 The inverse compositional algorithm

Preprocess:
1.Calculate image gradient $\nabla I(x_i)$;
2.Calculate Jacobian matrix $J = \nabla I(x_i)\frac{\partial \omega(x_i;p,d_i)}{\partial p}$ at $(x_i; 0, d_i)$;
3.Calculate Hessian matrix: $H = J^T J$;
Iteration optimization:
1.Warp I with $\omega(x_i; p, d_i)$ to compute $I^{'}(\omega(x_i; p, d_i))$;
2.Compute the error image $\sum_{xi} r$;
3.Compute $\nabla p = H^{-1}J^T \sum_{xi} r$, where $H^{-1}J^T$ is differentiating the matrix pseudoinverse;
4.Update $p_{s+1} \leftarrow p_s + \nabla p$;
Repeat till converge

The projected coordinates are obtained by:

$$I'_{l-t1} = f(K, T_{t_2 \to t_1}, D) \cdot I_{l-t1}, \tag{5}$$

where I'_{l-t1} is the reconstructed target view of I_{l-t2} based on the pose prediction function f, K is the camera's intrinsic matrix; D is the depth value of the pixel in the target view I_{l-t2}, and $T_{t2\to t1}$ is the camera coordinate transformation matrix from the target view to the reference view. We can synthesize the target view I_{l-t2} from the reference view I_{l-t1} using the estimated pose and spatial transform [49]. Therefore, the photometric consistency loss between the monocular image sequences is:

$$\mathcal{L}_t = \sum\nolimits_{x_i} \{\mathcal{L}_{s-l1}[I_{l-t2}(x_i), I'_{l-t1}(x_i)]\}. \tag{6}$$

Eq. (6)implies that the reconstruction accuracy is highly reliant on the depth map and the pose between two views. Since the depth map obtained is accurate with absolute scale in this paper, and thus, the temporal reconstruction error is backpropagated mainly to adjust the pose between two views in a self-supervised manner. However, the obtained pose may not be accurate since the self-supervised learning process learns the approximation of transformation with no geometry information involved.

Inspired by the recent advance in DVO, we use DVO as an explicit temporal geometric constraint for ego-motion estimation. DVO is a generalized image registration method based on Gauss-Newton minimization to accelerate convergence, and uses the inverse compositional algorithm to improve computational efficiency [50], [51]. DVO takes the target view I_{l-t2}, the corresponding depth map D and the adjacent reference view I_{l-t1} as inputs, and aims to find an optimal camera pose P to minimize the photometric error between the warped reference image and the target image in an iterative manner; the objective of DVO is:

$$\mathcal{L}_{t-vo} = min_p \sum\nolimits_{x_i} [(W(x_i, p, D)) - I_{l-t2}(x_i)]^2, \tag{7}$$

where W is the warping transformation, and p is the parameter of pose T. This nonlinear least square problem can be well solved by the Gauss-Newton method. To improve computational efficiency, instead we use the inverse compositional algorithm [51].

Since the Jacobian matrix and the Hessian matrix in the inverse compositional algorithm do not need to be recalculated in each iteration, the inverse compositional algorithm has faster convergence than the gradient descent method. It is also worthy of noting that DVO is convergent only when the initial error r is close to zero; thus, a good initial pose is a prerequisite for obtaining an optimal solution. Therefore, we jointly train depth-net and pose-net in the first stage to provide pose initialization for DVO, and the DVO provides a temporal geometry constraint to refine the pose-net. For the initial error caused by excessive motion, we use the image pyramid to downsample the input image and use the depth map to ensure the convergence of the algorithm.

C. Joint Training With Spatial-Temporal Geometrical Constraints

In our proposed joint training pipeline, the depth and ego-motion estimations are optimized with spatial and temporal geometrical constraints. Meanwhile, the depth and pose networks provide a credible initial pose and a depth estimate without scale ambiguity through pretraining, respectively. The depth and ego-motion estimation are optimized in an iterative manner, which means the depth map with absolute scale improves the ego-motion estimation based on DVO, while the DVO errors are backpropagated to the depth network to improve the estimation accuracy. More specifically, for the photometric consistency loss of the monocular image sequence, our reconstruction error$\mathcal{L}$ is related to only the depth D and the pose T; thus, the training objective can be described as:

$$\mathcal{L}_{dvo} = argmin_\tau min_p \mathcal{L}\{f_D(\tau), f_T(p)\}, \tag{8}$$

where τ is the parameter of depth-net,and p is the parameter of camera pose T. Given a depth D, minimizing $\mathcal{L}\{f_T(p)\}$over camera pose p can be viewed as the function

$$argmin_p \mathcal{L}\{f_T(p)\} = f_T(D, I_{l-t1}, I_{l-t2}) \tag{9}$$

where D is $f_D(\tau)$. Our solution to the pose predictor f_T can be expressed as the best pose by minimizing the photometric consistency loss, such that by substituting Eq. (9) into Eq. (8), we can derive:

$$\mathcal{L}_{dvo} = argmin_\tau \mathcal{L}\{f_D(\tau), f_T[f_D(\tau)]\}. \tag{10}$$

Therefore, in the differentiable DVO process [50], the effect of depth on the photometric consistency loss mainly derives from two aspects: the partial derivative of loss over depth and pose.

$$\frac{d\mathcal{L}_{dvo}}{df_D} = \frac{\partial \mathcal{L}_{dvo}}{\partial f_D} + \frac{\partial \mathcal{L}_{dvo}}{\partial f_T}\frac{\partial f_T}{\partial f_D}. \tag{11}$$

Eq. (11) shows that our depth-net can obtain additional backpropagation information from the pose prediction. In view of this, the temporal geometry constraint establishes a direct relationship between the depth map and the pose in the DVO process, and provides additional error gradients to the depth-net to optimize the network, resulting in a more accurate depth map and fine-grained ego-motion.

Moreover, in order to obtain smooth depth predictions, our approach encourages depth local smoothing by introducing edge-aware smoothing terms in the joint optimization process. Edge smoothness loss is based on the work of Godard *et al.* [15], expressed as:

$$\mathcal{L}_{sm} = \sum_{x_i} \left|\partial_x D_{x_i}\right| e^{-\left|\partial_x I_{x_i}\right|} + \left|\partial_y D_{x_i}\right| e^{-\left|\partial_y I_{x_i}\right|}. \quad (12)$$

Our total loss function for view synthesis based on geometric constraints is:

$$\mathcal{L}_{gc} = \alpha \mathcal{L}_s + \beta(\mathcal{L}_{t-v_0} + \mathcal{L}_t) + \gamma\, \mathcal{L}_{sm}. \quad (13)$$

Since the loss is defined on top of spatial geometric constraints and temporal geometric constraints, rather than utilizing labeled data, our proposed network is an self-supervised learning network. Depth-net uses the spatial geometric constraint between stereo pairs to restore the absolute scale of the depth map, while the two consecutive monocular images use temporal geometric constraints to estimate the camera pose and introduce DVO to provide additional backpropagated error information to optimize the depth map, which in turn refines the ego-motion. Finally, spatial-temporal geometry constraints are used to reconstruct the temporal image-pairs to achieve joint optimization of the depth network and the pose network. Although our system is trained with binocular video sequences, only a monocular camera is used in deployment, thus, our proposed network is a monocular system for depth and ego-motion estimation based on monocular video sequences.

D. Global Optimization With Adversarial Training

In addition to the photometric error between the reconstructed view and target views, which is essentially the pixel-level difference between two views, the reconstructed view tends to be blurred by using photometric error. The beauty of the adversarial learning is to force the network to generate a high-quality reconstructed view instead of a blurred image, and the error signal from adversarial learning will be backpropagated to the depth estimation network (part of the generator) to refine the depth map. In adversarial training, we use a joint view synthesis network as a generator, which is a depth-pose prediction network with geometric optimization. Under the supervision of the synthetic view, through the feedback adjustment, the networks learn the correlation between the overall structure and the depth-pose of the view, and eventually it generates reliable depth and pose transformation information. To better evaluate the difference between the synthetic view and the real target view, the synthetic view and the target view are further fed into the discriminator, aiming to distinguish the distributions of latent representations of real and synthetic views, while the generator is trained to force these two distributions resemble each other.

In our adversarial training framework, the potential distribution of the high-dimensional real or reconstructed views from the discriminator stabilizes the training of the GAN based on WGAN-GP loss [42]. On the other hand, if the training is based on the standard cross-entropy loss, it can cause the GAN to collapse easily.

The generator is first trained to minimize the geometric constraint-based loss function $\mathcal{L}_{gc}$, as a content loss to synthesize the reconstructed view. In adversarial learning, the reconstructed image is generated by using the loss function $\mathcal{L}_g$ as follows

$$\mathcal{L}_g = \mathcal{L}_{gc} + \delta(-E[D(y')]), \quad (14)$$

where the real view and the reconstructed view are identified by the loss function $\mathcal{L}_d$.

$$\mathcal{L}_d = E[D(y')] - E[D(x)] + \lambda E[(\left\|\nabla_{y^*} D(y^*)\right\|_2 - 1)^2], \quad (15)$$

where $y^* = \epsilon x + (1-\epsilon)y'$, $\epsilon \sim U[0,1]$, x is the real view and y' is the reconstructed view. A detailed definition and derivation of the loss function can be found in the related works [41], [42].

In this paper, minimizing $\mathcal{L}_{gc}$ tends to synthesize a reasonable reconstruction view; in adversarial training, minimizing $\mathcal{L}_g$ tends to have a more similar data distribution for the synthesized view than the target view. By minimizing $\mathcal{L}_d$, we can improve the discriminator's ability to distinguish between a target view and a reconstructed view.

IV. Experiments

In this section, performance of the proposed joint learning and estimation pipeline is validated on both monocular and stereo sequences. We show the qualitative and quantitative evaluation results on the benchmark KITTI [52] dataset and compare with state-of-the-art monocular depth estimators and VO methods. In addition, we also test the generalization ability of the proposed method on the Make3D [53] and NYUDv2 [54] datasets.

The system is trained with the KITTI dataset, which contains 61 video sequences, including 42382 rectified stereo pairs. The original image resolution is downsampled from 1242×375 size to 640×192 size for computational efficiency. Two different data split methods on the KITTI dataset are adopted to evaluate the depth and ego-motion estimation performance, respectively. For monocular depth estimation, we use the Eigen-split scheme [5]. For the quantitative comparison, we employ several metrics that have been used in prior works [5]:

Abs Relative difference: $\frac{1}{T}\sum_{p\in T}\left|d_p - d_p^{gt}\right| / d_p^{gt}$.

Squared Relative difference: $\frac{1}{T}\sum_{p\in T}\left\|d_p - d_p^{gt}\right\|^2 / d_p^{gt}$.

RMSE: $\sqrt{\frac{1}{T}\sum_{p\in T}\left\|d_p - d_p^{gt}\right\|^2}$.

RMSE (log): $\sqrt{\frac{1}{T}\sum_{p\in T}\left\|\log d_p - \log d_p^{gt}\right\|^2}$.

Threshold: % of dp $s.t.$ $\max\left(\frac{d_p}{d_p^{gt}}, \frac{d_p^{gt}}{d_p}\right) = \delta < thr$,

where T is the number of pixels with ground-truth in the test set, d_p^{gt} and d_p are the ground truth and predicted depths, respectively, at a pixel indexed by p. Regarding the evaluation of VO performance, the official KITTI Odometry set is used to train and test our proposed networks.

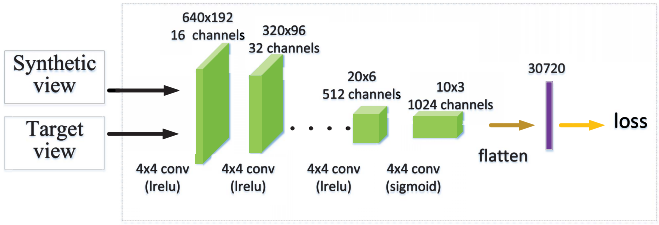

Fig. 3. Adversarial-net structure, where the adversarial loss drives the synthetic view toward the real view, producing perceptually more convincing solutions.

A. Network Architecture

1) Depth-Net: Our depth-net is composed of an encoder and a decoder, which is similar to the network structure in [15]. The skip-connections scheme is adopted to fuse the features from different lower layers of the encoder [55], [56]. The rectified linear unit (ReLU) is used as the activation function after the last prediction layer in the depth-net. The output of our depth-net is the inverse of depth D_{inv}, instead of the depth. Since ReLU might produce a zero prediction, i.e., for an infinite depth output in the network, we use the formula $D = 1/\left(D_{inv} + 10^{-4}\right)$ to convert the inverse of depth to depth.

2) Pose-Net: Our pose-net has a similar network structure to the one described in [34]. The input of the pose-net is two consecutive monocular frames, and the six degrees of freedom (DoF) pose matrix is the regression output, which can be converted to a 4×4 transformation matrix. The pose-net contains seven convolutional layers and three fully connection layers.

3) Adversarial-Net: For the adversarial network, we design a unique fully convolutional network (FCN) that is combined with the flatten operation. The details of the network are shown in Figure 3. A total of seven convolutional layers are used, and every convolution layer in the adversarial network are activated with a leaky ReLU except for the last layer, which uses the sigmoid function instead. Finally, the flattening operation is performed on the convolved features. Our discriminator is designed to output the corresponding long latent vector which represents distributions of the real view and the synthetic view.

B. Training Hyper-Parameters

We train our CNNs in the TensorFlow framework [57], and the Adam optimizer [58] is used for parameter optimization, and set $[\beta_1, \beta_2, \epsilon] = \left[0.9, 0.999, 10^{-8}\right]$. The network contains 50 million trainable parameters, and takes approximately 13 hours to train using a single NVIDIA RTX 2080ti. The training typically converges after approximately 200K iterations. The initial learning rate of all training networks is set to be 0.001. When the network iterates to 20% of the total iteration number, the learning rate is set to be half of the original value. The weights in the geometric constraint loss function is set to $\left[\alpha, \beta, \gamma\right] = [1, 1, 0.1]$, which empirically ensures a high stability in our training process.

For the WGAN-GP, λ is set as 10 for the gradient penalty as in [42]. In adversarial training, we set the weight δ to 0.01. It should be noted that in the mixed loss function $\mathcal{L}_{s-l1}$, the ratio of SSIM to $L1$ is 0.85:0.15.

C. Depth Estimation Results

1) Results on KITTI: The Eigen-spilt method selects 697 images from the 28 sequences in KITTI as the test datasets for monocular depth estimation. The rest of the 33 scenes containing 23,488 stereo pairs are used for training. We use the same settings as configured in [33], resulting in 23,455 temporal stereo pairs for training and 697 images from the images split for testing.

To obtain a fair comparison, the 80 meters is used as the maximum depth threshold values for metric evaluation. Table I and Figure 4 present the error measurements of different methods and the visualization results of estimated depth maps, respectively. As shown in Table 1, the methods trained on the KITTI raw dataset are denoted by K. D denotes depth supervision, S denotes stereo input pairs, M denotes the monocular video sequence and MS means monocular video sequences combined with stereo pairs. Models with additional training data from CityScapes [60]are denoted by CS+K. Our approach performs much better than state-of-the-art depth estimators in terms of lower estimation errors and higher accuracy.

For the algorithms trained with depth ground truth [5], [7], the direct learning from the input image to depth depends entirely on the fitting ability of the neural network. However, the depth samples are limited and there is no additional geometric constraint; therefore, the generalization ability of supervised methods is not satisfied.

Unlike the algorithms for training monocular sequences [2], [50], which use only temporal domain information, our proposed method makes full use of the geometric constraints of temporal-spatial domain information to achieve better results.

Compared with the binocular sequence algorithms, we use the geometric constraints in the traditional DVO algorithm to optimize the pose estimation in the spatial domain, and provide additional error propagation for the depth estimation network to achieve better poses.

In addition, in order to better integrate the temporal constraint with the spatial constraint, we concatenate the synthetic views of the spatial domain and the temporal domain, and optimize with the adversarial training to approximate the distribution of the target view, resulting in favorable performances for depth estimation. As seen in Figure 4, we obtain better fine-grained depth estimation than that of works in [2], [15] and [33], and the object boundaries are preserved much better.

2) Ablation Studies: Table II shows an ablation study on depth estimation for our method, which illustrate the importance of each component of the proposed system. Among them, *'temporal'* denotes training process with only monocular video sequences. *'spatial'* denotes training with binocular video sequences, where stereo pairs provide baseline approximations. *'direct-vo'* denotes the use of DVO in the framework with geometric constraints imposed on the temporal domain, and *'adv-learn'* denotes the adversarial training of synthetic views and target views.

TABLE I

QUANTITATIVE COMPARISONS OF OUR METHOD WITH THE METHODS REPORTED IN THE LITERATURE ON THE TEST SET OF THE KITTI RAW DATASET USED BY THE EIGEN-SPILT

Method	dataset	Supervision	Error Metric (lower,better)				Accuracy Metrics (higher,better)		
			Abs Rel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Depth: cap 80m									
Eigen *et al.* [5] (Fine)	K	D	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al.* [7]	K	D	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Xu *et al.* [13]	K	D	0.122	0.897	4.677	-	0.818	0.954	0.985
Kuznietsov *et al.* [19]	K	D	0.113	0.741	4.621	0.189	0.862	0.960	0.986
Zhou *et al.* [2]	K	M	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Yin *et al.* [26]	K	M	0.164	1.303	6.090	0.247	0.765	0.919	0.968
Wang *et al.* [50]	K	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Garg *et al.* [14]	K	S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Pilzer *et al.* [44]	K	S	0.152	1.388	6.016	0.247	0.789	0.918	0.965
Godard *et al.* [15]	K	S	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Li *et al.* [34]	K	MS	0.183	1.730	6.570	0.268	-	-	-
Zhan *et al.* [33]	K	MS	0.144	1.391	5.869	0.241	0.803	0.928	0.969
Yang *et al.* [27]	K	MS	0.127	1.239	6.247	0.214	0.847	0.926	0.969
Godard *et al.* [59]	K	MS	0.127	1.031	5.266	0.221	0.836	0.943	0.974
Ours	K	MS	**0.115**	**1.019**	**5.121**	**0.213**	**0.843**	**0.945**	**0.976**
Yin *et al.* [26]	CS+K	M	0.153	1.328	5.737	0.232	0.802	0.934	0.972
Wang *et al.* [50]	CS+K	M	0.148	1.187	5.496	0.226	0.812	0.938	0.975
Godard *et al.* [15]	CS+K	S	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Aleotti *et al.* [46]	CS+K	S	0.124	1.055	5.289	0.220	0.847	0.942	0.973
Yang *et al.* [27]	CS+K	MS	0.114	1.074	5.836	0.208	0.856	0.939	0.976
Ours	CS+K	MS	**0.108**	**0.901**	**5.023**	**0.206**	**0.851**	**0.949**	**0.978**

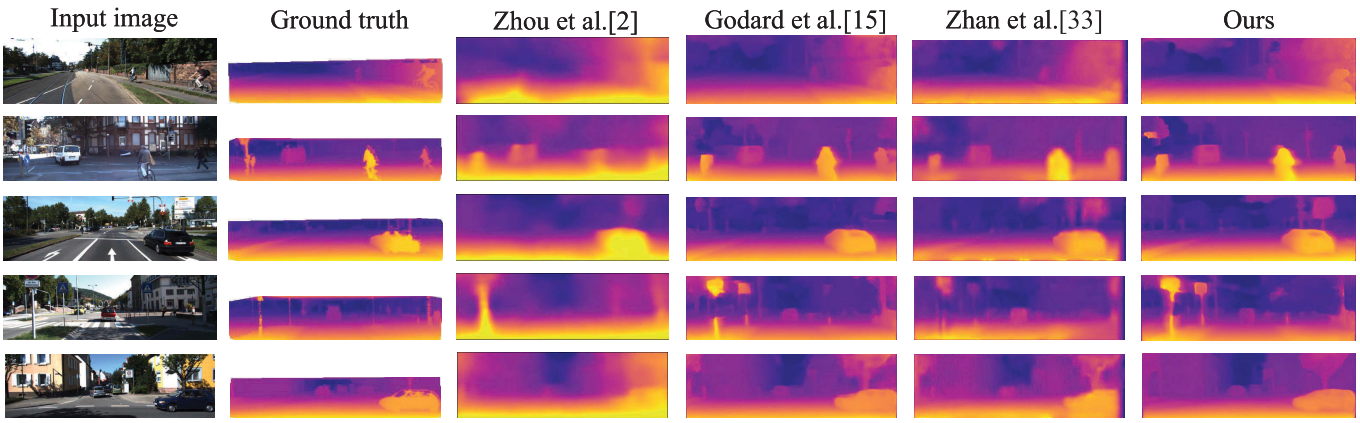

Fig. 4. Qualitative comparison results on the KITTI dataset. Here, ground truth depth maps are interpolated for visualization purposes. Compared to Zhou *et al.* [2], Godard *et al.* [15] and Zhan *et al.* [33], our depth map prediction preserves more details and provides a more precise reconstruction of objects, such as the van, person, guidepost and tree.

The comparison of the experimental results of the two methods demonstrates that each component is beneficial for improving the depth estimation performance. The comparison of Method T and Method TS shows that the geometric constraints from spatial domain are more powerful than the geometric constraints of the temporal one, since the baseline of the stereo pair provides accurate spatial geometric constraints, while the temporal domain geometric constraints suffer from the mutual coupling of depth and pose.

A comparison of Method TS and Method TS+Adv shows that adversarial training can learn the relevant structural information between the synthesis view and the target view and provide effective feedback for the deep network, similar to the conclusion reached by Kumar *et al.* [45]. The difference is that we use a more advanced adversarial loss function and a more effective self-supervised learning optimization scheme, without using the depth ground truth information for training. Method TS+DVO shows that the joint usage of spatial-temporal geometric constraints can provide better feedback for the depth estimator and further improve performance.

Method TS+DVO+Adv demonstrates that the best performance of depth estimation is achieved when using the

TABLE II

ABLATION STUDY ON THE TEST SET OF THE KITTI RAW DATASET USED BY THE EIGEN-SPILT SCHEME. THE RESULTS ARE CAPPED AT 80 M DEPTH. MEANS THAT THE CONSTRAINT IS USED, ADV* REPRESENTS THE RESULT OF USING VANILLA GAN

	Method	adv-learn	direct-vo	spatial	temporal	Error Metric (lower,better)				Accuracy Metrics (higher,better)		
						Abs Rel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Mono	T(Temporal)				✓	0.185	1.612	6.563	0.272	0.721	0.902	0.961
Mono + Stereo	TS(Temporal+Spatial)			✓	✓	0.139	1.153	5.562	0.232	0.810	0.929	0.969
	TS+Adv	✓		✓	✓	0.131	1.139	5.481	0.227	0.830	0.935	0.971
	TS+DVO		✓	✓	✓	0.126	1.078	5.363	0.218	0.836	0.943	0.973
	TS+DVO+Adv*	✓	✓	✓	✓	0.125	1.081	5.352	0.216	0.835	0.943	0.973
	TS+DVO+Adv	✓	✓	✓	✓	0.115	1.019	5.121	0.213	0.843	0.945	0.976

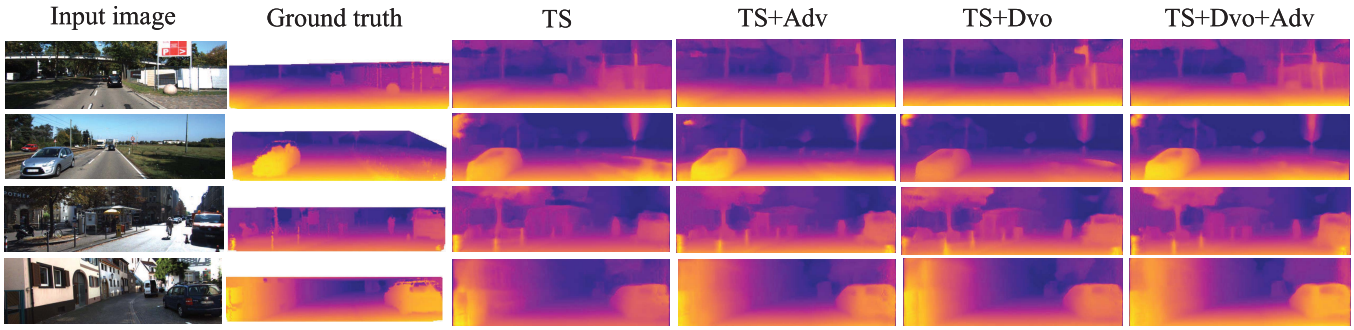

Fig. 5. Qualitative ablation study. We can see that our model with all added components results in better depth artifacts.

combination of all the components. Furthermore, we are not the first to use generative adversarial ideas for such tasks. In [46], the authors tried to use a vanilla GAN in a similar framework; however, the improvement in the results shown in the paper was not significant. Initially, we also used a vanilla GAN in our framework, and similarly, the effect showed a limited improvement. As it does not always obtain better results, we analyzed the reasons for these findings.

The discriminator aims to distinguish whether the warped images are plausible or not. If we use 0/1 to judge, there is no way to give the generator an accurate signal, the reason is that both the target image and reconstruction images are high-dimensional distributions. To naively classify it as only two categories would fail to capture geometric details of the object. Alternatively, our discriminator is designed to output the corresponding long latent vector that represents the distributions of the target image and reconstruction images, and we use WGAN-GP as loss functions for our framework. Therefore, our discriminator is employed to distinguish the distributions of latent representations of the target image and reconstruction image, while the generator is trained to make the two distributions resemble each other.

In addition, in the ablation studies, we have expanded the qualitative comparison, as shown Figure 5, and we can see that our model with all the added components results in better depth artifacts.

3) Results on Other Datesets: To evaluate the generalization ability of our monocular depth estimation model, we further test the model trained by KITTI on the test dataset Make3D [53] directly. Quantitative evaluation results are provided in Table III, from which we can state that our model obtains

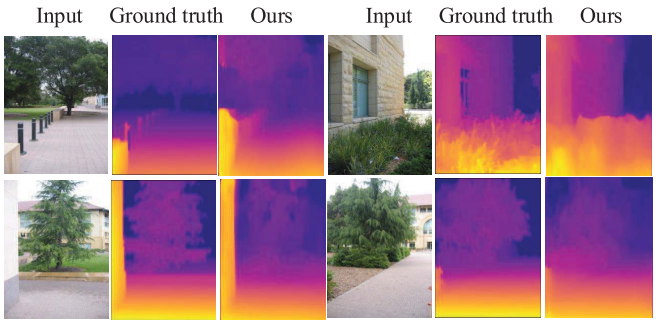

Fig. 6. Qualitative results on the Make3D dataset.

comparable and even better performance than most of the other methods. It is worth noting that the models used as comparisons in Table III are not trained on the Make3D dataset; analogous to the proposed approach,they use only Make3D to demonstrate the generalization capabilities. From the experiments, we can say that the global scene structures are well captured by our model. As shown in Figure 6, object details in the image can be estimated precisely. In addition, we tested the model trained by KITTI directly on the widely used authoritative indoor dataset NYU-Dv2 [54], and the results are shown in 7. We also obtained reasonable depth predictions on the indoor dataset.

D. Pose Estimation Results

To obtain a fair comparison both qualitatively and quantitatively, we used the same KITTI Odometry dataset as in Sfm [2] for training. The KITTI Odometry-split provides only ground

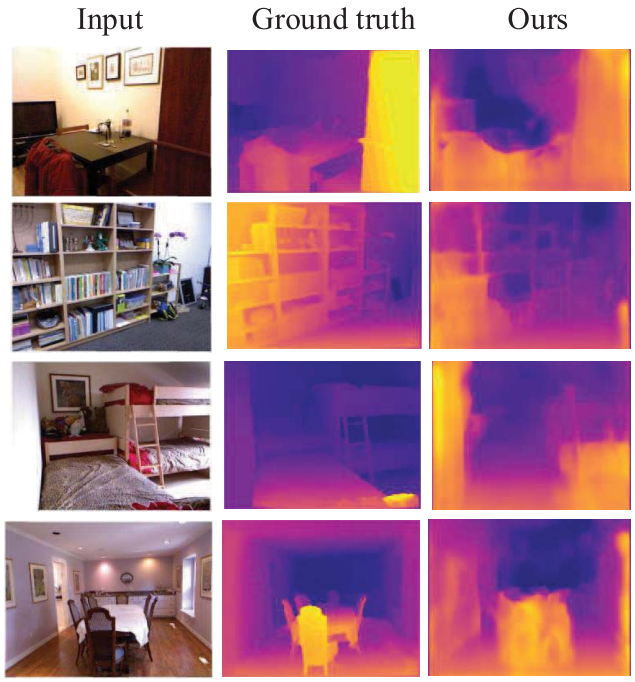

Fig. 7. Qualitative results on the NYU-Dv2 dataset.

TABLE III

RESULTS ON THE MAKE3D DATASET. THE ERRORS ARE COMPUTED ONLY FOR PIXELS IN A CENTRAL IMAGE CROP WITH GROUND TRUTH DEPTH LESS THAN 70 METERS. ALL OF MODELS ARE NOT TRAINED ON THE MAKE3D DATASET AND USE ONLY MAKE3D TO EVALUATE THE GENERALIZATION CAPABILITIES

Method	Supervision	Abs Rel	SqRel	RMSE	RMSE (log)
Train set mean	depth	0.876	13.98	12.27	0.377
Liu *et al.* [7]	depth	0.475	6.562	10.05	0.165
Laina *et al.* [8]	depth	0.204	1.840	5.683	0.084
Wang *et al.* [50]	none	0.387	4.72	8.09	0.204
Zhou *et al.* [2]	none	0.383	5.321	10.47	0.478
Godard *et al.* [15]	stereo	0.544	10.94	11.76	0.193
Ours	stereo	0.306	4.681	7.712	0.126

truth camera poses of 11 videos indexed from 00 to 10. We use videos indexed from 00 to 08 for training, and the rmaining are used for testing.

We use monocular images to test all the methods under comparison. Since SfM [2], VISO2-M [61] and ORB-SLAM [62] have the scale ambiguity problem, the absolute scale of camera poses and depth maps cannot be recovered. Hence, we conduct postprocessing to align their results with the ground truth. Referring to DFR-F [33], the reconstructed map by ORB-SLAM and VISO2-M are rescaled to match the ground truth map according to the standard protocol. The pose estimation between frames in SfM [2] works only in short videos that last for five frames. By tuning the scale factor, each of the short videos is aligned to the ground truth independently. Therefore, the measured error for [2] represents only the relative translation error for short sequences.

TABLE IV

VISUAL ODOMETRY RESULTS EVALUATED ON SEQUENCE 09 AND 10 OF THE KITTI ODOMETRY DATASET. LC DENOTES LOOP CLOSURE. t_{err} IS AVERAGE TRANSLATIONAL RMSE DRIFT (%). r_{err} IS AVERAGE ROTATIONAL RMSE DRIFT (°/100M)

Method	Seq. 09		Seq. 10	
	t_{err} (%)	r_{err} ($^\circ/100m$)	t_{err} (%)	r_{err} ($^\circ/100m$)
VISO2-M. [61]	9.60	1.63	14.62	4.39
ORB-SLAM. [62]	15.30	0.26	3.69	0.48
Sfm. [2]	10.72	4.83	13.16	3.78
DFR-F [33](Temporal)	11.93	3.91	12.55	3.45
DFR-F [33] (Full)	11.92	3.62	12.67	3.40
Ours(TS+Dvo)	10.68	3.72	12.32	4.21
Ours(TS+Adv)	9.16	3.18	10.56	3.98
Ours	7.62	1.13	4.58	1.62

TABLE V

QUANTITATIVE EVALUATION OF THE ODOMETRY TASK USING THE METRIC OF THE ABSOLUTE TRAJECTORY ERROR

Method	frames	Seq.09	Seq.10
Yin *et al.* [26]Geo-net	5	0.012 ± 0.007	0.012 ± 0.009
Zou *et al.* [25]DF-net	5	0.017 ± 0.007	0.015 ± 0.009
Zhou *et al.* [2]	5	0.021 ± 0.017	0.020 ± 0.015
Mahjourian *et al.*(no ICP) [21]	3	0.014 ± 0.010	0.013 ± 0.011
Mahjourian *et al.*(with ICP) [21]	3	0.013 ± 0.010	0.012 ± 0.011
Ours(FULL)	5	0.012 ± 0.007	0.013 ± 0.008

Since our approach jointly optimizes DVO and the absolute scale from stereo images, we do not need any postprocessing for scale alignment in evaluation. We obtain the estimated framewise camera pose by applying the proposed approach to the entire sequences. Based on the evaluation protocol of the KITTI VO dataset, we use possible sub-sequences with the length (100, 200, …, 800). Table IV shows the average translation error and the rotation error for testing the sequences 09 and 10. As seen from Table IV, our stereo vision-based VO learning approach obtains superior results to those of the monocular learning method in SfM [2], VISO2-M [61] and ORB-SLAM [62]. Compared with the other stereo vision-based method DFR-F [33], our performance is the best. Note that, we do not have the loop closure step in our framework. In Table IV, we also show the effect of adding various optimization modules on the pose estimation through the ablation study. It can be seen that the performance improvement of the joint optimization of all modules is the most significant.

In addition, we evaluate our pose estimation performance using the average of the absolute trajectory error. The results are shown in Table V. Compared to other deep learning methods, we have achieved similar or better results.

Since the KITTI sequence is recorded by cameras that are mostly mounted on moving cars, the camera coordinate y representing the camera height is almost constant. Thus, only the z, x components in the trajectories generated by the different methods are displayed, as shown in Figure 8. All methods are tested with monocular images. The purple title is

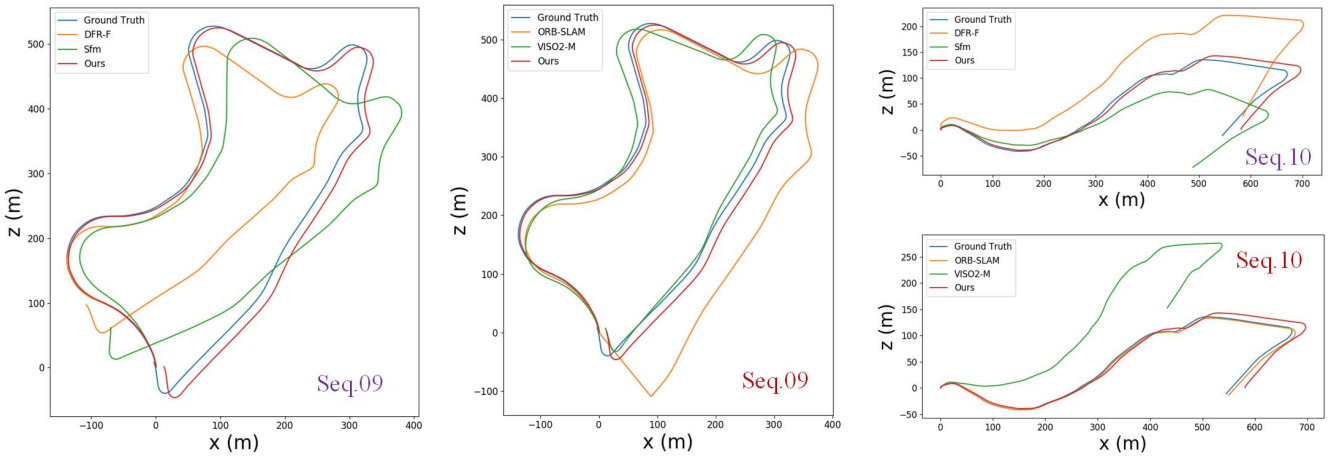

Fig. 8. Qualitative result on visual odometry. Full trajectories on the testing sequences (09, 10) are plotted.

the deep learning method, and the red title is the traditional algorithm, our method is the better among all methods.

V. Conclusion

We presented a self-supervised learning framework for monocular depth estimation and monocular VO measurement. To obtain the depth estimation of images without the scale ambiguity, stereo image pairs with known transform poses are used for training. When the absolute scale is determined, VO can be computed by minimizing the photometric warp error between consecutive frames in sequences. We note that DVO can calculate the camera pose matrix between frames without learning the pose transform by CNN models. By incorporating DVO, our learning strategy requires fewer parameters and promising for many applications. In addition, our pipeline uses a view synthesis framework based on spatial and temporal geometric constraints as a generator, and a discriminator network is used to distinguish the synthetic view from the real target view, while a global joint optimization is performed by adversarial learning. The experimental results show that our model obtains a more accurate pose reference and a fine-grained dense depth map estimation.

In the future, we will consider solving the problems of occlusions, nonrigid objects, and dynamic scenarios in the framework. The SfM technique can be incorporated in the pipeline. Moreover, since no bundle adjustment (BA) is implemented in the proposed system, our VO estimation performance is worse than the state-of-the-art SLAM systems. Thus, integrating BA into our pipeline is another plan.

References

[1] J. Xie, R. Girshick, and A. Farhadi, "Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 842–857.

[2] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.

[3] F. Steinbrucker, J. Sturm, and D. Cremers, "Real-time visual odometry from dense RGB-D images," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 719–722.

[4] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[5] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2366–2374.

[6] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.

[7] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[8] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 239–248.

[9] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1119–1127.

[10] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2800–2809.

[11] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, and A. L. Yuille, "Surge: Surface regularized geometry estimation from a single image," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 172–180.

[12] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale Continuous CRFs as sequential deep networks for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5354–5362.

[13] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3917–3925.

[14] R. V. K. B. G. Garg and G. I. Carneiro Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 740–756.

[15] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.

[16] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 817–833.

[17] Y. Luo *et al.*, "Single view stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 155–163.

[18] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 484–500.

[19] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6647–6655.

[20] A. Wang, Z. Fang, Y. Gao, X. Jiang, and S. Ma, "Depth estimation of video sequences with perceptual losses," *IEEE Access*, vol. 6, pp. 30536–30546, 2018.

[21] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5667–5675.

[22] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.

[23] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2800–2810.

[24] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia, "Learning monocular depth estimation infusing traditional stereo knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9799–9809.

[25] Y. Zou, Z. Luo, and J.-B. Huang, "DF-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 36–53.

[26] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1983–1992.

[27] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Every pixel counts: Unsupervised geometry learning with holistic 3D motion understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 691–709.

[28] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single RGB images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3372–3380.

[29] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 53–69.

[30] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2938–2946.

[31] S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 2043–2050.

[32] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "VINet: Visual-inertial odometry as a sequence-to-sequence learning problem," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 3995–4001.

[33] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. M. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 340–349.

[34] R. Li, S. Wang, Z. Long, and D. Gu, "UnDeepVO: Monocular visual odometry through unsupervised deep learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 7286–7291.

[35] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2016, pp. 1–16.

[36] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Controllable text generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 1587–1596.

[37] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 702–716.

[38] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[39] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 2172–2180.

[40] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 2539–2547.

[41] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1–32.

[42] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5767–5777.

[43] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham, and N. Trigoni, "3D object reconstruction from a single depth view with adversarial learning," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 679–688.

[44] A. Pilzer, D. Xu, M. Puscas, E. Ricci, and N. Sebe, "Unsupervised adversarial depth estimation using cycled generative networks," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 587–595.

[45] A. C. Kumar, S. M. Bhandarkar, and M. Prasad, "Monocular depth prediction using generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 300–308.

[46] F. Aleotti, F. Tosi, M. Poggi, and S. Mattoccia, "Generative adversarial networks for unsupervised monocular depth prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 337–354.

[47] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deep monocular depth estimation via integration of global and local predictions," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4131–4144, Aug. 2018.

[48] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[49] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 2017–2025.

[50] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2022–2030.

[51] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, Feb. 2004.

[52] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[53] A. Saxena, M. Sun, and A. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.

[54] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.* Florence, Italy: Springer, 2012, pp. 746–760.

[55] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.

[56] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 447–456.

[57] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: https://arxiv.org/abs/1603.04467

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[59] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3828–3838.

[60] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[61] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3D reconstruction in real-time," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 963–968.

[62] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[63] A. Wang, Y. Gao, Z. Fang, X. Jiang, S. Wang, S. Ma, and J. Hwang, "Unsupervised learning of depth and ego-motion with spatial-temporal geometric constraints," in *Proc. IEEE Conf. Multimedia Expo. (ICME)*, Jul. 2019, pp. 1798–1803.

Anjie Wang received the master's degree from the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China, in 2019. He is currently a Visiting Scholar with the Institute of Digital Media, Peking University, Beijing, China. His research interests include SLAM, computer vision, and machine learning.

Zhijun Fang (Senior Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China. He is currently a Professor and the Dean with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science. His current research interests include image processing, video coding, and pattern recognition. He was the General Chair of the Joint Conference on Harmonious Human Machine Environment (HHME) 2013 and the General Co-Chair of the International Symposium on Information Technology Convergence (ISITC) 2014, 2015, 2016, 2017. He received the GanPo 555 Talents Program, the One-Hundred, the One-Thousand, and the Ten-Thousand Talent Project Award of Jiangxi province.

Yongbin Gao received the Ph.D. degree from Chonbuk National University, South Korea. He is currently a Faculty Member with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. He has published numerous SCI articles in prestigious journals, such as *Information Science* and *Pattern Recognition Letters*, in the area of image processing, patter recognition, and computer vision.

Songchao Tan received the B.S. and Ph.D. degrees in computer science and technology from the Dalian University of Technology, Dalian, China. He currently holds a postdoctoral position with Peking University. His research interests include video coding and quality assessment.

Shanshe Wang received the B.S. degree from the Department of Mathematics, Heilongjiang University, Harbin, China, in 2004, the M.S. degree in computer software and theory from Northeast Petroleum University, Daqing, China, in 2010, and the Ph.D. degree in computer science from the Harbin Institute of Technology. He held a postdoctoral position at Peking University, Beijing, from 2016 to 2018. He joined the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, where he is currently a Research Assistant Professor. His current research interests include video compression and image and video quality assessment.

Siwei Ma (Member, IEEE) received the B.S. degree from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. From 2005 to 2007, he held a postdoctoral position with the University of Southern California, Los Angeles, USA. Then, he joined the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, where he is currently a Professor. He has authored more than 100 technical articles in refereed journals and proceedings in the areas of image and video coding, video processing, video streaming, and transmission.

Jenq-Neng Hwang (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree from the University of Southern California. In 1989, he joined the Department of Electrical and Computer Engineering (ECE), University of Washington, Seattle, WA, USA, where he has been promoted to as a Full Professor, in 1999. He was the Associate Chair for Research, from 2003 to 2005 and from 2011 to 2015. He is currently the Associate Chair for Global Affairs and International Development with the ECE Department, University of Washington. He is also the Founder and the Co-Director of the Information Processing Laboratory, which received several AI City Challenges Awards. He has written more than 330 journals, conference articles, and book chapters in the areas of machine learning, multimedia signal processing, and multimedia system integration and networking. He has authored a textbook *Multimedia Networking: From Theory to Practice* (Cambridge University Press). He has close working relationship with the industry on multimedia signal processing and multimedia networking.

Dr. Hwang is currently a Founding Member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society. He is also a member of the Multimedia Technical Committee of the IEEE Communication Society and the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society. He received the 1995 IEEE Signal Processing Society's Best Journal Paper Award. He was the Program Co-Chair of the ICASSP 1998 and the ISCAS 2009. He has served as the Program Co-Chair of the IEEE ICME 2016. He is currently on the Editorial Board of the ZTE Communications, ETRI, IJDMB, and JSPS journals. He was the Society's Representative of the IEEE Neural Network Council, from 1996 to 2000. He has served as an Associate Editor for the IEEE T-SP, T-NN, T-CSVT, T-IP, and the *IEEE Signal Processing Magazine*.