SDOF-GAN: Symmetric Dense Optical Flow Estimation With Generative Adversarial Networks

Tongtong Che, Yuanjie Zheng, *Member, IEEE*, Yunshuai Yang, Sujuan Hou, *Member, IEEE*, Weikuan Jia, Jie Yang, *Member, IEEE*, and Chen Gong, *Member, IEEE*

***Abstract*— There is a growing consensus in computer vision that symmetric optical flow estimation constitutes a better model than a generic asymmetric one for its independence of the selection of source/target image. Yet, convolutional neural networks (CNNs), that are considered the *de facto* standard vision model, deal with the asymmetric case only in most cutting-edge CNNs-based optical flow techniques. We bridge this gap by introducing a novel model named SDOF-GAN: symmetric dense optical flow with generative adversarial networks (GANs). SDOF-GAN realizes a consistency between the forward mapping (source-to-target) and the backward one (target-to-source) by ensuring that they are inverse of each other with an inverse network. In addition, SDOF-GAN leverages a GAN model for which the generator estimates symmetric optical flow fields while the discriminator differentiates the "real" ground-truth flow field from a "fake" estimation by assessing the flow warping error. Finally, SDOF-GAN is trained in a semi-supervised fashion to enable both the precious labeled data and large amounts of unlabeled data to be fully-exploited. We demonstrate significant performance benefits of SDOF-GAN on five publicly-available datasets in contrast to several representative state-of-the-art models for optical flow estimation.**

***Index Terms*— Symmetric optical flow estimation, generative adversarial networks, forward-backward consistency, semi-supervised learning.**

Manuscript received December 18, 2019; revised January 3, 2021 and April 6, 2021; accepted May 10, 2021. Date of current version July 6, 2021. This work was supported by in part by the National Natural Science Foundation of China under Grant 81871508, Grant 61773246, Grant 61973162, and Grant 62072289; in part by the Major Program of Shandong Province Natural Science Foundation under Grant ZR2019ZD04 and Grant ZR2018ZB0419; in part by the Taishan Scholar Program of Shandong Province of China under Grant TSHW201502038; in part by the Fundamental Research Funds for the Central Universities under Grant 30920032202; in part by the National Key Research and Development Program of China under Grant 2019YFB1311503; and in part by the Committee of Science and Technology, Shanghai, China, under Grant 19510711200. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Denis Kouame. *(Corresponding author: Yuanjie Zheng.)*

Tongtong Che, Yuanjie Zheng, Yunshuai Yang, Sujuan Hou, and Weikuan Jia are with the School of Information Science and Engineering, Shandong Normal University, Jinan 250300, China, also with Key Laboratory of Intelligent Computing and Information Security in Universities of Shandong, Shandong Normal University, Jinan 250300, China, and also with the Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Shandong Normal University, Jinan 250300, China (e-mail: tong.che@qq.com; zheng.vision@gmail.com; 1520119848@qq.com; hsj1985@126.com; jwk_1982@163.com).

Jie Yang is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jieyang@sjtu.edu.cn).

Chen Gong is with the PCA Laboratory, Nanjing University of Science and Technology, Nanjing 210094, China, and also with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: chen.gong@njust.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3084073

I. Introduction

OPTICAL flow estimation is a challenging task and has been an active field of research in computer vision. It establishes a meaningful spatial mapping between a pair of input images (the source image and the target image) [1]–[3] to facilitate the subsequent tasks such as motion estimation, object tracking, video compression, self-localization, etc. The general approaches of dense optical flow estimation result in an asymmetric estimation of flow fields due to their dependence on the choice of the target image. Asymmetric optical flow estimation only considers the flow field along a single direction, namely from image I_1 to image I_2 or from image I_2 to image I_1. However, the results will differ if a unidirectional flow field is applied from images I_1 to I_2 and from images I_2 to I_1, which is obviously unreasonable. To address this drawback, some symmetric optical flow estimation methods have been proposed to jointly estimate the flow fields in both forward and backward directions while constraining that these bidirectional flow fields are inverse of each other [4]–[6] (see Fig. 1). Symmetric optical flow estimation techniques have been widely validated to be capable of improving flow estimation accuracy. However, the goal of existing symmetric methods is to optimize an expensive non-convex objective function, and solving this optimization is hence unavoidably time-consuming.

With the surge of deep learning applied to resolve the problems in different fields, a variety of convolutional neural networks (CNNs) or generative adversarial networks (GANs) have also been proposed for solving optical flow estimation problems. CNNs for optical flow estimation can be divided into three categories: supervised learning approaches, unsupervised learning approaches, and semi-supervised learning approaches. Supervised learning approaches exploit a large-scale ground-truth flow dataset to train a deep CNN in an end-to-end manner [7]–[11]. They are usually optimized by minimizing the differences between the predicted

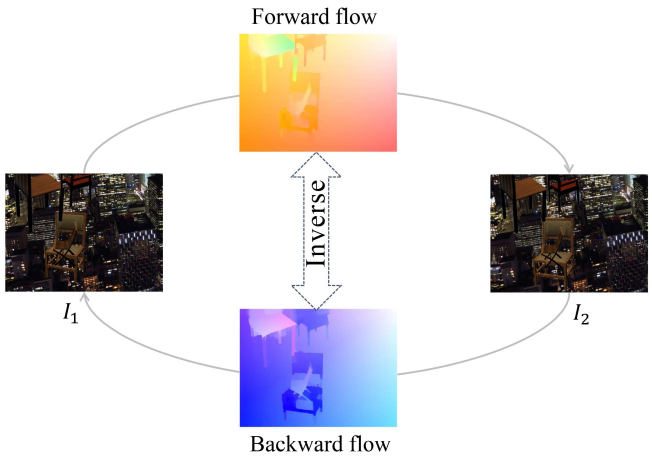

Fig. 1. A schematic diagram of symmetric dense optical flow estimation. Given a pair of input images, the goal is to estimate the forward flow and backward flow jointly while enforcing a consistency constraint, *i.e.*, these bidirectional transformations are inverse of each other.

flow and the ground-truth flow. For these methods, synthetic datasets are commonly used to avoid the high labor cost in creating ground-truth flow fields. In contrast, unsupervised learning methods are free from the requirement of supervision information (*i.e.*, the ground-truth flow), and the training process is driven by measuring brightness constancy and spatial smoothness of the estimated flow fields [12]–[16]. However, the assumption of brightness constancy clearly lacks the robustness to complex motion, especially on the boundaries of objects in motion. Although the unsupervised learning algorithms show a potential to outperform supervised learning approaches in certain conditions (*e.g.*, the images with simple scenarios), they still cannot exceed the performance of the classical optical flow estimation method in complex motion scenarios. In contrast, semi-supervised learning techniques [17]–[19] are proposed to take full advantages of both labeled data and unlabeled data. Besides CNNs, GANs have also been used to resolve the optical flow estimation [18], [20]–[22] or the closely-related image registration tasks [23], [24] (as to be detailed in Sec. II). They achieve superior performances to conventional CNNs. However, it is worthy to point out that the training of GANs is generally unstable and complex, as they cannot control the reciprocity of the forward and backward flows.

Therefore, in this paper, we propose a novel algorithm for estimating symmetric dense optical flow with generative adversarial networks, which is termed "SDOF-GAN". SDOF-GAN model benefits from the symmetric constraint that realizes a consistency between the forward mapping (source-to-target) and the backward one (target-to-source) by ensuring that they are inverse of each other with an inverse network. This significantly reduces the function searching space, so the optimal solution can be found in a stable and efficient way. In addition, we exploit the adversarial mechanism [25] to learn the pattern of error map between the target image and the one warped from the source image using the "real" flow field or a "fake" estimation. This replaces the brightness constancy widely used in general CNN-based optical flow estimation methods, allowing for an end-to-end estimation of the flow field, and eliminating the requirement of the time-costing energy function minimization process. Moreover, we formulate our SDOF-GAN for symmetric optical flow estimation in a semi-supervised learning manner. The generator of our SDOF-GAN is trained by making use of supervised information as well as unsupervised cues to incorporate the best of both worlds. This is better than the traditional supervised methods [7]–[9] as the synthetic datasets used by supervised methods do not match the characteristics of real data, and is also better than the unsupervised methods [12]–[16] as the training process is completely driven by measuring brightness constancy and spatial smoothness of the estimated flow fields. Compared with the general GAN for dealing with various unsupervised tasks, our SDOF-GAN is designed under the semi-supervised setting which introduces scarce yet valuable supervision information. This also helps to make the training of SDOF-GAN stable and efficient. The discriminator is trained in a supervised learning manner that distinguishes the pattern of the error map between the target image and the one warped from the source image using the ground-truth flow or the estimated flow. The experimental results on five publicly available datasets show that the SDOF-GAN outperforms several representative state-of-the-art methods.

The major contributions of our work are as follows:

- We introduce an early deep learning framework for accomplishing symmetric optical flow estimation. It is realized by an inverse network with a forward-backward consistency constraint and can improve the estimation accuracy by eliminating the estimation bias of single-directional optical flow estimation techniques.
- We bring in a GAN structure targeting a more accurate estimation of symmetric optical flow. It replaces the brightness constancy assumption used commonly in traditional optical flow methods with a discrimination capability to capture the spatial structure of flow-warp-error map.
- We equip our symmetric optical flow estimation with a semi-supervised learning strategy which harnesses the precious labeled data as well as large amounts of unlabeled data.

II. Related Work

In this part, we review the works that are related to our research from the following three aspects.

A. Dense Optical Flow Estimation

Traditionally, the variational approach [2] is widely adopted for dense optical flow estimation by minimizing an objective function which captures the optical flow constraint and maintains smoothness constraint. For example, Weinzaepfel *et al.* [26] proposed a descriptor matching algorithm that blends a matching algorithm with a variational approach for estimating optical flow. Besides, a multi-resolution framework [27] is designed for the accurate optical flow estimation. One serious drawback of these

methods lies in the high computational cost. To overcome time-consuming estimation of features, an effective method [28] leveraging coded motion information has been proposed to obtain fast and high-quality motion field estimation. Moreover, CNN-based framework such as FlowNet [7] is proposed, which learns to estimate optical flow by training on ground-truth data and a large synthetic data in a supervised manner. Subsequently, an end-to-end learning method of optical flow, *i.e.*, FlowNet2.0 [8], has been proposed of which the quality and speed of the optical flow estimation are improved when compared with FlowNet. These supervised methods require a lot of data with ground-truth that are manually synthesized. Unfortunately, most synthetic data do not reflect the complexity of real data, such as brightness differences, occlusion, noise, etc. Besides, the flow learned by these supervised methods might be biased because of the selected ground-truth flow. To overcome this problem, the unsupervised learning framework was proposed to exploit the unlabeled data. For example, Jason *et al.* [13] presented an end-to-end CNN framework for optical flow prediction by measuring photometric constancy over time, and models the expected variation of flow across the two images. However, most of the unsupervised learning approaches require the assumptions of brightness uniformity and spatial smoothness as well as a coarse-to-fine image alignment loss, which leads to high computational complexity. Since both supervised learning and unsupervised learning methods have their own shortcomings in the training stage, some semi-supervised methods that combine their advantages have emerged. In [17], the sparse ground-truth data are used for supervised depth estimation on single image, while the deep network is developed to predict dense depth maps through an unsupervised image alignment loss. Besides, Lai *et al.* [18] leveraged both the labeled and unlabeled data for motion analysis on optical flow, which does not rely on the assumptions of brightness constancy and spatial smoothness.

In summary, existing deep-learning methods are typically used to estimate the optical flow in a single direction and ignore the inverse-consistent property of flow between two images. Inspired by traditional inverse-consistent approaches [4]–[6], we propose to estimate the flow fields from two reciprocal directions simultaneously and enforce the consistency constraint to ensure that these bidirectional flow fields are inverse mappings of one another.

B. Symmetric Transformation Estimation

Symmetric transformation estimation can help to deal with the problems caused by the bias in generic directional transformation estimation. Up to now, a variety of symmetric transformation estimation algorithms have been devised to calculate the displacement or flow between two images in the field of computer vision. In terms of image registration, the work of [29] jointly estimates the forward and reverse transformations between two images by linear-elasticity, and it gives satisfactory registration results. Zhang [30] proposed an inverse-consistent deep network (ICNet) for unsupervised image registration, which enforces that the two images are symmetrically deformed toward one another. For optical flow estimation, a symmetric optical flow method [4] has been proposed to estimate the flow field symmetrically via using a combination of the flows in both directions. Moreover, Hur and Roth [6] utilized forward-backward consistency term for symmetric optical flow estimation, which jointly estimates optical flow in both forward and backward directions. Similarly, Alvarez *et al.* [5] provided a symmetric variational approach to recover a dense flow field map from two images. These methods effectively solve the symmetric transformation estimation and avoid the paranoia of unidirectional deformation. However, most of these methods rely on the assumption on brightness constancy or spatial smoothness, which are usually not satisfied in complex scenarios.

C. Generative Adversarial Networks

GAN was initially developed by Goodfellow *et al.* [25], which contains a discriminator and a generator with adversarial losses. The goal of the generator network is to map random vectors to real images, and the goal of the discriminator is to distinguish the generated images from the real images. In computer vision, GANs have been widely used in various fields such as image synthesis [31]–[33], image translation [34]–[36], and super-resolution [37], [38]. In recent years, the frameworks of GAN have been successfully applied to optical flow [18], [20]–[22] because GAN-based architecture can replace the brightness constancy widely used in general CNN-based optical flow estimation methods, allowing for an end-to-end estimation of the flow field, and eliminating the requirement of the time-costing energy function minimization process. For example, Thakur and Mukherjee [21] proposed a conditional adversarial network for estimating scene flow from stereo images obtained at different time instances. A conditional GAN based semi-supervised single-directional optical flow estimation was proposed [18] using both labeled and unlabeled data. Although these GAN-based models have achieved considerable results, they do not take into account the symmetric properties of the optical flow. Consequently, motivated by [18], here we propose an end-to-end symmetric optical flow estimation method based on GAN framework and enforce the motions of corresponding pixels in the forward flow map and backward flow map to be inverse of one another.

D. Semi-Supervised Learning

Semi-supervised learning is a learning strategy which refers to use readily available unlabeled data to improve supervised learning tasks when the labeled data are scarce or expensive. Chen *et al.* [39] proposed a semi-supervised framework consisted of unsupervised pretraining, supervised fine-tuning on ImageNet, outperforming standard supervised or unsupervised methods. Lai *et al.* [18] leveraged both the labeled and unlabeled data for motion analysis on optical flow, which does not rely on the assumptions of brightness constancy and spatial smoothness. Sedghi *et al.* [40] used a semi-supervised learning strategy for multi-modal image registration, which reduced the requirement for well-registered training data. Hering *et al.* [41] proposed a weakly-supervised approach, which combines the prior information of segmentation labels

with an energy-based distance metric for deformable cardiac image registration. Semi-supervised learning has tremendous practical value because it can alleviate the limitations of costly data annotation and comprehensively utilize both labeled and unlabeled data to achieve better performance than supervised learning with only a handful of labeled data.

III. Problem Definition

The problem of optical flow estimation is traditionally defined as: given two images $\{I_1, I_2\}$, then the goal is to find the flow mapping ϕ that maps one image I_1 into the second image I_2. The mapping of flow field ϕ is optimized by minimizing the following loss function:

$$\min_{\phi} \mathcal{L}_{data}(I_1, I_2, \phi) + \mathcal{L}_{smooth}(\phi), \tag{1}$$

where $\mathcal{L}_{data}$ measures the photometric differences between the warped image according to the predicted flow and its corresponding reference image, and $\mathcal{L}_{smooth}$ is used to control the spatial smoothness of the flow field.

For general symmetric optical flow estimation methods, they aim to find the inverse results by forcing the data term $\mathcal{L}_{data}$ and the regularization term $\mathcal{L}_{smooth}$ to be symmetric if we exchange source images. The design of the network relies on the losses of brightness constancy and spatial smoothness. Mathematically, symmetric optical flow estimation can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{data}(I_1, I_2, \phi^f) &= \mathcal{L}_{data}(I_2, I_1, \tilde{\phi}^f), \\ \mathcal{L}_{data}(I_2, I_1, \phi^b) &= \mathcal{L}_{data}(I_1, I_2, \tilde{\phi}^b), \end{aligned} \tag{2}$$

$$\begin{aligned} \mathcal{L}_{smooth}(\phi^f) &= \mathcal{L}_{smooth}(\tilde{\phi}^b), \\ \mathcal{L}_{smooth}(\phi^b) &= \mathcal{L}_{smooth}(\tilde{\phi}^f), \end{aligned} \tag{3}$$

where ϕ^f is the dense forward flow field from I_1 to I_2, ϕ^b is the dense backward flow field from I_2 to I_1, $\tilde{\phi}^f$ denotes the inverse flow of flow ϕ^f, and $\tilde{\phi}^b$ denotes the inverse flow of flow ϕ^b.

A. Data Term $\mathcal{L}_{data}$

The goal of the data term is to measure the photometric differences between the warped image according to the predicted forward/backward flow and its corresponding reference image. By utilizing the generalized Charbonnier penalty function [13] as the photometric difference metric, the loss of data term can be defined as follows:

$$\mathcal{L}_{data} = \rho(I_1 - \mathcal{T}(I_2, \phi^f)) + \rho(I_2 - \mathcal{T}(I_1, \phi^b)), \tag{4}$$

where $\rho(\cdot)$ denotes a robust difference measurement function to penalize photometric differences. The function $\rho(\cdot)$ is imposed on $I_1 - \mathcal{T}(I_2, \phi^f)$ and $I_2 - \mathcal{T}(I_1, \phi^b)$ that represent forward and backward flow warping errors, respectively. The function $\mathcal{T}(\cdot)$ is a spatial transformation operation to warp I_1 with ϕ^b or warp I_2 with ϕ^f by using the bilinear interpolation [42]. The general deep learning-based methods that aim to minimize $I_1 - \mathcal{T}(I_2, \phi^f)$ and $I_2 - \mathcal{T}(I_1, \phi^b)$ will lead the $\mathcal{L}_{data}$ to gradually approach to zero.

B. Regularization Term $\mathcal{L}_{smooth}$

To ensure the spatial smoothness of flow field between two images, the regularization term $\mathcal{L}_{smooth}$ controls the continuity in space, which is expressed as:

$$\mathcal{L}_{smooth} = \lambda_1 \|\nabla^2 \phi^f\|_2^2 + \lambda_2 \|\nabla^2 \phi^b\|_2^2, \tag{5}$$

where ∇^2 represents the Laplacian operator, and λ_1, λ_2 are the nonnegative weighting parameters.

IV. The Proposed Method

This section introduces our proposed method. First, we give a structure overview of the proposed SDOF-GAN and describe each of its components in detail. Then, we will introduce the network with forward-backward consistency loss, supervised loss, and an adversarial loss, which constitute the objective function of our network.

A. Algorithm Overview

Our goal is to train a symmetric optical flow estimation network in a semi-supervised manner. We achieve this by embedding the forward-backward consistency constraint into GAN, which does not require making the assumption of brightness constancy. Given two images I_1 and I_2, our proposed SDOF-GAN simultaneously estimates two transformations, namely ϕ^f: from I_1 to I_2, and ϕ^b: from I_2 to I_1. As shown in Fig. 2, our proposed SDOF-GAN is mainly trained in two phases, namely generators updating and discriminators updating.

1) Generators Updating: For the stage of updating generator networks, we design two generators G^f and G^b and an inverse network, to jointly estimate optical flow fields in both directions.

The parameters of generators G^f and G^b are shared, and G^f and G^b are trained to generate more realistic flows that are difficult for the discriminators to distinguish between true and false. We exploit a semi-supervised learning strategy that uses both labeled data with the ground-truth flow and unlabeled data without any ground-truth information. For labeled data, we learn flow fields using labeled image pairs by optimizing a supervision loss $\mathcal{L}_{sup}$ (as described in Sec. IV-B). The ground-truth information provides a definite cue for measuring the correctness of flow prediction during training. Notably, the ground-truth forward flow ϕ^{gf}, and the ground-truth backward flow ϕ^{gb} are inverse to each other. The unsupervised flow estimation complements to the ground-truth by a large number of unlabeled training images, which ensures low training costs and algorithm robustness. For unlabeled data, the networks are optimized by minimizing the adversarial loss $\mathcal{L}_{GAN}^{G}$ (as described in Sec. IV-B) based on discriminator networks. The discriminators can automatically judge whether the flow is positive by learning the pattern of flow warping error.

The inverse network is designed to generate inverse flow (*e.g.*, $\tilde{\phi}^f$) of each flow (*e.g.*, ϕ^f). Our goal is to ensure that the bidirectional flows are inverse to each other, that is to say, $\tilde{\phi}^f$ is equal to ϕ^b. The forward-backward consistency is implemented via an inverse network and a symmetry loss $\mathcal{L}_{sym}$ (as described in Sec. IV-B), which can ensure that

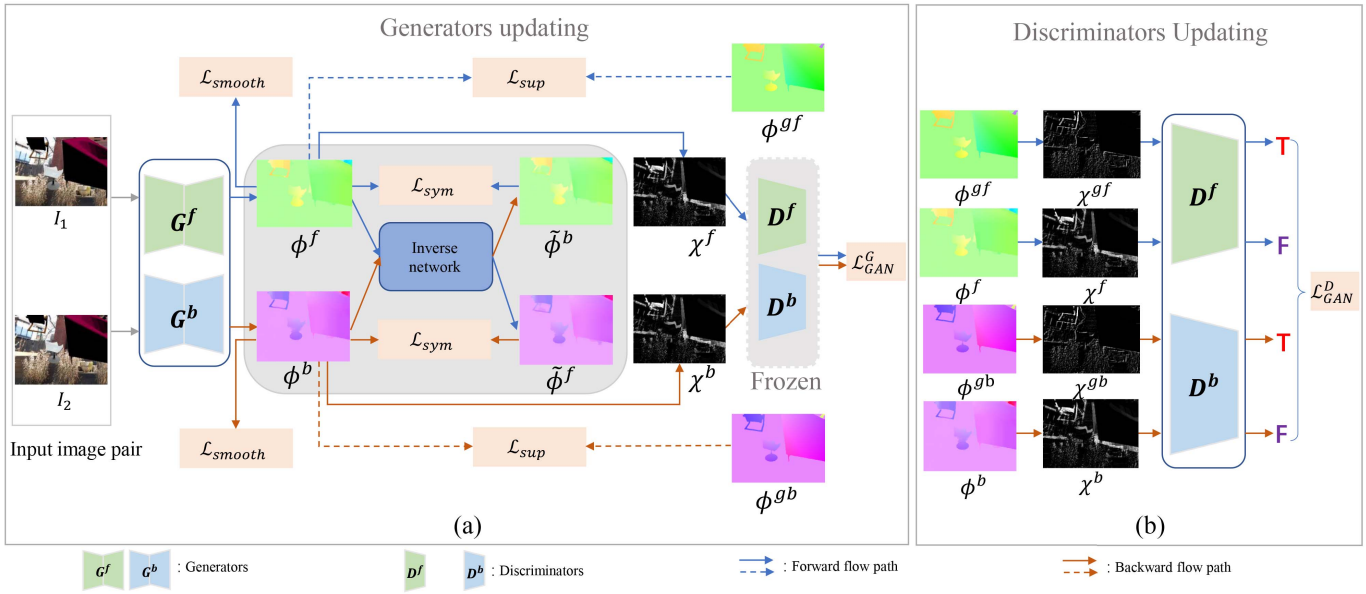

Fig. 2. Schematic illustration of our method, which includes two stages of adversarial training. (a) The updating of generators G^f and G^b with a pair of labeled images or unlabeled images. The supervised loss $\mathcal{L}_{sup}$ indicated by the dotted line only acts when the labeled images are taken as inputs. The parameters of two generators G^f and G^b are shared. The discriminators are fixed when the generators are updated. (b) The updating of discriminators D^f and D^b based on the ground-truth flow warping error and the flow warping error from the estimated flow. The two discriminators D^f and D^b also share parameters and the generators are fixed when the discriminators are updated.

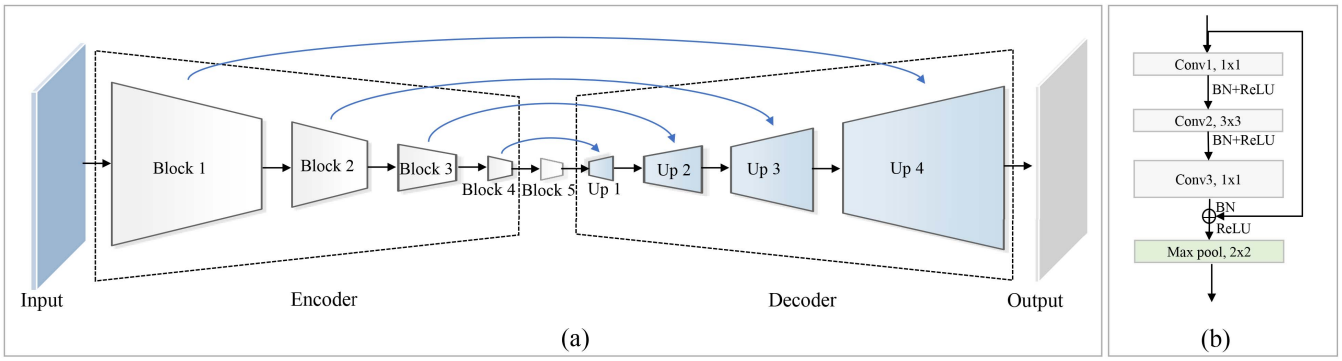

Fig. 3. The detailed network structure of the generator G. (a) The whole architecture of G consists of an encoder with five residual blocks and a decoder with four times upsampling operation ("Up" is short for "Upsampling"). (b) Schematic illustration of the residual block. The residual is obtained from three successive convolutions, and each convolution is followed by ReLU and BN. In particular, the 5th residual block does not contain a pooling operation.

the bidirectional flow fields are inverse to each other. The generator networks for bidirectional flows are correlated by enforcing a forward-backward consistency constraint over their outputs.

2) Discriminators Updating: For the stage of updating discriminator networks, we design two discriminators D^f and D^b to distinguish whether the estimated flow is "real" or "fake". Two cases are fed into the discriminator alternatively, namely the positive case where the flow warping error is generated by ground-truth flow, and the negative case where the flow warping error is generated by the predicted flow. Since the training of the discriminator requires ground-truth information, only the labeled image pairs are involved in this step. These two discriminator modules (D^f and D^b) also share network parameters, and they are updated alternately with the generators.

B. Loss Functions and Objective

This part introduces the loss functions and the resultant objective function for our method.

1) Symmetry Loss (Forward-Backward Consistency Loss): In order to achieve a symmetric estimation of the bidirectional optical flow, the symmetry loss is proposed to penalize the motion differences between two flow fields from the corresponding inverse mappings. It can enforce the forward flow ϕ^f (from I_1 to I_2) and the backward flow ϕ^b (from I_2 to I_1) to be inverse to one another. Besides, we consider that occlusion occurs when multiple objects in a complex scene move with different displacements and overlap each other in consecutive frames. The deformations of the occluded pixels are difficult to maintain inverse consistency. Thus, we exploit the bidirectional inconsistency check to handle the region of occlusion [16], [43]. We mark one pixel as occluded when

the mismatch between the forward flow ϕ^f and the backward flow ϕ^b is too large. Taking forward occlusion map O^f as an example, a pixel is considered to be occluded whenever it violates the following constraint:

$$|\phi^f(x)+\phi^b(x+\phi^f(x))|^2 < \alpha_1(|\phi^f(x)|^2+|\phi^b(x+\phi^f(x))|^2)+\alpha_2. \quad (6)$$

The occluded pixel is set to 1, and 0 otherwise. We obtain the backward occlusion map O^b in the same way by simply changing ϕ^f and ϕ^b in Eq. (6). We set $\alpha_1 = 0.01$ and $\alpha_2 = 0.5$ in all of our experiments. Therefore, the proposed symmetry loss at all non-occluded pixels x can be defined as:

$$\mathcal{L}_{sym} = (1-O^f)\cdot\|\phi^f-\tilde{\phi}^b\|_2^2+(1-O^b)\cdot\|\phi^b-\tilde{\phi}^f\|_2^2, \quad (7)$$

where $\tilde{\phi}^f$ and $\tilde{\phi}^b$ are generated by inverse network which will be detailed in Sec. V-B.

2) Supervised Loss: Similar to the End Point Error (EPE) in [18], we use supervised loss term to measure the shifts of the predicted flow field of pixels from the ground-truth flow field, namely:

$$\mathcal{L}_{sup} = \|\phi^f-\phi^{gf}\|_2^2+\|\phi^b-\phi^{gb}\|_2^2, \quad (8)$$

where ϕ^{gf}, ϕ^{gb} denote the ground-truth forward flow and the ground-truth backward flow, respectively. Therefore, SDOF-GAN aims to minimize the $\mathcal{L}_{sup}$ for labeled images with the ground-truth flow.

3) Adversarial Loss: To train the generator G and discriminator D with semi-supervised learning on both labeled images and unlabeled images, an adversarial loss [25], [34] is applied to force the generated flow to be indistinguishable from the ground-truth flow. For our generator $G = \{G^f, G^b\}$ and discriminator $D = \{D^f, D^b\}$, the adversarial loss for training GAN can be expressed as:

$$\mathcal{L}_{GAN}(G, D, \chi^g, \chi) = \mathbb{E}_{\chi^g}[\log D(\chi^g)]+\mathbb{E}_{\chi}[\log(1-D(\chi))], \quad (9)$$

where $\chi = \{\chi^f, \chi^b\}$ denotes the generated flow warping error and $\chi^g = \{\chi^{gf}, \chi^{gb}\}$ is the ground-truth flow warping error. Note that the $\chi^f = I_1 - \mathcal{T}(I_2, \phi^f)$ can be expressed as the generated forward flow warping error and $\chi^b = I_2 - \mathcal{T}(I_1, \phi^b)$ can be expressed as generated backward flow warping error. Similarly, $\chi^{gf} = I_1 - \mathcal{T}(I_2, \phi^{gf})$ and $\chi^{gb} = I_2 - \mathcal{T}(I_1, \phi^{gb})$ can be understood as the ground-truth forward flow warping error and the ground-truth backward flow warping error, respectively.

4) Objective Function: The proposed SDOF-GAN is optimized by minimizing the loss of the generator and meanwhile maximizing the loss of the discriminator. Consequently, our full objective is defined as:

$$\mathcal{L}(G, D) = \mathcal{L}_{GAN}(G, D, \chi^g, \chi)+\lambda_1\mathcal{L}_{smooth}(G) +\lambda_2\mathcal{L}_{sym}(G, \tilde{\phi})+\lambda_3\mathcal{L}_{sup}(G, \phi^g), \quad (10)$$

where $\phi^g = \{\phi^{gf}, \phi^{gb}\}$ denotes ground-truth flow, and $\tilde{\phi} = \{\tilde{\phi}^f, \tilde{\phi}^b\}$ is the inverse flow of each estimated flow. Three parameters λ_1, λ_2 and λ_3 are used to control the relative importance of all terms. The goal of SDOF-GAN can be described as a maximizing minimization problem, namely:

$$\min_G \max_D \mathcal{L}(G, D). \quad (11)$$

C. Adversarial Training

Our SDOF-GAN model is implemented in a semi-supervised strategy with alternate training of the generator G and the discriminator D (*e.g.*, fix D when training G). Convergence occurs when the discriminator cannot distinguish the positive case and the negative case. For both the labeled and unlabeled images, training G is driven by one symmetry loss $\mathcal{L}_{sym}$ controlling the consistency of the forward and backward optical flows, a spatial smoothness loss $\mathcal{L}_{smooth}$ characterizing local smoothness of the deformation field, and an adversarial loss $\mathcal{L}^G_{GAN}$. Our objective for updating G using unlabeled images is then written as:

$$\mathcal{L}^u_G = \mathcal{L}^G_{GAN}+\lambda_1\mathcal{L}_{smooth}(G)+\lambda_2\mathcal{L}_{sym}(G, \tilde{\phi}), \quad (12)$$

where $\mathcal{L}^G_{GAN} = -D(\chi)$. For the labeled images, we additionally use a supervision loss $\mathcal{L}_{sup}$ with which the ground-truth flow participates in the training process. The objective for updating G using labeled images can thus be expressed as:

$$\mathcal{L}^s_G = \mathcal{L}^G_{GAN}+\lambda_1\mathcal{L}_{smooth}(G)+\lambda_2\mathcal{L}_{sym}(G, \tilde{\phi}) +\lambda_3\mathcal{L}_{sup}(G, \phi^g). \quad (13)$$

We train D to determine whether the flow is "real" or "fake" by maximizing the adversarial loss $\mathcal{L}^D_{GAN}$. Consequently, the objective for updating D using labeled images can be expressed as:

$$\mathcal{L}^D_{GAN} = -\log D(\chi^g)-\log(1-D(\chi)). \quad (14)$$

V. Network Structure and Implementation

This section introduces the employed network architecture and implementation details for predicting the high-precision optical flow.

A. Generator

As shown in Fig. 3, our generator G takes full advantage of the residual block [44] and U-Net architecture [45]. In the encoder, we exploit a ResNet-style network with five blocks that combine the features from shallow and deep layers by using the deformed residual unit of ResNet. The network can avoid vanishing gradient problem by the bypass connections in the residual block during backpropagation. The residual block of ResNet50 is illustrated in the right panel of Fig. 3. There are three successive convolution layers with stride 1. The first convolution layer and the third convolution layer use 1×1 convolutional filters, and the size of convolutional filters in the second convolution layer is 3×3. Except for the last residual block, three convolutions in each residual block are followed by a 2×2 max-pooling layer with stride 2. Additionally, we exploit ReLU activation function [46] and batch normalization (BN) [47] at the output of the convolutions. Note that ReLU and BN are not used at the inputs to the sum operation of the residual blocks. Instead, ReLU comes after the

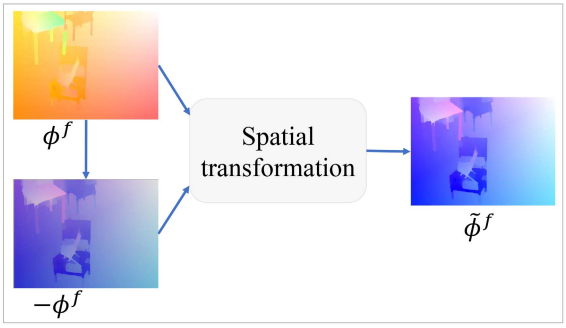

Fig. 4. Illustration of the inverse network. The generation process of the inverse flow is explained by taking the forward optical flow as an example.

sum operation. The output of the encoder acts as an input to the decoder. The decoder contains four upsampling operations (deconvolution layers) and two successive convolution layers (the same as the convolution layers of blocks without sum operation). Furthermore, to recover the flow details, skip connections are utilized between corresponding layers in encoder and decoder to combine the high-frequency features from the contracting path with up-sampled output.

B. Inverse Network

We use an inverse network to compute the inverse flow $\tilde{\phi}$ of flow ϕ based on [30] as shown in Fig. 4. Taking the forward optical flow ϕ^f as an example, we first calculate the negative flow $-\phi^f$ of flow ϕ^f. Next, the inverse network takes the flow ϕ^f (in the I_2 space) and the negative flow $-\phi^f$ (in the I_2 space) as inputs. We finally obtain the standard inverse flow $\tilde{\phi}^f$ (in the I_1 space) via a differentiable transformation operation which is similar to [13], [42]. Specifically, the localization step of Spatial Transformer Network [42] is not required in our work, because the flow prediction, ϕ^f, provides the necessary parameters for the mapping between bidirectional flow fields. Spatial Transformer Network [42] is applied to perform two steps here, namely sampling grid generation and differentiable image sampling. In this way, the differences of flow ϕ^b and the inverse flow $\tilde{\phi}^f$ can be obtained by forward-backward consistency loss that further guides the network optimization.

C. Discriminator

Our discriminator D is trained to solve the maximization problem in Eq. (11), which leverages the patchGAN [34] to distinguish whether an overlapping patch is "real" or "fake", rather than to distinguish the whole image. Such an approach has fewer parameters compared with the entire image as input, and it is suitable for the images with arbitrary size. Both the ground-truth flow warping error and the flow warping error from the predicted flow are taken as inputs to D, and the output is a binary discriminant value.

D. Implementation Details

The entire network containing the generators and the discriminators is implemented via using Pytorch [48]. The input to the network is in the form of six-channel images which are obtained by directly concatenating two source images. Each network is trained on the NVIDIA Tesla V100 GPUs, using Adam optimization [49]. In order to improve the stability of the model, we train the network using a strategy that gradually decays the learning rate. We set the initial learning rate as 0.0001 and then multiply by 0.5 every 100k iterations after the first 200k iterations. We train the network for 50 epochs in total. We tune the weighting parameters λ_1, λ_2 and λ_3 based on grid search in the range [0.001, 0.01, 0.1, 1], and finally choose the best performing combination $\lambda_1 = 0.01$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$.

VI. Experiments

To evaluate the performance of our proposed framework, we compare our SDOF-GAN with state-of-the-art optical flow estimation algorithms in this section. Then the ablation studies are conducted to verify the effectiveness of our contributions. The computational costs of different methods are also investigated.

A. Experimental Settings

We exploit several publicly available datasets to train and evaluate our network, which include FlyingChairs [7], FlyingThings3D [50], KITTI raw dataset [51], KITTI 2012 [52], KITTI 2015 [53], and MPI-Sintel [54]. The FlyingChairs dataset is a synthetic dataset with the dense ground-truth forward flow by applying affine transformations to images. It is collected from Flickr and a publicly available set of renderings for 3D chair models. It contains 22232 training image pairs and 640 test image pairs. The FlyingThings3D can be seen as a three-dimensional version of FlyingChairs. It is a synthetic scene dataset with ground-truth flow in forward and backward directions and contains 25000 stereo frames. Compared with the images in FlyingChairs which only have plane transformation, the images in FlyingThings3D have real 3D motion and brightness changes, and also contain richer motion transformation information. The KITTI-related datasets are collected from real-world driving scenarios. The KITTI raw dataset is divided into the categories 'Road', 'City', 'Residential', 'Campus', and 'Person'. It contains about 48000 frames. The KITTI 2012 consists of 194 training image pairs and 195 test image pairs, while KITTI 2015 consists of 200 training image pairs and 200 test image pairs. They also provide ground-truth flow by 3D laser measurements from a Velodyne laser scanner, and the density per-frame is about 50%, resulting in sparse optical flow ground-truth. The MPI-Sintel dataset contains 1064 training image pairs and 564 test image pairs with dense ground-truth flow. There are two versions in the MPI-Sintel dataset, namely the clean version containing small motion, and the final version containing large motion blur.

1) Data Division: Since our model is semi-supervised, we use the raw dataset of KITTI as unlabeled data for training. For the selection of labeled data, we propose three different ways, namely only using FlyingChairs ("Chairs"), only using FlyingThings3D ("Things3D"), and the mixture of the two datasets ("Mixed"). The ground-truth backward

flows in FlyingChairs and MPI-Sintel are synthesized by affine transformations, which are inverse to the ground-truth forward flows. The 1041 training image pairs from the MPI-Sintel dataset are used to fine-tune our network. We evaluate the performance of SDOF-GAN on KITTI 2012, KITTI 2015, MPI-Sintel, and the test image pairs of FlyingChairs.

2) Data Augmentation: To avoid overfitting, we follow the semi-supervised optical flow estimation method [18] and do data augmentation from different ways on the images in the training datasets. We randomly rotate them with the angles in $(-17°, 17°)$, re-scaling them with the size in [1, 2], adding white Gaussian noise with a standard deviation from [0, 0.01], and changing the brightness, contrast, and saturation by using $ColorJitter$ with the random factor in [0, 0.04].

3) Fine-Tuning: Notably, the object types and motions contained in different datasets vary broadly, and a single dataset has a very limited number of data for training our network. To solve this problem, we follow [7] and train the proposed SDOF-GAN on FlyingChairs, FlyingThings3D, KITTI raw dataset, and then fine-tune the network on the training set of MPI-Sintel.

4) Evaluation Metrics: We use two metrics for evaluating the performance of SDOF-GAN, which are described as follows:

- End point error (EPE) [18]: EPE is defined as the distance between the endpoints of the predicted flow and ground-truth flow. It is suitable for evaluating the performance of a method on the labeled data. In our experiments, we compute average EPE over all pixels by Eq. (8) for the images with the ground-truth flow.
- Percentage of erroneous pixels (Fl) [53]: Fl is the ratio of erroneous pixels averaged over all ground truth pixels of test images. We consider a pixel to be incorrectly estimated if the disparity or flow EPE is larger than 3 pixels or 5% of the ground-truth value. As mentioned, we denote the percentage of erroneous pixels over all pixels as "Fl-all", the error rate on foreground objects flow as "Fl-f", and the error rate on background motion as "Fl-b". By following [16], Fl is employed as an algorithm evaluation metric on KITTI benchmarks.

B. Comparisons With the State-of-the-Art Methods

We report the results of our network on the FlyingChairs, KITTI 2012, KITTI 2015, and MPI-Sintel benchmark datasets. We compare our method with two CNN-based supervised asymmetric methods (*i.e.*, FlowNet2.0 [8], SpyNet [9]), two CNN-based unsupervised asymmetric methods (*i.e.*, SelFlow [16], DDFlow [15]), a GAN-based semi-supervised asymmetric method (*i.e.*, SemiFlowGAN [18]), and a traditional symmetric optical flow prediction method (*i.e.*, MirrorFlow [6]). The SpyNet [9] used five pyramid levels with a mini-batch size of 32 across all networks. For SelFlow [16], the supervised fine-tuned model is achieved by pre-training the unsupervised model in [16]. Note that SelFlow utilizes three frames, while all other methods use only two consecutive frames. Additionally, we fine-tune our network on MPI-Sintel dataset and compare the results with other fine-tuned baseline networks such as FlowNet2.0 [8], SpyNet [9], DDFlow [15], SelFlow [16], and SemiFlowGAN [18]. The fine-tuned SpyNet on the Driving dataset [50], while other methods are fine-tuned on the MPI-Sintel dataset. It is worth noting that for all asymmetric methods, the forward and backward flows are predicted separately.

We use different labeled datasets as training data, denoted as 'SDOF-GAN(Chairs)', 'SDOF-GAN(Things3D)' and 'SDOF-GAN(Mixed)'. From Table I, we can find that the SDOF-GAN shows a large performance difference under different labeled datasets. The network model training on a mixture of two labeled datasets is more robust because the FlyingChairs has plane transformation and the FlyingThings3D has realistic 3D motion and brightness changes. We finally select the pre-trained model with the best performance which is trained on both datasets for further analysis. That is to say, the SDOF-GAN mentioned below is trained on a mixture of two labeled datasets as well as the KITTI raw dataset.

The detailed comparison results of different methods are shown in Table I. From this table, we have four important findings. Firstly, the SDOF-GAN outperforms asymmetric prediction methods on almost all datasets. Our method without fine-tuning reduces the previous best EPE from 1.68 to 1.58 on the FlyingChairs dataset and achieves 23% relative improvement on the Sintel-Final dataset. This is benefited from the symmetric property of SDOF-GAN, which simultaneously predicts optical flow in both directions. Secondly, by comparing with the symmetric prediction method MirrorFlow, our SDOF-GAN without fine-tuning achieves state-of-the-art results on all three datasets, with EPE $=$ 3.46 on Sintel-Clean, EPE $=$ 4.61 on Sintel-Final, and Fl-all $=$ 8.22% on KITTI 2015. The superior performances of our method to MirrorFlow indicate that GAN-based adversarial training can effectively improve the precision of optical flow estimation. Thirdly, our SDOF-GAN outperforms all unsupervised flow estimation methods on all datasets. Furthermore, the fine-tuned SDOF-GAN achieves EPE $=$ 3.24 on Sintel-Clean and EPE $=$ 4.08 on Sintel-Final, which is even better than the results of supervised methods FlowNet2.0 [8], SpyNet [9] and fine-tuned SelFlow [16]. This shows that the semi-supervised strategy combined with the adversarial training effectively improves the performance of the algorithm by learning from the labeled and unlabeled data. Finally, we see an improvement of EPE by fine-tuning the trained models, which means that fine-tuning can improve the generalization ability and robustness of the model.

In addition, since our SDOF-GAN runs in a semi-supervised way, we randomly select 50% and 20% of the labeled data from the original labeled datasets and the remaining 50% and 80% image pairs are left as unlabeled data. We also train the semi-supervised method SemiFlowGAN [18] with the same amount of labeled data for comparison. For the supervised methods such as FlowNet2.0 [8] and SpyNet [9], only the labeled image pairs are used for training. In particular, we train supervised networks with the same amount of labeled data for comparisons.

As shown in Table II and Table III, the proposed SDOF-GAN based on semi-supervised strategy achieves the

TABLE I

THE RESULTS OF VARIOUS METHODS ON THE FIVE DATASETS, WHERE "FT" REPRESENTS THE METHOD WITH FINE-TUNING

	Methods	Chairs	Sintel-Clean		Sintel-Final		KITTI 2012			KITTI 2015		
		EPE test	EPE train	EPE test	EPE train	EPE test	EPE train	EPE test	Fl-all test	Fl-f test	Fl-b test	Fl-all test
	MirrorFlow [6]	-	-	3.32	-	6.07	-	-	-	17.07%	8.93%	10.29%
Supervised	FlowNet2.0 [8]	**1.68**	2.02	3.96	3.14	6.02	4.09	-	-	10.75%	8.75%	10.41%
	FlowNet2.0-ft [8]	-	**1.45**	4.16	2.01	5.74	3.61	-	-	-	-	-
	SpyNet [9]	2.63	4.12	6.69	5.57	8.43	9.12	-	-	-	-	-
	SpyNet-ft [9]	3.07	3.17	6.64	4.32	8.36	8.25	10.10	20.97%	36.96%	32.25%	35.07%
	SelFlow-ft [16]	-	1.68	**3.74**	**1.77**	**4.26**	**0.76**	**1.50**	**6.19%**	**8.02%**	9.07%	**8.42%**
Unsupervised	SelFlow [16]	-	**2.88**	6.56	**3.87**	**6.57**	**1.69**	**2.20**	**7.68%**	**13.24%**	14.75%	**14.19%**
	DDFlow [15]	**2.97**	3.83	-	4.85	-	8.27	-	-	-	-	-
	DDFlow-ft [15]	3.46	2.92	**6.18**	3.98	7.40	2.35	3.00	8.86%	14.51%	**13.76%**	14.29%
Semi-supervised	SemiFlowGAN [18]	1.95	3.30	6.28	4.68	7.61	7.16	7.50	17.21%	-	-	39.71%
	SemiFlowGAN-ft [18]	2.41	2.41	6.27	3.16	7.31	5.23	6.80	16.30%	-	-	31.01%
	SDOF-GAN (Chairs)	1.63	3.25	6.22	3.46	6.50	4.12	3.57	13.26%	18.74%	15.37%	17.12%
	SDOF-GAN (Things3D)	1.61	2.71	5.01	3.04	5.68	3.68	2.82	10.31%	15.71%	13.13%	14.58%
	SDOF-GAN (Mixed)	**1.58**	1.75	3.46	2.70	4.61	1.86	2.57	7.52%	11.99%	9.12%	8.22%
	SDOF-GAN(Mixed)-ft	1.65	**1.33**	**3.24**	**1.75**	**4.08**	**1.66**	**1.57**	**5.56%**	**12.07%**	**8.05%**	**8.03%**

TABLE II

THE RESULTS OF VARIOUS METHODS ON THE FIVE DATASETS WHEN 50% OF LABELED DATA FROM TRAINING DATASETS

	Methods	Chairs	Sintel-Clean		Sintel-Final		KITTI 2012			KITTI 2015		
		EPE test	EPE train	EPE test	EPE train	EPE test	EPE train	EPE test	Fl-all test	Fl-f test	Fl-b test	Fl-all test
Supervised	FlowNet2.0 [8]	2.06	2.54	4.21	3.32	6.19	4.25	-	-	17.88%	9.33%	11.27%
	SpyNet [9]	2.81	4.73	6.80	5.69	8.57	9.33	-	-	-	-	-
Semi-supervised	SemiFlowGAN [18]	2.30	3.41	6.37	4.75	7.68	7.24	7.59	18.02%	-	-	39.82%
	SDOF-GAN	1.81	1.93	3.60	2.76	4.91	2.03	2.68	7.94%	12.38%	10.11%	8.30%

TABLE III

THE RESULTS OF VARIOUS METHODS ON THE FIVE DATASETS WHEN 20% OF LABELED DATA FROM TRAINING DATASETS

	Methods	Chairs	Sintel-Clean		Sintel-Final		KITTI 2012			KITTI 2015		
		EPE test	EPE train	EPE test	EPE train	EPE test	EPE train	EPE test	Fl-all test	Fl-f test	Fl-b test	Fl-all test
Supervised	FlowNet2.0 [8]	2.35	2.60	4.33	3.40	6.28	4.36	-	-	17.97%	9.54%	11.71%
	SpyNet [9]	2.97	4.84	6.87	5.76	8.69	9.45	-	-	-	-	-
Semi-supervised	SemiFlowGAN [18]	2.37	3.46	6.42	4.83	7.74	7.30	7.64	18.67%	-	-	39.93%
	SDOF-GAN	2.09	2.06	3.64	2.82	5.03	2.09	2.74	7.99%	12.53%	10.20%	8.36%

top level performance when the networks are trained with partially labeled images. First, the proposed semi-supervised method consistently outperforms the previous best-supervised method FlowNet2.0 and achieve 2.97%, 2.35% relative improvement on the KITTI 2015 dataset in Table II and Table III, respectively. Besides, by comparing Table II and Table III, we can see that the SDOF-GAN achieves a greater improvement on all datasets than previous supervised methods. This shows that the semi-supervised approach is superior to the supervised approach as the amount of labeled images decreases. Therefore, in the real-world situations where the amount of labeled data is limited, semi-supervised strategy can improve the performance of the algorithm. Furthermore, our proposed SDOF-GAN is consistently better than the previous semi-supervised method SemiFlowGAN on partially labeled data. This is due to the fact that SDOF-GAN uses a forward-backward consistency constraint to ensure consistent bidirectional optical flow estimation.

The visualization of optical flow estimation obtained by different methods is shown in Fig. 5 and Fig. 6. For all asymmetric optical flow estimation methods (*i.e.*, FlowNet2.0 [8], SpyNet [9], DDFlow [15], SelFlow [16], and SemiFlow-GAN [18]), forward flow and backward flow are obtained by swapping the source image and target image. We can clearly see that the SDOF-GAN is better than all the other methods on Sintel-Final. Our method produces the smoother flow fields compared with MirrorFlow and SemiFlowGAN. This means that combining the symmetric properties of optical flow and the adversarial training mechanism can effectively improve the performance of the algorithm. Apparently, Fig. 6 indicates that the SDOF-GAN can do better than previous methods on moving object boundaries. Besides, on the difficult KITTI

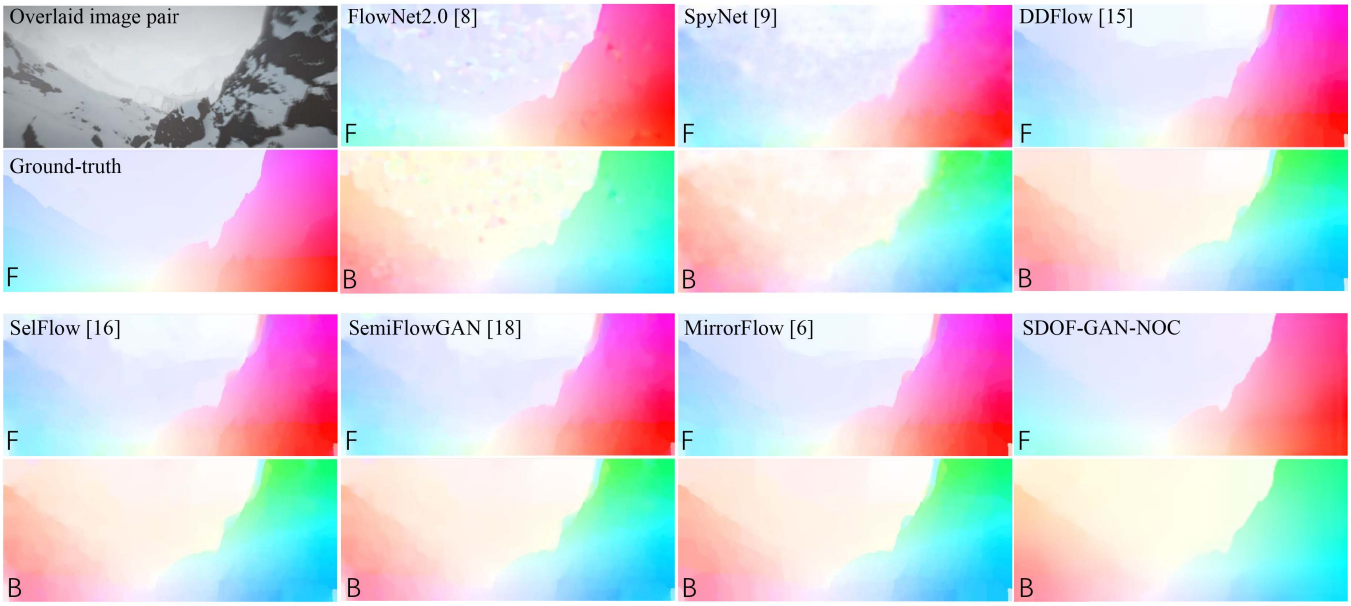

Fig. 5. Visualization of results on Sintel-Final dataset. The overlaid image pair on the upper left is the input image pair. In the lower left corner of each flow map: 'F' represents forward flow and 'B' represents backward flow.

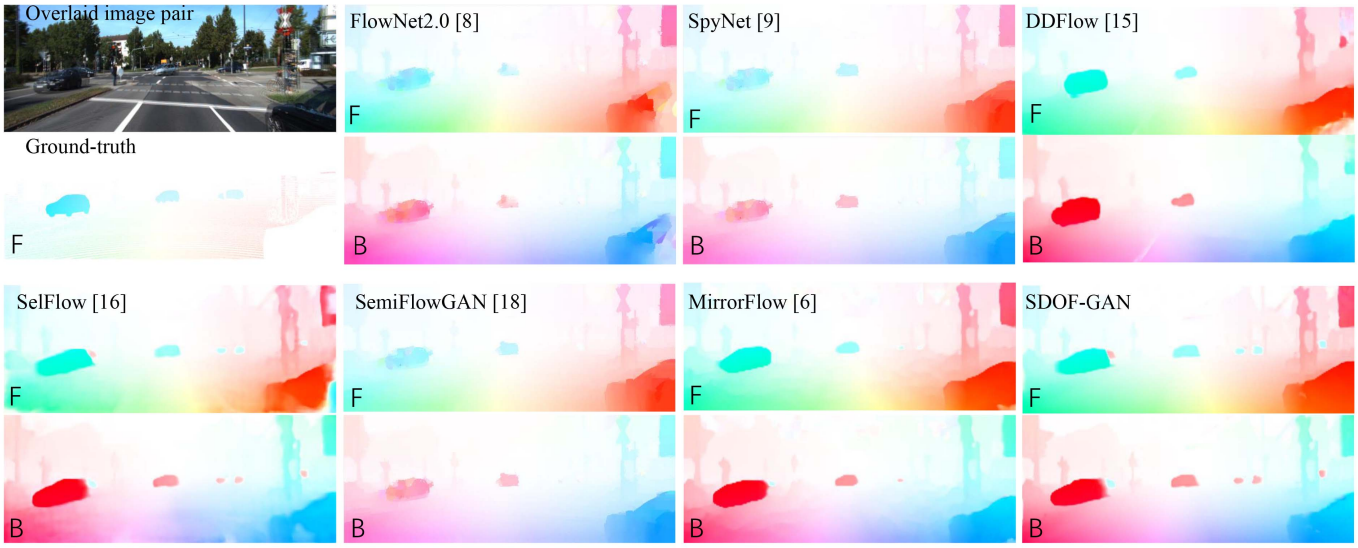

Fig. 6. Visualization of results on KITTI 2015 dataset. The overlaid image pair on the upper left is the input image pair. In the lower left corner of each flow map: 'F' represents forward flow and 'B' represents backward flow.

2015 dataset, SDOF-GAN is even slightly better than the classic symmetric method MirrorFlow. CNN-based asymmetric methods perform unsatisfactorily in object boundaries. In contrast, our GAN-based symmetric method further improves the precision of motion estimation through the forward-backward consistency constrain under a semi-supervised strategy.

C. Ablation Study

In addition, to emphasize the importance of our introduced symmetric constraint (forward-backward consistency) and adversarial loss, we design four variants of SDOF-GAN and compare them with the completed SDOF-GAN on Sintel-Clean, Sintel-Final, and KITTI 2015 datasets. Specifically, we first remove the symmetry loss (forward-backward consistency loss) from SDOF-GAN to form a variant named "DOF-GAN", which learns the bidirectional flow fields independently. Then, we remove the adversarial loss of SDOF-GAN and degenerate our model to a fully supervised setting (denoted as "SDOF-S"). SDOF-S is carried out by minimizing the squared error between the estimated flow and the ground-truth. Besides, we use the data term $\mathcal{L}_{data}$ in Eq. (4) to replace the adversarial loss and arrive at an unsupervised learning setting denoted as "SDOF-U". Finally, we train the generator of SDOF-GAN to a semi-supervised deep network (denoted as "SDOF-Semi").

1) Symmetry Analysis: Table IV shows the quantitative evaluation of SDOF-GAN and its asymmetric variant DOF-GAN.

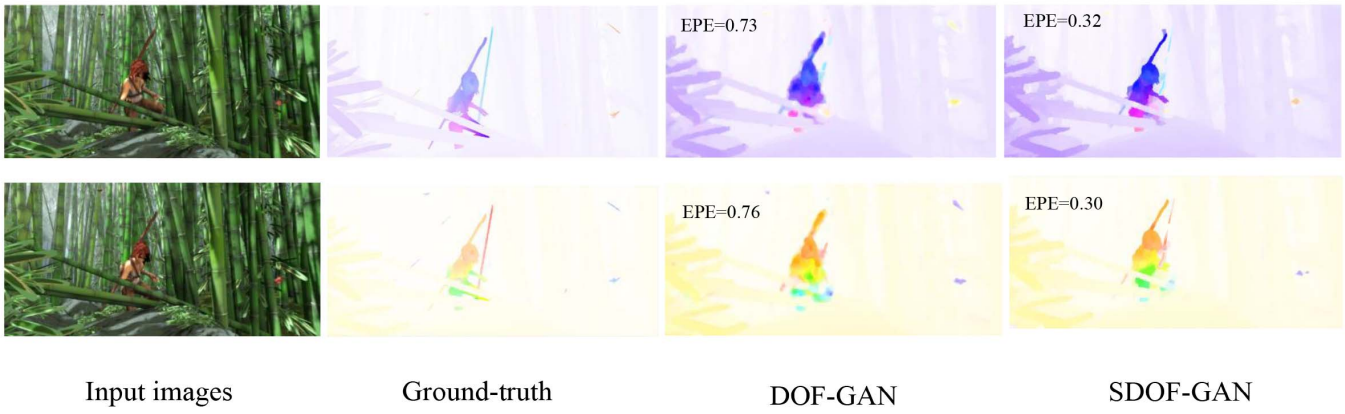

Fig. 7. Visual comparison of our symmetric (SDOF-GAN) and asymmetric (DOF-GAN) models on Sintel-Clean dataset.

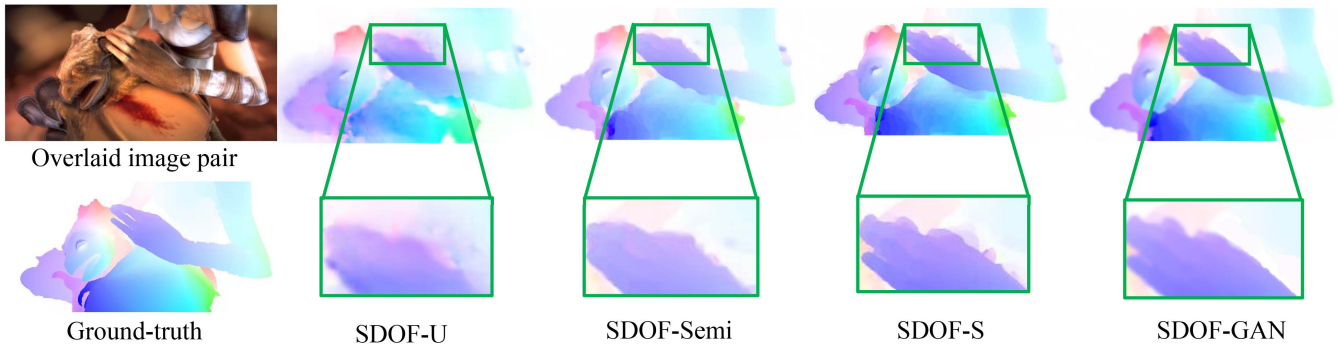

Fig. 8. Visual comparison of our models with and without GAN on Sintel-Final dataset. Here we only show the flow field in one direction.

TABLE IV

THE RESULTS OF THREE VARIANTS OF SDOF-GAN. "DOF-GAN" DENOTES THE PROPOSED FRAMEWORK WITHOUT SYMMETRIC CONSTRAINT, "SDOF-S" DENOTES THE PROPOSED FRAMEWORK THAT IS TRAINED IN A SUPERVISED FASHION, "SDOF-U" REPRESENTS THAT THE PROPOSED METHOD IS TRAINED IN AN UNSUPERVISED FASHION, AND "SDOF-SEMI" IS THE SEMI-SUPERVISED DEEP MODEL WITHOUT GAN ARCHITECTURE

Methods	Sintel-Clean	Sintel-Final	KITTI 2015		
	EPE	EPE	Fl-f	Fl-b	Fl-all
DOF-GAN	3.80	5.56	14.82%	11.95%	11.96%
SDOF-S	3.70	5.03	12.32%	9.29%	9.45%
SDOF-U	4.21	5.59	14.51%	12.03%	12.07%
SDOF-Semi	3.77	5.11	12.21%	10.18%	10.42%
SDOF-GAN	3.46	4.61	11.99%	9.12%	8.22%

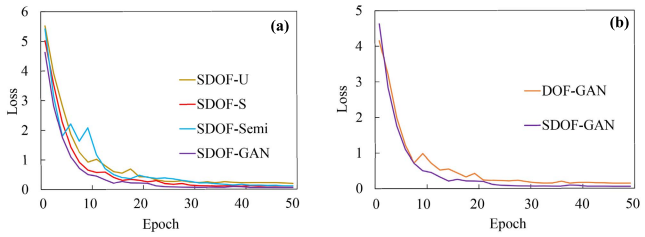

Fig. 9. The subgraph (a) presents the training curves for the model with GAN structure (*i.e.*, SDOF-GAN) and without GAN structure (*i.e.*, SDOF-S, SDOF-U and SDOF-Semi) during training. The subgraph (b) presents the training curves for the model with symmetry constraint (*i.e.*, SDOF-GAN) and without symmetry constraint (*i.e.*, DOF-GAN) during training. The loss values of different models have been rescaled to the same range.

We can find that the error produced by the symmetric method decreases substantially by about 3.7% on KITTI 2015 dataset. Furthermore, the symmetric method achieves better performance than the asymmetric method on the MPI-Sintel dataset, especially on "Final" data. This implies that the symmetry constraint makes significant contribution to the high accuracy of optical flow estimation. As shown in Fig. 7, an asymmetric approach may produce bias and cause object boundaries to be ambiguous. In addition, the curves of adversarial loss are shown in Fig. 9 (b). One could observe that the value of loss in SDOF-GAN drops faster than the loss in DOF-GAN. The underlying reason may be that the symmetry constraint controls the consistency of the bidirectional flows, therefore the decreasing speed of the loss value is accelerated.

2) Adversary Analysis: Optical flow estimation traditionally relies on the assumptions of brightness constancy and spatial smoothness. Arguably, this produces limited precision in predicting the optical flow in the boundary region and in the presence of complex motion. From Table IV, we can find that the flow accuracy of our SDOF-GAN significantly surpasses the three variants (*i.e.*, SDOF-S, SDOF-U, and SDOF-Semi). From the visualization results in Fig. 8, we see that SDOF-GAN yields higher accuracy in predicting optical

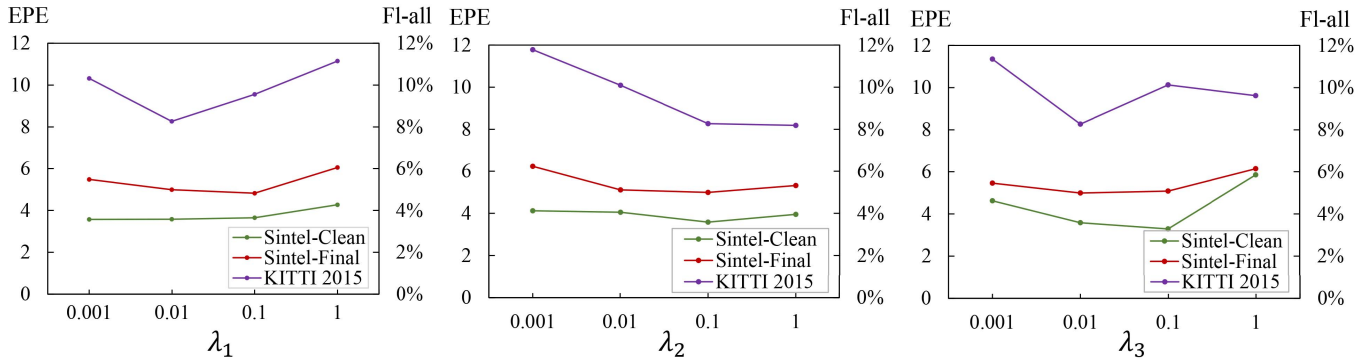

Fig. 10. The results produced by the SDOF-GAN under different weighting parameters. For Sintel-Clean and Sintel-Final, the vertical axis represents EPE. For KITTI 2015, the vertical axis represents Fl-all. (a) presents the results by changing λ_1 when fixing λ_2 to 0.1 and λ_3 to 0.01; (b) presents the results by changing λ_2 when fixing λ_1 to 0.01 and λ_3 to 0.01; (c) presents the results by changing λ_3 when fixing λ_1 to 0.01 and λ_2 to 0.1.

flow at motion boundaries than all the other variants, especially the unsupervised method "SDOF-U". This is due to the brightness constancy used in unsupervised method does not hold in the occluded regions and motion boundaries. Moreover, the convergence speed of the models with and without GAN (*i.e.*, SDOF-S, SDOF-U, and SDOF-Semi) is also compared. The convergence curves of different models are shown in Fig. 9 (a). For SDOF-GAN, we show the loss curve of the generator model during training. We can find that the GAN-based model performs a faster convergence rate than the model without GAN. Therefore, it can be seen that the use of adversarial loss term improves the performance when compared with the variants that do not contain a discriminator.

3) Parameter Sensitivity: For the proposed SDOF-GAN, the weighting parameters of different losses may significantly affect optical flow estimation performance. In this section, we evaluate how the selections of each of the parameters λ_1, λ_2, and λ_3 influence the model performance. For the three weighting parameters, we fix two of them and vary the value of the other one. Fig. 10 shows how the performance of optical flow estimation on the test set is influenced by the weighting parameters in the proposed SDOF-GAN. From the results, we can see that when a smaller or larger weight is assigned to the loss, the performance of the network will degrade. It is difficult to find the optimal parameter combination for all evaluated datasets based on grid search. Therefore, we choose a combination $\lambda_1 = 0.01$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$ that performs well in most cases.

D. Computational Speed

Here we compare the testing time of our SDOF-GAN with other baseline methods. For the asymmetric methods (*i.e.*, FlowNet2.0 [8], SpyNet [9], DDFlow [15], and SemiFlowGAN [18]), we only calculate the time of the forward flow prediction. For SelFlow [16], we count the total time of predicting bidirectional flows from the current image to the previous image and from the current image to the next image.

As shown in Table V, we compute the run-time of different methods in seconds. From the experimental results, we find that our method is more efficient than the traditional symmetric flow estimation method (*i.e.*, MirrorFlow) and the CNN-based unsupervised approaches (*i.e.*, DDFlow, SelFlow). That is to say, our GAN-based symmetric flow estimation method is superior to all other bidirectional flows prediction methods in terms of computational efficiency. The calculation speed of SDOF-GAN is comparable to the unidirectional flow estimation methods (*i.e.*, FlowNet2.0 and SemiFlowGAN). The speed of our method is slightly slower than the currently published most efficient method (*i.e.*, SpyNet) for optical flow estimation. This is mainly because that the SpyNet only estimates the flow in one direction and the network structure of SpyNet is relatively simple. In general, our proposed symmetric optical flow estimation method SDOF-GAN effectively overcomes the time-consuming defect and slow convergence by using a forward-backward consistency constraint and the adversarial training mechanism. Therefore, SDOF-GAN makes a good balance between efficiency and accuracy.

TABLE V

COMPUTATIONAL COST OF SEVEN DIFFERENT METHODS IN THE TESTING STAGE. ALL REPORTED TIME CONSUMPTIONS ARE MEASURED ON THE KITTI 2015 DATASET EXCLUDING THE IMAGE LOADING TIME. NOTE THAT THE LAST ROW SHOWS THE TIME OBTAINED BY UNIDIRECTIONAL FLOW OR BIDIRECTIONAL FLOWS. "F" DENOTES THE FORWARD FLOW ESTIMATION, AND "B" DENOTES THE BACKWARD FLOW ESTIMATION

Methods	FlowNet 2.0 [8]	SpyNet [9]	DDFlow [15]	SelFlow [16]	SemiFlow GAN [18]	Mirror Flow [6]	SDOF -GAN
Hardware	GPU	GPU	GPU	GPU	GPU	CPU	GPU
Time(sec)	0.18	0.1	0.47	0.43	0.19	1172	0.27
Direction	F	F	F	F&B	F	F&B	F&B

VII. CONCLUSION

In this paper, we have proposed SDOF-GAN for achieving symmetric optical flow estimation with generative adversarial networks. Specifically, it achieves symmetric effect in optical flow estimation by jointly estimating the forward and backward flows while enforcing consistency between them. In addition, it is trained adversarially to learn the pattern of the

error map between the target image and the one obtained by warping the source image. Moreover, SDOF-GAN is trained in a semi-supervised fashion with which both the labeled and unlabeled data can be fully exploited. Experimental results show the superior performances of SDOF-GAN to several other representative state-of-the-art techniques. Our future work would focus on incorporating occlusion-disocclusion symmetry as well in SDOF-GAN.

REFERENCES

[1] L. Alvarez, J. Weickert, and J. Sánchez, "Reliable estimation of dense optical flow fields with large displacements," *Int. J. Comput. Vis.*, vol. 39, no. 1, pp. 41–56, 2000.

[2] E. Mémin and P. Pérez, "Dense estimation and object-based segmentation of the optical flow with robust techniques," *IEEE Trans. Image Process.*, vol. 7, no. 5, pp. 703–719, May 1998.

[3] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-preserving interpolation of correspondences for optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1164–1172.

[4] L. Alvarez *et al.*, "Symmetric optical flow," in *Proc. Int. Conf. Comput. Aided Syst. Theory*. Springer, 2007, pp. 676–683.

[5] L. Alvarez, R. Deriche, T. Papadopoulo, and J. Sánchez, "Symmetrical dense optical flow estimation with occlusions detection," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2002, pp. 721–735.

[6] J. Hur and S. Roth, "MirrorFlow: Exploiting symmetries in joint optical flow and occlusion estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 312–321.

[7] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2758–2766.

[8] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2462–2470.

[9] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4161–4170.

[10] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 402–419.

[11] S. Zhao, Y. Sheng, Y. Dong, E. I.-C. Chang, and Y. Xu, "MaskFlownet: Asymmetric feature matching with learnable occlusion mask," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6278–6287.

[12] A. Ahmadi and I. Patras, "Unsupervised convolutional neural networks for motion estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1629–1633.

[13] J. Y. Jason, A. W. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 3–10.

[14] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4884–4893.

[15] P. Liu, I. King, M. R. Lyu, and J. Xu, "DDFlow: Learning optical flow with unlabeled data distillation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8770–8777.

[16] P. Liu, M. Lyu, I. King, and J. Xu, "SelFlow: Self-supervised learning of optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019.

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

[18] W.-S. Lai, J.-B. Huang, and M.-H. Yang, "Semi-supervised learning for optical flow with generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 354–364.

[19] Y. Yang and S. Soatto, "Conditional prior networks for optical flow," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 271–287.

[20] A. Odena, "Semi-supervised learning with generative adversarial networks," 2016, *arXiv:1606.01583*. [Online]. Available: https://arxiv.org/abs/1606.01583

[21] R. K. Thakur and S. Mukherjee, "A conditional adversarial network for scene flow estimation," in *Proc. 28th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Oct. 2019, pp. 1–6.

[22] K. Ohnishi, S. Yamamoto, Y. Ushiku, and T. Harada, "Hierarchical video generation from orthogonal information: Optical flow and texture," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1–8.

[23] P. Yan, S. Xu, A. R. Rastinehad, and B. J. Wood, "Adversarial image registration with application for MR and TRUS image fusion," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Springer, 2018, pp. 197–204.

[24] J. Fan, X. Cao, Z. Xue, P.-T. Yap, and D. Shen, "Adversarial similarity network for evaluating image alignment in deep learning based registration," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2018, pp. 739–746.

[25] I. J. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 3, 2014, pp. 2672–2680.

[26] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1385–1392.

[27] P. Zille, T. Corpetti, L. Shao, and X. Chen, "Observation model based on scale interactions for optical flow estimation," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3281–3293, Aug. 2014.

[28] D. Rufenacht and D. Taubman, "HEVC-EPIC: Fast optical flow estimation from coded video via edge-preserving interpolation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3100–3113, Jun. 2018.

[29] G. E. Christensen and H. J. Johnson, "Consistent image registration," *IEEE Trans. Med. Imag.*, vol. 20, no. 7, pp. 568–582, Jul. 2001.

[30] J. Zhang, "Inverse-consistent deep networks for unsupervised deformable image registration," 2018, *arXiv:1809.03443*. [Online]. Available: https://arxiv.org/abs/1809.03443

[31] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9242–9251.

[32] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2794–2802.

[33] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: https://arxiv.org/abs/1701.07875

[34] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.

[35] L. Wang *et al.*, "SAR-to-optical image translation using supervised cycle-consistent adversarial networks," *IEEE Access*, vol. 7, pp. 129136–129149, 2019.

[36] K. Lata, M. Dave, and K. N. Nishanth, "Image-to-image translation using generative adversarial network," in *Proc. 3rd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Jun. 2019, pp. 186–189.

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

[38] H. Wang, W. Wu, Y. Su, Y. Duan, and P. Wang, "Image super-resolution using a improved generative adversarial network," in *Proc. IEEE 9th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2019, pp. 312–315.

[39] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," vol. 33, 2020, *arXiv:2006.10029*. [Online]. Available: https://arxiv.org/abs/2006.10029

[40] A. Sedghi *et al.*, "Semi-supervised deep metrics for image registration," 2018, *arXiv:1804.01565*. [Online]. Available: https://arxiv.org/abs/1804.01565

[41] A. Hering, S. Kuckertz, S. Heldmann, and M. P. Heinrich, "Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking," in *Bildverarbeitung für die Medizin 2019*. Springer, 2019, pp. 309–314.

[42] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[43] S. Meister, J. Hur, and S. Roth, "UnFlow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proc. AAAI*, New Orleans, LA, USA, Feb. 2018, pp. 1–9.

[44] G. Ma, Y. Zhu, and X. Zhao, "Learning image from projection: A full-automatic reconstruction (FAR) net for sparse-views computed tomography," 2019, *arXiv:1901.03454*. [Online]. Available: https://arxiv.org/abs/1901.03454

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.
[46] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015.
[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1505.00853*. [Online]. Available: https://arxiv.org/abs/1505.00853
[48] N. Ketkar, "Introduction to PyTorch," in *Deep Learning With Python*. Springer, 2017, pp. 195–208.
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[50] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.
[51] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
[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] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3061–3070.
[54] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.

Tongtong Che was born in Shandong, China, in 1995. She received the M.S. degree from the School of Information Sciences and Engineering, Shandong Normal University, China. She is currently pursuing the Ph.D. degree with the School of Biological Science and Medical Engineering, BUAA, China. Her research interests include medical image processing and deep learning.

Yuanjie Zheng (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, China, in 2006. He is currently a Full Professor with the School of Information Sciences and Engineering, Shandong Normal University, China, where he is serving as the Dean. He was a Senior Research Investigator with the Perelman School of Medicine, University of Pennsylvania, USA. His research interests include computer vision, artificial intelligence, medical image analysis, and translational medicine.

Yunshuai Yang was born in Shandong, China, in 1997. He is currently pursuing the M.S. degree with the School of Information Sciences and Engineering, Shandong Normal University, China. His research interests include computer vision and medical image analysis.

Sujuan Hou (Member, IEEE) received the Ph.D. degree from Chongqing University, China, in 2016. She is currently an Associate Professor with the School of Information Sciences and Engineering, Shandong Normal University of China. Her research interests include computer vision, video data mining, and pattern recognition.

Weikuan Jia received the Ph.D. degree from Jiangsu University, Zhenjiang, China, in 2016. He is currently an Associate Professor with the School of Information Science and Engineering, Shandong Normal University, Jinan, China. His research interests include artificial intelligence, smart agriculture, and computer vision.

Jie Yang (Member, IEEE) received the B.S. and M.S. degrees from Shanghai Jiao Tong University, China, in 1985 and 1988, respectively, and the Ph.D. degree from the University of Hamburg, Germany, in 1994. He is currently a Professor and the Director of the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China. His research interests include machine learning, image processing, and medical image analysis.

Chen Gong (Member, IEEE) received the B.E. degree from the East China University of Science and Technology in 2010 and the dual Ph.D. degree from Shanghai Jiao Tong University and the University of Technology Sydney in 2016 and 2017, respectively. He is currently a Full Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include machine learning, data mining, and learning-based vision problems.