IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 18, 2021, accepted January 28, 2021, date of publication February 1, 2021, date of current version February 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3055939

SS-SF: Piecewise 3D Scene Flow Estimation With Semantic Segmentation

CHENG FENG[1], LONG MA[1], CONGXUAN ZHANG[1,2], (Member, IEEE), ZHEN CHEN[1], LIYUE GE[1], AND SHAOFENG JIANG[1]

[1]Key Laboratory of Nondestructive Testing, Ministry of Education, Nanchang Hangkong University, Nanchang 330063, China
[2]Institute of Automation, Chinese Academy of Sciences, Beijing 100000, China

Corresponding author: Congxuan Zhang (zcxdsg@163.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFC2003800; in part by the National Natural Science Foundation of China under Grant 61772255, Grant 61866026, and Grant 61866025; in part by the Advantage Subject Team Project of Jiangxi Province under Grant 20165BCB19007; in part by the Outstanding Young Talents Program of Jiangxi Province under Grant 20192BCB23011; in part by the National Natural Science Foundation of Jiangxi Province under Grant 20202ACB214007; in part by the Aeronautical Science Foundation of China under Grant 2018ZC56008; in part by the China Postdoctoral Science Foundation under Grant 2019M650894; and in part by the Innovation Fund Designated for Graduate Students of Nanchang Hangkong University under Grant YC2019038 and Grant YC2020-S525.

ABSTRACT In order to address the issue of edge-blurring and improve the accuracy and robustness of scene flow estimation under motion occlusions, we in this article propose a piecewise 3D scene flow estimation approach with semantic segmentation, named SS-SF. First, we utilize the semantic optical flow to initialize the 3D plane and its rigid motion parameters, and then produce the initial mappings of pixel-to-segment and segment-to-plane of the input left and right image sequences. Second, we plan a novel energy function to optimize the initial mappings by using a semantic segmentation constraint term to regularize the classical scene flow model, which the optimized mappings are employed to update the assignment and motion parameters of each pixel. Third, we adopt the semantic label to extract the occlusion pixels and exploit an occlusion handling constraint to enhance the robustness of the scene flow estimation. Finally, we compare the proposed SS-SF model with several state-of-the-art approaches by using the KITTI and MPI-Sintel databases. The experimental results demonstrate that the proposed method has the advanced accuracy and robustness in scene flow estimation, especially owns the capacities of edge-preserving and occlusion handling.

INDEX TERMS Scene flow, optical flow, piece rigid, semantic segmentation, edge-preserving, occlusion handling.

I. INTRODUCTION

Dense motion estimation from consecutive frames is a focus of research in image processing and computer vision, with broad applications in human posture estimation and recognition [1], moving target segmentation and tracking [2], obstacle detection and identification [3], foreground prediction and navigation [4], facial expression recognition [5], video deblurring and coding [6], and many other fields [7], [8].

As the 3D extension of 2D optical flow, the scene flow corresponding to a dynamic scene is usually defined as a dense representation of the 3D motion field and shape structure [9], [10]. Thus, scene flow estimation can simultaneously recover dense 3D motion and geometry from stereoscopic image sequences, which generalizes the disparity and 2D optical flow computation. In spite of the fact that optical flow estimation has been rapidly advanced over the years, the progress in scene flow estimation has lacked significant achievements. The scene flow has many similarities with optical flow including constant assumptions, objective function and numerical computation scheme. These commonalities lead to the difficulties faced by the scene flow estimation, which are similar with the challenges for optical flow computation, such as illumination changing [11], large displacement [12] and motion occlusion [13].

In contrast to the classical scene flow approaches, the piecewise rigid scene flow method estimates the dense flow field by modeling the scene as a collection of planar

The associate editor coordinating the review of this manuscript and approving it for publication was Hongjun Su.

regions, which is robust to the motion occlusion [14]–[16]. However, the existing piecewise scene flow approaches may cause the issue of edge-blurring around image and motion boundaries because these models only use a random superpixel segmentation scheme to initialize the motion parameters and ignore the boundary differences of various objectives.

To address the abovementioned issue of edge-blurring and ensure the accuracy and robustness of scene flow estimation under motion occlusions, we present a piecewise rigid scene flow estimation with semantic segmentation optimization, named SS-SF. The experimental results indicate that the proposed method has high accuracy and good robustness in scene flow estimation, especially owns the benefits of edge-preserving and occlusion handling. Our main contributions are summarized in the following.

- First, we explore a semantic segmentation-based initialization framework of 3D plane and rigid motion parameters, which the proposed initialization scheme with semantic information is able to correct the initial assignment of pixels near image and motion boundaries.
- Second, we construct a novel energy function to optimize the mappings of pixel-to-segment and segment-to-plane by incorporating a semantic segmentation constraint into the classical piecewise scene flow model. The updated mappings are utilized to further optimize the pixel assignments and motion parameters, which promote the scene flow estimation to preserve image and motion boundaries.
- Third, we exploit an occlusion handling constraint by using the semantic labels of pixels to cope with the motion occlusions between the consecutive frames, by which the presented occlusion handling scheme can effectively develop the robustness of scene flow estimation in regions of occlusions.

The remainder of this article is organized as in the following. In section II, we reviewed the progress of optical flow and scene flow estimation. We then introduced a typical piecewise rigid scene flow estimation model and discussed the limitations of the traditional methods in section III. Section IV is devoted to describe the presented novel method of piecewise scene flow estimation method via semantic segmentation. In Section V, the experimental results and discussions are presented. Finally, we conclude the project in section VI.

II. RELATED WORKS

Optical Flow: After the remarkable contributions of Horn and Schunck [17] and Lucas and Kanade [18], a large number of studies led to significant development in improving the accuracy and robustness of optical flow estimation in the past decade. Despite massive reports of optical flow estimation in the recent years, it is beyond the scope of this report to review all past researches on the topic. To provide a straightforward presentation, we only discuss the most relevant publications that focused on the issues of edge-blurring and motion occlusion.

The original homogeneous regularization proposed by Horn and Schunck [17] used to blur the image and motion boundaries. Many past studies have modified the flow diffusion strategy to preserve the image or motion edges. For example, some image-driven flow diffusing models were exploited to reduce flow diffusion near image edges [19], [20]. In contrast, several reports recommended using flow-driven diffusion strategies to preserve motion boundaries caused by the over-segmentation of the image-driven models in textured regions [21], [22]. Moreover, several publications integrated the image- and flow-driven strategies in regularizing the flow field [23], [24], because not every image edge was coincided with a motion boundary. As an effective way to extract the image boundary, the semantic segmentation model has been utilized in optical flow estimation to preserve the image and motion edges [25], [26]. However, the most of current approaches were only suitable for rigid objectives.

In the past years, variety of studies focused on the issues of motion discontinuities caused by occlusions. For instance, the non-local constraint was utilized to remove the outliers of the flow field [27], in which the non-local constraint term imposes a particular smoothness assumption within a specified region of the flow field. Because it is difficult to directly minimize the total variation model with non-local constraint term, a common practice to replace the non-local constraint is by using a weighted median filter to optimize the flow field during the coarse-to-fine computing process [28]. Although the median filter can effectively enhance the robustness of optical flow estimation, it may generate biases in the occlusion regions because the occluded pixels are not always observable. To overcome the potential limitation, some studies utilized the robust optical flow models by checking the motion discontinuities in a post-processing scheme. For example, Zhang *et al.* [29] presented a dynamic regular triangulation-based occlusion detection model, and used the occlusion information to modify the weighted median filtering scheme. On the contrary, other publications [30], [31] recommended simultaneously estimating optical flow and occlusion by incorporating an occlusion constraint term into the optical flow objective function. For instance, Hur and Roth [32] exploited the occlusion-disocclusion symmetry in jointing optical flow and occlusion estimation, and presented a piecewise rigid formulation for optical flow computation.

Recently, deep learning models have being increasingly popular in optical flow computation. As a result, various convolutional neural network (CNN) frameworks were presented to improve the accuracy and efficiency of flow fields [33]–[36]. Despite that the CNN-based optical flow models have performed the superior performance on several public evaluation databases, there remain two issues for the CNN-based approaches. The first issue is that the CNN-based methods usually require a large number of training datasets,

limiting those models to be applied to real world data where ground truth is not easily accessible [37]. The other issue is that the CNN-based models are more prone to fall into the over-fitting problem, leading to obvious errors in the estimated flow field [38].

Scene Flow: Vedula *et al*. [9], [10] proposed the first 3D scene flow computation framework by utilizing optical flows of multi-views to estimate the scene motion and structure. Following their initial works, a large number of reports were published focusing on estimating scene flow with various data and systems. The existing scene flow estimation models can be roughly divided into two types: one type is the multiple-views-based approach, and the other is the RGBD-based method.

With the wide application of consumer depth sensors such as Microsoft Kinect, the RGBD scene flow estimation approach has been rapidly developed in the past years. For example, in order to gain a robust computation scheme, Gottfried *et al*. [39] investigated a variational framework for RGB-D scene flow estimation. Their model exploited a novel channel alignment algorithm to cope with the invalid and unstable depth regions. Because the depth map usually contains a large number of noises, Quiroga *et al*. [40] employed the combination of local and global constraints to overcome the influence of image noises. Herbst *et al*. [41] incorporated the color consistency into the RGBD scene flow framework to restore the image edges, because the boundaries in depth map are usually blurred. To address the issue of motion occlusion, Sun *et al*. [42] explored a layered RGBD scene flow computation scheme. Their method segmented the depth map using depth information and detected the occlusion boundaries, which is robust to complex scene and occlusion. To improve computational efficiency, Jaimez *et al*. [43] presented a GPU-based RGBD scene flow framework, which produced the real-time computation.

Despite that the RGBD-based approach provides a straightforward link to scene flow estimation, the low resolution and confidence of the depth map may limit its further development. While development in dense binocular stereo and optical flow has been both steady and significant over the years, the multiple-views-based scene flow method is becoming the focus of research. For instance, Huguet and Devernay [44] presented the first variational model for stereo scene flow computing. Their method can recover the scene flow by coupling the optical flow estimation in both cameras with dense stereo matching between the images. Because the coupling scene flow estimation may increase the computing complexity, Wedel *et al*. [45] recommended a decoupling variational framework by splitting scene flow computation into the disparity and optical flow estimation, and then utilized an optimal technique to solve the two sub-problems. Their method significantly improved both the computation accuracy and efficiency. In order to overcome the negative influence of large displacement, Basha *et al*. [46] incorporated the coarse-to-fine computation scheme into the variational scene flow framework, and improved the accuracy and reliability of scene flow estimation under large displacement motion. In order to ensure the computing efficiency, some publications recommended utilizing GPU and parallel computing strategies to achieve the real-time scene flow estimation [47].

Although the variational computation framework is capable of producing dense scene flow field, the global smoothing assumption of the classical model may be sensitive to motion discontinuities and occlusions. Popham *et al*. [48] employed an interconnected patch model to modify the classical scene flow model, which estimates the accurate scene flow at each region. Bleyer *et al*. [49] recommended a soft constraint by using object-level color models to address the issue of motion occlusion. Their method can recover the surfaces which are severely occluded between consecutive frames. Because the global regularization may blur image and motion boundaries, Vogel *et al*. [50] planned a piecewise rigid scene flow estimation model by assuming the rigid motion to be consistent over time, and then they utilized multiple frames to improve the accuracy of scene flow estimation. Furthermore, Menze *et al.* [51] exploited a slanted-plane scene flow method by modeling 3D scene as a collection of planar patches. Their method significantly improves the performance of scene flow in textureless or ambiguous regions. Afterwards, Schuster *et al*. [52] explored a multi-frame based scene flow estimation method based on pixel-wise matching and sparse-to-dense interpolation, and the presented method performs a competitive result in KITTI benchmark. Besides, a large number of publications had been presented to improve the accuracy and robustness of scene flow estimation [53]–[56].

In recent years, the CNN-based scene flow methods have shown the good performance on both computational accuracy and efficiency [57]–[59]. However, most of these CNN-based usually approaches require supervised training process and may have difficulty to be directly applied to real world data where ground truth is not easily accessible.

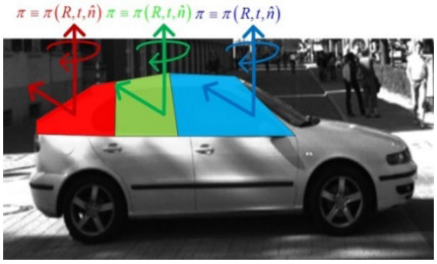


FIGURE 1. Illustration of the piecewise moving planes. one plane of a car is modeled by using a set of rigidly moving planar segments.

III. FORMULATION OF PIECEWISE SCENE FLOW MODEL

A. DEFINITION OF THE PIECEWISE MOVING PLANES

To implement the piecewise 3D scene flow estimation, Vogel *et al*. [14] described a dynamic scene as a set of piecewise planar regions moving rigidly over time. As shown in Fig. 1, each moving plane π ($\mathbf{R}$, $\mathbf{t}$, $\bar{\mathbf{n}}$) is governed by nine

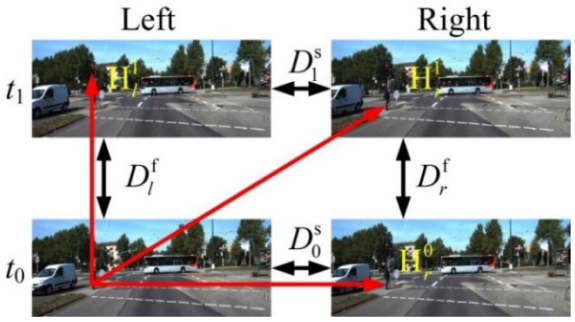


FIGURE 2. Schematic of single reference-view model.

parameters including a rotation matrix $\mathbf{R}$, a translation vector $\mathbf{t}$ and a scaled normal $\bar{\mathbf{n}}$, each with three degrees of freedom.

Assuming that the left and right cameras are calibrated; we use subscripts l and r to denote the pictures of left and right cameras, and employ superscripts $t \in T = \{0, 1, \ldots\}$ to indicate the photographing time. The stereo camera system simultaneously records two sequences from the left and right cameras and therefore provides four views for estimating the scene flow. In order to simplify the computational program, a common practice is to utilize a single reference view to model multiple views. As shown in Fig. 2, let the left view I_l^0 at time $t = 0$ denote the reference view, the transformations from the reference view to other views can be described as following:

$$\begin{cases} \mathbf{H}_r^0(\pi) = \left(\mathbf{M} - \mathbf{m}\bar{\mathbf{n}}^{\mathbf{T}}\right)\mathbf{K}^{-1} \\ \mathbf{H}_l^1(\pi) = \mathbf{K}\left(\mathbf{R} - \mathbf{t}\bar{\mathbf{n}}^{\mathbf{T}}\right)\mathbf{K}^{-1} \\ \mathbf{H}_r^1(\pi) = \left(\mathbf{MR} - (\mathbf{Mt} + \mathbf{m})\,\bar{\mathbf{n}}^{\mathbf{T}}\right)\mathbf{K}^{-1} \end{cases} \tag{1}$$

where notations $\mathbf{H}_r^0(\pi)$, $\mathbf{H}_l^1(\pi)$ and $\mathbf{H}_r^1(\pi)$ respectively denote the homographic transformations between the reference view and other three views. These definitions lead to the projection matrices $(\mathbf{K}|0)$ for the left camera and $(\mathbf{M}|\mathbf{m})$ for the right camera, where the notation $\mathbf{K}$ represents the calibration matrix of the left camera, the symbols $\mathbf{M}$ and $\mathbf{m}$ respectively denote the camera matrix of the right view and the translation vector between left and right cameras. For simplicity, the calibration matrix $\mathbf{K}$ is used to be identical for both left and right cameras [14].

B. ENERGY FUNCTION OF PIECEWISE SCENE FLOW ESTIMATION

To determine the 3D motion and depth of each pixel of the reference view, Vogel *et al.* [14] firstly defined two mappings f and g as following. Mapping f: assigns each pixel $\mathbf{p}$ of the reference view I_l^0 to a segmented region $s \in \mathbf{S}$. Mapping g: gather each segment s to a 3D moving plane $\pi \in \prod$. The symbols $\mathbf{S}$ and $\prod$ respectively denote a set of superpixel segments and a set of moving planes of reference view.

In order to determine the scene flow for each pixel of the reference view I_l^0, a global energy function was used to plan for optimizing the defined two mappings f and g, as shown in the following:

$$E(f, g) = E_D(f, g) + \lambda E_R(f, g), \tag{2}$$

where $E_D(f, g)$ and $E_R(f, g)$ denote the data term and regularization term, respectively. The notations λ is a weight of the regularization term.

To implement the piecewise scene flow estimation by minimizing the Eq. (2), the parameters of a moving 3D plane $\bar{\mathbf{n}}$ and its rigid motion $(\mathbf{R}, \mathbf{t})$ for each superpixel of the initial segmentation are estimated by using the following equations:

$$\sum_{\mathbf{p}\in s} \phi\left(\left\| T\left(H_r^0(\bar{\mathbf{n}})\,\mathbf{p}\right) - \mathbf{p}'\right\|^2\right) \to \min_{\bar{\mathbf{n}}}, \tag{3}$$

$$\sum_{\mathbf{p}\in s} \phi\left(\left\| T\left(H_l^1(\mathbf{R}, \mathbf{t})\,\mathbf{p}\right) - \mathbf{p}'\right\|^2\right) \to \min_{\mathbf{R},\mathbf{t}}, \tag{4}$$

where T denotes the conventional projection operator. The notation $\phi(x) = \log\left(1 + \frac{x}{2\sigma^2}\right)$ denotes the Lorentzian penalty function, which is used to remove outliers in the results of stereo and flow estimation. After minimizing the Eq. (3) and (4), each 3D pixel $\mathbf{p}$ of a segment $s \in \mathbf{S}$ can be matched to its 2D pixel $\mathbf{p}'$, and then the rigid motion parameters $(\mathbf{R}, \mathbf{t})$ and normal $\bar{\mathbf{n}}$ of the segment can be determined.

Given the initial parameters of segment and rigid motion, a further optimization by minimizing the Eq. (2) is operated to update the mappings f and g. The optimized motion parameters of each superpixel are fixed when the mappings f and g are optimization. Finally, the dense scene flow and disparity can be derived from the estimated rigid motion parameters.

C. DISCUSSION ON THE TRADITIONAL PIECEWISE RIGID MODEL

It is undoubted that the current piecewise rigid scene flow estimation model has performed a competitive performance on real-world datasets such as KITTI benchmark. However, the traditional piecewise rigid model may produce edge- blurring around image and motion boundaries. Fig. 3 respectively displays a reference image and the estimated flow field, disparity and scene flow error of the traditional piecewise rigid scene flow model from the KITTI online test database. For a clear presentation of the issue of edge- blurring caused by the traditional piecewise model, the close-up views of the moving vehicle region surrounded by a yellow square and the corresponding areas in the estimated results are shown at the bottom of Fig. 3. It is noticeable that the classical piecewise model yielded the edge-blurring around the boundaries of the moving car, which may be raised by motion occlusions.

To plan an accurate and robust scene flow estimation program requires a consideration of gaining an edge-preserving performance near the image and motion boundaries. In the traditional piecewise rigid model, scene flow optimization includes two stages, as summarized in the following: First,

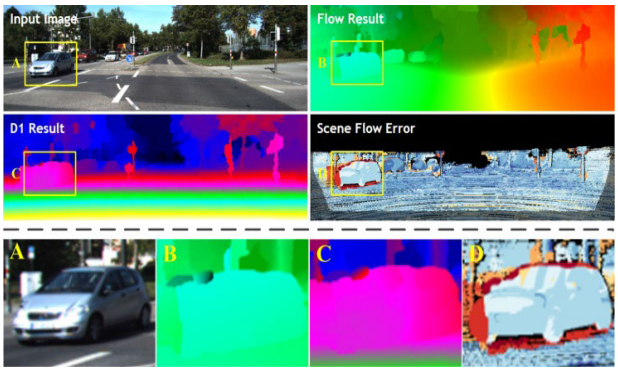


FIGURE 3. Illustration of the issue of edge-blurring caused by the tradition piecewise rigid scene flow estimation. the top respectively shows the reference image, optical flow result, disparity result and scene flow error (the yellow square indicate the moving vehicle region). the bottom displays the close-up views of the yellow squares.

after an initial superpixel segmentation, the shape and motion parameters of all segments were estimated by assigning each pixel to a segment. Second, given the initially estimated shape and motion parameters of segments, the mappings of pixel-to-segment and segment-to-plane were updated alternately, and the shape and motion parameters are optimized accordingly. Because the updated mappings may assign a pixel near image and motion edges to an incorrect segmented region or incorporate a segment around image and motion boundaries into an improper moving plane. The optimized shape and parameters of the pixels near the image and motion boundaries may be inaccurate due to the incorrect assignment of pixels and segments. As a result, the issue of edge-blurring is likely to appear around image and motion boundaries, especially under motion occlusion.

To address the issue of edge-blurring in traditional piecewise rigid scene flow estimation, the assignment of the pixels and segments should be determined cautiously. In particular, the pixels around the image and motion boundaries require more reliable segmentation, because the motion occlusion occurs near image and motion discontinuities. In order to gain accurate scene flow estimation especially at image and motion boundaries, we propose in this report a novel piecewise rigid scene flow computation method by using semantic segmentation. The detailed description of the proposed model is followed in the next Section.

IV. PIECEWISE RIGID SCENE FLOW ESTIMATION USING SEMANTIC SEGMENTATION

A. SEMANTIC OPTICAL FLOW BASED SUPERPIXEL SEGMENTATION AND MOTION PARAMETER INITIALIZATION

As a requisite input of the piecewise rigid scene flow model, optical flow plays an important role to access the superior performance of scene flow estimation because it is directly employed to compute the superpixel segmentation and initial motion parameters. The more accurate optical flow leads to better scene flow computational results. For a straightforward initialization of scene flow estimation, the traditional piecewise scene flow approach [16], [53] utilized the rigidly spatial regularization model to initialize motion parameters, and then applied the superpixel segmentation to gain initial mappings of pixel-to-segment and segment-to-plane. However, the rigidly spatial smoothing and randomized superpixel segmentation may result in edge-blurring flow field estimation.

To achieve accurate and robust scene flow estimation with additional benefit of edge-preserving, we in this report recommend the use of a semantic optical flow (SOF) model [25] to initialize superpixel segmentation and motion parameters. Although the SOF model proposed in literature [25] did not produce the top performance compared with other state-of-the-art approaches at present, it offers an available route to gain edge-preserving flow field estimation. Fig. 4 displays the semantic segmentation and flow field computation results by using SOF model. With the object segmented from background, the SOF method acquired the better flow results around image and motion boundaries compared with those of traditional piecewise rigid scene flow shown in Fig. 3.

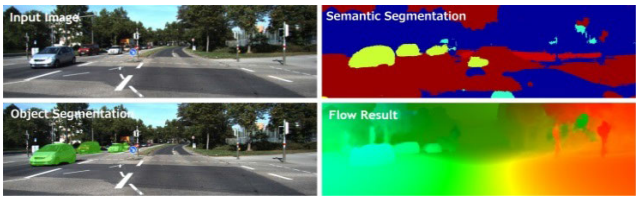


FIGURE 4. Optical flow estimation with semantic segmentation by using SOF model presented in reference [26].

By using the semantic flow to initialize motion parameters and superpixel segmentation, we firstly categorize the scene in the image into three classifications followed the SOF model, as summarized in the below:

Things: Things are corresponding to objects which are independent from the background in the scene. This category includes person, animal, airplane, bicycle, boat, bus, car, and other independently moving objects.

Planes: Planes are defined as the regions which have a broad spatial extent and are typical in the background. This category mainly contains sky, road, water and other elements with a planar shape.

Stuff: Stuff is generalized as the classes which have a complicated 3D shape and complex motion representation. Stuff usually includes building, vegetation, and other unknown elements that can't be categorized into the above two classifications.

Given the definitions of various classifications from an input image, we utilized a full-fledged semantic segmentation model by DeepLab [60] to predict the scene semantic segmentation result, which is able to achieve a satisfactory performance on image boundary segmentation. To gain the semantic flow field, we compute an initial dense flow field

by using the DiscreteFlow model [61], and then construct the flow fields of various segmented regions as following:

For the Planes region $\mathbf{R}_{Planes}$, the flow field $\mathbf{w}_{Planes}$ is modeled as $\mathbf{w}_{Planes}(\mathbf{x}_{Planes}, h_{Planes})$. The symbol $\mathbf{x}_{Planes}$ denotes all pixels located in the Planes region and the notation h_{Planes} indicates the parameter of homography [25].

For the Stuff region $\mathbf{R}_{Stuff}$, the flow field $\mathbf{w}_{Stuff}$ is created by directly using the initial optical flows in the Stuff region.

For the Things region $\mathbf{R}_{Things}$, the flow field $\mathbf{w}_{Things}$ can be computed by minimizing the following global energy function:

$$E_{Things}\left(\mathbf{w}_{Things}\right) = E_{data}\left(\mathbf{w}_{Things}\right) + \lambda_{motion}E_{motion}\left(\mathbf{w}_{Things}\right) + \lambda_{time}E_{time}\left(\mathbf{w}_{Things}\right), \quad (5)$$

where E_{data}, E_{motion} and E_{time} denote the data term, motion term and time term, respectively. The symbols λ_{motion} and λ_{time} are weights of the motion and time terms.

With the estimated flow fields of various segmented regions, the integrated semantic flow field can be obtained by compositing the flow fields of different regions [25], [60]. We then utilize the semantic flow fields to initialize superpixel segmentation and motion parameters. A description of the proposed refinement strategy is summarized in the following:

First, we adopt the semantic segmentation model to segment the input image into various semantic areas. Second, we utilize a random superpixel segmenting scheme to gain the initial superpixel regions in each semantic area of the input image. Specifically, the superpixel segmenting procedure is restrictedly implemented in every individual semantic area, allowing no any superpixel region to cross semantic boundaries. Third, we compute the initial 3D plane $\bar{\mathbf{n}}$ and its rigid motion $(\mathbf{R}, \mathbf{t})$ of each superpixel region by using Eq. (3) and (4). Thus, the initial superpixel segmentation and motion parameters are strictly coincide with the semantic segmented areas, to preserve the objective boundaries and avoid any incorrect assignment of segmented superpixel regions.

B. PIECEWISE RIGID ENERGY FUNCTION USING SEMANTIC SEGMENTATION CONSTRAINT

Despite that the dense scene flow can be directly calculated by using initial superpixel motion parameters $(\mathbf{R}, \mathbf{t})$ and scaled normal $\bar{\mathbf{n}}$ of each 3D plane, the solution would often fall into a local optimum because the initial random superpixel segmentation regions are usually not well aligned with depth and motion discontinuities of various objectives in the input images. As shown in Fig. 5, the subgraph (b) shows an initial superpixel segmenting result, in which the superpixel segmentation procedure separates a rigid plane into several trivial regions. As a result, the flow field of one rigid plane may be composed of several disparate scene flows of the superpixel segmenting regions. Although the classic piecewise rigid scene flow scheme is able to improve accuracy and robustness by using a global energy function to optimize the mappings of pixels-to-segments and segments-to-planes, the issue of edge-blurring may arise because the traditional optimization program ignores boundary differences of various objectives. As shown in subgraph (c) of Fig. 5, despite that the trivial superpixel segments have been aggregated to larger planes in the areas of car and bus, there are evident errors at object boundaries because some planes contain different objective areas. To achieve an accurate scene flow estimation, we exploit an improved energy function by incorporating a semantic segmentation constraint term to the classic model.

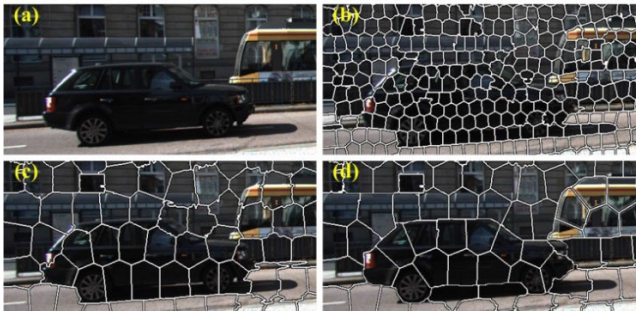


FIGURE 5. Illustration of the segmentation results of different optimization strategies. (a) reference image, (b) initial superpixel segmentation result, (c) segmentation result of traditional piecewise rigid model, (d) segmentation result of our SS-SF model.

The two major components of the optimization energy function shown in Eq. (2), the data term $E_D(f, g)$ and regularization term $E_R(f, g)$, are usually defined upon geometry and motion-based assumptions. For the data term, it is usually represented using the stereo and optical flow constraints as following:

$$E_D(f, g) = D_0^{\mathrm{s}} + D_1^{\mathrm{s}} + D_l^{\mathbf{f}} + D_r^{\mathbf{f}}, \quad (6)$$

where D_0^{s} & D_1^{s} and $D_l^{\mathbf{f}}$ & $D_r^{\mathbf{f}}$ represent respectively the stereo and optical flow constraints, as defined in Eq. (7).

$$\begin{cases} D_i^{\mathrm{s}} = \sum_{\mathbf{p}\in I_l^0} \rho\left(H_l^i(\pi_{\mathbf{p}})\,\mathbf{p}, H_r^i(\pi_{\mathbf{p}})\,\mathbf{p}\right), i \in \{0, 1\} \\ D_i^{\mathbf{f}} = \sum_{\mathbf{p}\in I_l^0} \rho\left(H_j^0(\pi_{\mathbf{p}})\,\mathbf{p}, H_j^1(\pi_{\mathbf{p}})\,\mathbf{p}\right), j \in \{l, r\} \end{cases}, \quad (7)$$

where $\pi_{\mathbf{p}} = g(f(\mathbf{p}))$ denotes the 3D moving plane at a pixel $\mathbf{p}$. The notation ρ indicates the census transform over a specific neighboring region, which enables the stereo and optical flow constraints to be robust.

By assuming that the 3D geometry and motion are piecewise smoothing, the regularization term is usually constituted by a geometric term $E_R^G(f, g)$ and a motion term $E_R^M(f, g)$ as following:

$$E_R(f, g) = E_R^G(f, g) + E_R^M(f, g), \quad (8)$$

where:

$$E_R^G(f, g) = \sum_{(\mathbf{p},\mathbf{q})\in\mathbf{N}} \omega_{\mathbf{p},\mathbf{q}}\psi\left(\left\|\mathbf{d}_1^G\right\|^2 + \left\|\mathbf{d}_2^G\right\|^2 + \left\langle\mathbf{d}_1^G, \mathbf{d}_2^G\right\rangle + \gamma\left\|\mathbf{d}_n^G\right\|^2\right). \quad (9)$$

$$E_R^M(f,g) = \sum_{(\mathbf{p},\mathbf{q})\in\mathbf{N}} \omega_{\mathbf{p},\mathbf{q}}\psi\left(\left\|\mathbf{d}_1^M\right\|^2 + \left\|\mathbf{d}_2^M\right\|^2 + \left\langle\mathbf{d}_1^M, \mathbf{d}_2^M\right\rangle + \gamma\left\|\mathbf{d}_n^M\right\|^2\right). \quad (10)$$

In Eq. (9) and (10), the pixels $\mathbf{p}$ and q are adjacent in a specific region $\mathbf{N}$, however they are assigned to independent segmented planes $\pi_{\mathbf{p}} = g(f(\mathbf{p}))$ and $\pi_{\mathbf{q}} = g(f(\mathbf{q}))$. The symbol $\omega_{\mathbf{p},\mathbf{q}}$ indicates a weight for determining the length of the common edge of adjacent pixels $\mathbf{p}$ and $\mathbf{q}$ in the region $\mathbf{N}$, and the notations $\left(\mathbf{d}_1^G, \mathbf{d}_2^G\right)\&\left(\mathbf{d}_1^M, \mathbf{d}_2^M\right)$ represent the endpoint distances of the 3D geometry and motion between adjacent pixels $\mathbf{p}$ and $\mathbf{q}$.

In order to address the issue of over-smoothing at image and motion boundaries, we present a semantic segmentation term to regularize the data and regularization terms, as shown in the following:

$$E_S(f) = \sum_{\substack{(\mathbf{p},\mathbf{q})\in \mathrm{N}_{se},\mathrm{N}_{se}\in I_l^0 \\ f(\mathbf{p})\neq f(\mathbf{q})}} \exp\left(\frac{-\delta\left|I_l^0(\mathbf{p}) - I_l^0(\mathbf{q})\right|}{\sigma(\mathbf{p},\mathbf{q}) + \varepsilon}\right) + \sum_{\mathbf{p}\in I_l^0} \begin{cases} 0, \exists \mathbf{e}\in\xi(i) : \|\mathbf{e}-\mathbf{p}\|_\infty < N_S \\ \infty, else. \end{cases}, \quad (11)$$

In Eq. (11), the first item is the semantic segmentation constraint which is employed to optimize the mapping of pixels-to-segments in a common semantic area. The notation $\mathbf{p}$ and $\mathbf{q}$ denote the adjacent pixels in a common semantic area N_{se}, but located in different superpixel segments. The notation σ indicates standard deviation, and symbols δ & ε are adjustment coefficients. To prevent superpixel segments from becoming overly large, we utilize a spatial segmentation constraint as a supplement item for the semantic segmentation constraint, as shown in Eq. (11). The spatial segmentation constraint offers a link between a segment and its seed point $\mathbf{e}\in\xi(i)$, which can restrict the number of candidate segments for a pixel and limit the maximum size of a segment.

To make an overview presentation, we combine the presented semantic segmentation term with basic data and regularization terms to construct an optimization energy function as following:

$$E(f,g) = E_D(f,g) + \lambda E_R(f,g) + \mu E_S(f), \quad (12)$$

where $E_D(f,g)$, $E_R(f,g)$ and $E_S(f)$ respectively denote the data, regularization and semantic segmentation terms. The symbols λ and μ are weights of the regularization and semantic segmentation terms. To illustrate the benefit of our model in segmenting optimization, Fig. 5(d) displays the output segmentation results. The presented model presents an accurate segmentation result that the segmented rigid planes are well coincident with the objective areas, especially a superior performance of edge-preserving at objective boundaries.

C. SEMANTIC SCENE FLOW ESTIMATION WITH OCCLUSION-AWARE CONSTRAINT

It is undoubted that occlusion is an awful challenge for most existing scene flow computational models, because most of basic geometry and motion constant assumptions will be invalid under occlusions. In order to ensure the robustness of scene flow estimation, we present in this section an occlusion-aware constraint term to cope with the occlusions by using the semantic segmentation information.

To simplify the presentation of the occlusion-aware constraint for scene flow, we redefine the basic data term of Eq. (12) by using the form of pseudo-Boolean function [62], as shown in the following:

$$D(\mathbf{x}) = \sum_{\mathbf{p}\in I_l^0} u_{\mathbf{p}}^0\left(1 - x_{\mathbf{p}}\right) + u_{\mathbf{p}}^1 x_{\mathbf{p}}, \quad (13)$$

where $x_{\mathbf{p}} \in \{0, 1\}$ indicates the segment assignment of pixel $\mathbf{p}$, and the symbol $\mathbf{x}$ denotes all segment assignments of the reference frame I_l^0. When $x_{\mathbf{p}} = 0$, the pixel $\mathbf{p}$ retains the previous segment assignment; In contrast, the pixel $\mathbf{p}$ switches to another segment assignment. The notation u_p^0 and $u_{\mathbf{p}}^1$ respectively represent data penalties if the pixel $\mathbf{p}$ belongs to the previous segment or switches to another segment.

Because the semantic segmentation result provides a semantic label for each pixel, it prompts us to check whether a pixel is occluded or not between the input frames. By utilizing the semantic labels of pixels, the occlusion-aware constraint term can be expressed as following:

$$D(\mathbf{x}) = \sum_{\mathbf{p}\in I_l^0}\left(\begin{array}{c}\theta_{occ}\Gamma\left(g_0^{\mathbf{p}} \neq g_1^{\mathbf{p}}\right) + \\ \left[u_{\mathbf{p}}^0\left(1 - x_{\mathbf{p}}\right) + u_{\mathbf{p}}^1 x_{\mathbf{p}}\right]\Gamma\left(g_0^{\mathbf{p}} = g_1^{\mathbf{p}}\right)\end{array}\right), \quad (14)$$

where $\theta_{occ} \in (0, 1)$ denotes a constant penalty [63], and the notations $g_0^{\mathbf{p}}$ and $g_1^{\mathbf{p}}$ respectively indicate the semantic labels of pixel $\mathbf{p}$ at the reference and next frames. Since the semantic label of a pixel should be constant between the input frames if the pixel is non-occluded, an indicator function $\Gamma(\cdot)$ with binary outputs is employed to denote the status of the pixel $\mathbf{p}$. When $g_0^{\mathbf{p}} = g_1^{\mathbf{p}}$, the pixel $\mathbf{p}$ is non-occluded, the indicator function $\Gamma\left(g_0^{\mathbf{p}} = g_1^{\mathbf{p}}\right) = 1$ and $\Gamma\left(g_0^{\mathbf{p}} \neq g_1^{\mathbf{p}}\right) = 0$; Otherwise, the pixel $\mathbf{p}$ is occluded, the indicator function $\Gamma\left(g_0^{\mathbf{p}} = g_1^{\mathbf{p}}\right) = 0$ and $\Gamma\left(g_0^{\mathbf{p}} \neq g_1^{\mathbf{p}}\right) = 1$.

To implement the semantic scene flow estimation with occlusion-aware constraint, we replace the basic data term of Eq. (12) by the presented occlusion-aware constraint term of Eq. (14). As a result, the scene flow estimation of the non-occluded pixel still relies on the geometry and motion constant assumptions because the indicator function of non-occluded pixel leads the occlusion-aware constraint term to return to the basic data term. On the contrary, the scene flow of the occluded pixels will be generated by flow diffusion from the neighboring pixels, because the indicator function of occluded pixel guides the occlusion-aware constraint term to a penalty constant.

For a clear presentation of the proposed piecewise scene flow estimation with semantic segmentation, we briefly summarize implementation steps as following:

Step.1 Segment input frames into variously semantic image areas using the semantic segmentation model [60].

Step.2 Compute an initially semantic flow field by using the DiscreteFlow model [61], and estimate an initial disparity result via a semiglobal matching model [64].

Step.3 Apply a random superpixel segmenting to the input reference frame to produce an initial superpixel segmentation field.

Step.4 Initialize the rigid motion parameters $(\mathbf{R}, \mathbf{t})$, normal $\bar{\mathbf{n}}$ of superpixel segments, and mapping of pixel-to-segment by using the initial semantic flow field and stereo disparity to minimize the Eq. (3) and (4).

Step.5 Determine the occluded and non-occluded pixels in the input frames by checking the semantic label of each pixel.

Step.6 Minimize the energy function in Eq. (12) to optimize mappings of pixel-to-segment and segment-to- plane.

Step.6.1 Update the mapping of segment-to-plane by fixing the mapping f of pixel-to-segment.

Step.6.2 Update the mapping f of pixel-to-segment by fixing the mapping g of segment-to-plane.

Step.7 Update the rigid motion parameters $(\mathbf{R}, \mathbf{t})$, normal $\bar{\mathbf{n}}$ of each pixel by using optimized mappings f and g.

Step.8 Output the piecewise scene flow result using the refined rigid motion parameters $(\mathbf{R}, \mathbf{t})$ and normal $\bar{\mathbf{n}}$.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. ERROR MEASUREMENTS

Because a scene flow connects directly with optical flow and disparity, a better performance on optical flow and disparity indicates a superior scene flow result. The KITTI benchmark recommends primarily using the metrics of optical flow and stereo matching to indicate performance of scene flow, as shown in the following:

$$Fl - all = \frac{P_1}{all} \times 100\%, \tag{15}$$

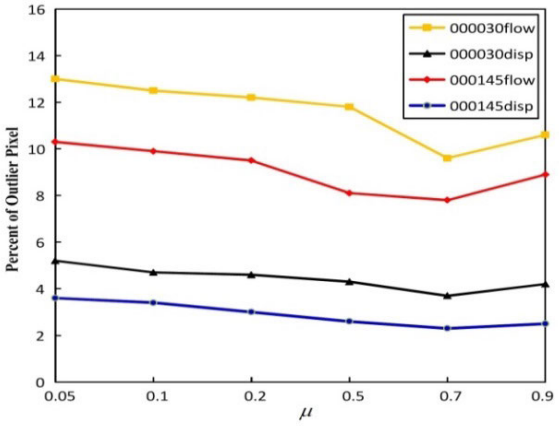
$$D1 - all = \frac{P_2}{all} \times 100\%, \tag{16}$$

where P_1 and P_2 respectively denote the number of outlier pixels (EPE>3) in the flow field and disparity map, the symbol *all* represents the entire image. Thus, the metrics of $Fl - all$ and $D1 - all$ indicate the percent of outlier pixels in the flow field and disparity maps, respectively.

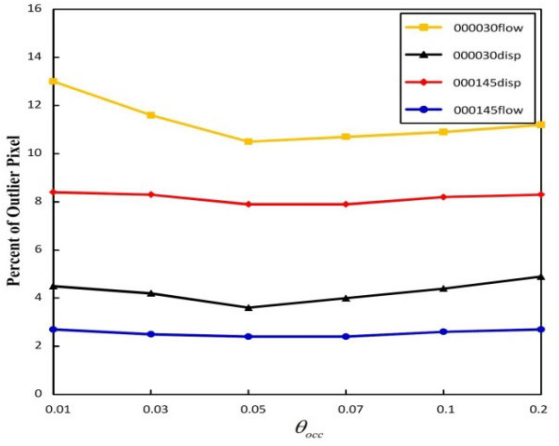
Additionally, for a straightforward online evaluation of scene flow, the KITTI benchmark counts the outliers in either optical flow or disparity results to indicate the performance of scene flow as following:

$$SF - all = \frac{P_1 \cup P_2}{all} \times 100\%, \tag{17}$$

where metric of $SF - all$ indicates the percent of outlier pixels in scene flow field.



(a) Variation of percent of outlier pixels respect to different values of weight μ .



(b) Variation of percent of outlier pixels respect to different values of constant penalty θ_{occ} .

FIGURE 6. The variation of the percent of outlier pixels in optical flow and disparity with different values of free parameters.

B. DISCUSSION OF FREE PARAMETERS

In the presented piecewise scene flow estimation method with semantic segmentation, there are several free parameters that deserve a careful consideration including the weight λ of the regularization term, weight μ of the semantic segmentation term and constant penalty θ_{occ} of the occlusion-aware constraint term. Because the regularization is used to blur image and motion edges, we set the weight $\lambda = 0.1$ to produce a slight smoothing diffusion by referring reference [14].

To choose reasonable values for the other two parameters, we run our model on the KITTI training sets including 000030 and 000145 with different values of μ(0.05, 0.1, 0.2, 0.5, 0.7 and 0.9) or θ_{occ}(0.01, 0.03, 0.05, 0.07, 0.1 and 0.2), and recorded results of percent of outlier pixels in the estimated optical flows and disparities for each value of μ or θ_{occ} (Fig. 6 (a) and (b)). As shown in Fig. 6, different choices of the free parameters can influence the accuracy of scene flow estimation significantly. In the following experiments, we set weight $\mu = 0.7$ to ensure the effect of semantic segmentation constraint, and fixed the constant penalty $\theta_{occ} = 0.05$ to apply strict penalty to occluded pixels.

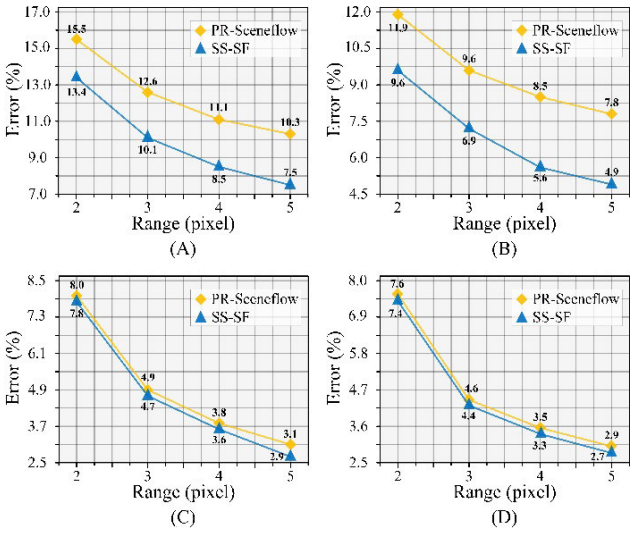


FIGURE 7. The quantitative comparison results between the SS-SF and PR-Sceneflow methods on KITTI training datasets. (A) The error of optical flow over entire image. (B) The error of optical flow over non-occlusion area. (C) The error of stereo matching over entire image. (B) The error of stereo matching over non-occlusion area.

C. COMPARISON WITH THE BASELINE METHOD

Because the PR-Sceneflow model [14] is a representative of the piecewise scene flow estimation method that uses the superpixel segmentation and traditional energy function to optimize piecewise scene flow. In order to demonstrate benefits of the proposed SS-SF model in edge-preserving and occlusion handling, we conduct a comparison experiment between the proposed SS-SF method and PR-Sceneflow approach, which is taken as a baseline method, by using the KITTI training datasets.

For a detailed evaluation, we respectively summarize results of the percent of outlier pixels in the entire image and non-occlusion areas of estimated optical flow and disparity with various outlier thresholds of 2, 3, 4, and 5 pixels. The quantitative comparison between the proposed SS-SF model and PR-Sceneflow method on KITTI training datasets are reported in Fig. 7. It is undisputed that the proposed SS-SF model performed superior results on scene flow estimation, as shown in the significantly decreased statistical errors compared with those of the PR-Sceneflow method.

For a visual comparison, we respectively display the ground truths, estimated optical flows and disparity maps of the SS-SF and PR-Sceneflow methods in Fig. 8. The proposed SS-SF model evidently produced the better performance on both optical flows and disparity maps because its estimated flow fields and disparities are more coincident with the ground truths, especially gain the satisfied results in the regions of motion boundaries and occlusions.

The quantitative results and visual comparison of KITTI training datasets between the two methods demonstrate that the proposed SS-SF method achieved more accurate and robust performance on scene flow computation, probably due to its better edge -preserving and occlusion handling.

D. ABLATION EXPERIMENT

In order to validate the benefit of each module of the proposed SS-SF method, we utilize the KITTI training sets to conduct an ablation experiment. Table 1 summarizes the results of the percent of outlier pixels (EPE $>$ 3pixels) in the entire image (**all**) and non-occlusion areas (**noc**) of the estimated optical flow and disparity of SS-SF method with different modeling choices, where SS-SF-OA model denotes the SS-SF method without the occlusion-aware constraint and SS-SF-SS model represents that replacing the original semantic segmentation DeepLab model [60] used in the proposed SS-SF method by an improved segmentation DeepLabV3+ model [65].

TABLE 1. Comparison results of optical flow and disparity errors of SS-SF method with different modeling choices.

Method	Optical Flow		Disparity	
	all	**noc**	**all**	**noc**
SS-SF	10.10	6.86	4.71	4.39
SS-SF-OA	10.54	6.94	4.73	4.43
SS-SF-SS	10.14	6.89	4.75	4.42

As shown in Table 1, the comparison results between different modeling choices indicate that SS-SF method without occlusion aware constraint leads to significant degradation in performance of scene flow estimation, because the optical flow errors of SS-SF-OA model are significantly increased compared with those of SS-SF method. Although the improved DeepLabV3+ model performs a better performance compared with the original DeepLab model on some specialized benchmarks of semantic segmentation, the optical flow and disparity errors of SS-SF-SS model are slightly increased compared with the SS-SF method. This is because the DeepLabV3+ model classifies the small objects into the background regions, which may lead to an inaccurate optimization in piecewise scene flow estimation.

For a visual comparison and discussion, we display the flow field and disparity results of SS-SF method with different modeling choices in Fig. 9, where the red and white squares respectively indicate some areas of motion boundaries and occlusions in flow fields and disparities. As can be seen from Fig. 9, either removing the occlusion-aware constraint or replacing the original semantic segmentation model by an improved method leads to the issue of edge-blurring. The comparison results between the SS-SF method and the different modeling choices indicate that the proposed occlusion-aware constraint and semantic segmentation scheme are beneficial for improving the performance of scene flow estimation, especially in regions of occlusions and motion boundaries.

E. COMPARISON RESULTS FROM KITTI TEST DATASETS

In recent years, the KITTI benchmark has being increasingly popular in evaluating accuracy and robustness of various vision related tasks such as stereo matching, optical flow and scene flow, because it was produced using a moving

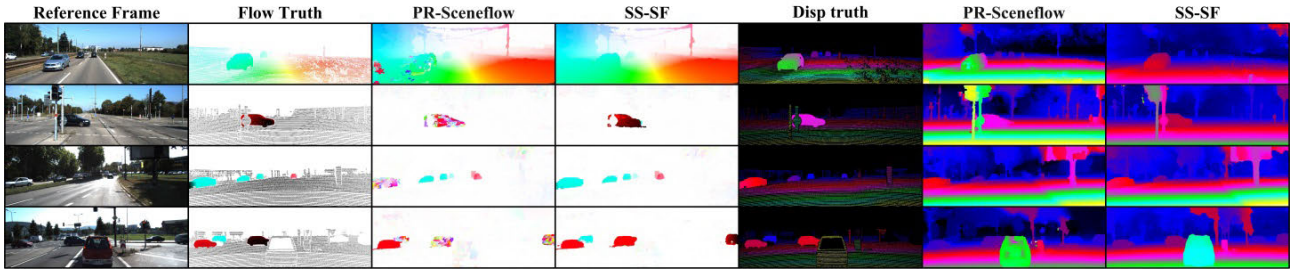
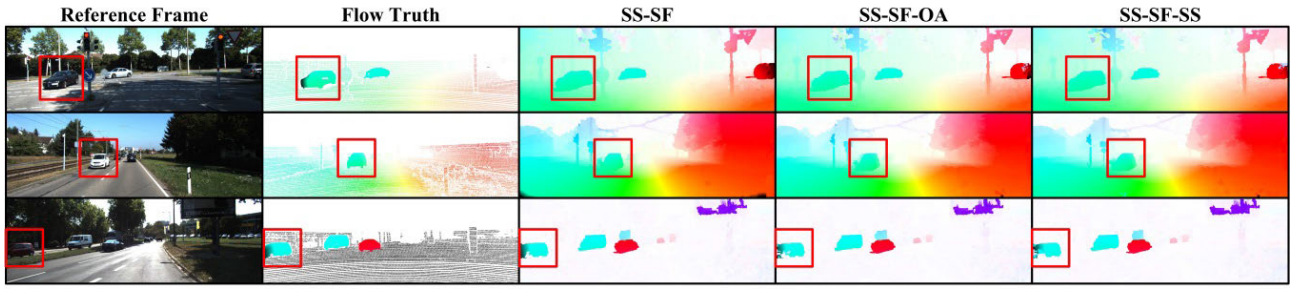
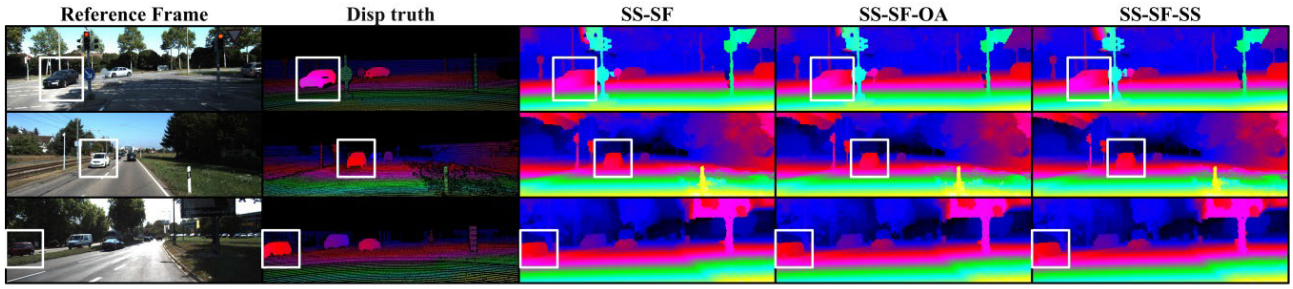


FIGURE 8. The visual comparison between the SS-SF and PR-Sceneflow methods on KITTI training datasets. From top to bottom: the datasets of 000041, 000044, 000096 and 000112.



(a) Optical flow results of SS-SF method with different modeling choices



(b) Disparity results of SS-SF method with different modeling choices

FIGURE 9. Optical flow and disparity results of SS-SF method with different modeling choices. From top to bottom: the datasets of 000011, 000036 and 000089. The red and white squares respectively indicate some areas of motion boundaries and occlusions in flow fields and disparities.

vehicle. In order to examine the accuracy and robustness of the proposed SS-SF approach, we run our SS-SF model on KITTI online test datasets to conduct a comprehensive comparison with several state-of-the-art scene flow methods including PCOF-LDOF [47], PR-Sceneflow [14], PRSM [50], DWBSF [54], CSF [55], SceneFFields [56], FSF+MS [53], OSF [51], SFF++ [52], PWOC-3D [57], Self-Mono-SF-ft [59] and Stereo expansion [66], in which the PCOF-LDOF, PR-Sceneflow, DWBSF, CSF, SceneFFields, and OSF methods are the dual-frame-based classical scene flow approach, the PRSM, FSF+MS and SFF++ methods are the multi-frame-based classical approach, and the PWOC-3D, Self-Mono-SF-ft and Stereo expansion methods are the CNN-based scene flow approach.

TABLE 2 respectively lists the quantitative comparison results of the various methods evaluated on KITTI 2015 test datasets, where the error metrics include disparity errors for two frames (D1, D2), optical flow error (Fl) for the reference frame and scene flow error (SF) for foreground pixels (fg), background pixels (bg), and all pixels (all). Based on the evaluation criteria of the KITTI benchmark, we rank various evaluated models in TABLE 2 according to the result of SF-all measurement. As can be seen from Table 2, the proposed SS-SF approach performs the fourth-best results among all the evaluated approaches and it achieves the second-best performance in the dual-frame-based classical methods. The comparison results demonstrate that the proposed SS-SF method performs a competitive performance on the KITTI 2015 test datasets.

Because KITTI 2015 test datasets contain 195 image sequences which have the different motion scenes, we run our SS-SF method and several state-of-the-art approaches on some test sequences including complex scenes and occlusions to make a further comparison. Table 3 summarizes the comparison results of scene flow errors of the various methods tested on image 1 and image 11. It is noticeable that the proposed SS-SF method respectively achieves the second-best result and the third-best result of metric SF-fg

TABLE 2. The quantitative comparison results of KITTI test datasets.

Rank	Method	Dual-frame	CNN	D1-bg	D1-fg	D1-all	D2-bg	D2-fg	D2-all	Fl-bg	Fl-fg	Fl-all	SF-bg	SF-fg	SF-all
1	Stereo expansion[66]	√	√	1.48	3.46	1.81	3.39	8.54	4.25	5.83	8.66	6.30	7.06	13.44	8.12
2	PRSM[50]			3.02	10.52	4.27	5.13	15.11	6.79	5.33	13.40	6.68	6.61	20.79	8.97
3	OSF[51]	√		4.11	11.12	5.28	5.01	17.28	7.06	5.38	17.61	7.41	6.68	24.59	9.66
4	SS-SF	√		3.59	13.11	5.18	7.50	21.79	9.87	8.17	25.20	11.00	9.64	32.88	13.51
5	SFF++[52]			4.27	12.38	5.62	7.31	18.12	9.11	10.63	17.48	11.77	12.44	25.33	14.59
6	FSF+MS[53]			5.72	11.84	6.74	7.57	21.28	9.85	8.48	25.43	11.30	11.17	33.91	14.96
7	PWOC-3D[57]	√	√	4.19	9.82	5.13	7.21	14.73	8.46	12.40	15.78	12.96	14.30	22.66	15.69
8	CSF[55]	√		4.57	13.04	5.98	7.92	20.76	10.06	10.40	25.78	12.96	12.21	33.21	15.71
9	SceneFFields[56]	√		5.12	13.83	6.57	8.47	21.83	10.69	10.58	24.41	12.88	12.48	32.28	15.78
10	PR-Sceneflow[14]	√		4.74	13.74	6.24	11.14	20.47	12.69	11.73	24.33	13.83	13.49	31.22	16.44
11	PCOF-LDOF[47]	√		6.31	19.24	8.46	19.09	30.54	20.99	14.34	38.32	18.33	25.26	49.39	29.27
12	Self-Mono-SF-ft[59]	√	√	20.72	29.41	22.16	23.83	32.29	25.24	15.51	17.96	15.91	31.51	45.77	33.88
13	DWBSF[54]	√		19.61	22.69	20.12	35.72	28.15	34.46	40.74	31.16	39.14	46.42	40.76	45.48

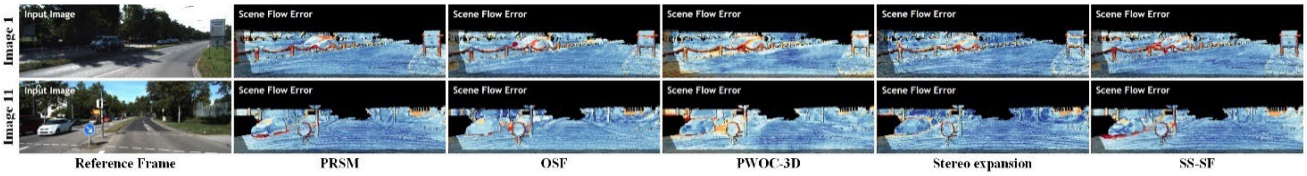


FIGURE 10. Scene flow error maps of PRSM OSF, PWOC-3D, Stereo expansion and SS-SF approaches tested on the image 1 and image 11.

TABLE 3. The comparison results of scene flow errors on some KITTI datasets.

Method	Test image 1			Test image 11		
	SF-bg	SF-fg	SF-all	SF-bg	SF-fg	SF-all
Stereo expansion[66]	3.26	14.47	4.51	3.43	4.33	3.59
PRSM[50]	3.96	17.43	5.46	3.31	5.22	3.65
OSF[51]	4.13	25.68	6.54	3.04	9.16	4.14
SS-SF	4.68	14.97	5.83	4.25	6.56	4.67
SFF++[52]	21.30	38.22	23.19	25.32	15.05	23.48
FSF+MS[53]	4.82	21.89	6.72	5.51	16.15	7.42
PWOC-3D[57]	9.34	28.65	11.50	5.82	8.18	6.24
CSF[55]	6.22	29.66	8.84	5.45	13.83	6.95
SceneFFields[56]	5.82	26.91	8.18	4.92	15.70	6.86
PR-Sceneflow[14]	5.74	26.97	8.12	5.24	10.37	6.16
PCOF-LDOF[47]	14.91	35.62	17.22	13.61	23.85	15.45
Self-Mono-SF-ft[59]	21.69	53.48	25.25	8.94	18.21	10.61
DWBSF[54]	31.90	34.82	32.23	34.84	31.09	34.17

on test image 1 and image 11. This demonstrates that the SS-SF method performs a good performance in the foreground areas. Because the objects are usually classified into the foreground regions, a good result on metric SF-fg indicates a good performance of scene flow estimation in object areas. Fig. 10 respectively displays the scene flow error maps of PRSM, OSF, PWOC-3D, Stereo expansion and SS-SF approaches tested on the image 1 and image 11, which indicates the proposed SS-SF method performs a good performance in the object areas.

To show the benefit of the proposed SS-SF method in coping with edge-blurring under motion occlusions, Fig. 11 lists the optical flow and disparity results of the various comparison approaches evaluated on test image 1 and image 11, where the black boxes indicate some regions including objects and occlusions. For a specific visual comparison, we display the close-up views within the black squares in Fig. 12. As can be seen from Fig. 12, the OSF and stereo expansion methods result in the over-segmentation and the PWOC-3D approach causes the over-smoothing around the object edges and motion boundaries. The PRSM method performs better results compared with the OSF and PWOC-3D approaches, however it blurs some edges and boundaries in the textureless and occluded areas. The proposed SS-SF method achieves a good performance because the object edges and motion boundaries are undamaged and distinct in both optical flow and disparity results, due to its significant benefit of edge-preserving.

F. COMPARISON RESULTS FROM MPI-SINTEL DATASETS

To further examine its capability in dealing with edge-blurring and occlusions, we evaluate our SS-SF model on the MPI-Sintel database because the MPI-Sintel datasets include large displacements, non-rigid deformation, motion occlusions, atmospheric effects and complex scenes.

For a quantitative evaluation, we measure the percentage of outlier pixels (EPE $>$ 3) for optical flow and disparity according to the KITTI benchmark. Because the MPI-Sintel database has not published the estimated results and ranks for any scene flow methods, we compare the

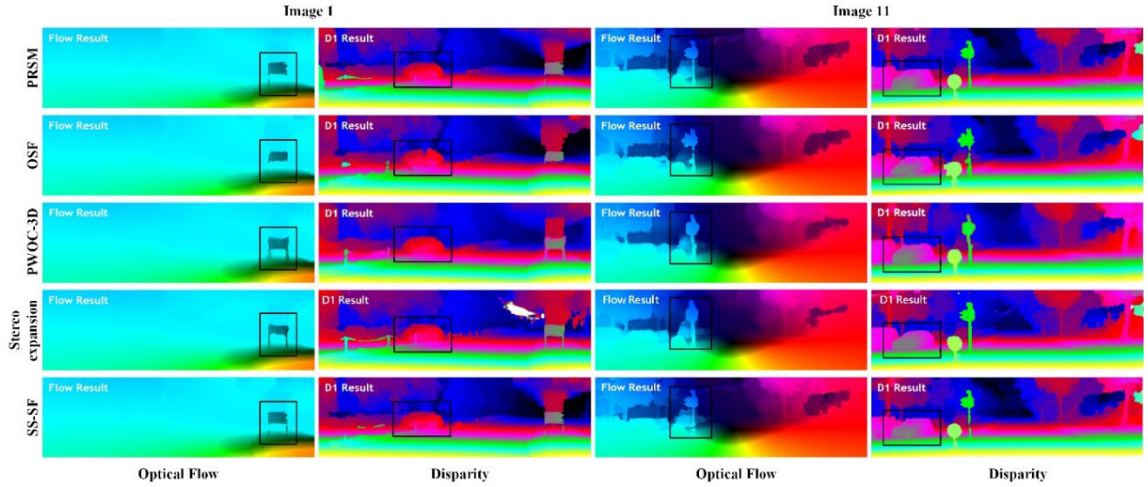


FIGURE 11. Optical flow and disparity results of the SS-SF and several comparison methods tested on image 1 and image 11.

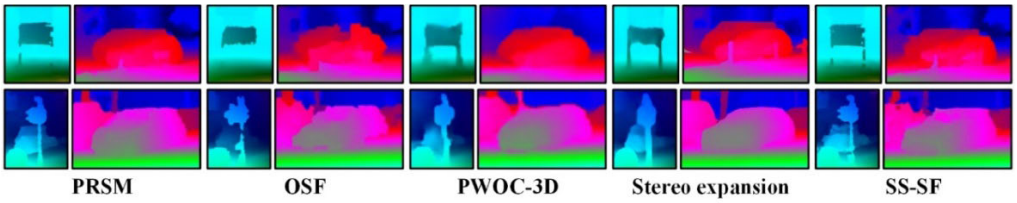


FIGURE 12. The close-up views within the black squares in the optical flow and disparity results.

optical flow errors of SS-SF models on MPI-Sintel datasets with those of some state-of-the-art methods, including PR-Sceneflow [14], FSF+MS [53], SceneFFields [56], OSF [51] and Stereo expansion [66] by using the results published in reference [53] and the published open-source code. TABLE 4 summarizes the comparison results, indicating that the proposed SS-SF method achieved the best performance among the evaluated models on average value of optical flow indicators. Although the OSF and Stereo expansion methods produced better results on KITTI test database, the presented SS-SF model won the competition on MPI-Sintel benchmark. Since the optical flow is a direct indicator for scene flow, the comparison results of optical flow on MPI-Sintel datasets indicate that the proposed SS-SF method has the advanced accuracy and robustness in scene flow estimating, particularly is capable of edge-preserving and occlusion handling.

TABLE 4. Comparison results of optical flow from MPI-Sintel datasets.

Sequence	PR-Sceneflow [14]	FSF+MS [53]	SceneFFields [56]	OSF [51]	Stereo expansion [66]	SS-SF
alley_1	2.05	2.11	5.94	7.33	3.02	2.04
alley_2	1.62	1.20	2.85	1.44	1.71	1.61
ambush_2	66.22	72.68	90.92	87.37	76.19	66.33
ambush_4	48.60	45.23	60.03	49.16	54.34	48.55
ambush_5	30.77	24.82	46.92	44.70	39.53	30.35
ambush_6	49.77	44.05	57.06	54.75	59.45	49.37
ambush_7	3.92	27.87	13.66	22.47	7.11	3.97
bamboo_1	2.86	4.11	6.11	4.04	4.03	2.86
bamboo_2	5.05	3.65	5.84	4.86	6.33	5.06
bandage_1	4.72	4.00	3.82	18.40	7.47	4.75
bandage_2	5.33	4.76	10.72	13.12	4.67	5.32
cave_4	16.79	14.62	15.63	33.94	20.76	16.78
market_2	5.81	5.17	7.11	10.08	6.06	5.79
market_5	41.33	26.31	40.77	29.58	39.83	41.36
market_6	22.84	13.13	28.92	16.39	16.97	22.87
moutain_1	5.12	17.05	90.60	88.60	3.86	5.05
shaman_2	1.66	0.56	8.85	1.67	0.93	1.67
shaman_3	2.93	1.31	15.91	11.45	2.43	2.99
sleeping_2	0.02	0.02	0.61	0.01	0.04	0.02
temple_2	12.61	9.66	29.58	10.52	13.70	12.57
temple_3	40.98	62.34	72.28	81.39	48.52	41.00
Average	17.67	18.32	29.24	28.16	19.85	17.63

G. RUNTIMES

To make a comprehensive comparison between the proposed SS-SF and the other state-of-the-art methods, Table 5 summarizes the average runtimes of the various evaluated approaches tested on the KITTI 2015 test datasets.

As can be seen from Table 5, the PWOC-3D, Self-Mono-SF-ft and Stereo expansion methods achieve the best performance on computational efficiency because the CNN-based approaches have the significant benefit of real-time computation. However these CNN-based methods usually require a large number of datasets to train the network parameters and may have difficulty to be directly applied to real world data. Though the PRSM

TABLE 5. The quantitative comparison results of KITTI Test datasets.

Method	Dual-frame	CNN	Runtime (unit: second)
Stereo expansion[66]	√	√	2
PRSM[50]			300
OSF[51]	√		390
SS-SF	√		180
SFF++[52]			78
FSF+MS[53]			2.7
PWOC-3D[57]	√	√	0.13
CSF[55]	√		80
SceneFFields[56]	√		60
PR-Sceneflow[14]	√		150
PCOF-LDOF[47]	√		50
Self-Mono-SF-ft[59]	√	√	0.09
DWBSF[54]	√		420

and OSF methods performed the better performance on the metrics of scene flow error compared with those of the proposed SS-SF method, they cost much more time consumption. The proposed SS-SF method is implemented by MATLAB2010 using a Lenovo computer equipped with an Intel Core I7-6700K CPU. Because we use a large number of iterations to optimize the energy function of piecewise mapping, the proposed SS-SF method requires more time consumption than that of some dual-frame based approaches including CSF, SceneFFields, PR-Sceneflow and PCOF-LDOF. Nevertheless, the proposed SS-SF performs the significantly better results in computation accuracy than those of the abovementioned dual-frame approaches, especially gains the capacity of edge-preserving and occlusion handling.

VI. CONCLUSION

In this report, we started by reviewing the progress and several previous approaches in the fields of scene flow and optical flow estimation. Then we presented the traditional piecewise scene flow computation model and discussed its limitations in edge-preserving and occlusion handling.

To deal with the issues of edge-blurring and motion occlusions, we proposed a novel piecewise rigid scene flow estimation method by using the semantic segmentation, named SS-SF. We first adopted the semantic flow field to initialize the 3D plane, motion parameters and mappings of pixel-to-segment and segment-to-plane. Second, we exploited an improved energy function for optimizing the mappings by incorporating a semantic segmentation constraint term, in which the assignment and motion parameters of each pixel can be optimized by using updated mappings. Third, we explored an occlusion handling constraint to cope with the motion occlusion, in which the presented occlusion handling scheme was able to improve the robustness of scene flow estimation. At last, we tested the proposed SS-SF method on KITTI and MPI-Sintel databases to conduct convincing comparisons with some of state-of-the-art approaches. The evaluation results indicate that the presented SS-SF method presented the excellent performance on both accuracy and robustness, especially showed significant benefits of edge-preserving and occlusion handling.

REFERENCES

[1] T. Pfister, J. Charles, and A. Zisserman, "Flowing ConvNets for human pose estimation in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1913–1921.

[2] L. Chen, J. Shen, W. Wang, and B. Ni, "Video object segmentation via dense trajectories," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2225–2234, Dec. 2015.

[3] K. McGuire, G. de Croon, C. De Wagter, K. Tuyls, and H. Kappen, "Efficient optical flow and stereo vision for velocity estimation and obstacle avoidance on an autonomous pocket drone," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 1070–1076, Apr. 2017.

[4] Y. Wan, Z. Miao, X.-P. Zhang, Z. Tang, and Z. Wang, "Illumination robust video foreground prediction based on color recovering," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 637–652, Apr. 2014.

[5] M. H. Kabir, M. S. Salekin, M. Z. Uddin, and M. Abdullah-Al-Wadud, "Facial expression recognition from depth video with patterns of oriented motion flow," *IEEE Access*, vol. 5, pp. 8880–8889, 2017.

[6] L. Pan, Y. Dai, M. Liu, and F. Porikli, "Simultaneous stereo video deblurring and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6987–6996.

[7] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2480–2492, Jun. 2016.

[8] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 416–425.

[9] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," in *Proc. IEEE Conf. Comput. Vis.*, Sep. 1999, pp. 722–729.

[10] S. Vedula, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 475–480, Mar. 2005.

[11] P. F. U. Gotardo, T. Simon, Y. Sheikh, and I. Matthews, "Photogeometric scene flow for high-detail dynamic 3D reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 846–854.

[12] A. Zanfir and C. Sminchisescu, "Large displacement 3D scene flow with occlusion reasoning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4417–4425.

[13] J. Quiroga, T. Brox, F. Devernay, and J. Crowley, "Dense semi-rigid scene flow estimation from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 567–582.

[14] C. Vogel, K. Schindler, and S. Roth, "Piecewise rigid scene flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1377–1384.

[15] M. Jaimez, M. Souiai, J. Stuckler, J. Gonzalez-Jimenez, and D. Cremers, "Motion cooperation: Smooth piece-wise rigid scene flow from RGB-D images," in *Proc. Int. Conf. 3D Vis.*, Oct. 2015, pp. 64–72.

[16] V. Golyanik, K. Kim, R. Maier, M. NieBner, D. Stricker, and J. Kautz, "Multiframe scene flow with piecewise rigid motion," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 273–281.

[17] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.

[18] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.

[19] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, "Anisotropic Huber-L1 optical flow," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 108.1–108.11.

[20] M. Drulea and S. Nedevschi, "Total variation regularization of local-global optical flow," in *Proc. 14th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 318–323.

[21] M. J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Comput. Vis. Image Understand.*, vol. 63, no. 1, pp. 75–104, Jan. 1996.

[22] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1744–1757, Sep. 2012.

[23] H. Zimmer, A. Bruhn, and J. Weickert, "Optic flow in harmony," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 368–388, Jul. 2011.
[24] N. Monzon, A. Salgado, and J. Sanchez, "Regularization strategies for discontinuity-preserving optical flow methods," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1580–1591, Apr. 2016.
[25] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black, "Optical flow with semantic segmentation and localized layers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3889–3898.
[26] M. Bai, W. Luo, K. Kundu, and R. Urtasun, "Exploiting semantic information and deep matching for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, vol. 2016, pp. 154–170.
[27] M. Werlberger, T. Pock, and H. Bischof, "Motion estimation with non-local total variation regularization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2464–2471.
[28] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 115–137, Jan. 2014.
[29] C. Zhang, Z. Chen, M. Wang, M. Li, and S. Jiang, "Robust non-local TV-L^1 optical flow estimation with occlusion detection," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 4055–4067, Aug. 2017.
[30] 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.
[31] C. Bailer, B. Taetz, and D. Stricker, "Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4015–4023.
[32] J. Hur and S. Roth, "MirrorFlow: Exploiting symmetries in joint optical flow and occlusion estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 312–321.
[33] 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. (CVPR)*, Jul. 2017, pp. 2462–2470.
[34] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.
[35] 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. (CVPR)*, Jun. 2019, pp. 4571–4580.
[36] J. Hur and S. Roth, "Iterative residual refinement for joint optical flow and occlusion estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5754–5763.
[37] N. Mayer, E. Ilg, P. Fischer, C. Hazirbas, D. Cremers, A. Dosovitskiy, and T. Brox, "What makes good synthetic training data for learning disparity and optical flow estimation?" *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 1–19, 2018.
[38] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Models matter, so does training: An empirical study of CNNs for optical flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1408–1423, Jun. 2020.
[39] J.-M. Gottfried, J. Fehr, and C. S. Garbe, "Computing range flow from multi-modal Kinect data," in *Proc. Int. Symp. Vis. Comput.*, 2011, pp. 758–767.
[40] J. Quiroga, F. Devernay, and J. Crowley, "Local/global scene flow estimation," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 3850–3854.
[41] E. Herbst, X. Ren, and D. Fox, "RGB-D flow: Dense 3-D motion estimation using color and depth," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 2276–2282.
[42] D. Sun, E. B. Sudderth, and H. Pfister, "Layered RGBD scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 548–556.
[43] M. Jaimez, M. Souiai, J. Gonzalez-Jimenez, and D. Cremers, "A primal-dual framework for real-time dense RGB-D scene flow," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 98–104.
[44] F. Huguet and F. Devernay, "A variational method for scene flow estimation from stereo sequences," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–7.
[45] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers, "Efficient dense scene flow from sparse or dense stereo data," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 739–751.
[46] T. Basha, Y. Moses, and N. Kiryati, "Multi-view scene flow estimation: A view centered variational approach," *Int. J. Comput. Vis.*, vol. 101, no. 1, pp. 6–21, Jan. 2013.
[47] M. Derome, A. Plyer, M. Sanfourche, and G. Le Besnerai, "A prediction-correction approach for real-time optical flow computation using stereo," in *Proc. German Conf. Pattern Recognit.*, 2016, pp. 365–376.
[48] T. Popham, A. Bhalerao, and R. Wilson, "Estimating scene flow using an interconnected patch surface model with belief-propagation inference," *Comput. Vis. Image Understand.*, vol. 121, pp. 74–85, Apr. 2014.
[49] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha, "Object stereo—Joint stereo matching and object segmentation," in *Proc. CVPR*, Jun. 2011, pp. 3081–3088.
[50] C. Vogel, K. Schindler, and S. Roth, "3D scene flow estimation with a piecewise rigid scene model," *Int. J. Comput. Vis.*, vol. 115, no. 1, pp. 1–28, Oct. 2015.
[51] M. Menze, C. Heipke, and A. Geiger, "Object scene flow," *ISPRS-J. Photogramm. Remote Sens.*, vol. 140, pp. 60–76, Jun. 2018.
[52] R. Schuster, O. Wasenmüller, C. Unger, G. Kuschk, and D. Stricker, "SceneFlowFields++: Multi-frame matching, visibility prediction, and robust interpolation for scene flow estimation," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 527–546, Feb. 2020.
[53] T. Taniai, S. N. Sinha, and Y. Sato, "Fast multi-frame stereo scene flow with motion segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6891–6900.
[54] C. Richardt, H. Kim, L. Valgaerts, and C. Theobalt, "Dense wide-baseline scene flow from two handheld video cameras," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 276–285.
[55] Z. Lv, C. Beall, P. F. Alcantarilla, F. Li, Z. Kira, and F. Dellaert, "A continuous optimization approach for efficient and accurate scene flow," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 757–773.
[56] R. Schuster, O. Wasenmuller, G. Kuschk, C. Bailer, and D. Stricker, "SceneFlowFields: Dense interpolation of sparse scene flow correspondences," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1056–1065.
[57] R. Saxena, R. Schuster, O. Wasenmuller, and D. Stricker, "PWOC-3D: Deep occlusion-aware End-to-End scene flow estimation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 324–331.
[58] W.-C. Ma, S. Wang, R. Hu, Y. Xiong, and R. Urtasun, "Deep rigid instance scene flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3614–3622.
[59] J. Hur and S. Roth, "Self-supervised monocular scene flow estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7396–7405.
[60] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[61] M. Menze, C. Heipke, and A. Geiger, "Discrete optimization for optical flow," in *Proc. German Conf. Pattern Recognit.*, 2015, pp. 16–28.
[62] H. Y. Jung, K. M. Lee, and S. U. Lee, "Window annealing for pixel-labeling problems," *Comput. Vis. Image Understand.*, vol. 117, no. 3, pp. 289–303, Mar. 2013.
[63] C. Vogel, K. Schindler, and S. Roth, "3D scene flow estimation with a rigid motion prior," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1291–1298.
[64] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.
[65] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 833–851.
[66] G. Yang and D. Ramanan, "Upgrading optical flow to 3D scene flow through optical expansion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1331–1340.

CHENG FENG received the bachelor's degree in automation from the Wuchang University of Technology, Wuhan, in 2016. He is currently pursuing the master's degree in instrumentation engineering with Nanchang Hangkong University, China. His current research interests include image processing and computer vision.

LONG MA received the bachelor's degree in electronic information engineering from the Hubei Institute of Engineering, Xiaogan, in 2016. He is currently pursuing the master's degree in instrument and meter engineering with Nanchang Hangkong University, China. His current research interests include image processing and computer vision.

CONGXUAN ZHANG (Member, IEEE) received the Ph.D. degree in measurement technology and instruments from the Nanjing University of Aeronautics and Astronautics, Nanjing, in 2014. From 2018 to 2019, he was a Visiting Scholar with the Department of Biomedical Engineering, The University of Kansas. He is currently an Assistant Professor with the School of Measuring and Optical Engineering, Nanchang Hangkong University, China. His current research interests include image processing and computer vision.

ZHEN CHEN received the Ph.D. degree in mechanical design and theory from Northwestern Polytechnical University, Xi'an, in 2003. From 2006 to 2007, he was a Visiting Scholar with the Department of Biomedical Engineering, The University of Kansas. He is currently a Professor with the School of Measuring and Optical Engineering, Nanchang Hangkong University, China. His current research interests include image understanding and measurement.

LIYUE GE received the master's degree in instrument and meter engineering from Nanchang Hangkong University, Nanchang, China, in 2019. He is currently a Lecturer with the College of Information Engineering, Nanchang Hangkong University. His current research interests include image processing and computer vision.

SHAOFENG JIANG received the Ph.D. degree in biomedical engineering from Southern Medical University, Guangzhou, in 2008. He is currently a Professor with the Biomedical Engineering Department, Nanchang Hangkong University, Nanchang, China. His research interests include medical image processing, pattern recognition, and artificial intelligence.