IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 20, 2019, accepted November 10, 2019, date of publication November 21, 2019,
date of current version December 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954917

Direct Optical-Flow-Aware Computational Framework for 3D Reconstruction

HUIJUAN HU AND PEI CHEN
School of Data Science and Computer, Sun Yat-sen University, Guangzhou 521055, China
Guangdong Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou 521055, China

Corresponding author: Pei Chen (chenpei@mail.sysu.edu.cn)

ABSTRACT In this paper, a direct computational method is presented which combines optical flow and structure from motion (SfM) by putting the SfM problem in the framework of optical flow estimation. In other word, the optical flow is reparametrized in term of the camera's motion and scene's depth, resulting in a similar variation optimization as in optical flow estimation. Meanwhile, three techniques are proposed to improve the accuracy and robustness of the direct approach, including the fast guided interpolation (FGI), the left-right consistency constraint and the soft segment constraint. Experimental results on the Middlebury dataset and KITTI2012 dataset show that the proposed approach can achieve highly-accurate 3D reconstruction with the dense and smooth surface which results in a state-of-the-art performance in optical flow.

INDEX TERMS Optical flow, structure from motion, fast guided interpolation, the left-right consistency constraint, the soft segment constraint.

I. INTRODUCTION

Optical flow [1], [2] has been extensively investigated for many years. Since Horn and Schunck's work [1], the estimation of optical flow has been fully extended in several aspects, including the extension to large-displacement cases by using a from-coarse-to-fine technique [3], robust techniques to deal with discontinuity in flow field [4], gradient constancy assumption [4] and a multigrid framework [5] for speeding up optical flow's computation. Contributions on optical flow are described in the comprehensive reviews in [3]–[8]. Recently, with the advent of end-to-end deep learning methods, a few CNN based models have been proposed and achieved state of the art performance [9]–[12].

Closely related to optical flow, SfM has been extensively investigated in the past 30+ years, especially in the last 20 years of the 20th century, since Longuet-Higgins [13] presented the work on 3D scene reconstruction from two calibrated images. Most representative work on SfM can be found in two monographs [14] and [15]. Most work on SfM follows the pattern used in Longuet-Higgins's work [13]: sparse and two-stage, with feature detection (including matching) in the first stage and 3D estimation of detected features in the second stage. Another active direction in SfM is dense SfM. Much work, for example [16]–[18] has been done for dense or quasi-dense SfM.

Though optical flow and SfM are closely related and have already been fully investigated, separately, they were rarely investigated in an interactively beneficial way, usually unidirectionally. A typical application is to use the estimated optical flow to further estimate the depth of the scene [19]. Such an application intrinsically falls into Longuet-Higgins's pattern, by regarding estimated optical flow as features. On the other hand, the knowledge of motion for example the fundamental matrix, is beneficial to the estimation of optical flow [20]. Valgaerts [21] built a direct link, in a bidirectionally beneficial way, between optical flow estimation and SfM problem of fundamental matrix estimation. However, 3D structure estimated in [21] does not exactly comply with the estimated fundamental matrix and extra error would be introduced, as in traditional SfM approaches. Lately, Becker *et al.* [22] proposed to jointly estimate camera motion and dense structure of a static scene in terms of depth maps from monocular image sequences. But their approach does not refine the optical flow based on the epipolar geometry. Aubry *et al.* [23] presented a spatially dense variational approach which estimated the calibration of multiple cameras in the context 3D reconstruction. To address the issue that the resulting depth (and disparity map) is highly dependent on the correct pose estimation, Roxas and Oishi [24] proposed

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

to slightly decouple the correspondence problem and the depth estimation by imposing the epipolar geometry as a soft constraint. However, for all approaches above, the optical flow estimated in such brute ways is two-stages and would introduce extra error when being used to further estimate 3D structure.

In this paper, we propose a direct optical-flow-aware computational framework for static 3D reconstruction, [1] in which optical flow can be calculated from estimated camera motion and 3D structure. The solution can be obtained by minimizing an optical-flow-alike objective function. In order to improve the direct approach's accuracy and robustness, three techniques are employed, including the fast guided interpolation (FGI), the left-right consistency constraint and the soft segment constraint. The FGI is used as depth initialization in the coarse-to-fine interpolation process. FGI is a hierarchical, cascaded WLS-optimization based technique [26] that handles low-resolution, noisy depth upsampling and sparse motion match densification in a unified manner. Consequently, it can validly recover accurate depth estimation in homogeneous regions and along depth boundaries, while preserving thin structures by alternating the color image and an interpolated intermediate depth as the guidance. Meanwhile, the left-right motion consistency constraint and the soft segment constraint are used to deal with occlusion and textureless problem to estimate the optical flow more accurately and 3D reconstruction more smoothly.

The rest of the paper is organized as follows. Section II presents our variational model. Section III describes the numerical solution of the proposed model. Section IV shows the experiment results of our methods on some publicly available datasets.

II. FRAMEWORK OVERVIEW

Giving image pairs $I = \{I_L, I_R\} : \Omega \to \mathbb{R}^2$, the forward optical flow from I_L to I_R of a pixel $\mathbf{x}$ in the image domain $\Omega \in \mathbb{R}^2$ is defined as w. Suppose $\mathbf{x} = (x, y, t)$ and $w = (u, v, 1)$ denote the image lattice and flow field. The 3D point corresponding to each pixel $\mathbf{x}$ is denoted as $X \in \mathbb{R}^3$. A basic aim addressed by this paper is to find dense correspondence between those two images by estimating the scene structure and the camera matrices. The value constancy assumption states that the gray value of a "point" in two views is not changed by the displacement. Then, the gray constancy assumption is usually expressed as

$$I(\mathbf{x}) = I(\mathbf{x} + w) \tag{1}$$

Brox *et al.* [4] introduced the gradient constancy assumption into optical flow estimation:

$$\nabla I(\mathbf{x}) - \nabla I(\mathbf{x} + w) = 0 \tag{2}$$

with $\nabla = (\partial x, \partial y)^T$. By including a smoothness constraint, an energy function as follows is used in [4]:

$$\begin{aligned} E(u, v) &= E_{data}(w) + E_{smooth}(w) \\ &= \int_{\Omega} d\mathbf{x}\Psi\left(|I(\mathbf{x} + w) - I(\mathbf{x})|^2 \right. \\ &\quad \left. +\gamma\, |\nabla I(\mathbf{x} + w) - \nabla I(\mathbf{x})|^2\right) \\ &\quad +\lambda_s \Psi\left(|\nabla u|^2 + |\nabla v|^2\right) \end{aligned} \tag{3}$$

where the modified L_1 norm $\Psi\left(s^2\right) = \sqrt{s^2 + \epsilon}$ with a small positive constant $\epsilon = 0.001$ is used to deal with discontinuities in the flow field and λ_s, γ are some constants. Usually, the first part in (3) is referred to as data term f_{data}, and smoothness term f_{smooth} for the second part.

The minimization of the variation problem (3) relies on the Euler-Lagrange equation. A numerical algorithm was carefully designed in [4] to fulfill the Euler-Lagrange equation.

In order to avoid the complicated mathematics with the Euler-Lagrange equation, a discrete form of the objective function is used in [27]. In the iteration, a current w has been iteratively estimated, and the increment $\delta w' = (\delta u, \delta v)$ needs to be determined so that the following locally approximation function achieves the minima at $\widehat{\delta w'}$:

$$\begin{aligned} E(\delta u, \delta v) = \sum \Psi &\left(\left|I\left(\mathbf{x} + w + \delta w'\right) - I(\mathbf{x})\right|^2 \right. \\ &\left. +\gamma \left|\nabla I\left(\mathbf{x} + w + \delta w'\right) - \nabla I(\mathbf{x})\right|^2\right) \\ &+\lambda_s \Psi\left(\left|\nabla(u + \delta u)\right|^2 + \left|\nabla(v + \delta v)\right|^2\right) \end{aligned} \tag{4}$$

In [27], an Iterative Reweighted Least Square (IRLS) approach [28] was utilized to minimize the objective function.

A. PARAMETERIZATION

In an un-calibrated or calibrated setting, suppose that camera calibration matrix encapsulates the internal camera parameters as $K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$, where $f = (f_x, f_y)$ is the focal length and $c = (c_x, c_y)$ is the offset of the principal point. Suppose that two camera projection matrices are factored as $P = K\left[I_3\ 0\right] \in R^{3,4}$ and $P' = K\left[R\ t\right] \in R^{3,4}$. $\left[R\ t\right]$ describes the pose of the second camera in terms of a rotation matrix R characterized by three rotation angles $\theta = \left(\theta_x\ \theta_y\ \theta_z\right)$ and a translation vector $t = \left(t_x\ t_y\ t_z\right)^T$ between two cameras, respectively.

$$\begin{aligned} R = &\begin{bmatrix} cos(\theta_z) & -sin(\theta_z) & 0 \\ sin(\theta_z) & cos(\theta_z) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos(\theta_y) & 0 & sin(\theta_y) \\ 0 & 1 & 0 \\ -sin(\theta_y) & 0 & cos(\theta_y) \end{bmatrix} \\ &\times \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta_x) & -sin(\theta_x) \\ 0 & sin(\theta_x) & cos(\theta_x) \end{bmatrix} \end{aligned} \tag{5}$$

For projective reconstruction, the 3D point corresponding to pixel (x, y) in the first view is determined by its depth $d_{x,y}$:

$$\left(d_{x,y} \times x, d_{x,y} \times y, d_{x,y}\right) \tag{6}$$

[1] The original method [25] has been patented in China, USA and Japan by P. Chen.

and its projection on the second view is $\left(x'_{x,y}, y'_{x,y}\right)$:

$$x'_{x,y} = \frac{P'_1 \left[d_{x,y} \times x \; d_{x,y} \times y \; d_{x,y} \; 1\right]}{P'_3 \left[d_{x,y} \times x \; d_{x,y} \times y \; d_{x,y} \; 1\right]}$$
$$y'_{x,y} = \frac{P'_2 \left[d_{x,y} \times x \; d_{x,y} \times y \; d_{x,y} \; 1\right]}{P'_3 \left[d_{x,y} \times x \; d_{x,y} \times y \; d_{x,y} \; 1\right]} \tag{7}$$

where P'_i is the i^{th} row vector of the second camera projection matrix P'. The disparity for pixel (x, y) in the first view can be expressed as

$$u = x'_{x,y} - x$$
$$v = y'_{x,y} - y \tag{8}$$

Put simply, the new approach is to estimate two cameras and a smooth surface that match two views as best as possible.

If a projective transformation of $H = \begin{bmatrix} K^{-1} & 0 \\ 0 & 1 \end{bmatrix}$ is applied (i.e., rightâĂ"multiplying both camera projection matrices with H) so that the first camera projection matrix is still the default form $\left[I_3 \; 0\right]$ and consequently the second camera projection matrix is re-parameterised as $P' = K\left[RK^{-1} \; t\right]$, according to projective ambiguity in reference [29]. For a concise representation, suppose the parameters P' and d are encapsulated in a vector ϑ which can be written as $\vartheta\,(\theta, t, K, d)$.

Furthermore, when two cameras are calibrated, the binocular depth estimation is reduced to minimize an objective function of $\vartheta\,(\theta, t, d)$.

Suppose $\vartheta\,(\theta, t, K, d)$ is a vector that the parameters P' and d are encapsulated in. The problem of 3D reconstruction can be solved by minimizing an objective function in the framework of optical flow, so that two views fit each other optimally. By embedding the SfM problem in the framework of Brox's optical flow seamlessly, the flow field w is defined as the function with ϑ as their variables: $w\,(\vartheta)$. The energy function can be reparametrized in terms of ϑ as

$$E_{data}\,(w\,(\vartheta)) = \int_{\Omega} d\mathbf{x}\Psi \left(|I\,(\mathbf{x} + w\,(\vartheta)) - I\,(\mathbf{x})|^2 \right.$$
$$\left. + \gamma\,|\nabla I\,(\mathbf{x} + w\,(\vartheta)) - \nabla I\,(\mathbf{x})|^2\right) \tag{9}$$

The smooth constraint term is expressed as

$$E_{smooth}\,(w\,(\vartheta)) = \lambda_s \int_{\Omega} d\mathbf{x}\Psi \left(|\nabla w\,(\vartheta)|^2\right) \tag{10}$$

Put together, the objective function becomes, with ϑ as its parameters

$$E = E_{data}\,(w\,(\vartheta)) + E_{smooth}\,(w\,(\vartheta)) \tag{11}$$

B. THE LEFT-RIGHT CONSISTENCY CONSTRAINT

Occlusion handling is a critical issue in optical flow estimation [20], [30]. A reconstruction pipeline that relies on optical flow needs to be robust to these errors, because it violates the one-to-one-mapping assumption between two views. Moreover, depth discontinuity exists wherever there is occlusion. Then, the reconstruction accuracy around occlusion/discontinuity is usually unsatisfactory. Therefore, the display of optical flow will be affected, alternatively. In order to obtain better optical flow and 3D reconstruction, more care is need to be taken to specifically deal with occlusion, such as occlusion detection or segmentation. In addition, for SfM, 3D structure can't be estimated for occluded parts. Thus, we adopt the strategy of explicitly handling occlusion. The left-right consistency constraint based on the left view is defined as

$$E_{L(consistency)}\,(w_L\,(\vartheta_L), w_R\,(\vartheta_R)) = \lambda_{co} \int_{\Omega_L} \Psi$$
$$\times \left(|w_L\,(\vartheta_L) - w_R\,(\mathbf{x}_L + w_L\,(\vartheta_L))|^2\right) \tag{12}$$

The constraint term based on the right view can be defined as the same way. λ_{co} is a weighting factor. Here, we detect the occlusion regions by set the thresholds T_L and T_R as follows:

$$|w_L\,(\vartheta_L) - w_R\,(\mathbf{x}_L + w_L\,(\vartheta_L))| \leqslant T_L$$
$$|w_R\,(\vartheta_R) - w_L\,(\mathbf{x}_R + w_R\,(\vartheta_R))| \leqslant T_R \tag{13}$$

T_L and T_R are two different small values. The left-right consistency constraint makes that pixels are forced to be either visible and satisfy the bi-directional flow consistency, or are identified as occlusions [31].

C. SOFT SEGMENT CONSTRAINT

Recently, segmentation-based stereo approaches (e.g. [32], [33]) have been proposed and demonstrated that the difficulties and ambiguities caused by texturelessness or occlusion can be handled by using groups of pixels with similar colors. Plane fitting from the initial disparities in a segment is discussed in details by Tao *et al.* [34]. Use a plane that is fitted for each color segment to model the inverse depth values, and then our soft segmentation term can be defined as

$$E_{segment}\,(d) = \sum_{x,y} \rho \left| \frac{1}{d_{x,y}} - \left(a_{x,y}x + b_{x,y}y + c_{x,y}\right)\right|^2 \tag{14}$$

where (x, y) is an image pixel, and $d_{x,y}$ is its depth value. Then ρ controls the strength of segment constraint and $a_{x,y}$, $b_{x,y}$, $c_{x,y}$ are the 3D plane parameters which are the least square solution of a linear system about inverse depth values. Segmented 3D plane parameters for each region are estimated by the RANSAC-based algorithm [35].

In this paper, we employ the commonly-used mean-shift segmentation algorithm [36] with default parameters. In the iterative process, the reliable pixels are selected by robust 3D plan fitting algorithm and subsequently used for 3D plane estimation. Note that the regions with too small (<500 visible pixels) are ignored for reliable 3D plane parameter estimation. Segmentation errors are ignored and thus can't be propagated to the depth processing stage. The inverse depth value is changed within a reasonable range of the fitted plane and the 3D plane parameters are updated based on the modified inverse depths accordingly. Although some depth accuracy may be lost due to the limitation of the plane approximation, the quality of depth estimation still can be improved to some extent.

D. GUIDED DEPTH INTERPOLATION

As in most optical flow method, the Gaussian pyramid is used in the proposed method to overcome the difficulty of large displacement. However, it has a major drawback of error propagation. Errors at coarser levels can propagate across scales especially for large displacements and motion discontinuities. In this paper, we propose a guided coarse-to-fine strategy. In the proposed coarse-to-fine framework for depths estimation with a ratio $\sigma < 1$ of scaling, we use FGI [37] to interpolate a parse set of depths of the coarse level to initiate the depths estimation, then use this estimation to initialize a one-level energy minimization, and obtain the final depths estimation at the current level.

As in our model, the interpolation of depths is very important to the effect of the scheme. From level n to $n-1$, initial depth can be obtained by the interpolation. For a progressively densified input data $d^{(n)}$ at a certain level n, two cascaded WLS are performed by alternating the color image $I^{(n)}$ and an intermediate interpolated depth d_* as the guidance.

1st WLS:

Using $I^{(n)}$as the Guidance. Knowing the sparse input data $d^{(n)}$ and the guidance color image $I^{(n)}$at the n^{th} level, minimize the following objective function

$$\varepsilon\left(d_*\right)=\left(d_*-d^{(n)}\right)^T M^{(n)}\left(d_*-d^{(n)}\right)+\lambda_1 d_*^T A_{I(n)} d_* \tag{15}$$

where $A_{I(n)}$ denotes the spatially varying Laplacian matrix defined by the guidance image $I^{(n)}$. $M^{(n)}$ is a diagonal matrix with its elements given by the mask map $m^{(n)}$. $m^{(n)}$ denotes the constraint map for the sparse data input whose elements are 1 for pixels with valid data and 0 otherwise. An intermediate dense output d_* is computed with

$$d_*\left(\mathbf{x}\right)=\frac{\left(\left(E+\lambda A_{I(n)}\right)^{-1} d^{(n)}\right)\left(\mathbf{x}\right)}{\left(\left(E+\lambda A_{I(n)}\right)^{-1} \mathbf{m}^{(n)}\right)\left(\mathbf{x}\right)} \tag{16}$$

$\mathbf{m}^{(n)}$ denotes the corresponding vectored form of $m^{(n)}$.

2nd WLS:

Using d_* as the Guidance. Suppose the input data $d^{(0)}$ as the bicubic interpolated data $d^{(n)}$ and the intermediate interpolated data d_* as the guidance signal. The similar objective is minimized as

$$\varepsilon\left(\widetilde{d^{(n)}}\right)=\left(\widetilde{d^{(n)}}-d^{(0)}\right)^T\left(\widetilde{d^{(n)}}-d^{(0)}\right)+\lambda_2 \widetilde{d^{(n)}}^T A_{d_*} \widetilde{d^{(n)}} \tag{17}$$

where A_{d_*} denotes the Laplacian matrix defined by d_*. Over the iterations, both intermediate guidance signal d_* and the 2nd WLS output $\widetilde{d^{(n)}}$ are progressively improved, and the final output $\widetilde{d^{(0)}}$ preserves important depth or motion structures.

In this paper, the FGI method with default parameters [37] is chosen to deal with depth up sampling which builts on a WLS formulation with its recent fast solver-fast global smoothing technique. It densifies the input data set by efficiently performing the cascaded, global interpolation (or smoothing) with alternating guidance. This scheme can effectively address the potential structure inconsistency between the sparse input data and the guidance image, while preserving depth or motion boundaries.

E. VARITIONAL MODEL

Finally, the left-right consistency constraint (12) and the segment constraint (14) are incorporated into (11) to obtain the finial energy function,

$$\begin{aligned} E\left(w\left(\vartheta\right)\right)=&E_{data}\left(w\left(\vartheta\right)\right)+E_{smooth}\left(w\left(\vartheta\right)\right)\\ &+E_{consistency}\left(w\left(\vartheta\right)\right)+E_{segment}\left(d\right) \end{aligned} \tag{18}$$

where E_{data}, E_{smooth}, $E_{consistency}$ and $E_{segment}$ are denoted as the corresponding energy function in a single view either left or right.

III. NUMERICAL SOLUTION

A. A NEW INTERPRETATION IN NONLINEAR LS

Here, we reformulate the minimization problem (3) as a general nonlinear least squares problem. Specifically here, the Gauss-Newton technique [38] is utilized for minimization. First, consider the data term:

$$\begin{aligned} E_{data}\approx\sum\Psi\Big(&\left(I_x\delta u+I_y\delta v+I_z\right)^2+\gamma\left(I_{xx}\delta u+I_{xy}\delta v\right.\\ &\left.+I_{xz}\right)^2+\gamma\left(I_{xy}\delta u+I_{yy}\delta v+I_{yz}\right)^2\Big) \end{aligned} \tag{19}$$

where the following first-order Taylor expansions have been used

$$\begin{aligned} I\left(\mathbf{x}+w+\delta w'\right)-I\left(\mathbf{x}\right)&\approx I_x\delta u+I_y\delta v+I_z\\ I_x\left(\mathbf{x}+w+\delta w'\right)-I_x\left(\mathbf{x}\right)&\approx I_{xx}\delta u+I_{xy}\delta v+I_{xz}\\ I_y\left(\mathbf{x}+w+\delta w'\right)-I_y\left(\mathbf{x}\right)&\approx I_{xy}\delta u+I_{yy}\delta v+I_{yz}\\ I_z&=I\left(\mathbf{x}+w\right)-I\left(\mathbf{x}\right)\\ I_{xz}&=I_x\left(\mathbf{x}+w\right)-I_x\left(\mathbf{x}\right)\\ I_{yz}&=I_y\left(\mathbf{x}+w\right)-I_y\left(\mathbf{x}\right) \end{aligned} \tag{20}$$

I_x, I_y and their second-order partial derivatives are calculated at $(x+w)$.

Note that the first order Taylor expansion in (19) is just the technique in used Guess-Newton approximation [38].

Taking partial derivatives of data over δw,

$$\begin{aligned} \frac{\partial E_{data}}{\partial\delta w}=&\,2\sum\dot{\Psi}_d\times\left(\begin{bmatrix} I_x^2 & I_xI_y\\ I_xI_y & I_y^2\end{bmatrix}+\gamma\begin{bmatrix} I_{xx}^2 & I_{xx}I_{xy}\\ I_{xx}I_{xy} & I_{xy}^2\end{bmatrix}\right.\\ &\left.+\gamma\begin{bmatrix} I_{xy}^2 & I_{xy}I_{yy}\\ I_{xy}I_{yy} & I_{yy}^2\end{bmatrix}\right)+2\sum\dot{\Psi}_d\times\left(I_z\begin{bmatrix} I_x\\ I_y\end{bmatrix}\right.\\ &\left.+\gamma I_{xz}\begin{bmatrix} I_{xx}\\ I_{xy}\end{bmatrix}+\gamma I_{yz}\begin{bmatrix} I_{xy}\\ I_{yy}\end{bmatrix}\right)\\ =&\,\mathbf{H}_{data}\delta w+\mathbf{b}_{data} \end{aligned} \tag{21}$$

where Ψ'_d(similarly Ψ'_s for the smoothness term) takes the simplified form of $\Psi'_d=\Omega'\left(I_z^2+\gamma\left(I_{xz}^2+I_{yz}^2\right)\right)$. The $\mathbf{H}_{data}$ and $\mathbf{b}_{data}$ in (21) act the Gauss-Newton Hessian matrix and gradient vector with the data term, in Gauss-Newton algorithm [38]. (Similarly for $\mathbf{H}_{smooth}$ and $\mathbf{b}_{smooth}$ in (23).)

Similarly, due to the involvement of its neighbors in each pixel, the smoothness term is,.

$$\begin{aligned} E_{smooth} \approx \alpha \sum \Psi \Bigg(& \left(u^{i,j} + \delta u^{i,j} - u^{i,j-1} - \delta u^{i,j-1}\right)^2 \\ & + \left(u^{i,j} + \delta u^{i,j} - u^{i-1,j} - \delta u^{i-1,j}\right)^2 \\ & + \left(v^{i,j} + \delta v^{i,j} - v^{i,j-1} - \delta v^{i,j-1}\right)^2 \\ & + \left(v^{i,j} + \delta v^{i,j} - v^{i-1,j} - \delta v^{i-1,j}\right)^2 \Bigg) \end{aligned} \tag{22}$$

where the gradient vector ∇u in (22) is replaced by the difference of between its value at (i, j) and two neighbors at $(i-1, j)$ and $(i, j-1)$: $\left(u^{i,j} - u^{i-1,j}, u^{i,j} - u^{i,j-1}\right)$ and similarly for v. Note that the superscript here denotes the image lattice system.

Taking partial derivatives of f_{smooth} over δw, we have

$$\begin{aligned} \frac{\partial E_{smooth}}{\partial \delta w} &= 2\alpha \sum \Psi'_s \times \left(\mathbf{H}^{i,j} \delta^{i,j} + \boldsymbol{b}^{i,j}\right) \\ &= \mathbf{H}_{smooth} \delta w + \mathbf{b}_{smooth} \end{aligned} \tag{23}$$

where $\delta^{i,j} = \left[\delta u^{i-1,j}, \delta u^{i,j}, \delta u^{i,j-1}, \delta v^{i-1,j}, \delta v^{i,j}, \delta v^{i,j-1}\right]^T$, $\mathbf{H}^{i,j} = \begin{bmatrix} \mathbf{C} & 0 \\ 0 & \boldsymbol{C} \end{bmatrix}$ with $\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$ and $\boldsymbol{b} = \left[u_x^{i,j} + u_y^{i,j}, -u_x^{i,j}, -u_y^{i,j}, v_x^{i,j} + v_y^{i,j}, -v_x^{i,j}, -v_y^{i,j}\right]^T$ with $u_x^{i,j} = u^{i,j} - u^{i-1,j}$ and $u_y^{i,j} = u^{i,j} - u^{i,j-1}$ as the first order differences of u along x and y coordinates, respectively (similar for v).

Putting (21) and (23) together, conceptually, we have

$$\frac{\partial E}{\partial \delta w} = \mathbf{H} \delta w + \mathbf{b} \tag{24}$$

where $\mathbf{H}$ and $\mathbf{b}$ do not have a simple and compact form due to the Laplacian operator. (In [27], an explicit formula of $\mathbf{H}$ and $\mathbf{b}$ is given by introudcing a few extra symbols.) The incremental estimate of δw in each iteration is solved as

$$\widehat{\delta w} = -\mathbf{H}^{-1} \mathbf{b} \tag{25}$$

From the above equation, the solution of (25) is reformulated in the framework of Gauss-Newton approximation [38]. Though ending with a same solution as in [27], the reformulation in non LS has the following advantages. First, the extension to a new reparameterization can be easily obtained, in the framework of Gauss-Newton approximation, as will be seen in the next section. Second, the new interpretations help understanding the implementation of the Euler-Lagrange equation in optical flow. In the algorithms for nonlinear LS, such as Newton and Guess-Newton algorithms, the Hessian matrix (or Guess-Newton matrix) and the gradient vector are updated around the current estimate of w, in each iteration. This incremental method is just the essence of the warping technique in optical flow.

B. DIRECT FORM OF SOLUTION

The Newton algorithm, Gauss-Newton algorithm and their alike solve the optimization problem by iteratively minimizing the second order Taylor expansion or a modified one [38]:

$$-\frac{1}{2} \delta w \mathbf{H} \delta w + \delta w^T \mathbf{b} \tag{26}$$

In an optical flow problem, the flow field u and v are encapsulated in w. Now, in the optical flow based on SfM problem, u and v are functions of ϑ. By using the first order difference, δw can be represented as, with a general matrix $\mathbf{J}$ as the Jacobian matrix

$$\delta w = \mathbf{J} \delta \vartheta \tag{27}$$

Substituting (27) into (26), we have

$$-\frac{1}{2} \delta \vartheta^T \mathbf{J}^T \mathbf{H} \mathbf{J} \delta \vartheta + \delta \vartheta^T \mathbf{J}^T \mathbf{b} \tag{28}$$

Then, the flow field w and ϑ can be explicitly present themselves by minimizing the objective function within the framework nonlinear LS. The Gauss-Newton technique [38] is utilized for minimization. Thus, the incremental estimate of $\delta \vartheta$ is as

$$\widehat{\delta \vartheta} = -\left(\mathbf{J}^T \mathbf{H} \mathbf{J}\right)^{-1} \mathbf{J}^T \mathbf{b} \tag{29}$$

where $\mathbf{J} = \frac{\partial w}{\partial \vartheta}$, $\mathbf{H}$ and $\mathbf{b}$ act as the Gauss-Newton Hessian matrix and gradient vector for optical flow. Within the framework of optical flow, only the extra computation of the two Jacobian matrices are needed, so that this jointly estimation of optical flow and SfM algorithm can be solved.

C. THE MULTI-RESOLUTION FRAMEWORK

An iterative optimization algorithm is adopted to minimize the energy of objective function (18) with optical flow and occlusions iteratively. It is difficult to minimize the energy of parameters ϑ and occlusions O in (18) simultaneously. The Euler-Lagrange equation according to the objective function will have multiple solutions. In general, we adopt a multi-resolution framework to reduce the risk of being tracked in local minima. The solution computed from the coarse scale is considered as initialization for fine scale.

The optimization process can be summarized as: starting from an initial solution $\left(\vartheta_L^0, \vartheta_R^0\right)$ at each level, the energy function is minimized based on alternating the Euler-Lagrange partial differential equations corresponding to the energies E_L and E_R, then $\left(\vartheta_L^{l+1}, \vartheta_R^{l+1}\right)$ is obtained from $\left(\vartheta_L^l, \vartheta_R^l\right)$ as

$$E\left(\left(\vartheta_L^{l+1}, \vartheta_R^{l+1}\right)\right) = E_L\left(\left(\vartheta_L^{l+1}, \vartheta_R^{l}\right)\right) + E_R\left(\left(\vartheta_L^{l}, \vartheta_R^{l+1}\right)\right) \tag{30}$$

l is the iterations per resolution level. Notice that the optical flow is estimated at the same time with the camera projection matrixes and depths.

As shown in Algorithm 1, all optimization scheme is initialized with the solution of function (11) at the coarse level and refined with the solution of the function (18) at the

Algorithm 1 Our Iterative Optimization Algorithm

```
At the coarse level:
Estimate $\vartheta_L^l$ and $\vartheta_R^l$ (the inner-product $w_L^l$ and $w_R^l$, simultaneously ) independently using the function of (11);
At the finest level:
Estimate $O_L$ and $O_R$ respectively from computed $\vartheta_L^l$ and $\vartheta_R^l$ using the function (13);
Estimate $\vartheta_L^{l+1}$ one iteration of the process of minimizing $E_L\left(\vartheta_L^l, \vartheta_R^l\right)$ with respect to $\vartheta_L^l$ to obtain $\vartheta_L^{l+1}$ using the function (18);
Estimate $\vartheta_R^{l+1}$ one iteration of the process of minimizing $E_R\left(\vartheta_L^l, \vartheta_R^l\right)$ with respect to $\vartheta_R^l$ to obtain $\vartheta_R^{l+1}$ using the function (18).
```

finest level. Applying the extension to the new reparametrized framework of Gauss-Newton approximation, the incremental estimation of parameterizations in each iteration can be obtained.

IV. EXPERIMENTS

In this section, we will show the performance and results of our method and compare with existing state-of-the-art methods for optical flow, variational depth estimation and camera pose estimation. To evaluate the performance of our algorithm in the un-calibrated settings, we use the Middlebury dataset [7], which contains ground truth optical flow. For the calibrated settings, we use the KITTI2012 dataset [39], which contains both ground truth optical flow and depth map, and the official KITTI visual odometry split, which contains 11 driving sequences with ground truth odometry. For all the experiments, we choose the same parameter set: $\{\gamma, \lambda_s, \lambda_{co}, \rho\} = \{1, 1/25, 2/255, 1/255\}$. The maximum errors T_L and T_R are respectively 0.7 and 1.3. The truncations T_L and T_R are set to relatively small values to increase the robustness in occlusion. In addition, the from-coarse-to-fine technique with a scale factor of $\sqrt{3}$ is employed in our implementation to overcome the difficulty of being trapped in local minima. The iteration number is set as 15 at each resolution level. With these techniques, the algorithm is strictly implemented in the sense of nonlinear LS.

As for calibrated or uncalibrated cameras with a small baseline, the depth for all pixels is initially set as 1 in the coarsest level of the pyramid framework. For other levels, it is obtained by using the FGI interpolation. The rotation matrix is initially set as the identity matrix, i.e., the rotation angles are set as 0. The translation vector is set as 0. For uncalibrated cameras, the offset of the calibration matrix is initially set as 0. The focal length is initially set as 10 in all experiments.

For the KITTI dataset with a large baseline, the algorithm 2 below is used to initialize the depths and the camera pose of our method. To estimation the initialize camera matrix P', we use the SIFT features as input to the fundamental matrix estimation. With the given intrinsic camera parameters K, we solve the essential matrix and decompose it to get the relative pose.

Algorithm 2 The Initialization of Our Algorithm on KITTI2012 Dataset [39]

```
1. Extracting SIFT features from images and matching;
2. Estimating three-dimensional information of feature points and camera projection matrices of the cameras based on the extracted features;
3. Realizing dense depth information by utilizing interpolation based on the first step and the second step.
```

A. RECONSTRUCTION

In the un-calibrated setting, for the Middlebury dataset [7], we mainly evaluate the performance in terms of optical flow and present the visual quality of the reconstruction. In the calibrated setting, we are mainly concerned about the depth estimation and optical flow on the KITTI2012 dataset [39], while presenting the visual quality of the reconstruction.

The Middlebury dataset [7] contains complex motion, but displacements are limited to a few pixels. When the internal camera parameters are supposed unknown on Middlebury dataset [7], it is impossible to evaluate the quality of computed camera matrix and estimated depth. To conclude, we just obtain a so-called range image [21] which is associated with each pixel of the image for sample pairs. Figure 4 presents the visual quality of the recovered 3D points using COLMAP [18] [2] and our method. COLMAP is a general-purpose Structure-from-Motion and Multi-View Stereo pipeline and produces a sparse and dense point cloud of the scene from multiple images. However, due to a small baseline, COLMAP does not work in two consecutive frames. Instead, an image sequence up to 8 frames is used in COLMAP. We can see that the scene reconstruction from our method is very smooth and accurate, and even more dense than that from COLMAP with more images.

KITTI2012 dataset [39] was created from a driving platform and contains images of city streets. It contains complex lighting conditions and large displacements. The internal camera parameters are supposed known on the KITTI2012 dataset [39]. To evaluate the depth estimation, we show the qualitative depth results in Figure 1 for sample pairs as far as the estimated pose is concerned. More specifically, the effect of the FGI and soft segment constraint is shown in Figure 3. Table 1 gives detailed numbers for sample pairs with the error metric measuring the percentage of erroneous pixels $\tau > 3$ units in non-occluded areas (Out-Noc). The color coding visualizes outliers (>3 px EPE) in red and inliers (<3 px EPE) in blue on a logarithmic color scale. Pixels without ground truth value are shown in black. The depth results in Table 1 show that our method outperforms [24] and [41]. The result is still good even without FGI initialization or soft segment constraint. Our approach is significantly superior in all sample pairs. FGI as depth initialization can effectively address the potential structure inconsistency and preserve depth boundaries to improve the

[2] https://github.com/colmap/colmap

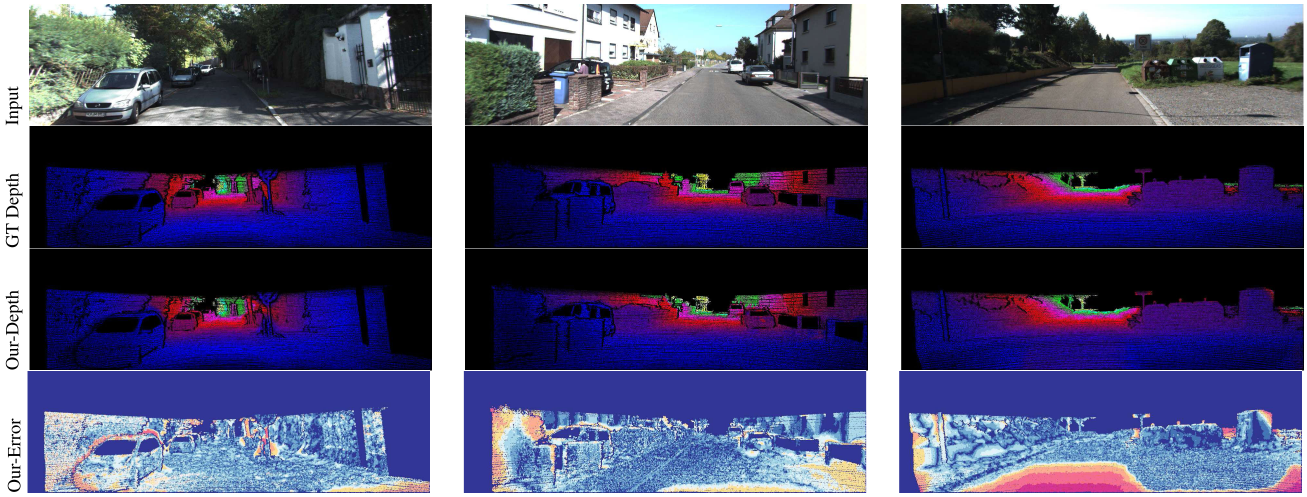

FIGURE 1. Qualitative results of depth (normalized color) and Out-Noc (percentage of erroneous pixels in non-occluded areas) results for our method using estimated pose on sample pairs of KITTI2012.

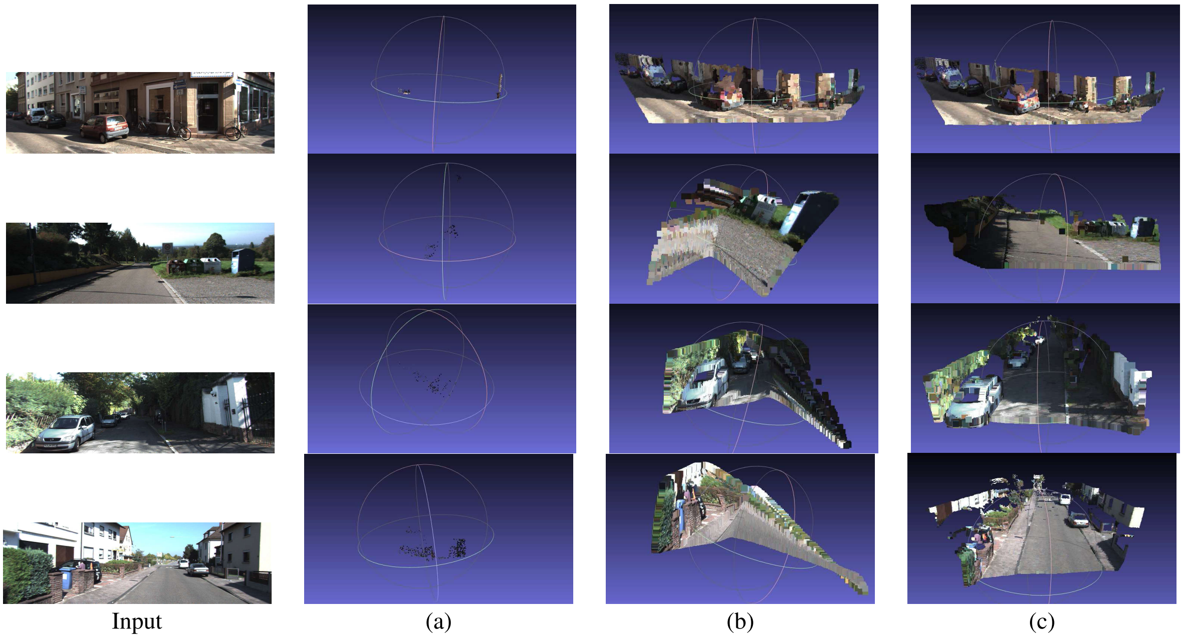

FIGURE 2. Sample frames of KITTI2012 and their reconstructed 3D point (Point Cloud). (a) COLMAP (b) FlowNet2-CSS [40] (c) our method.

accuracy of the estimated depths. Soft segment constaint can effectively handle error propagate problem in the depth processing stage and improve the performance of our model.

To further evaluate the visual 3D structure quality of our approach, the actual reconstruction results for some sample pairs are shown in Figure 2. We compare our method with reconstruction from the dense correspondences obtained from the-state-of-the-art optical flow algorithm FlowNet2-CSS [40] and COLMAP. The FlowNet2-CSS [40] as a dense correspondence can be used to estimate the foundamental matrix using Least Median Square (LMeS) method [42]. With the given intrinsic camera parameters K, we can estimate the camera matrix P' and the 3D points. Figure 6 gives 3D points for some sample pairs of COLMAP, FlowNet2-CSS [40] and our algorithm, respectively. The 3D points reconstructed from COLMAP are very sparse and disordered, and 3D structure estimated from FlowNet2-CSS is even distorted on some samples. Compared with two methods above, the surface of the 3D reconstruction of our method is more dense and smooth.

B. OPTICAL FLOW

The Middlebury dataset [7] has been extensively used for evaluating optical flow methods. On this benchmark, we can

TABLE 1. Comparison of depth results from [24], [41] and our method on selected KITTI2012 showing Out-Noc metric $\tau > 3$.

	Graber	Roxas	No FGI	No soft	Our method
	$\tau > 3$	$\tau > 3$	$\tau > 3$	$\tau > 3$	$\tau > 3$
000068	92.72%	11.42%	4.74%	3.78%	2.87%
000081	54.58%	16.13%	15.46%	14.97%	14.86%
000109	19.18%	13.51%	10.76%	9.37%	8.51%
000134	29.90%	12.43%	10.23%	8.89%	8.53%

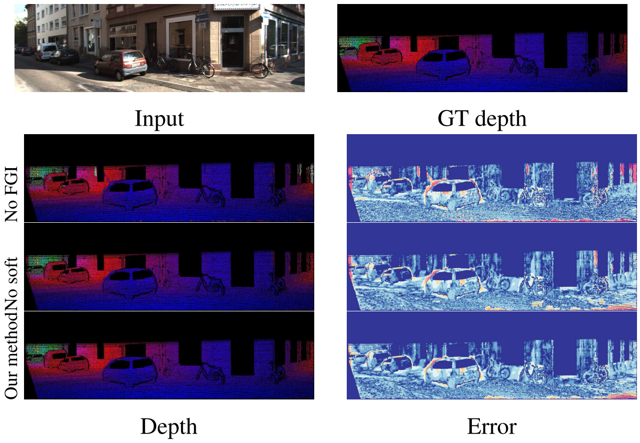

FIGURE 3. Comparison of depth estimation between our method, our method without FGI (No FGI) and our method without soft segment constraint (No soft).

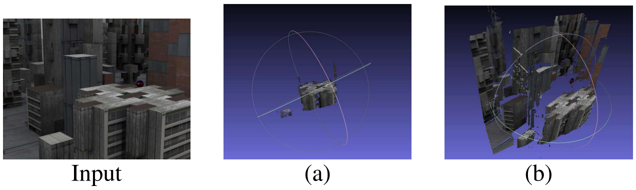

FIGURE 4. Sample frames and the reconstructed 3D points (Point Cloud) from (a) COLMAP and (b) our method.

TABLE 2. Middlebury dataset: Accuracy in terms of endpoint error (EPE) on Grove2, Grove3, Urban2 compared to some traditional algorithms.

Methods	Grove2	Grove3	Urban2
Brox [4]	0.19	0.69	0.32
Zimmer [6]	0.16	0.59	0.26
Valgaerts [21]	0.14	0.57	0.24
DeepFlow [11]	0.14	0.57	0.30
Our method	0.14	0.56	0.25

see that our algorithm based on the un-calibrated setting has good performance. It outperforms Brox's HS method on all three example datasets. We compare our result with some traditional optical flow algorithms and present a subset in Table 2. Figure 5 shows qualitative results. To evaluate the optical flow based on the calibrated setting, our simulation results on KITTI2012 dataset [39] will be compared with some top ranked start-of-the-art optical flow algorithms.

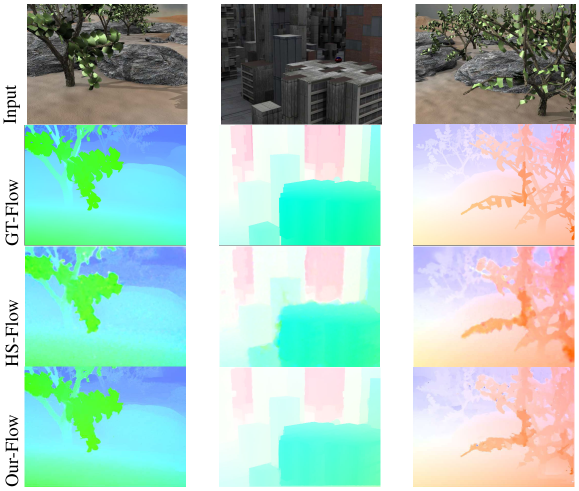

FIGURE 5. Qualitative results (from left to right) on some sample pairs of Middlebury dataset. From top to bottom, the initial first frame, GT-flow, HS-flow and Our-Flow, respectively.

TABLE 3. Endpoint error (EPE) on KITTI2012 comparison of the optical flow results among some top ranked methods.

Methods	EPE
DeepFlow [11]	4.48
FlowNet2-CSS [40]	3.55
Roxas [24]	4.21
Our method	4.02

TABLE 4. Absolute Trajectory Error (ATE) on the KITTI odometry dataset averaged over all multi-frame snippets (lower is better). Our method significantly outperforms the baselines, but falls short of ORB-SLAM (full).

	Seq.09	Seq.10
ORB-SLAM (full)	0.014 ± 0.008	0.012 ± 0.011
ORB-SLAM (short)	0.064 ± 0.141	0.064 ± 0.130
Mean Odom	0.032 ± 0.026	0.028 ± 0.023
Our method	0.017 ± 0.010	0.015 ± 0.011

We evaluate the computed optical flow by means of EPE in Table 3. The result shows that our method has superior performance in both accuracy and efficiency. Figure 6 depicts qualitative results of our method on a subset of KITTI2012 with the error metric (Out-Noc) measuring the percentage of erroneous pixels (> 3). The same color coding is used as with reconstruction.

C. CAMERA POSE ESTIMATION

To evaluate the performance of our camera pose estimation, the official KITTI odometry split 09-10 sequences are used for test in the experiment. The length of image frames is fixed to two frames in our system. We compare our pose estimation with a traditional representative SLAM framework: ORB-SLAM [43]. For the ORB-SLAM, it is a well-established SLAM system. Here we present two versions: 1) ORG-SLAM (full), which recovers odometry using all

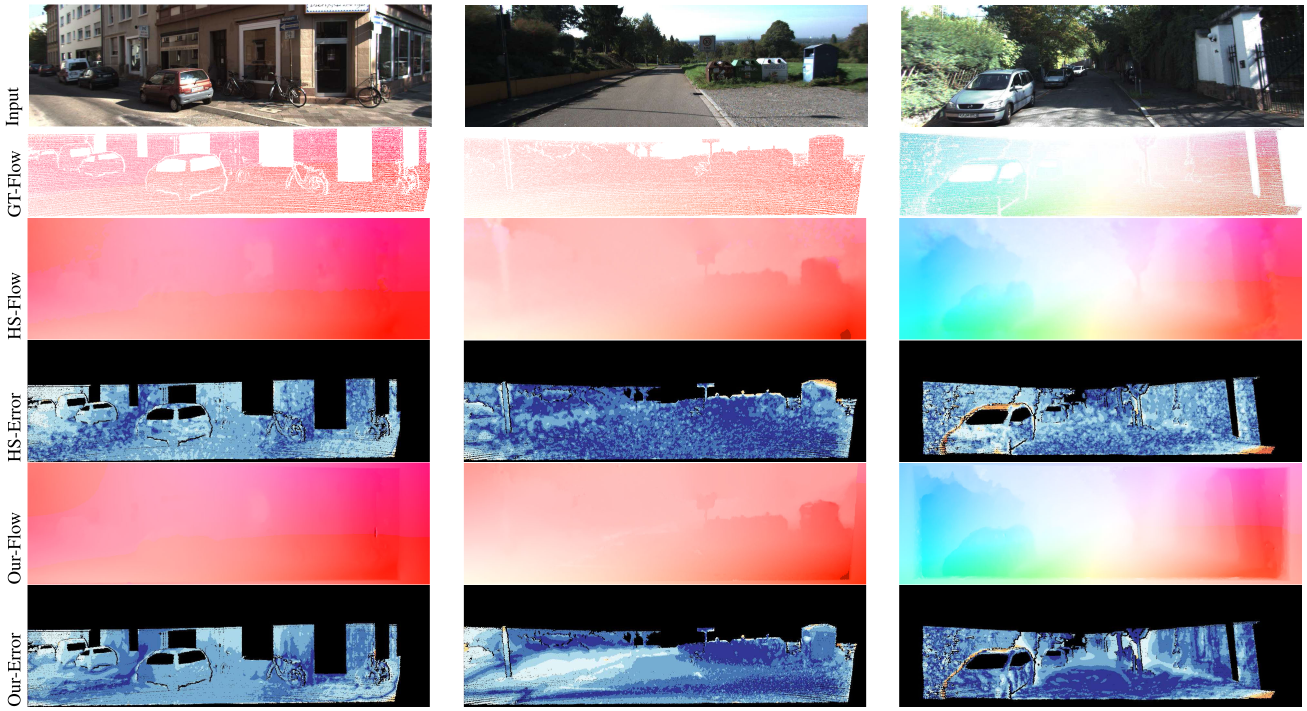

FIGURE 6. Qualitative results of Optical flow and Out-Noc (percentage of erroneous pixels in non-occluded areas) results on some sample pairs of KITTI2012 for our method and HS algorithm.

frames of the driving sequence (i.e. allowing loop closure and re-localization), and 2) ORB-SLAM (short), which runs on 5-frame snippets. Table 4 reports the estimated pose accuracy of our model over two sample sequences from the KITTI odometry dataset. It shows that our method significantly outperforms both baselines (mean odometry) and ORB-SLAM (short), but falls short of ORB-SLAM (full) across the entire spectrum.

V. CONCLUSION

In this paper, a direct optical-flow-aware computational framework for 3D reconstruction is presented, by jointly employing the theory of multi-view geometry initialized by Longuet-Higgins. It has the following characteristics: markless, dense and direct. It is achieved by putting the SfM problem in the framework of optical flow estimation. Because the reconstruction process is implemented in a direct optimization way, the solution is optimal if not falling into a local minimum. In addition, FGI approach is utilized in our from-coarse-to-fine process for depth initialization, and new regularization strategies about the left-right consistency constraint and soft segment strategy are used to deal with disparity discontinuity and outlier. The experimental results not also show good performance of the algorithm in estimation of optical flow, but also result in prefect depth estimation and a rather dense and smooth 3D reconstruction.

REFERENCES

[1] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185–310, Aug. 1981.

[2] R. M. Haralock and L. G. Shapiro, *Computer and Robot Vision*. 1991.

[3] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *Int. J. Comput. Vis.*, vol. 2, no. 3, pp. 283–310, 1989.

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. ECCV*, vol. 3024, no. 10. May 2004, pp. 25–36.

[5] A. Bruhn, J. Weickert, T. Kohlberger, and C. Schnörr, "A multigrid platform for real-time motion computation with discontinuity-preserving variational methods," *Int. J. Comput. Vis.*, vol. 70, no. 3, pp. 257–277, 2006.

[6] H. Zimmer, A. Bruhn, and J. Weickert, "Optic flow in harmony," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 368–388, 2011.

[7] S. Baker, S. Roth, D. Scharstein, M. J. Black, J. P. Lewis, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[8] 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, 2014.

[9] M. Menze, C. Heipke, and A. Geiger, "Discrete optimization for optical flow," in *Proc. German Conf. Pattern Recognit.*, 2015, pp. 16–28.

[10] J. Wulff, L. Sevilla-Lara, and M. J. Black, "Optical flow in mostly rigid scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4671–4680.

[11] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "DeepMatching: Hierarchical deformable dense matching," *Int. J. Comput. Vis.*, vol. 120, no. 3, pp. 300–323, 2016.

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

[13] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, no. 5828, pp. 133–135, 1981.

[14] G. F. Page, "Multiple view geometry in computer vision, by Richard Hartley and Andrew Zisserman, CUP, Cambridge, UK, 2003, vi+ 560 pp., ISBN 0-521-54051-8. (Paperback £44.95)," *Robotica*, vol. 23, no. 2, p. 271, 2005.

[15] P. N, "The geometry of multiple images: The laws that govern the formation of multiple images of a scene and some of their applications," *Ind. Robot*, vol. 29, no. 3, pp. 1721–1727, 2001.

[16] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2008.

[17] M. Lhuillier and L. Quan, "A quasi-dense approach to surface reconstruction from uncalibrated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 418–433, Mar. 2005.

[18] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.

[19] B. Shahraray and M. K. Brown, "Robust depth estimation from optical flow," in *Proc. Int. Conf. Comput. Vis.*, Dec. 1988, pp. 641–650.

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

[21] L. Valgaerts, A. Bruhn, M. Mainberger, and J. Weickert, "Dense versus sparse approaches for estimating the fundamental matrix," *Int. J. Comput. Vis.*, vol. 96, no. 2, pp. 212–234, 2012.

[22] F. Becker, F. Lenzen, J. H. Kappes, and C. Schnörr, "Variational recursive joint estimation of dense scene structure and camera motion from monocular high speed traffic sequences," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1692–1699.

[23] M. Aubry, K. Kolev, B. Goldluecke, and D. Cremers, "Decoupling photometry and geometry in dense variational camera calibration," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1411–1418.

[24] M. Roxas and T. Oishi, "Real-time simultaneous 3D reconstruction and optical flow estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2018, pp. 885–893.

[25] P. Chen, "Non-feature extraction-based dense SFM three-dimensional reconstruction method," U.S. Patent 9 686 527 B2, Jun. 20, 2017.

[26] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski, "Edge-preserving decompositions for multi-scale tone and detail manipulation," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–67, 2008.

[27] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.

[28] A. Choukroun and V. Charvillat, "Bucketing techniques in robust regression for computer vision," in *Proc. Scandin. Conf. Image Anal. (SCIA)*, Halmstad, Sweden, Jun. 2003, pp. 609–616.

[29] J. Oliensis, "A new structure-from-motion ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 7, pp. 685–700, Jul. 2000.

[30] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2001, pp. 508–515.

[31] S. Ince and J. Konrad, "Occlusion-aware optical flow estimation," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1443–1451, Aug. 2008.

[32] M. H. Lin and C. Tomasi, "Surfaces with occlusions from layered stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 1073–1078, Aug. 2004.

[33] Y. Wei and L. Quan, "Region-based progressive stereo matching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2004, p. 1.

[34] H. Tao, H. S. Sawhney, and R. Kumar, "A global matching framework for stereo computation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2002, pp. 532–539.

[35] M. A. Fischler, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," in *Readings in Computer Vision: Issues, Problem, Principles, and Paradigms*. 1987, pp. 726–740.

[36] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[37] L. Yu, D. Min, M. N. Do, and J. Lu, "Fast guided global interpolation for depth and motion," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 717–733.

[38] P. Chen, "Hessian matrix vs. Gauss–Newton hessian matrix," *SIAM J. Numer. Anal.*, vol. 49, no. 4, pp. 1417–1435, 2011.

[39] 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.

[40] 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.

[41] G. Graber, J. Balzer, S. Soatto, and T. Pock, "Efficient minimal-surface regularization of perspective depth maps in variational stereo," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 511–520.

[42] Z. Zhang, "Determining the epipolar geometry and its uncertainty: A review," *Int. J. Comput. Vis.*, vol. 27, no. 2, pp. 161–195, 1998.

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