

Received March 25, 2019, accepted April 4, 2019, date of publication April 16, 2019, date of current version April 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2911021

SemFlow: Semantic-Driven Interpolation for Large Displacement Optical Flow

XIANSHUN WANG^{1,2}, DONGCHEN ZHU¹, YANQING LIU^{1,2}, (Student Member, IEEE),
XIAOQING YE^{1,2}, JIAO LI^{1,2}, (Member, IEEE), AND XIAOLIN ZHANG^{1,2,3}, (Member, IEEE)

¹Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China

²School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

³School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China

Corresponding author: Jiaolin Li (jmlili@mail.sim.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61873255, in part by the Shanghai Municipal Science and Technology Major Project through Zhangjiang Laboratory under Grant 2018SHZDZX01, and in part by the Shanghai Science and Technology Committee, China, under Grant 16JC1420503.

ABSTRACT This paper presents a semantic-guided interpolation scheme (SemFlow) to handle motion boundaries and occlusions in large displacement optical flow. The basic idea is to segment images into superpixels and estimate their homographies for interpolation. In order to ensure each superpixel can be approximated as a plane, a semantic-guided refinement method is introduced. Moreover, we put forward a homography estimation model weighted by the distance between each superpixel and its K -nearest neighbors. Our newly-proposed distance metric combines the texture and semantic information to find proper neighbors. Our homography model performs better than the original affine model, since it accords with the real world projection relationship. The experiments on KITTI dataset demonstrate that SemFlow outperforms other state-of-the-art methods, especially in solving the problem of large scale motions and occlusions.

INDEX TERMS Homography transformation, interpolation-based, optical flow, superpixel, semantic-guided.

I. INTRODUCTION

Optical flow refers to the motion field of the observed scenes. Accurate optical flow estimation plays an important role in many applications including motion segmentation, driver assistance, object detection and augmented reality, *etc.* The state-of-the-art methods have been able to handle the case of small displacements well. However, the optical flow in large displacement is still a challenging problem because of motion discontinuities and occlusions, which are common in real-world.

The classic variational framework casts the optical flow estimation into an energy minimization problem. In order to cope with large displacements, a coarse-to-fine scheme [8], [9] are often used. Such methods work well when the motion structure is larger than the displacement, but this prerequisite often fails. Having considered the robustness of feature matching to large displacements and motion discontinuities, some methods [1], [4], [11] interpolate sparse

matches into dense optical flow directly. They outperform the traditional coarse-to-fine scheme by a large extent. However, the performance of these methods' interpolation models is often deteriorated by inappropriate matches, since the matches are selected only under the guidance of edge maps which just represent the boundaries of texture rather than motion. It causes these algorithms to perform poorly in occlusion and shadow regions. Besides, simplifying the interpolation model as an affine transformation may not be reasonable enough which reduces the overall performance of these methods.

To cope with above problems of existing interpolation-based methods, we propose a semantic-driven interpolation scheme (*SemFlow*) for optical flow estimation in this paper. We use semantic information and edge information together to enable our algorithm to choose more appropriate sparse matches to calculate the interpolation model. Additionally, the more reasonable homography transformation model is applied to improve the overall performance of our algorithm. The input images are over-segmented into superpixels leveraging the semantic information, and

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

the K -nearest neighbors for each superpixel are found guided by our designed distance metric. Then the proposed superpixel-weighted homography model is applied for interpolation. Given the same input matches as RicFlow [11], EpicFlow [1], and InterpoNet [4] used, *SemFlow* outperforms on KITTI benchmarks [15], [16], particularly at motion boundaries. To sum up, the contribution of this paper consists of:

- A novel interpolation-based method for large displacement optical flow is proposed, which is robust to tackle with motion boundaries and occlusions.
- A superpixels refinement scheme guided by semantic information is proposed, which makes sure each superpixel can be approximated as a plane.
- A homography estimation model is proposed, which is weighted by the distance between each superpixel and its K -nearest neighbors, and a distance metric between superpixels is designed to pick out more reasonable neighbors.
- Our method achieves the state-of-the-art interpolation results on KITTI dataset.

II. RELATED WORK

The traditional way to estimate dense optical flow is based on the variational framework which is introduced by Horn and Schunck [7]. It's a kind of energy-based method which has many limitations in practice, as then many different energy terms are proposed to improve and perfect this framework [8], [18], [19]. These methods have worked effectively in small displacements. However, because these energy terms is highly nonlinear, the minimization often get stuck in local minima when it comes to large displacements even though it is carried using a coarse-to-fine scheme [8]. Due to the robustness of descriptor matching in large displacements, some methods [2], [12], [20] integrate matching results into the existing energy terms and achieve better performance.

However, the above mentioned methods still fail when the motion structure is smaller than the displacements because of the limitation of the coarse-to-fine scheme. In order to fully exploit the robustness of descriptor matching to large displacement and occlusion, many methods [1], [4], [10], [11], [21] directly interpolate the sparse set of matches in dense optical flow. Leordeanu *et al.* [21] firstly obtain the sparse matches by minimizing their matching energy based on local feature descriptors and then estimating the interpolation models based on its matching results. Revaud *et al.* [1] propose an edge-preserving interpolation model and directly use the state-of-the-art matching methods as input [2]. Zweig and Wolf [4] design a interpolation network which can directly output the optical flow fields from the input matches. To cope with the input matching noise, Hu *et al.* [11] utilize a RANSAC-like technique to estimate the interpolation models. However, all these above methods except [4] interpolate the optical flow using an affine model which doesn't accord with the actual projection relationship. Additionally, the

distance metrics of above methods are mainly based on the edge maps which cannot reflect the real motion boundaries.

Actually, the precious transformation between matches in two different views is a little complicated. It not only depends on the relative motion of the camera but also the depth of the objects, thus for different matches there are different transformations. Fortunately, when the object points stay on the same plane, the transformations become simpler. For these matches, they will satisfy the same transformation, i.e., the homography transformation. Hereby, we propose a novel piecewise parametric flow based on homography model. Actually, the idea of piecewise flow has already appeared before [22], [23], [26]. However, most of these methods characterize the transformation using an affine model. Differently, Yang *et al.* [25], [27] firstly introduce the homography model to estimate the optical flow. But the segmentation and the transformation parameters are estimated by global energy minimization which don't take advantage of descriptor matching in large displacements and occlusions.

Besides, we also introduce the high-level semantic information in our interpolation scheme. Sevilla-Lara *et al.* [24] segment the scene into objects of different types and use different motion models to describe these objects. Similarly, Bai *et al.* [28] segment the scene into foreground and background and estimate the flow separately by a deep convolution neural network. In contrast, leveraging the semantic information, we segment the scenes into smaller patches rather than directly use the semantic segmentation results. This makes our method robust to inaccuracy of semantic segmentation. Moreover, in our method, the semantic information is also used to measure the distance between two superpixels which provides more reasonable results than [1], [10], [11], [21].

III. APPROACH

Our optical flow estimation system follows the pipeline based on interpolation, which consists of two steps: 1) obtainment of sparse correspondences; 2) dense flow estimation by interpolating sparse matches. There have been lots of works [2], [29], [30] that has contributed to the first issue. In this article, we focus on the second one.

Given two consecutive images I, I' and their corresponding matches $\mathcal{M} = \{(\mathbf{p}_m, \mathbf{p}'_m)\}$ where the pixel $\mathbf{p}_m \in I$ and $\mathbf{p}'_m \in I'$. Our purpose is to find a dense correspondence field $F: I \rightarrow I'$. An intuitive idea is to estimate a piecewise transformation model X leveraging the sparse matches and then calculate the optical flow at \mathbf{p} using the follow transformation relationship:

$$F(\mathbf{p}) = X\mathbf{p} - \mathbf{p} \quad (1)$$

In most of methods, X is characterized as an affine model but it doesn't accord with the actual projection. Instead, the homography model is used in our method. The outline of our approach is shown in Figure 1. Firstly, we use the semantic segmentation to refine the original superpixel results. In consequence, every segmentation can be approximated as

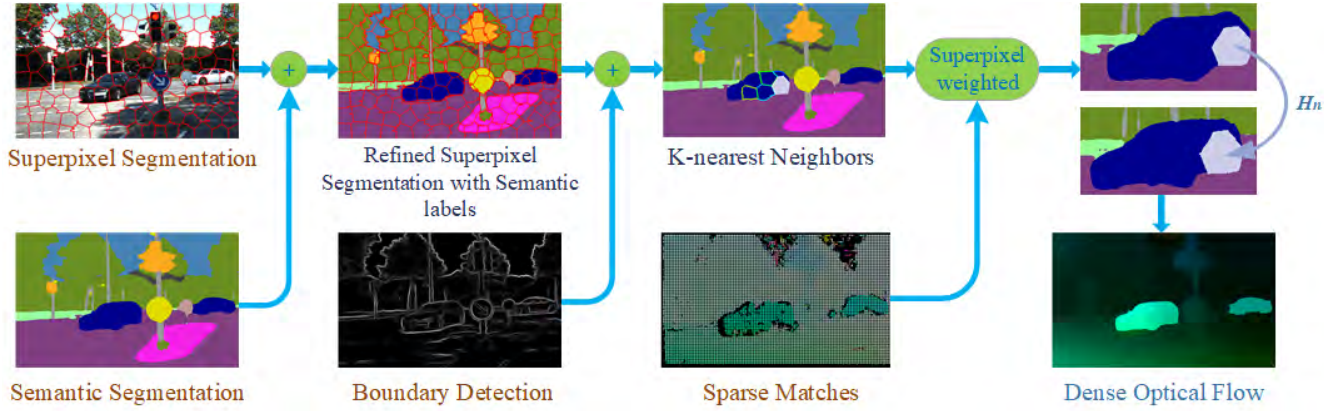


FIGURE 1. Overview of the proposed SemFlow.

a plane and has its own semantic label. Secondly, we find the K -nearest neighbors for each superpixel using a new proposed distance metric. Thirdly, a superpixel-weighted homography transformation model H_n is estimated for each segmentation s_n . Then given any pixel coordinate $p_i \in s_n$, its corresponding optical flow can be given by:

$$F(p_i) = T(H_n \tilde{p}_i) - p_i \quad (2)$$

where \tilde{p}_i is the homogeneous coordinate of p_i , $T(\cdot)$ is a transformation from homogeneous coordinates to inhomogeneous coordinates.

A. SUPERPIXEL SEGMENTATION REFINEMENT

The key assumption of *SemFlow* is that each superpixel can be approximated as a plane, which can only be satisfied when the superpixel size is small enough. However, if the size of superpixels is too small, there will not be enough matches to estimate homography robustly for each superpixel especially when the matches contain a lot of noise. In practice, we found that only those superpixels containing different objects violate our assumption, when the average superpixel size is set to a proper value. Thus the semantic information [3] is introduced to further segment the violative superpixels.

In specific, the basic superpixel segmentation SLIC [13] and semantic segmentation DeepLab [3] are first performed. The average superpixel size is set to σ . For those superpixels that contain two or more semantic objects, we label each connected component containing a single object as a new superpixel. For example, when we detect the connect components for semantic label l in superpixel s , we set all the pixels in s with label l as foreground and with other labels as background. Then a connect components detector is applied in this boolean area to get the connect components of label l . Consequently, each image is segmented into N non-overlapping superpixels s_n with semantic label l_n . Figure 2 shows an example of our refinement procedure.

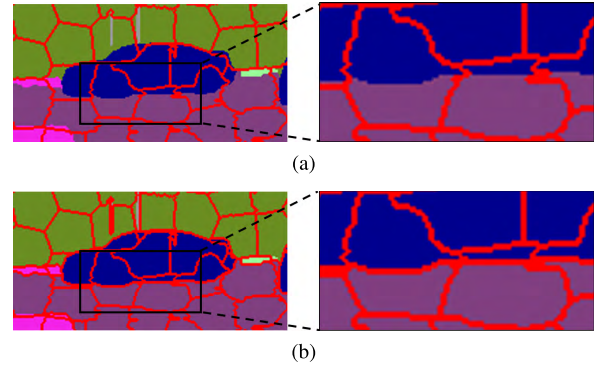


FIGURE 2. Superpixel refinement utilizing semantic information. Some original superpixels contain different objects which lie on different planes. Our approach can obtain more appropriate results. (a) Original superpixels and partial enlarged details. (b) Refined superpixels and partial enlarged details.

B. SUPERPIXEL-WEIGHTED HOMOGRAPHY ESTIMATION

Even though we have set the average superpixel size properly, there are still some superpixels that don't have enough matches within themselves. Besides, it is highly possible that the homography transformations between two neighboring superpixels change smoothly, and thus, flow values, but the flow values in many methods are often discontinuous due to the outliers. To this end, a novel superpixel-weighted model is designed to estimate H_n .

Considering a superpixel s_n and its K -nearest neighbors $\mathcal{N}_K(s_n)$, the homography transformation H_n is estimated by minimizing the following cost function:

$$C(H_n) = \sum_{s_i \in \mathcal{N}_K(s_n)} (k_D(s_i, s_n) \times \sum_{(p_i, p'_i) \in \mathcal{M}_{s_i}} \Psi(\|p'_i - T(H_n \tilde{p}_i)\|^2)) \quad (3)$$

where \mathcal{M}_{s_i} is the matches subset within superpixel s_i and $\mathcal{M} = \{\mathcal{M}_{s_i} \mid i = 1 \dots N \text{ and } \forall k \neq m, \mathcal{M}_{s_k} \cap \mathcal{M}_{s_m} = \emptyset\}$. $k_D(s_i, s_n) = \exp(-\alpha D(s_i, s_n))$ is the weight of s_i , which is a Gaussian kernel function with parameter α . $D(\cdot, \cdot)$ denotes



FIGURE 3. Comparison of K -nearest neighbors without and with semantic information. (a) shows the selected neighbors for the central superpixel (filled in white) which is only guided by the edge maps, and (b) shows the selected neighbors using our defined distance. The K -nearest superpixels on (a) lie on totally different planes. Our distance metric can find the suitable neighbors successfully.

the distance between two superpixels. In order to achieve robustness to outliers, the Cauchy loss $\Psi(s) = \log(1 + s)$ is applied in our cost function. $\mathbf{H}_n \in \mathbb{R}^{3 \times 3}$ has the following form:

$$\mathbf{H}_n = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix} \quad (4)$$

\mathbf{H}_n has 9 variables but only 8 degrees of freedom, thus we set $h_9 \equiv 1$. Note that, the cost function in formula 3 depends on not only matches from current superpixel s_n but also from its neighbors $\mathcal{N}(s_n)$. We believe the introduction of the matches in $\mathcal{N}(s_n)$ can keep the estimated results robust from outliers and keep flow values smooth along boundaries by adding additional consistency from neighborhood. In addition, for the superpixels with fewer matches, we take advantage of the matches from their neighbors to compute the homography transformation.

C. SEMANTIC-GUIDED K -NEAREST NEIGHBORS

The key to our weighted homography is to find proper K -nearest neighbors $\mathcal{N}(s_n)$ for each superpixel s_n . This requires us to find a reasonable solution to measure the distance between two superpixels. The distance metrics used in many other methods [1], [10], [11], [21] are mainly based on edge maps. However, edges only reflect boundaries of texture rather than motion, thus cannot provide accurate distance information. Here we further make use of the semantic information to define a novel distance.

Firstly, we define the distance between two adjacent superpixels (s_a, s_b). Given the edges [17] as our basic cost map, the distance between s_a and s_b is defined as:

$$D(s_a, s_b) = \frac{\omega(l_a, l_b)}{|\mathcal{A}|} \sum_{(p,q) \in \mathcal{A}} (B(p) + B(q)) + (1 - \delta(l_a - l_b))\gamma \quad (5)$$

where \mathcal{A} denotes the adjacent pixels set between s_a and s_b , and $B(\cdot)$ is the basic cost of the corresponding pixels. $\omega(l_a, l_b)$ is a weight determined by semantic label. l_a and l_b are the semantic labels of s_a and s_b . $\delta(l_a - l_b)$ is the Dirac delta function which equal to 0 everywhere except for $l_a = l_b$ where equal to 1. We assume that superpixels with different

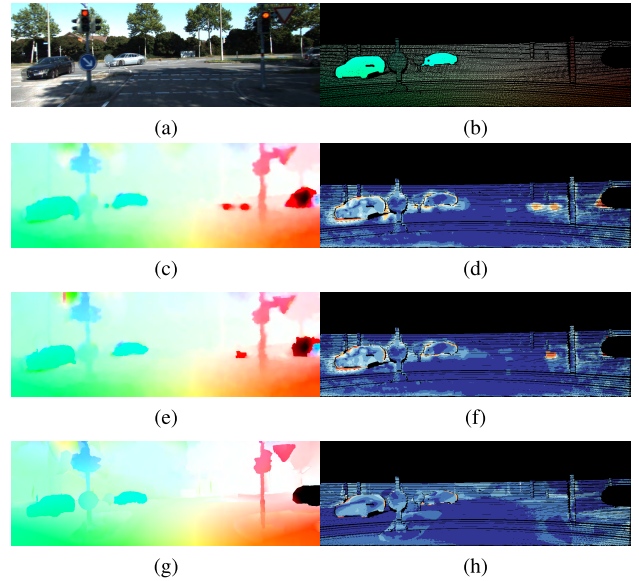


FIGURE 4. Compare our result with other state-of-the-art interpolation-based methods. (a) and (b) are the average images and the sparse ground truth. (c), (e), and (f) are the interpolation results of different methods. (d), (f), and (h) are the corresponding error maps. Notable improvements can be observed at the boundaries of different objects (correct estimates colored in blue and wrong estimates colored in red).

semantic labels are most likely to lie on different planes so we add an offset γ in this situation. Inspired by [24], we treat ‘roads’, ‘sky’ and ‘sidewalks’ as a special class of objects \mathcal{L}_s . Then $\omega(l_a, l_b)$ has the following form:

$$\omega(l_a, l_b) = \begin{cases} \beta & l_a = l_b \text{ and } l_a \in \mathcal{L}_s \\ 1 & \text{others} \end{cases} \quad (6)$$

Here we have $\beta < 1$ since superpixels with same special semantic labels are most likely to lie on the same plane.

To define the distance between any superpixels we construct a neighborhood graph $G = (\mathcal{V}, \mathcal{E})$. \mathcal{V} represents the set of nodes and is composed of all the superpixels. \mathcal{E} represents the set of edges which only connect adjacent superpixels. The distance between any superpixels, thus, can be calculated by Dijkstra’s algorithm and then the K -nearest neighbors of any superpixel. Note that, we set the number of neighbors to a larger value K_s for these superpixels with semantic labels in \mathcal{L}_s , because the objects in class \mathcal{L}_s all have a broad spatial extent and are roughly planar [24]. An example of selected K -nearest neighbors using different distance metrics(ours resp. edge maps) is shown in Figure 3.

D. SPECIAL CASE

In order to ensure that each superpixel has enough matches to estimate the homography, we may need to set K to a large value since the input matches may be very sparse in some regions. However, our assumption requires the superpixel and its neighbors lie on the same plane approximately which rejects those distant superpixels. A complementary solution is needed even if it’s seldom a problem. For superpixel s_n

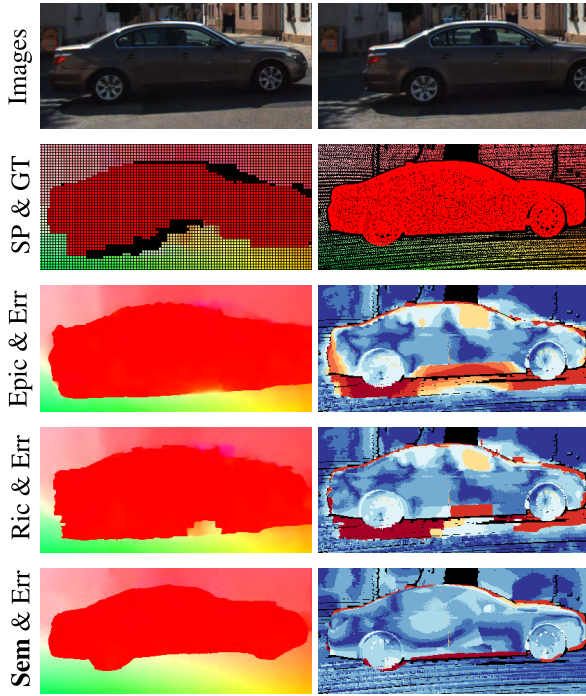


FIGURE 5. Comparison at motion boundaries. Our method can handle the motion boundaries quite well and suppress the input noise to some extent.

TABLE 1. Comparison of the outliers percentage($EPE > 3px$) of SemFlow with RicFlow and EpicFlow for different input matches. By comparison under KITTI-2012 and KITTI-2015, our method outperforms EpicFlow and RicFlow, especially for the more challenging KITTI-2015. In addition, the results with and without variational refinement are reported separately to be fair.

Method	Var	Matching	KITTI-2012(%)	KITTI-2015(%)
EpicFlow	-	DM	19.00	31.53
RicFlow	-	DM	13.78	25.15
SemFlow	-	DM	12.61	19.27
EpicFlow	-	CPM	13.32	24.66
RicFlow	-	CPM	10.15	20.31
SemFlow	-	CPM	9.37	12.18
EpicFlow	-	DF	15.04	22.71
RicFlow	-	DF	14.10	21.39
SemFlow	-	DF	11.28	14.32
EpicFlow	+	DM	16.63	28.03
RicFlow	+	DM	13.82	23.56
SemFlow	+	DM	13.66	20.39
EpicFlow	+	CPM	14.09	23.52
RicFlow	+	CPM	11.97	20.34
SemFlow	+	CPM	11.31	14.93
EpicFlow	+	DF	14.51	21.85
RicFlow	+	DF	14.08	21.04
SemFlow	+	DF	13.06	16.91

that meets:

$$\sum_{s_i \in \mathcal{N}_K(s_n)} |\mathcal{M}_{s_i}| < t \quad (7)$$

we keep increasing the number of neighbors until the matches are enough. In such cases, the homography model is no longer used due to our plane assumption. Instead, the original affine model is applied.

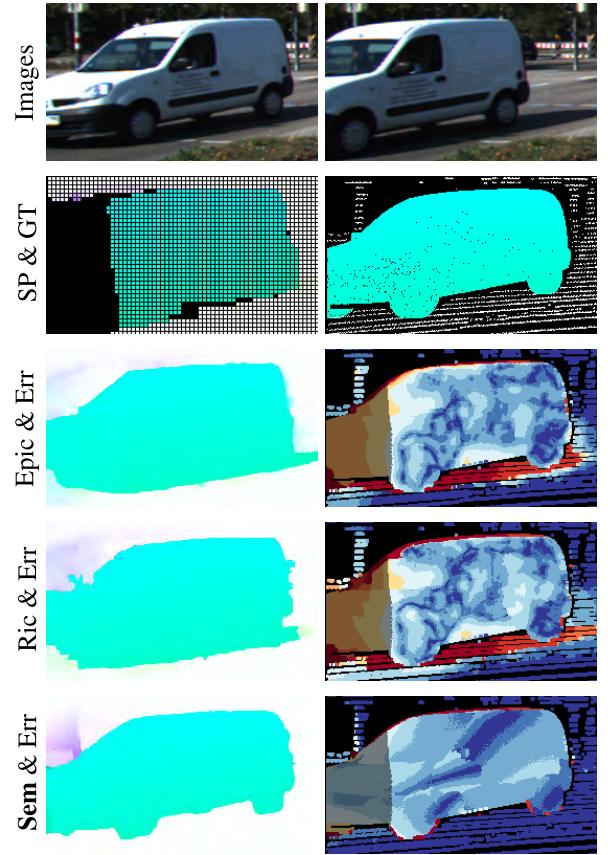


FIGURE 6. Comparison in occluded regions. Our method outperforms the others with the same input matches. Dark regions in the error images denote the occluded pixels.

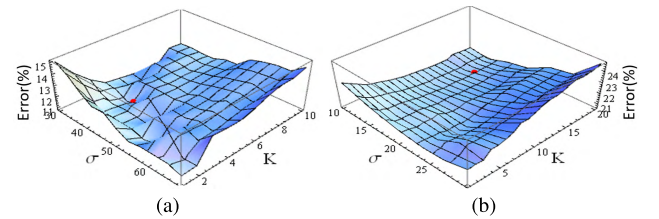


FIGURE 7. Performance of different combinations of K and σ on KITTI-2015. (a) and (b) show the results with and without semantic information respectively. The red points mark the best combinations which appear at $\{\sigma, K\} = \{40, 4\}$ with 10.88% outliers percentage in (a) and $\{\sigma, K\} = \{15, 16\}$ with 20.65% outliers percentage in (b).

IV. EXPERIMENTS

We evaluated our method on the KITTI dataset:

- *KITTI-2012 dataset* [15], it contains 194 training and 195 testing scenes of a static environment in real world, including large displacements, complex 3D objects, and lighting conditions which often happen in driving environment.
- *KITTI-2015 dataset* [5], it is similar to KITTI-2012 but more challenging. It consists of 200 dynamic scenes in the training set and 200 dynamic scenes in the testing set. As thus, compared to KITTI-2012, the optical flow estimation on KITTI-2015 is more difficult.

Similar to [1], [6], [11], the parameters in our method are optimized on a subset (20%) of the KITTI-2015 training set

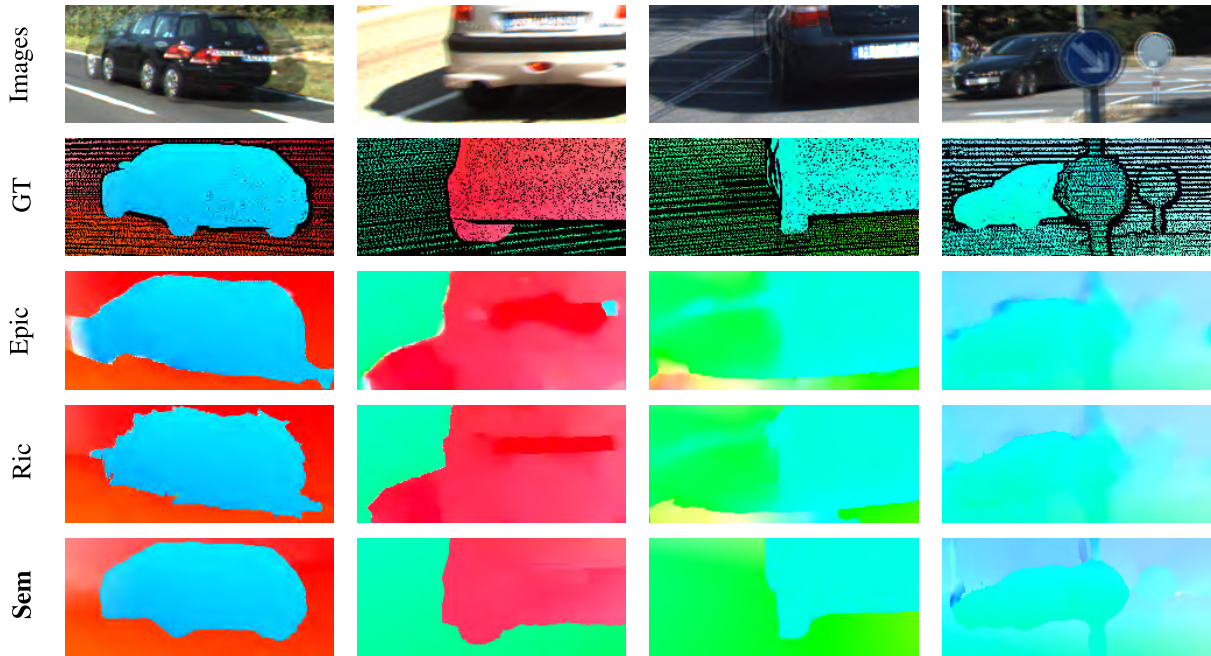


FIGURE 8. Details of different interpolation-based methods. Notice that how our method handle the boundary of the moving objects and the shadows.

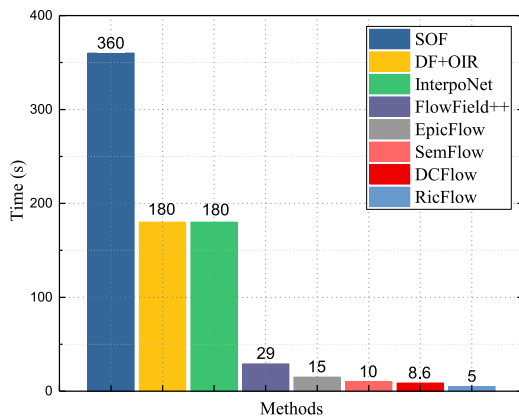


FIGURE 9. Running times of different algorithms on KITTI.

and then the percentages of outliers ($EPE > 3px$) over all pixels on the training set of KITTI-2015 and KITTI-2012 are reported. Here we set $K_s = 20K$ manually and then optimize the other parameters. The typical parameters we used in all evaluations are: $\{\sigma, K, \beta, \gamma\} = \{40, 4, 0.2, 2\}$. The variational refinement code¹ in Table 1 is exactly the same as EpicFlow [1].

A. INTERPOLATION-BASED METHODS COMPARISON

1) INPUT MATCHES

To demonstrate the robustness of our method, we compared the interpolation results with three popular matches:

¹<http://lear.inrialpes.fr/src/epicflow>

DM [2]², **CPM** [14]³ and **DF** [29]⁴. We directly used their code published online without any modification.

2) COMPARISON RESULTS

We report our method's performance along with other two popular interpolation schemes (RicFlow and EpicFlow) in Table 1 and a visualized comparison of these three algorithms is illustrated in Figure 4. For a fair comparison, the results with(+Var) and without(−Var) variational refinement are reported. We can see that our method outperforms the other two methods in all cases, especially in the more challenging KITTI-2015. Interestingly, we find that the post refinement step doesn't improve our results but leads them worse. Moreover, this phenomenon not only appear in our algorithm but also happen on RicFlow when the input matches come from **CPM**, see Table 1. This is due to the fact that our results already exceed the optimization ability of the variational framework used in EpicFlow. In terms of occlusions and motion boundaries, our SemFlow outperforms other interpolation-based methods as shown in Figure 5 and Figure 6 respectively.

As an interpolation-based method, our results are highly relative to the quality of input matches. Table 1 has shown **CPM**'s great performance among different interpolation schemes. As thus, all the following experiments are based on the **CPM** input matches.

²<http://lear.inrialpes.fr/src/deepmatching/>

³<https://github.com/YinlinHu/CPM>

⁴http://www.cvlibs.net/download.php?file=discrete_flow.zip

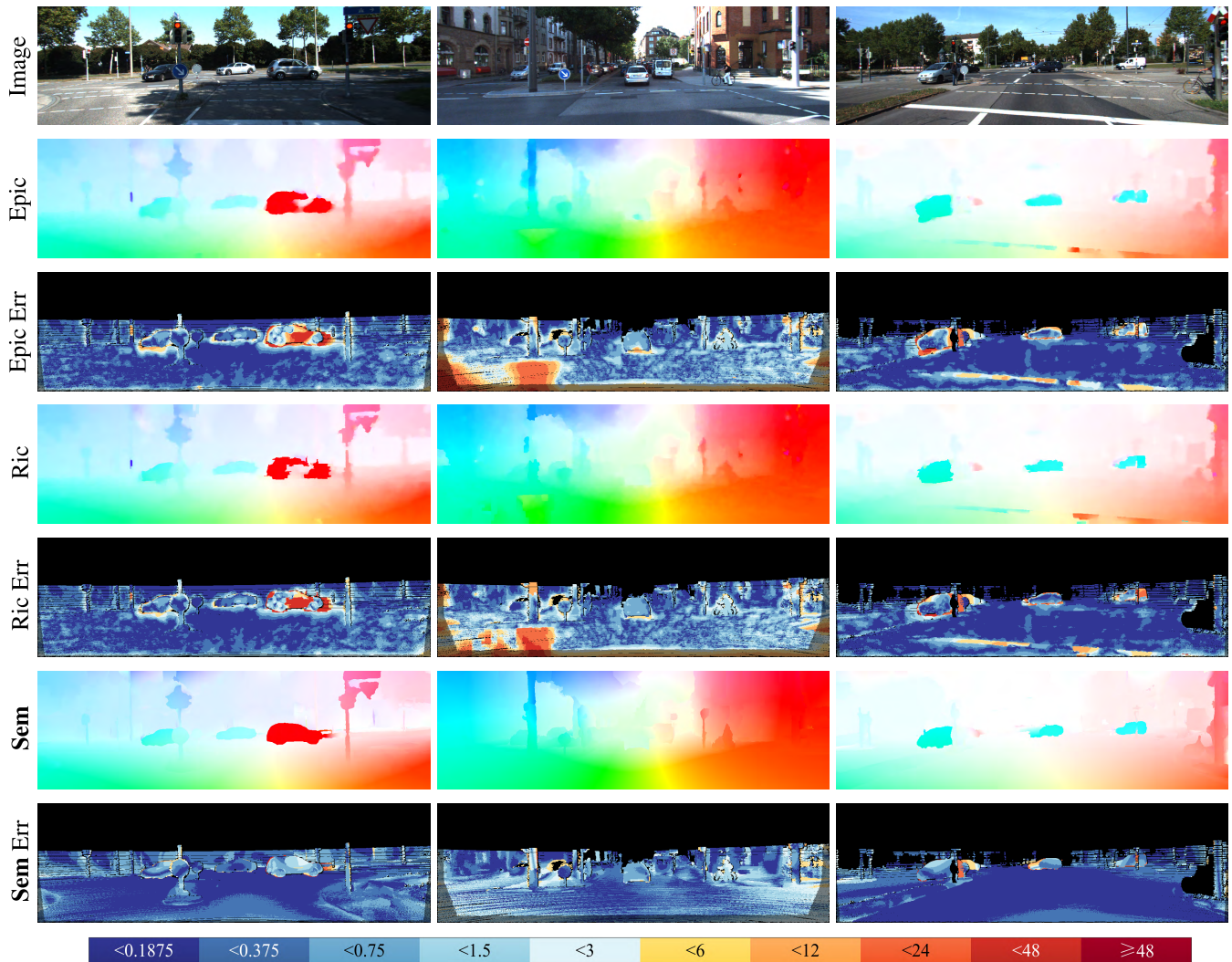


FIGURE 10. Qualitative results of different interpolation-based approaches on three images from KITTI-2015.

B. DIFFERENT PARAMETERS COMPARISON

1) PARAMETERS TRADE-OFF

On one hand, the core assumption of our method is that the segmentations must be small enough to be approximated as planes. On the other hand, however, we also need to ensure enough matches within every segmentation and its neighbors. Thus these two parameters, i.e., the average superpixel size σ and the number of the nearest neighbors K play an important role to our method's performance. To understand the impact of them, we test different combinations of K and σ within proper ranges on a subset of KITTI-2015 training set. The results are shown in Figure 7a. We can see that for a smaller σ , more neighboring superpixels are needed to achieve best performance and vice versa, since smaller superpixel usually needs more neighbors to get enough matches. However, we also find that the performance doesn't always get better with more neighbors, since the central superpixel and its distant neighbors under our distance metric can't be simply supposed to lie on the same plane. Our experiment shows that $\{\sigma, K\} = \{40, 4\}$ gives a good balance.

2) IMPACT OF SEMANTIC INFORMATION

In order to figure out the impact of semantic information on our algorithm. We also test our algorithm's performance under different combinations of K and σ when the semantic information is removed. This is achieved by skipping the superpixel segmentation refinement step and setting $\omega = 1, \gamma = 0, K_s = K$. The test result is shown in Figure 7b. We can observe that the best combination tends to appear at smaller superpixel size and more neighbors where $\{\sigma, K\} = \{15, 16\}$, since only small enough superpixel can be approximated as a plane without semantic information which needs more neighbors to find enough matches. Under this circumstance, the outliers percentage increased by 9% overall.

C. KITTI OPTICAL FLOW BENCHMARK

For fair comparison, we only compare our method with other top methods that do not use epipolar geometry or stereo vision.

Table 2 reports the result of our method along with other top methods on KITTI-2012 and KITTI-2015

TABLE 2. Results of top algorithms on KITTI-2015 and KITTI-2012 benchmarks. Out-Bg3 (resp. Out-Fg3 and Out-All3) represents the percentage of outliers (EPE > 3) in background regions (resp. foreground regions and all regions). InterpoNet and SDF are learning-based methods. Starred methods are based on interpolation.

Method	2015-%Out			2012-%Out
	Bg3	Fg3	All3	All3
SDF	8.61	23.01	11.01	7.69
SemFlow*	11.49	18.12	12.60	11.10
DCFlow [33]	13.10	23.70	14.86	-
FlowField++ [31]	14.82	17.77	15.31	-
SOF	14.63	22.83	15.99	-
DF+OIR [32]	15.11	23.45	16.50	10.43
RicFlow*	18.73	19.09	18.79	13.04
InterpoNet*	22.15	26.03	22.80	14.13
EpicFlow*	25.81	28.69	26.29	17.08

testing set. RicFlow, EpicFlow, and InterpoNet are three interpolation-based methods. Our method outperforms these methods significantly. In terms of Out-All3 on KITTI-2015, our method performs better by 6% at least. SOF [24] and SDF [28] are two semantic-driven methods. Our method performs better than SOF but less worse than SDF which is a deep learning based approach. Even so, our method outperforms SDF in the foreground regions.

Compared with the interpolation-based methods, the advantage of our method is more obvious on KITTI-2015. In the favor of the semantic information, our method can handle the boundaries between moving objects and background effectively. Furthermore, it also can suppress the noise of input matches caused by shadows. Figure 8 shows some comparison results at boundaries of moving objects and in the shaded areas. As illustrated in Figure 9, our SemFlow runs in 10 seconds for a KITTI image pair (1242×375) on Intel i7 at 2.4Ghz which is among the fast methods. More qualitative comparison results are shown in Figure 10.

V. CONCLUSION

This paper introduces a semantic-driven optical flow estimation method, named *SemFlow*, which produces a dense flow field by a sparse-to-dense interpolation. The core assumption is that the images can be segmented into distinct superpixels which can be approximated as planes independently. Starting from this assumption, two key issues need to be addressed. The first one is how to obtain the appropriate superpixels, and the second is how to handle the deficiency of matches within superpixels and the smoothness between adjacent superpixels. In practice, we found that most superpixels of SLIC [13] meet our segmentation requirement except those containing different objects. Consequently, we introduced semantic information to refine these superpixels. Moreover, we proposed a superpixel-weighted homography model for interpolation instead of the widely-used affine model, and designed a semantic-guided distance metric for weight estimation. Experiments on KITTI dataset show that *SemFlow* outperforms the state-of-the-art methods, especially in solving the problem of large scale motions and occlusions.

ACKNOWLEDGMENT

Project supported by Shanghai Municipal Science and Technology Major Project (Grant No. 2018SHZDZX01, ZHANGJIANG LAB).

REFERENCES

- [1] 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.
- [2] 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, Dec. 2016.
- [3] 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.*, Sep. 2018, pp. 801–818.
- [4] S. Zweig and L. Wolf, "InterpoNet, a brain inspired neural network for optical flow dense interpolation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4563–4572.
- [5] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3061–3070.
- [6] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2014, pp. 1385–1392.
- [7] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.
- [8] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. Eur. Conf. Comput. Vis.*, May 2004, pp. 25–36.
- [9] 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.
- [10] R. Schuster, O. Wasenmuller, G. Kuschik, C. Bailer, and D. Stricker, "SceneFlowFields: Dense interpolation of sparse scene flow correspondences," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2018, pp. 1056–1065.
- [11] Y. Hu, Y. Li, and R. Song, "Robust interpolation of correspondences for large displacement optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4791–4799.
- [12] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.
- [13] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
- [14] Y. Hu, R. Song, and Y. Li, "Efficient coarse-to-fine patch match for large displacement optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5704–5712.
- [15] 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.
- [16] M. Menze, C. Heipke, and A. Geiger, "Joint 3D estimation of vehicles and scene flow," in *Proc. ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, Aug. 2015, pp. 1–8.
- [17] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1841–1848.
- [18] A. Bruhn and J. Weickert, "Towards ultimate motion estimation: Combining highest accuracy with real-time performance," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 749–755.
- [19] J. Xiao, H. Cheng, H. Sawhney, C. Rao, and M. Isnardi, "Bilateral filtering-based optical flow estimation with occlusion detection," in *Proc. Eur. Conf. Comput. Vis.*, May 2006, pp. 211–224.
- [20] 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.
- [21] M. Leordeanu, A. Zanfir, and C. Sminchisescu, "Locally affine sparse-to-dense matching for motion and occlusion estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1721–1728.
- [22] L. Xu, J. Chen, and J. Jia, "A segmentation based variational model for accurate optical flow estimation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2008, pp. 671–684.

- [23] X. Ren, "Local grouping for optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
- [24] 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.*, Jun. 2016, pp. 3889–3898.
- [25] M. Hornáček, F. Besse, J. Kautz, A. Fitzgibbon, and C. Rother, *Highly Overparameterized Optical Flow Using PatchMatch Belief Propagation*. Springer, Sep. 2014, pp. 220–234.
- [26] M. J. Black, D. Sun, and E. B. Sudderth, "Layered segmentation and optical flow estimation over time," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1768–1775.
- [27] J. Yang and H. Li, "Dense, accurate optical flow estimation with piecewise parametric model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1019–1027.
- [28] M. Bai, W. Luo, K. Kundu, and R. Urtasun, "Exploiting semantic information and deep matching for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 154–170.
- [29] M. Menze, C. Heipke, and A. Geiger, "Discrete optimization for optical flow," in *Proc. German Conf. Pattern Recognit.*, Nov. 2015, pp. 16–28.
- [30] T. Kroeger, R. Timofte, D. Dai, and L. van Gool, "Fast optical flow using dense inverse search," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 471–488.
- [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.*, Dec. 2015, pp. 4015–4023.
- [32] D. Maurer, M. Stoll, and A. Bruhn, "Order-adaptive and illumination-aware variational optical flow refinement," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 9–26.
- [33] J. Xu, R. Ranftl, and V. Koltun. (Apr. 2017). *Accurate Optical Flow via Direct Cost Volume Processing*. [Online]. Available: <https://arxiv.org/abs/1704.07325>



XIANSHUN WANG received the B.S. degree in opto-electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2012. He is currently pursuing the Ph.D. degree in information and communication engineering with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences. He is also an intern student with Shanghai Eyevolution Technology Co., Ltd., Shanghai, China. His current research interests include 3D reconstruction, 3D scene understanding, and visual odometry.



DONGCHEN ZHU received the B.S. degree from Wuhan University, China, in 2013, and the Ph.D. degree in information and communication engineering from the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China, in 2018, where she is currently an Assistant Professor.



YANQING LIU received the B.E. degree in the Internet of Things engineering from Shandong University, Shandong, China, in 2014. He is currently pursuing the Ph.D. degree in communication and information system with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences. His research interests include robot vision, visual odometry, visual simultaneous localization and mapping, and robotics.



XIAOQING YE received the B.S. degree from Wuhan University, China, in 2014. She is currently pursuing the Ph.D. degree in information and communication engineering with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences. Her current research interests include stereo vision, 3-D reconstruction, and autonomous driving.



JIAMAOL I received the Ph.D. degree from the Tokyo Institute of Technology, Japan, in 2012. He is currently a Professor with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences. His current research interests include computer vision, machine vision, 3D micro-imaging, and artificial intelligence.



XIAOLIN ZHANG received the Ph.D. degree from Yokohama National University, in 1995. He was a Professor with the Tokyo Institute of Technology, Japan, from 2012 to 2013. He is currently a Professor with the Shanghai Institute of Microsystem and Information Technology, the Chinese Academy of Sciences, the University of Chinese Academy of Sciences, and ShanghaiTech University. His research interests include bionics, brain science, computer vision, and artificial intelligence.

...