

Joint Stereo Video Deblurring, Scene Flow Estimation and Moving Object Segmentation

Liyuan Pan, Yuchao Dai, Miaomiao Liu, Fatih Porikli, and Quan Pan

Abstract—Stereo videos for the dynamic scenes often show unpleasant blurred effects due to the camera motion and the multiple moving objects with large depth variations. Given consecutive blurred stereo video frames, we aim to recover the latent clean images, estimate the 3D scene flow and segment the multiple moving objects. These three tasks have been previously addressed separately, which fail to exploit the internal connections among these tasks and cannot achieve optimality. In this paper, we propose to jointly solve these three tasks in a unified framework by exploiting their intrinsic connections. To this end, we represent the dynamic scenes with the piece-wise planar model, which exploits the local structure of the scene and expresses various dynamic scenes. Under our model, these three tasks are naturally connected and expressed as the parameter estimation of 3D scene structure and camera motion (structure and motion for the dynamic scenes). By exploiting the blur model constraint, the moving objects and the 3D scene structure, we reach an energy minimization formulation for joint deblurring, scene flow and segmentation. We evaluate our approach extensively on both synthetic datasets and publicly available real datasets with fast-moving objects, camera motion, uncontrolled lighting conditions and shadows. Experimental results demonstrate that our method can achieve significant improvement in stereo video deblurring, scene flow estimation and moving object segmentation, over state-of-the-art methods.

Index Terms—Stereo deblurring, motion blur, scene flow, moving object segmentation, joint optimization.

I. INTRODUCTION

IMAGE deblurring aims at recovering the latent clean image from a single or multiple blurred images, which is a classic and fundamental task in image processing and computer vision. Image blur could be caused by various reasons, for example, optical aberration, medium perturbation, defocus, and motion [1], [2], [3], [4], [5]. In this work, we only focus on motion blur, which is widely encountered in real-world applications such as autonomous driving [6], [7]. The effects become more apparent when the exposure time increased due to low-light conditions.

Motion deblurring has been extensively studied and various methods have been proposed in the literature. It is common to model the blur effect using kernels [4], [12]. Early deblurring methods mainly focus on the blur caused by camera shake in constant depth or static scenes with moving objects [13], [14]. In this work, we focus on a more generalized motion blur

Liyuan Pan, Miaomiao Liu and Fatih Porikli are with Research School of Engineering, the Australian National University, Canberra, Australia.

Yuchao Dai is with School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. Yuchao Dai (daiyuchao@gmail.com) is the corresponding author.

Quan Pan is with School of Automation, Northwestern Polytechnical University, Xi'an, China.

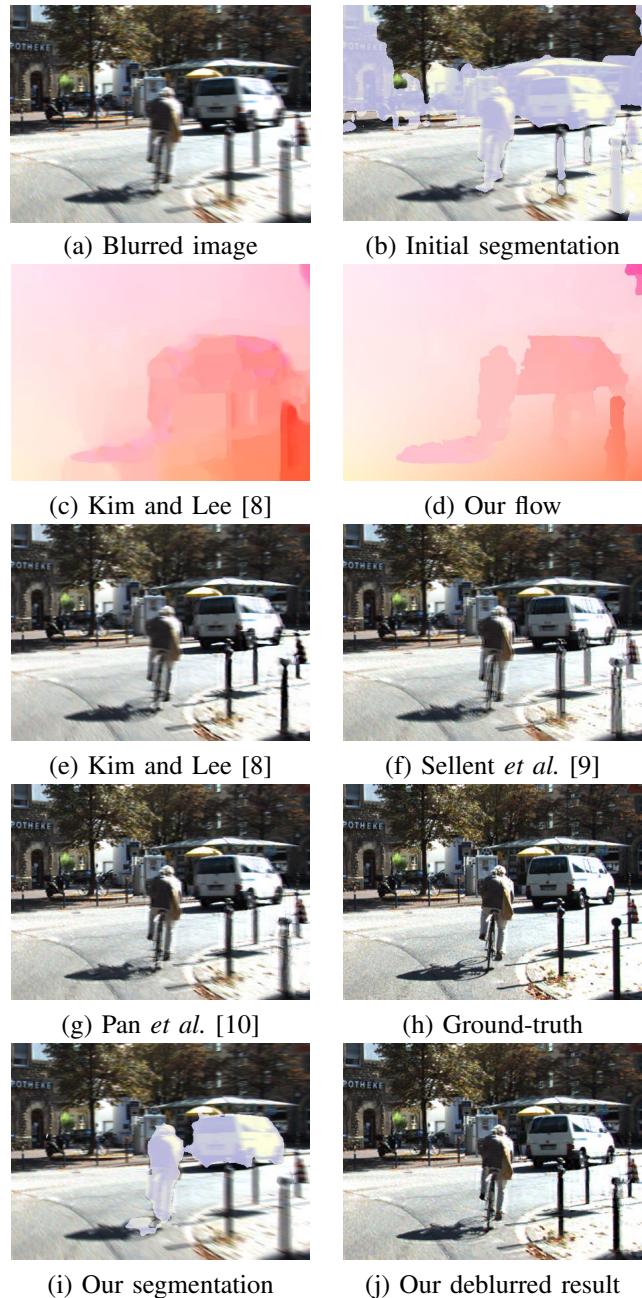


Fig. 1. Stereo deblurring, scene flow estimation and moving object segmentation results with (a) and (b) as input. (a) Blurred image. (b) Initial segmentation prior. (c) Flow estimation by [8]. (d) Our flow estimation result. (e) Deblurring results by [8]. (f) Stereo deblurring results by [9] which uses [11] to estimate scene flow. (g) Deblurring results by [10]. (h) Ground-truth latent image. (i) Our moving object segmentation result. (j) Our stereo deblurring result. Best viewed in colour on the screen.

caused by both camera motion and moving objects. Therefore, conventional blur removal methods, such as [3], [15] cannot be directly applied since they are restricted to a single or a fixed number of blur kernels, making them inferior in tackling general motion blur problems.

For a scenario where both camera motion and multiple moving objects exist, the blur kernel is, in principle, defined for each pixel individually. Recently, several researchers have studied to handle the blurred images with *spatially-variant blur* [8], [9], [10] which uses accurate motion estimation to model the blur kernel. The phenomenon around motion and blur can be viewed as a chicken-egg problem: effective motion blur removal requires accurate motion estimation. Yet, the accuracy of motion estimation highly depends on the quality of the images.

It is a problem for any of the algorithms exploiting motion information as the condition is a major challenge to reliable flow computation.

In this paper, we aim to tackle a ‘generalized stereo deblurring’ problem. The moving stereo cameras observe a dynamic scene with varying depth, and the moving objects’ boundaries are mixed with the background pixels. Thus we propose to utilize the motion boundary information provided by semantic segmentation [16]. In our approach, we jointly estimate scene flow, segment the moving objects and deblur the images under a unified framework. Using our formulation, we attain significant improvement in numerous real challenging scenes as illustrated in Fig. 1.

We would like to argue that, the scene flow estimation approaches that make use of colour brightness constancy may be hindered by the blurred images. Existing optical flow methods make generic, spatially homogeneous, assumptions about the spatial structure of the flow. Due to the inherent correlation between semantic segmentation and Moving object segmentation (for example, the movement of pixels a vehicle tends to be the same and be different from the background), semantic segmentation has been used to provide motion segmentation prior. Thus, we investigate the benefits of semantic grouping [16] which are more beneficial for the scene flow estimation task. Here, we only need a coarse and simple semantic segmentation prior to distinguish foreground and background. The more of the boundary information can be detected during the deconvolution process, the better quality of the estimated results [17], [18]. In Fig. 2, we compare the scene flow estimation results with the state-of-the-art solutions on different blurred images. It could be observed that the scene flow estimation performance deteriorates quickly w.r.t. the image blur because of the inaccuracy at boundaries.

On the other hand, motion segmentation or Moving object segmentation alone is also very challenging as the objects could be rigid, non-rigid, and deformable. How to unify these different scene models and achieve moving object segmentation is an active research direction. In this paper, we focus on outdoor traffic scenes with multiple moving objects, such as vehicles, cyclists, and pedestrians. Specifically, we exploit both the semantic cue and 3D geometry cue to better handle moving object segmentation together with scene flow estimation and stereo deblurring.

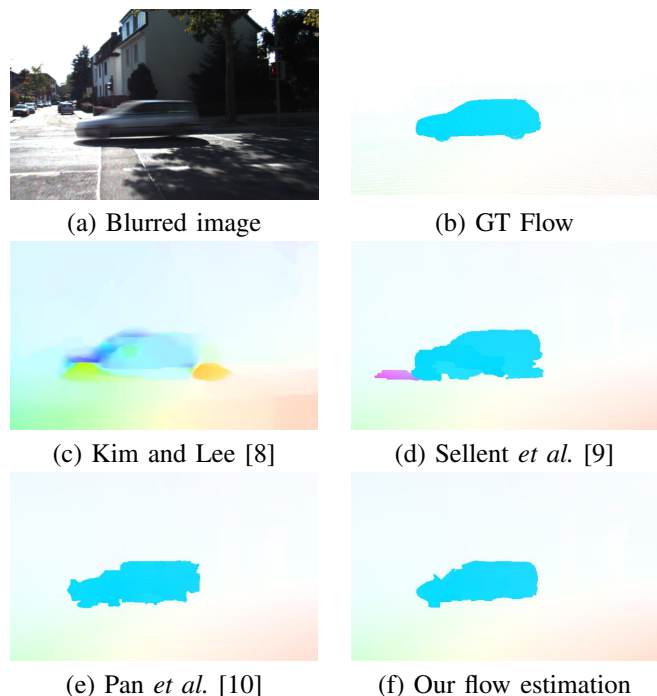


Fig. 2. Scene flow estimation results for an outdoor scene. (a) Blurred reference image from **BlurData-1**. (b) Ground truth optical flow for the scene. (c) Estimated flow by Kim and Lee [8]. (d) Estimated flow by Sellent *et al.* [9] which uses [11] to estimate scene flow. This approach ranks as one of the top 3 approaches on KITTI scene flow benchmark [19]. (e) Estimated flow by Pan *et al.* [10]. (f) Our flow estimation result. Compared with these state-of-the-art methods, our method achieves the best performance.

Furthermore, existing works fail to exploit the connections between stereo deblurring, scene flow estimation and Moving object segmentation, which actually are closely connected. Specifically, better scene flow estimation and Moving object segmentation will enable better stereo deblurring. Correspondingly, stereo deblurring and Moving object segmentation also help scene flow estimation. However, building their intrinsic connections is not easy as the dynamic scenes could be rather generic, from a static scene to a highly dynamic scene consisting of multiple moving objects (vehicles, pedestrians and etc). Having a unified formulation for the dynamic scenes is highly desired. We propose to exploit the piecewise plane model for the dynamic scene structure, and under this formulation, the joint task of scene flow estimation, stereo deblurring and moving object segmentation has been expressed as the parameter estimation for each planar, the camera motion and pixel labelling. Therefore, we put these three tasks in a loop under a unified energy minimization formulation in which the intra-relation has been effectively exploited.

In our previous work [10], we only consider the relationship between optical flow and deblurring without adding segmentation information. We extend the previous work significantly in the following ways:

- We propose a novel joint optimization framework to estimate the scene flow, segment moving objects and restore the latent images for generic dynamic scenes. Our deblurring objective benefits from improved boundaries information and the estimated scene structure.

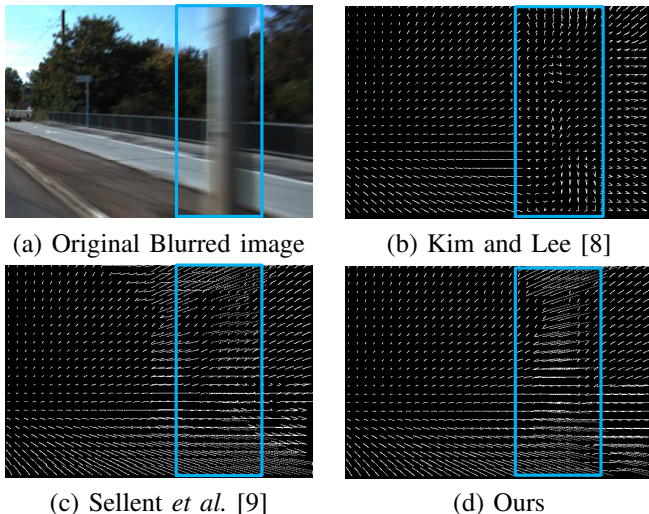


Fig. 3. Blur kernel estimation for an outdoor scene. (a) Blurred reference image from **BlurData-1**. (b) Blur kernel estimation by Kim and Lee [8]. (c) Blur kernel estimation by Sellent *et al.* [9]. (d) Our blur kernel estimation. Compared with these monocular and stereo deblurring methods, our method achieves more accurate blur kernel estimation.

- We integrate high-level semantic cues for camera motion and scene structure estimation by exploiting the intrinsic connection between semantic segmentation and Moving object segmentation.
- We propose a method to exploit motion segmentation information in aiding the challenging video deblurring task. Similarly, the scene flow and objects boundary objective allow deriving more accurate pixel-wise spatially varying blur kernels (see Section.III-B).
- Extensive experiments demonstrate that our method can successfully handle complex real-world scenes depicting fast-moving objects, camera motions, uncontrolled lighting conditions, and shadows.

II. RELATED WORK

Image deblurring (even under stereo configuration) is generally an ill-posed problem, thus certain assumptions or additional constraints are required to regularize the solution space. Numerous methods have been proposed to address the problem [8], [9], [10], [20], [13], [5], [18], [21], [22]. As per the system configuration, the methods can be roughly categorized into two groups: monocular based approaches and binocular or multi-view based approaches. We also briefly discuss recent efforts in deep learning-based deblurring, Moving object segmentation, semantic segmentation, and scene flow estimation.

A. Single view deblur

Monocular based deblurring approaches often assume that the captured scene is static or has uniform blur kernel [3], or need user interaction [18]. A series of widely-used priors and regularizers are based on image gradient sparsity, such as the total variational regularizer [24], the Gaussian scale mixture prior [25], the $l_1 \setminus l_2$ norm based prior [15], and the l_0 -norm regularize [14], [26]. Non-gradient-based priors have

also been proposed, such as the edge-based patch prior [27], the colour line based prior [28], and the dark/white channel prior [29], [30]. Hu *et al.* [13] proposed to jointly estimate the depth layering and remove non-uniform blur caused by the in-plane motion from a single blurred image. While this unified framework is promising, user input for depth layers partition is required, and potential depth values should be known in advance. Pan *et al.* [18] proposed an algorithm to jointly estimate object segmentation and camera motion by incorporating soft segmentation, but require user input. In practical settings, it is still challenging to remove strongly non-uniform motion blur captured in complex scenes.

Since blur parameters and a latent image are difficult to be estimated from a single image, the monocular based approaches are extended to video to remove blurs in dynamic scenes. In the work of Wulff and Black [31], a layered model is proposed to estimate the different motions of both foreground and background layers. Kim and Lee [32] proposed a method based on a local linear motion without segmentation. This method incorporates optical flow estimation to guide the blur kernel estimation and is able to deal with certain object motion blur. In [8], a new method is proposed to simultaneously estimate optical flow and tackle the case of general blur by minimization a single non-convex energy function. Park *et al.* [33] estimate camera poses and scene structures from severely blurred images and deblurring by using the motion information.

B. Multi-view deblur

As depth factor can significantly simplify the deblurring problem, multi-view deblurring methods have been proposed to leverage available depth information. Ezra and Nayar [34] proposed a hybrid imaging system, where a high-resolution camera captures the blurred frame and a low-resolution camera with faster shutter speed is used to estimate the camera motion. Xu *et al.* [35] inferred depth from two blurred images captured by a stereo camera and proposed a hierarchical estimation framework to remove motion blur caused by the in-plane translation. Sellent *et al.* [9] proposed a video deblurring technique based on a stereo video, where 3D scene flow is estimated from the blurred images using a piecewise rigid 3D scene representation. Along the same line, Ren *et al.* [21] proposed an algorithm where accurate semantic segmentation is known. In their work, they also used the pixel-wise non-linear kernel model to approximate motion trajectories in the video. While the performance of their experiments shows limited effective for images which included multiple types of moving objects. We [10] proposed a single framework to jointly estimate the scene flow and deblur the images in CVPR 2017, where the motion cues from scene flow estimation and blur information could reinforce each other. These two methods represent the state-of-the-art in multi-view video deblurring and will be used for comparisons in the experimental section.

C. Deep learning based deblurring methods

Recently, deep learning-based methods have been used to restore clean latent images. Gong *et al.* [22] estimated flow from a single blurred image caused by camera motion through

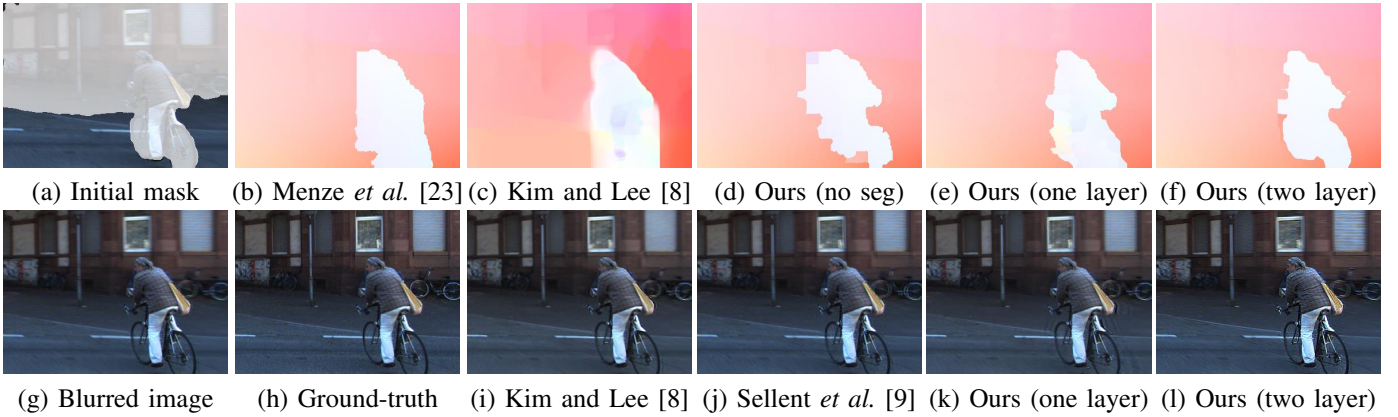


Fig. 4. Scene flow results for an outdoor scenario. (a) and (g) The initial segmentation and blurred reference image from **BlurData-1**. (b) Estimated flow by [23]. (c) Estimated flow by [8]. (d)-(f) Our flow estimation result. (d) Without semantic segmentation. (e) With semantic segmentation, one layer StereoSLIC. (f) With semantic segmentation, two layer StereoSLIC. (h) The ground-truth latent image. (i) Deblurred result by [8]. (j) Deblurred result by [9]. (k) and (l) Our deblurred result. (k) Without semantic segmentation. (l) With semantic segmentation. The results show that, our two layer StereoSLIC could preserve edge information. Compared with both these state-of-the-art methods, our method achieves competitive performance. Best viewed in colour on the screen.

a fully convolutional deep neural network and recovered a clean image from the estimated flow. Su *et al.* [36] introduced a deep learning solution to video deblurring, where a CNN is trained end-to-end to learn how to accumulate information across frames. However, they aimed to tackle motion blur from camera shake. Nah *et al.* [37] proposed a multi-scale convolutional neural network that restores latent images in an end-to-end manner without assuming any restricted blur kernel model. Kim *et al.* [38], [39] proposed a novel network layer that enforces temporal consistency between consecutive frames by dynamic temporal blending which compares and adaptively shares features obtained at different time steps. Kupyn *et al.* [40] presented an end-to-end learning approach for motion deblurring. The model they used is Conditional Wasserstein GAN with gradient penalty and perceptual loss based on VGG-19 activations. Tao *et al.* [41] propose a light and compact network, SRN-DeblurNet, to deblur the image. Jin *et al.* [42] proposed to restore a video with fixed length from a single blurred image. However, deep deblurring methods generally need a large dataset to train the model and usually require sharp images provided as supervision. In practice, blurred images do not always have corresponding ground-truth sharp images.

D. Moving object segmentation

According to the level of supervision required, video segmentation techniques can be broadly categorized as unsupervised, semi-supervised and supervised methods. Unsupervised methods [43] use a rapid technique to produce a rough estimate of which pixels are inside the object based on motion boundaries in pairs of subsequent frames. Then automatically bootstraps an appearance model based on the initial foreground estimate, and uses it to refine the spatial accuracy of the segmentation and to also segment the object in frames where it does not move. The works [44], [45], [46] extend the concept of salient objects detection [47] as prior knowledge to infer the objects. Semi-supervised video segmentation, which also refers to label propagation, is usually

achieved via propagating human annotation specified on one or a few key-frames onto the entire video sequence [48], [49], [50]. The idea of combining the best from both CNN model and MRF/CRF model is not new. A video object segmentation method by Jang and Kim [51] performs MRF optimization to fuse the outputs of a triple-branch CNN. However, the loosely-coupled combination cannot fully exploit the strength of MRF/CRF models. Supervised methods require tedious user interaction and iterative human corrections. These methods can attain high-quality boundaries while needing human supervision [52], [53]. Yan [54] proposed a multi-task ranking model for the higher-level weakly-supervised actor-action segmentation task.

E. Semantic segmentation

Another crucial factor to compute latent clean image is detecting moving objects boundaries. A general problem is that the object boundaries with mixed foreground and background pixels can lead to severe ringing artifacts. Semantic segmentation can help to provide objects information as initialization. He *et al.* [55] proposed the ResNets to combat the vanishing gradient problem in training very deep convolutional networks. [16] obtain the semantic segmentation masks with the ResNet-38 network. Lin *et al.* [56] present RefineNet with multi-resolution fusion (MRF) to combine features at different levels, chained residual pooling (CRP) to capture background context, and residual convolutional units (RCUs) to improve end-to-end learning. Tsai *et al.* [57] first generated the object-like tracklets and then adopted a sub-modular function to integrate object appearances, shapes and motions to co-select tracklets that belong to the common objects. Taking one step further, the Deep Parsing Network (DPN) [58] is designed to approximate the mean-field inference for MRFs in one pass.

F. Optical flow estimation

Menze *et al.* [23] proposed a novel model and dataset for 3D scene flow estimation with an application to autonomous driving. Pan *et al.* [10] proposed a single framework to jointly

estimate the scene flow and deblur the images. Tanai *et al.* [59] presented a multi-frame method for efficiently computing scene flow (dense depth and optical flow) and camera ego-motion for a dynamic scene observed from a moving stereo camera rig. Yin *et al.* [60] proposed an unsupervised learning framework GeoNet for jointly estimating monocular depth, optical flow and camera motion from video. Gong *et al.* [22] directly estimates the motion flow from the blurred image through a fully-convolutional deep neural network (FCN) and recovers the unblurred image from the estimated motion flow. PWC-Net [61] uses the current optical flow estimate to warp the CNN features of the second image. It then uses the warped features and features of the first image to construct a cost volume, which is processed by a CNN to estimate the optical flow. The FlowNet by Dosovitskiy *et al.* [62] represented a paradigm shift in optical flow estimation. The work shows the feasibility of directly estimating optical flow from raw images using a generic U-Net CNN architecture. FlowNet 2.0 [63] develop a stacked architecture that includes warping of the second image with the intermediate optical flow which decreases the estimation error by more than 50% than the original FlowNet.

III. PROBLEM FORMULATION

In this paper, we propose to solve the challenging and practical problem of stereo deblurring by using consecutive stereo image pairs of a calibrated camera in complex dynamic environments, where the blur is caused by the camera motion and the objects' motion. Under the problem setup, stereo deblurring and the scene flow estimation is already deeply coupled, *i.e.*, stereo deblurring depends on the solution of the scene flow estimation while the scene flow estimation also needs the solution of stereo deblurring. In addition, with the multiple moving objects representation of the observed scene, Moving object segmentation also closely relates to both scene flow estimation and stereo deblurring, *i.e.*, improper Moving object segmentation could result in dramatical changes in scene flow estimation and stereo deblurring especially along the object boundaries [64]. Therefore, we could conclude that the scene flow estimation, Moving object segmentation and video deblurring are deeply coupled under our problem setup.

To better exploit the deeply coupling nature of the problem, we propose to formulate our problem as a joint estimation of scene flow, Moving object segmentation and stereo image deblurring for complex dynamic scenes. In particular, we rely on the assumptions that the scene can be well approximated by a collection of 3D planes [65] belonging to a finite number of objects¹ performing rigid motions individually [23]. Therefore, the problem of scene flow estimation can be reformulated as the task of geometric and motion estimation for each 3D plane. The rigid motion is defined for each moving object, which naturally encodes the Moving object segmentation information. The blurred stereo images are generated due to the camera motion, multiple moving objects motion and the 3D scene structure, which are all characterized by the scene flow

¹The background is regarded as a single 'object' due to the camera motion only.

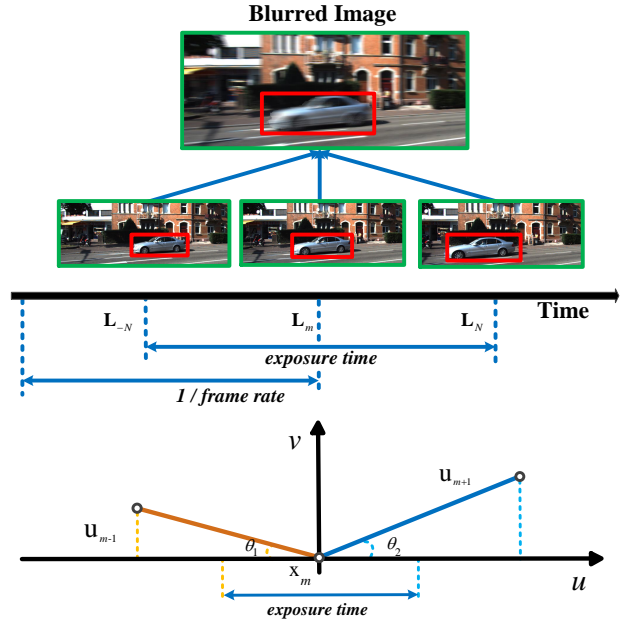


Fig. 5. The pipeline of generating blurred images. We approximate the motion blur kernel as a piece-wise linear function based on bi-direction optical flows and generate blurred images by averaging consecutive frames whose relative motions between two neighbouring frames are known. Notably, ground truth sharp image is chosen to be the middle one.

estimation and the Moving object segmentation. Specifically, our structured blur kernels are expressed with the geometry and motion of each 3D plane.

A. Blurred Image Formation based on the Structured Pixel-wise Blur Kernel

Blurred images are formed by the integration of light intensity emitted from the dynamic scene over the aperture time interval of the camera. We assume that the blurred image \mathbf{B} can be generated by the integral of the latent high frame-rate image sequence $\{\mathbf{L}_n\}$ during the exposure time. This model follows by [32], [3], [66], [67], which supposes the integration of light intensity happens in pixel colour space over the shutter time of the camera.² This defines the blurred image frame in the video sequence as

$$\mathbf{B}_m = \frac{1}{2N+1} \sum_{n=-N}^N \mathbf{L}_n, \quad (1)$$

where \mathbf{B}_m is the m^{th} blurred image in the video sequence, \mathbf{L}_n , $n \in [-N, N]$ denotes latent frames that generate the blurred image. The middle frame \mathbf{L}_m among the latent frames is defined as the deblurred image, which associated with \mathbf{B}_m . This integration model has been widely used in the image/video deblurring literature [12], [32], [22], which has also been used in [37], [36], [38] to generate realistic blurred images from high frame-rate videos. With optical flow, we can transform \mathbf{L}_n with \mathbf{L}_m . Thus, the blur can be modelled

²We notice that several methods model the integration in the raw sensor value and consider the effects of CRFs (camera response function) on motion deblurring. These yield a slightly different solution for deblurring [37], [68].

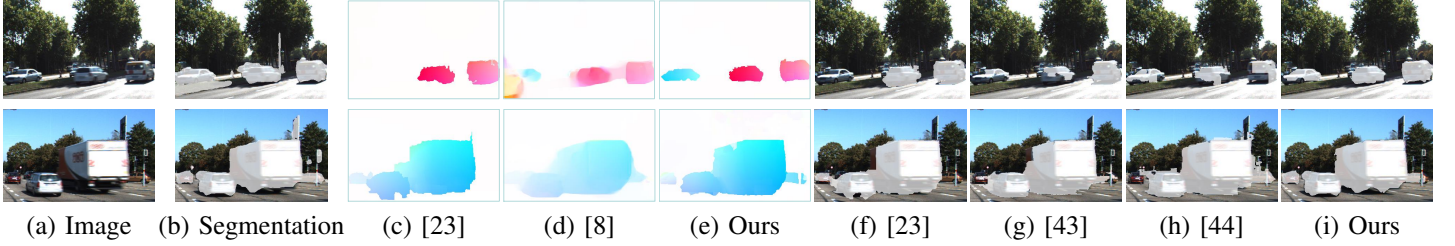


Fig. 6. Scene flow and Moving object segmentation results for an outdoor scenario from **BlurData-1**. (a) Input blurred image. (b) Input semantic segmentation. (c) Estimated flow by [23]. (d) Estimated flow by [8]. (e) Our flow estimation result. (f) Segmentation result by [23]. (g) Segmentation result by [43]. (h) Segmentation result by [44]. (i) Our segmentation result. Compared with both these state-of-the-art methods, our method achieves competitive performance. Best viewed in colour on the screen.

by bi-directional optical flows. We approximate the kernel as piece-wise linear using bidirectional optical flows, where the kernel $\mathbf{A}_m^{\mathbf{x}}$ is spatially varying for each pixel.

$$\mathbf{B}_m(\mathbf{x}) = \text{vec}(\mathbf{A}_m^{\mathbf{x}})^T \text{vec}(\mathbf{L}_m), \quad (2)$$

where $\mathbf{x} \in \mathbb{R}^2$ denotes the pixel location in the image domain, vec denotes the vectorization operator, $\mathbf{A}_m^{\mathbf{x}} \in \mathbb{R}^{h \times w}$ is the blur kernel for each pixel \mathbf{x} , where h , w are the image size. In order to handle multiple types of blurs, we assumed that the blur kernel $\mathbf{A}_m^{\mathbf{x}}$ can be linearized in terms of a motion vector, which can be expressed as [32]:

$$\mathbf{A}_m^{\mathbf{x}}(\tilde{u}, \tilde{v}) = \begin{cases} \frac{\delta(\tilde{u}v_{m+1} - \tilde{v}u_{m+1})}{\tau \|\mathbf{u}_{m+1}\|}, & \text{if } \tilde{\mathbf{u}} \in [0, \tau \mathbf{u}_{m+1}], \\ \frac{\delta(\tilde{u}v_{m-1} - \tilde{v}u_{m-1})}{\tau \|\mathbf{u}_{m-1}\|}, & \text{if } \tilde{\mathbf{u}} \in [0, \tau \mathbf{u}_{m-1}], \\ \mathbf{0}, & \text{otherwise,} \end{cases} \quad (3)$$

where $\tau = \frac{1}{2} \times \text{exposure time} \times \text{frame rate}$, δ denotes the Kronecker delta function, \mathbf{u}_{m+1} and \mathbf{u}_{m-1} are the bidirectional optical flows at frame m . In particular, $\tilde{\mathbf{u}} = (\tilde{u}, \tilde{v})$ which denotes the motion between exposure time, the kernel model is shown in Fig. 5. We obtain the blur kernel matrix $\mathbf{A}_m \in \mathbb{R}^{(h \times w) \times (h \times w)}$ by stacking $\text{vec}(\mathbf{A}_m^{\mathbf{x}})$ over the whole image domain. This leads to the blur model for the image as

$$\text{vec}(\mathbf{B}_m) = \mathbf{A}_m \text{vec}(\mathbf{L}_m). \quad (4)$$

We omit the vectorize symbol in the following sections. We can cast the kernel estimation problem as a motion estimation problem.

In our setup, the stereo video provides the depth information for each frame. Based on our piece-wise planar assumptions on the scene structure, optical flows for pixels lying on the same plane are constrained by a single homography. In particular, we represent the scene in terms of superpixels and finite number of objects with rigid motions. We denote \mathcal{S} and \mathcal{O} as the set of superpixels and moving objects, respectively. Each superpixel $i \in \mathcal{S}$, is associated with a region \mathcal{S}_i in the image, each region is denoted by a plane variable $\mathbf{n}_{i,k_i} \in \mathbb{R}^3$ in 3D ($\mathbf{n}_{i,k_i}^T \mathbf{x} = 1$ for $\mathbf{x} \in \mathbb{R}^3$), where $k_i \in \{1, \dots, |\mathcal{O}|\}$ denotes the i^{th} superpixel associated with the k_i^{th} object. Object inheriting its corresponding motion parameters $\mathbf{o}_{k_i} = (\mathbf{R}_k, \mathbf{t}_k) \in \mathbb{SE}(3)$, where $\mathbf{R}_k \in \mathbb{R}^{3 \times 3}$ is the rotation matrix and $\mathbf{t}_k \in \mathbb{R}^3$ is the translation vector. Note that (\mathbf{n}, \mathbf{o}) encodes the scene flow

information [23], where $\mathbf{n} = \{\mathbf{n}_{i,k_i} | i \in \mathcal{S}\}$ and $\mathbf{o} = \{\mathbf{o}_{k_i} | k_i \in \mathcal{O}\}$. Given the motion parameters \mathbf{o}_{k_i} , we can obtain the homography defined by superpixel i as

$$\mathbf{H}_i = \mathbf{K}(\mathbf{R}_k - \mathbf{t}_k \mathbf{n}_{i,k_i}^T) \mathbf{K}^{-1}, \quad (5)$$

where $\mathbf{K} \in \mathbb{R}^{3 \times 3}$ is the camera calibration matrix. We note that, \mathbf{H}_i relates corresponding pixels across two frames.

The optical flow is then defined as

$$\mathbf{u}_i = \mathbf{x} - \pi(\mathbf{H}_i \mathbf{x}), \quad (6)$$

where we denote $\mathbf{x}^* = \pi(\mathbf{H}_i \mathbf{x})$. $\pi(\cdot)$ is the perspective division such that $\pi([x, y, z]^T) := [x/z, y/z]^T$. This shows that the optical flows for pixels in a superpixel are constrained by the same homography. Thus, it leads to a structured version of blur kernel defined in Eq. (3).

B. Moving object segmentation

Semantic segmentation breaks the image into semantically consistent regions such as road, car, person, sky, *etc.*. Our algorithm computes each region independently based on the semantic class label, resulting in more precise Moving object segmentation and flow estimation, particularly at object boundaries. The provided additional information about object boundaries contributes to avoiding ringing and boundary artifact.

A general problem in motion deblurring is that the moving object boundaries with mixed foreground and background pixels can lead to severe ringing artifacts (see Fig. 1 for details). Most motion deblurring methods address this problem by segmenting blurred images into regions or layers where different kernels are estimated and applied for image restoration [69], [31], [18]. Segmentation on blurred images is difficult due to ambiguous pixels between regions, but it plays an important role in motion deblurring.

In our formulation, we use ResNet38 [16] to predict the semantic label map $\mathbf{M} \in \mathbb{N}^{w \times h}$ as initialization for our ‘‘generalized stereo deblur’’ model. This approach ranks higher on Cityscapes [70] where the image is captured on an urban street. A \mathbf{M} determines the predicted semantic instance label for each pixel in each frame, which provides strong prior for boundary detection, motion estimation, and label classification for superpixels.

We first set roads, sky and trees are static background layer, and assume other things have a higher moving possibility to

be the foreground layer. Here, a convincing background layer will provide the inline feature points on the background for ego-motion estimation. Then, we can estimate the disparity map and the 6-DOF camera motion using stereo matching and visual odometry with coarse background segmentation. We identify regions inconsistent with the estimated camera motion and estimate the motion at these regions separately. Each motion parameter \mathbf{o} is generated by moving clusters from sparse features points. In particular, the motion hypothesis is then generated using the 3-point RANSAC algorithm implemented in [71]. These inconsistent regions can match with our prior \mathbf{M} . This helps to maintain the boundary information for moving objects and avoid ringing artifacts (see Fig. 4 for details).

Each slanted plane in the image is labelled as moving or static according to the ego-motion estimation. With the semantic segmentation masks, we can give each superpixel an additional label, foreground or background. We then use the label map to initialize object label k_i for each superpixel i . If most pixels' semantic label in i^{th} superpixel are fore/background, the superpixel is more likely to belongs to the fore/background.

$$k_i(\mathbf{x}) \in \begin{cases} \{1\} & , \text{if } \mathbf{M}(\mathbf{x}) = \text{Background} \\ \{2, \dots, |\mathcal{O}|\} & , \text{if } \mathbf{M}(\mathbf{x}) = \text{Foreground}. \end{cases} \quad (7)$$

Although we provide over segmentation initially as shown in Fig. 1(a), our algorithm can precisely segment the moving objects after optimization (Fig. 1(b)) and provide more accurate motion boundaries information for optical flow estimation (Fig. 1(d)), and thereby facilitates stereo video deblurring (Fig. 1(h)).

With the semantic segmentation prior, we label each superpixel and objects more accurately, our approach obtains superior results in Moving object segmentation and scene flow estimation (see Fig. 6 for details).

In the optimization part, instead of giving sample k_i for every superpixel randomly, we use the semantic segmentation prior \mathbf{M} to give a more reliable sample for each superpixel (see Section IV-A for detail).

C. Energy Minimization

We formulate the problem in a single framework as a discrete-continuous optimization problem to jointly estimate the scene flow, Moving object segmentation and deblur the stereo images. Specifically, our model is defined as

$$\mathbf{E}(\mathbf{n}, \mathbf{o}, \mathbf{L}) = \underbrace{\sum_{i \in \mathcal{S}} \phi_i(\mathbf{n}_i, \mathbf{o}, \mathbf{L})}_{\text{data term}} + \underbrace{\sum_{i, j \in \mathcal{S}} \phi_{i, j}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{o})}_{\text{scene flow smoothness term}} + \underbrace{\sum_m \psi_m(\mathbf{L})}_m}_{\text{latent image regularisation}} \quad (8)$$

where i, j denotes the set of adjacent superpixels in \mathcal{S} . The function consists of a data term, a smoothness term for scene flow, and a spatial regularization term for latent images. Our model is initially defined on three consecutive pairs of stereo video sequences. It can also allow the input with two pairs of frames. Details are provided in Section V. The energy terms

are discussed in Section III-D, Section III-E, and Section III-F, respectively.

In Section IV, we perform the optimization in an alternative manner to handle mixed discrete and continuous variables, thus allowing us to jointly estimate scene flow, Moving object segmentation and deblur the images.

D. Data Term

Our data term involves mixed discrete and continuous variables, and are of three different kinds. The first kind encodes the fact that the corresponding pixels across the six latent images should have a similar appearance, *i.e.*, brightness constancy. This lets us write the term as

$$\phi_i^1(\mathbf{n}_i, \mathbf{o}, \mathbf{L}) = \theta_1 \sum_{\mathbf{x} \in \mathcal{S}_i} |\mathbf{L}(\mathbf{x}) - \mathbf{L}^*(\mathbf{x}^*)|_1, \quad (9)$$

where \mathbf{L} denotes the reference image, \mathbf{L}^* denotes the target image, the superscript $* \in \{\text{stereo}, \text{flow}_{f,b}, \text{cross}_{f,b}\}$ denote the warping direction to other images and $(\cdot)_{f,b}$ denotes the forward and backward direction, respectively (see Figure 7). The terms is defined by summing the matching costs of all pixels inside superpixel i . We adopt the robust ℓ_1 norm to enforce its robustness against noise and occlusions.

Our second potential, similar to one term used in [23], is defined as

$$\phi_i^2(\mathbf{n}_i, \mathbf{o}) = \begin{cases} \theta_2 \sum_{\mathbf{x} \in \mathcal{S}_i} \rho_{\alpha_1}(\|\mathbf{x} - \mathbf{x}^*\|_2) & , \text{if } \mathbf{x} \in \Pi_{\mathbf{x}}, \\ 0 & , \text{otherwise,} \end{cases} \quad (10)$$

where $\rho_{\alpha}(\cdot) = \min(|\cdot|, \alpha)$ denotes the truncated l_1 penalty function. More specifically, it encodes the information that the warping of feature points $x \in \Pi_x$ based on \mathbf{H}^* should match its extracted correspondences \mathbf{x}^* in the target view. In particular, Π_x is obtained in a similar manner as [23].

The third data term, making use of the observed blurred images, is defined as

$$\phi_i^3(\mathbf{n}_i, \mathbf{o}, \mathbf{L}) = \theta_3 \sum_m \sum_{\partial} \|\partial \mathbf{A}_m(\mathbf{n}_i, \mathbf{o}) \mathbf{L}_m - \partial \mathbf{B}_m\|_2^2, \quad (11)$$

where ∂ denotes the Toeplitz matrices corresponding to the horizontal and vertical derivative filters. This term encourages the intensity changes in the estimated blurred image to be close to that of the observed blurred image.

E. Smoothness Term for Scene Flow

Our energy model exploits a smoothness potential that involves discrete and continuous variables. It is similar to the ones used in [23]. In particular, our smoothness term includes three different types.

The first one is to encode the compatibility of two superpixels that share a common boundary by respecting the depth discontinuities. We define our potential function as

$$\phi_{i, j}^1(\mathbf{n}_i, \mathbf{n}_j) = \theta_4 \sum_{\mathbf{x} \in \mathcal{B}_{i, j}} \rho_{\alpha_2}(\omega_{i, j}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{x})), \quad (12)$$

where $d(\mathbf{n}_i, \mathbf{x})$ is the disparity of pixel \mathbf{x} in superpixel i in the reference disparity map, $\omega_{i, j}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{x}) = d(\mathbf{n}_i, \mathbf{x}) - d(\mathbf{n}_j, \mathbf{x})$

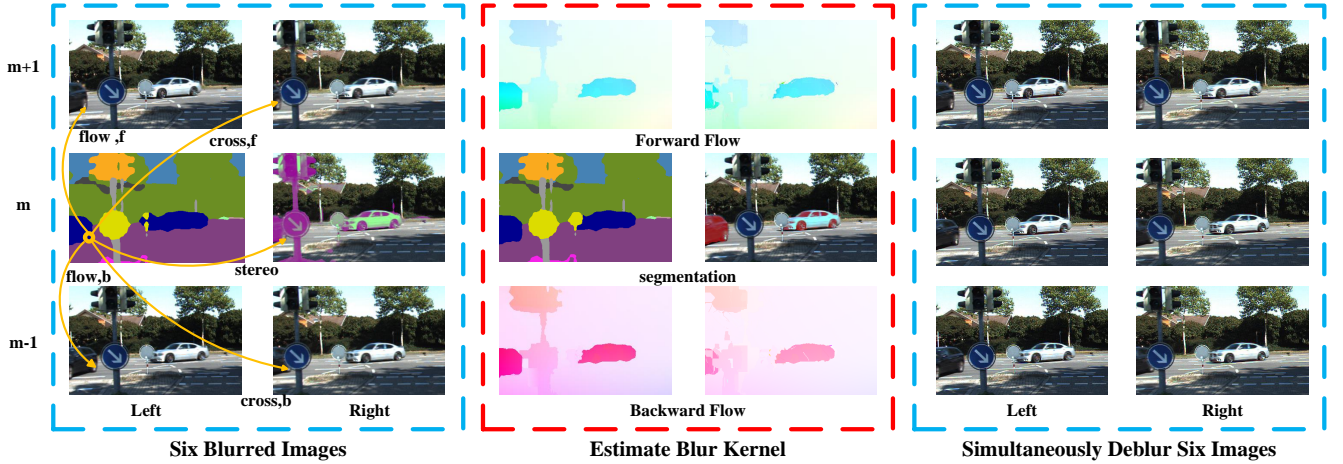


Fig. 7. Illustration of our ‘generalized stereo deblurring’ method. We simultaneously compute four scene flows (in two directions and in two views), Moving object segmentation and deblur six images. In case the input contains only two images, we use the reflection of the flow forward as the flow backward in the deblurring part.

are the dissimilarity value of disparity for pixel $\mathbf{x} \in \mathcal{B}_{i,j}$ on the boundary.

The second potential is to encourage the neighbouring superpixels to orient in similar directions. It is expressed as

$$\phi_{i,j}^2(\mathbf{n}_i, \mathbf{n}_j) = \theta_5 \rho_{\alpha_3} \left(1 - \frac{|\mathbf{n}_i^T \mathbf{n}_j|}{\|\mathbf{n}_i\| \|\mathbf{n}_j\|} \right). \quad (13)$$

The shadows of moving objects have motion boundaries but no disparity discontinuities. However, the motion boundaries are co-aligned with disparity discontinuities in general. Thus, we use the third and fourth potential encodes these discontinuities. This potential can be expressed as

$$\begin{aligned} \phi_{i,j}^3(\mathbf{n}_{i,k_i}, \mathbf{n}_{j,k_j}) = \\ \exp \left\{ -\frac{\lambda}{|\mathcal{B}_{i,j}|} \sum_{\mathbf{x} \in \mathcal{B}_{i,j}} \omega_{i,j}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{x})^2 \right\} \times \frac{|\mathbf{n}_i^T \mathbf{n}_j|}{\|\mathbf{n}_i\| \|\mathbf{n}_j\|} \times [k_i \neq k_j], \end{aligned} \quad (14)$$

where $|\mathcal{B}_{i,j}|$ denotes the number of pixels shared along boundary between superpixels i and j .

$$\begin{aligned} \phi_{i,j}^4(\mathbf{n}_{i,k_i}, \mathbf{n}_{j,k_j}, \mathbf{o}_{k_i}, \mathbf{o}_{k_j}) = \\ \exp \left\{ -\frac{\lambda}{|\mathcal{B}_{i,j}|} \sum_{\mathbf{x} \in \mathcal{B}_{i,j}} G(\mathbf{o}_{k_i}, \mathbf{o}_{k_j}) \right\} \times \frac{|\mathbf{n}_{i,k_i}^T \mathbf{n}_{j,k_j}|}{\|\mathbf{n}_{i,k_i}\| \|\mathbf{n}_{j,k_j}\|} \times [k_i \neq k_j], \end{aligned} \quad (15)$$

$$G(\mathbf{o}_{k_i}, \mathbf{o}_{k_j}) = \theta_r (\text{trace}(\mathbf{R}_{k_i}^T \mathbf{R}_{k_j}) - 1) / 2 + \theta_t (\exp(-\|\mathbf{t}_{k_i} - \mathbf{t}_{k_j}\|)),$$

where $[\cdot]$ denotes the Iverson bracket. This encodes our belief that motion boundaries are more likely to occur at 3D folds or discontinuities than within smooth surfaces.

F. Regularization Term for Latent Images

Several works [72], [15] have studied the importance of spatial regularization in image deblurring. In our model, we

use a total variation term to suppress the noise in the latent image while preserving edges, and penalize spatial fluctuations. Therefore, our potential takes the form

$$\psi_m = \sum_{\mathbf{x}} |\nabla \mathbf{L}_m|. \quad (16)$$

Note that the total variation is applied to each colour channel separately.

IV. SOLUTION

The optimization of our energy function defined in Eq.(8), involving discrete and continuous variables, is very challenging to solve. Recall that our model involves two set of variables, namely scene flow variables and latent clean images. Fortunately, given one set of variables, we can solve the other efficiently. Therefore, we perform the optimization iteratively by the following steps,

- Fix latent clean image \mathbf{L} , solve scene flow by optimizing Eq.(17) (See Section IV-A).
- Fix scene flow parameters, \mathbf{n} and \mathbf{o} , solve latent clean images by optimizing Eq.(18) (See Section IV-B).

In the following sections, we describe the details for each optimization step.

A. Scene flow estimation

We fix latent images, namely $\mathbf{L} = \tilde{\mathbf{L}}$. Eq.(8) reduces to

$$\min_{\mathbf{n}, \mathbf{o}} \sum_{i \in \mathcal{S}} \sum_{m=1}^3 \phi_i^m(\mathbf{n}_i, \mathbf{o}, \tilde{\mathbf{L}}_i) + \sum_{i,j \in \mathcal{S}} \sum_{m=1}^4 \phi_{i,j}^m(\mathbf{n}_i, \mathbf{n}_j, \mathbf{o}), \quad (17)$$

which becomes a discrete-continuous CRF optimization problem.

We use the sequential tree-reweighted message passing (TRW-S) method in [23] to find an approximate solution. Since the label k of \mathbf{n}_i of each superpixel is drawing randomly, we use the semantic segmentation prior \mathbf{M} to give a more reliable sample of each superpixel. We modify their sampling strategy as shown in Algorithm 1.

Algorithm 1: TRW-S Optimization

Input : $\tilde{\mathbf{L}}, \mathbf{M}, \mathbf{B}$.

- 1 Initialize \mathbf{n} and \mathbf{o} as described in ‘Initialization’.
- 2 Iteration times = 3
- 3 For all $i \in S$
- 4 Draw sample for \mathbf{n}_i (Gaussian)
- 5 Draw sample for $\mathbf{k}_i(\mathbf{M})$
- 6 For all $k \in \mathcal{O}$
- 7 Draw sample for \mathbf{o}_k (MCMC)
- 8 Run TRW-S [73] on discretized problem

Output: $\mathbf{n}_{i,k_i}, \mathbf{o}_{k_i}$

B. Deblurring

Given the scene flow parameters, namely $\mathbf{n} = \tilde{\mathbf{n}}$, and $\mathbf{o} = \tilde{\mathbf{o}}$, the blur kernel matrix, \mathbf{A}_m is derived based on Eq.(3), and Eq.(6). The objective function in Eq. (8) becomes convex with respect to \mathbf{L} and is expressed as

$$\min_{\mathbf{L}} \sum_{S_i \in S} \phi_i^1(\tilde{\mathbf{n}}_i, \tilde{\mathbf{o}}, \mathbf{L}) + \phi_i^3(\tilde{\mathbf{n}}_i, \tilde{\mathbf{o}}, \mathbf{L}) + \psi_m(\mathbf{L}). \quad (18)$$

In order to obtain sharp image \mathbf{L} , we adopt the conventional convex optimization method [74] and derive the primal-dual updating scheme as follows

$$\begin{cases} \mathbf{p}^{r+1} = \frac{\mathbf{p}^r + \gamma \nabla \mathbf{L}_m^r}{\max(1, \text{abs}(\mathbf{p}^r + \gamma \nabla \mathbf{L}_m^r))} \\ \mathbf{q}^{r+1} = \frac{\mathbf{q}^r + \gamma \theta_1 (\mathbf{L}_m^r - \mathbf{L}_*^r)}{\max(1, \text{abs}(\mathbf{q}^r + \gamma \theta_1 (\mathbf{L}_m^r - \mathbf{L}_*^r))} \\ \mathbf{L}_m^{r+1} = \arg \min_{\mathbf{L}_m} \sum_i \theta_3 \sum_{\partial} \|\partial \mathbf{A}_m \mathbf{L}_m - \partial \mathbf{B}_m\|_2^2 + \\ \frac{\|[\mathbf{L}_m - \eta((\nabla \mathbf{p}_m^{r+1})^T + \theta_1 (\mathbf{q}^{r+1} - \mathbf{q}_*^{r+1})^T)] - \mathbf{L}_m^r\|_2^2}{2\eta}, \end{cases} \quad (19)$$

where $\mathbf{p}_m, \mathbf{q}_m, *$ are the dual variables, γ and η are the step variants which can be modified at each iteration, and r is the iteration number.

Algorithm 2: Proposed deblurring system

Input : Stereo Blurred Image Sequences \mathbf{B} , Semantic Segmentation of Reference Image Pair.

- 1 Initialize \mathbf{n} and \mathbf{o} as described in ‘Initialization’.
- 2 Run Algorithm 1 minimize Eq. (17). Estimate scene flow and moving object segmentation map.
- 3 Run Primal-Dual [74] minimize Eq. (18). Restoration clean image.
- 4 Repeat steps 2,3 until reaches a preset iteration number (3 in our experiment).

Output: Latent Images \mathbf{L} , Moving object Segmentation Map, Scene Flow

V. EXPERIMENTS

To demonstrate the effectiveness of our method, we evaluate it based on two datasets: the synthetic chair sequence [9] and the KITTI dataset [19]. We report our results on both datasets in the following sections.

TABLE I
QUANTITATIVE COMPARISONS ON DISPARITY, OPTICAL FLOW AND DEBLURRING RESULTS ON THE KITTI DATASET (BLURDATA-1).

KITTI Dataset	Disparity		Flow		PSNR	
	m	m+1	Left	Right	Left	Right
Vogel <i>et al.</i> [11]	8.20	8.50	13.62	14.59	/	/
Kim and Lee [8]	/	/	38.89	39.45	28.25	29.00
Sellent <i>et al.</i> [9]	8.20	8.50	13.62	14.59	27.75	28.52
Kupyn <i>et al.</i> [40]	/	/	/	/	28.34	28.73
Tao <i>et al.</i> [41]	/	/	/	/	29.55	29.95
Pan <i>et al.</i> [10]	6.82	8.36	10.01	11.45	29.80	30.30
Ours	6.18	7.49	9.83	11.14	29.85	30.50
Baseline						
[11] and [8]	/	/	22.42	/	28.11	/

A. Experimental Setup

Initialization. Our model in Section III is formulated on three consecutive stereo pairs. In particular, we treat the middle frame in the left view as the reference frame. We adopt StereoSLIC [65] to generate superpixels. Given the stereo images, we apply the approach in [71] to obtain sparse feature correspondences. The traditional SGM [75] method is applied to obtain the disparity map. We further leverage the semantic segmentation results to provide priors for motion segmentation. In particular, we applied the pre-trained model from the high-accuracy method [16] on our blurred image. Based on the obtained semantics, we generate a binary map \mathbf{M} which indicates the foreground as 1 and background as 0 by grouping the estimated semantics (see Section III-B for details.) The motion hypotheses are first generated using RANSAC algorithm implemented in [71]. Regarding the model parameters, we perform grid search on 30 reserved images. In our experiments, we fix the model parameters as $\theta_1 = 0.7$, $\theta_2 = 5.5$, $\theta_3 = 0.7$, $\gamma = 250$, $\theta_4 = 0.37$, $\theta_5 = 17$, $\lambda = 0.13$, $\alpha_1 = 3.39$, $\alpha_2 = 2.5$, $\alpha_3 = 0.25$, $\theta_r = 0.05$, $\theta_t = 0.1$.

Evaluation metrics. Since our method estimates the scene flow, segments moving objects and deblurs images, we thus evaluate multiple tasks separately. As for the scene flow estimation results, we evaluate both the optical flow and disparity map by the same error metric, which is by counting the number of pixels having errors more than 3 pixels and 5% of its ground-truth. We adopt the PSNR to evaluate the deblurred image sequences for left and right view separately. We report precision (P), recall (R) and F-measure (F) for our motion segmentation results. Those metrics are defined as:

$$P = \frac{t_p}{t_p + f_p}, \quad R = \frac{t_p}{t_p + f_n}, \quad F = \frac{2R * P}{R + P}, \quad (20)$$

where the true positive t_p represents the number of pixels that have been correctly detected as moving objects; false positive f_p are defined as pixels that have been mis-detected as moving pixels; false negative f_n are denoted as moving pixels that have not been detected correctly. Thus, for each sequence, we report disparity errors for three stereo image pairs, flow errors in forward and backward directions, and PSNR values for six images, and precision, recall and F-measure for the Moving object segmentation results.

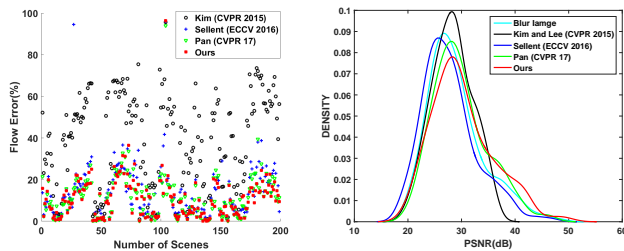


Fig. 8. Left: The flow estimation errors for 199 scenes in the KITTI dataset. Our method clearly outperforms the monocular and stereo video deblurring methods. Right: The distribution of the PSNR scores for 199 scenes in the KITTI dataset(BlurData-1). The probability distribution function for each PSNR was estimated using kernel density estimation with a normal kernel function. The heavy tail of our method means larger PSNR can be achieved using our method.

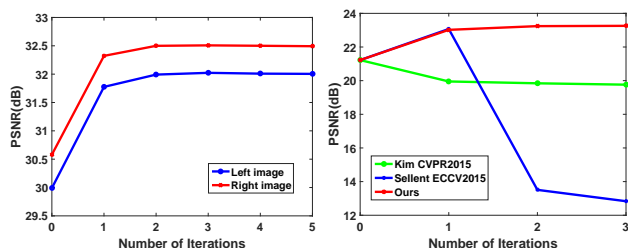


Fig. 9. The deblurring performance of our approach with respect to the number of iterations. (left) Our method on our dataset with the gap of 0.3 dB between the first and the last iteration. (right) Several baselines on 'Chair'.

Baselines. We first compare our scene flow results with piecewise rigid scene flow method (PRSF) [11], whose performance ranks as one of the top 3 approaches on KITTI optical flow benchmark [19]. We then compare our results with the state-of-the-art stereo deblurring approach [9], monocular deblurring approach [32] and deep-learning-based deblurring approaches [41], [40]. We compare our moving object segmentation results with the state-of-the-art approach using sharp stereo video sequences [76]. In addition, we further choose NLC [44] and FST [43] as baselines since they are more robust to occlusions, motion blur and illumination changes according to the comprehensive evaluations in [77]. We make the quantitative comparison of our model w/o explicitly imposing semantics priors for our flow and deblurring results in Fig 8. In addition, we compare with our previous method (Pan *et al.* CVPR 17) that has no semantics priors. The comparison clearly shows that the performance is improved significantly with the introduction of semantics as priors.

Runtime: In all experiments, we simultaneously compute two directions, namely forward and backward, scene flows, restore six blurred images and segment all moving objects. Our MATLAB implementation with C++ wrappers requires a total runtime of 35 minutes for processing one scene (6 images, 3 iterations) on a single i7 core running at 3.6 GHz.

B. Results on KITTI

To the best of our knowledge, currently, there are no realistic benchmark datasets that provide blurred images and corresponding ground-truth deblurring and scene flow. We take advantage of the KITTI dataset [19] to create a synthetic



Fig. 10. The moving object segmentation result with respect to the number of iterations

TABLE II
MOVING OBJECT SEGMENTATION EVALUATION ON THE KITTI DATASET
BLURDATA-1.

Methods	Recall(R)	Precision (P)	F-measure (F)
Menze <i>et al.</i> [23]	0.7995	0.5841	0.6045
Zhou <i>et al.</i> [76]	0.7641	0.6959	0.7284
Papazoglou <i>et al.</i> [43]	0.5945	0.3199	0.2938
Faktor <i>et al.</i> [44]	0.4761	0.3148	0.3339
Baseline	0.7633	0.6113	0.6789
Our	0.8520	0.7281	0.7426

blurry image dataset on realistic scenery, which contains 199 challenging outdoor sequences. Each sequence includes 6 images (375×1242). Our blurry image dataset is generated in two different ways. First, we follow the general practice in image deblurring and generate the blurry image dataset, referred to as **BlurData-1**, using the piecewise linear 2D kernel in Eq. 3 which is defined on the dense scene flow. We use method [23] to generate dense ground-truth flows. In addition, $\tau = 0.23$ and the number of frame is $N = 20$ (see Fig. 5 for details).

Second, we follow the way of generating blurry image in [39], by averaging the reference image together with its neighbouring frames. In particular, we average 7 frames in total (3 on either side of the reference frame). Note that the image sequence in KITTI, in general, has large relative motion. We therefore only choose 10 sequences to generate blurry images based on averaging, which is denoted as **BlurData-2**. In the following, we report results on our generated two synthetic datasets, respectively.

Deblurring and Scene Flow Results. We evaluated our approach by averaging errors and PSNR scores over m and $m+1$ stereo image pairs. Table I shows the PSNR values, disparity errors, and flow errors averaged over 199 test sequences on **BlurData-1**. Note that our method consistently outperforms all baselines. We achieve the minimum error scores of 9.83% for optical flow and 6.18% for the disparity in the reference view. Figure 8 and Figure 8 show the estimated flows and deblurring results of the KITTI stereo flow benchmark, which includes 199 scenes. Figure 9 (left) shows the performance of our deblurring stage with respect to the number of iterations. While we use 5 iterations for all our experiments, our experiments indicate that only 3 iterations are sufficient in most cases to reach an optimal performance under our model. In Figure 11, we show qualitative results.

Moving Object Segmentation Results. We report the quantitative comparison of our results with the baselines in Table II. It shows that our approach significantly outperforms the baselines by a large margin. Fig. 11(g-k) show the qualitative comparison of our approach with baselines. The results show



Fig. 11. Qualitative comparison of our approach with baselines for deblurring, Moving object segmentation, and flow estimations. Our method use (a) blurred image and (g) Initial semantic prior from **BlurData-1** as input. (b) Ground-truth latent image. (c) Deblurring results by Kim and Lee [8]. (d) Stereo deblurring results by Sellent *et al.* [9]. (e) and (f) show our deblurring results w/o imposing semantic priors, respectively; (h) Segmentation result by [43]. (i) Segmentation result by [44]. (j) Our segmentation result. (k) and (l) show the optical flow estimation results w/o imposing semantic priors. Best viewed in colour on the screen.

that our final segmentation follows the boundary of the moving objects very well. It further demonstrates that our approach can segment the moving objects more accurately than other approaches. Therefore, we can achieve a conclusion that joint scene flow estimation, deblurring, and moving object segmentation benefit each task.

C. Results on Other Dataset

In order to evaluate the generalization ability of our approach on different images, we use the datasets based on the 3D kernel model and average kernel model which is different from our **Blurred image dataset**. In order to compare our performance on images blurred by the 3D kernel model, we also use the data courtesy of Sellent [9]. Those sequences contain four real and four synthetic scenes and each of them have six blurred images with its sharp images. The synthetic

sequences are blurred by the 3D kernel model and have ground-truth for those sequences. Figure 9 (right) shows the performance of several baselines on synthetic dataset. This plot affirms our assumption that jointly and simultaneously solving scene flow and video deblur benefit each other. It also shows that a simple combination of two stages cannot achieve the targeted results. For real scenes, they use real images captured with a stereo camera which moves forward very slowly and attached to a motorized rail. By averaging the frames, they obtain motion blurred images where all objects in the scene are static and the camera moves toward the scene. For these reasons, we give the semantic segmentation map as all background (see Figure 12 1st and 2nd rows show the performance of the result of the real scene).

In Fig. 12(the 3rd and 4th rows.), we show qualitative



Fig. 12. Sample deblur results on the real image dataset from Sellent *et al.* [9] in 1st and 2nd row, and average model dataset in 3rd row. It shows that our 'generalized stereo deblur' model can tackle different kinds of motion blur model and get better results. Best viewed in colour on the screen.

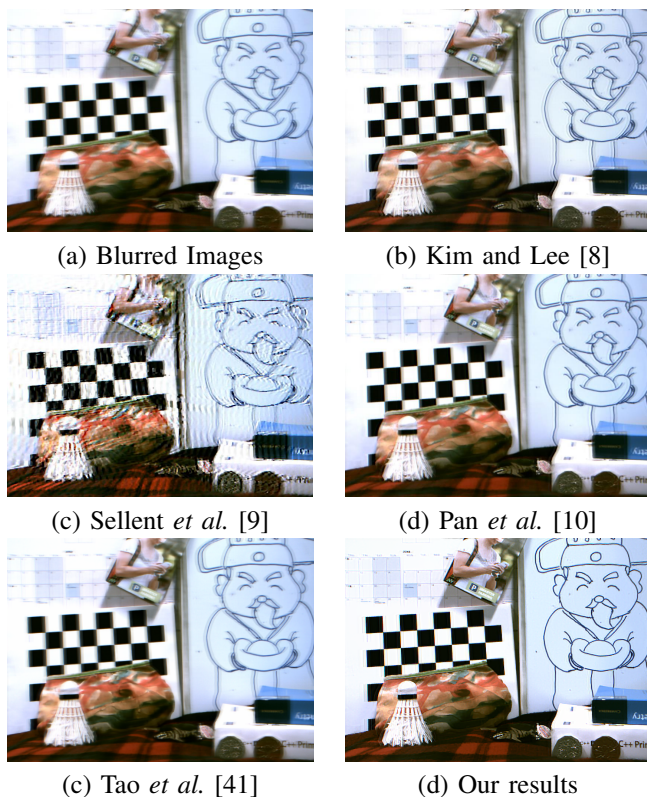


Fig. 13. Deblurring results on our Blur dataset. (a) The blurred image. (b) Deblurring results by Kim and Lee [8]. (c) Stereo deblurring results by Sellent *et al.* [9]. (d) Deblurring results by Pan *et al.* [10]. (e) Deblurring results by Tao *et al.* [41]. (f) Our result. It shows that our 'generalized stereo deblur' model can get competitive result compared with the state-of-the-art deblurring methods results. Best viewed in colour on the screen.

results of our method and other methods on sample sequences from this two datasets, where our method again achieves the best performance.

D. Limitations

Our method is based on calibrated stereo cameras which seem sometimes not convenient for routine application. The framework may fail in the texture-less case, the scene with strong reflection or under low lighting conditions. The occlusion will also reduce the accuracy of the segmentation boundaries. Our model cannot tackle defocus blur and scenery with transparency or translucency. Following the recent deblurring works such as [32], [3], [67], [66], we make the similar assumption that the intensity integral happens in colour space during the exposure time, while we are aware of several methods model the integration in the raw sensor value and consider the effects of CRFs on motion deblurring [37], [68]. We leave these limitations as future works.

VI. CONCLUSION

In this paper, we present a joint optimization framework to tackle the challenging task of stereo video deblurring where scene flow estimation, Moving object segmentation and video deblurring are solved in a coupled manner. Under our formulation, the motion cues from scene flow estimation and blur information could reinforce each other, and produce superior results than conventional scene flow estimation or stereo deblurring methods. We have demonstrated the benefits of our framework on extensive synthetic and real stereo sequences. In future, we plan to extend our method to deal with multiple frames to achieve better stereo deblurring.

ACKNOWLEDGEMENTS

This research was supported in part by Australia Centre for Robotic Vision, the Natural Science Foundation of China grants (61871325, 61420106007, 61671387, 61603303) and the Australian Research Council (ARC) grants (DE140100180, DE180100628, DP150104645).

REFERENCES

- [1] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf, "Blind correction of optical aberrations," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2012, pp. 187–200.
- [2] J. Shi, L. Xu, and J. Jia, "Just noticeable defocus blur detection and estimation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 657–665.
- [3] A. Gupta, N. Joshi, C. L. Zitnick, M. Cohen, and B. Curless, "Single image deblurring using motion density functions," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2010, pp. 171–184.
- [4] J. Jia, "Mathematical models and practical solvers for uniform motion deblurring," *Motion Deblurring: Algorithms and Systems*, p. 1, 2014.
- [5] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 769–777.
- [6] U. Franke and A. Joos, "Real-time stereo vision for urban traffic scene understanding," in *IEEE Intelligent Vehicles Symposium*, 2000.
- [7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012, pp. 3354–3361.
- [8] T. Hyun Kim and K. Mu Lee, "Generalized video deblurring for dynamic scenes," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 5426–5434.
- [9] A. Sellent, C. Rother, and S. Roth, "Stereo video deblurring," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2016, pp. 558–575.
- [10] L. Pan, Y. Dai, M. Liu, and F. Porikli, "Simultaneous stereo video deblurring and scene flow estimation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, July 2017.
- [11] C. Vogel, K. Schindler, and S. Roth, "3d scene flow estimation with a piecewise rigid scene model," *Int. J. Comp. Vis.*, vol. 115, no. 1, pp. 1–28, 2015.
- [12] H. Seok Lee and K. Mu Lee, "Dense 3d reconstruction from severely blurred images using a single moving camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 273–280.
- [13] Z. Hu, L. Xu, and M.-H. Yang, "Joint depth estimation and camera shake removal from single blurry image," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014, pp. 2893–2900.
- [14] L. Xu, S. Zheng, and J. Jia, "Unnatural l0 sparse representation for natural image deblurring," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 1107–1114.
- [15] D. Krishnan, T. Tay, and R. Fergus, "Blind deconvolution using a normalized sparsity measure," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2011, pp. 233–240.
- [16] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognition*, vol. 90, pp. 119–133, 2019.
- [17] Y. Zhou and N. Komodakis, "A map-estimation framework for blind deblurring using high-level edge priors," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2014, pp. 142–157.
- [18] J. Pan, Z. Hu, Z. Su, H.-Y. Lee, and M.-H. Yang, "Soft-segmentation guided object motion deblurring," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 459–468.
- [19] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, p. 0278364913491297, 2013.
- [20] L. Li, J. Pan, W.-S. Lai, C. Gao, N. Sang, and M.-H. Yang, "Learning a discriminative prior for blind image deblurring," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2018.
- [21] W. Ren, J. Pan, X. Cao, and M.-H. Yang, "Video deblurring via semantic segmentation and pixel-wise non-linear kernel," in *Proc. IEEE Int. Conf. Comp. Vis.*, Oct 2017.
- [22] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. v. d. Hengel, and Q. Shi, "From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.
- [23] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 3061–3070.
- [24] D. Perrone and P. Favaro, "Total variation blind deconvolution: The devil is in the details," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014, pp. 2909–2916.
- [25] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," in *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3. ACM, 2006, pp. 787–794.
- [26] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring text images via l0-regularized intensity and gradient prior," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014, pp. 2901–2908.
- [27] L. Sun, S. Cho, J. Wang, and J. Hays, "Edge-based blur kernel estimation using patch priors," in *IEEE International Conference on Computational Photography*. IEEE, 2013, pp. 1–8.
- [28] W.-S. Lai, J.-J. Ding, Y.-Y. Lin, and Y.-Y. Chuang, "Blur kernel estimation using normalized color-line prior," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 64–72.
- [29] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, "Blind image deblurring using dark channel prior," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 1628–1636.
- [30] Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao, "Image deblurring via extreme channels prior," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.
- [31] J. Wulff and M. J. Black, "Modeling blurred video with layers," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2014, pp. 236–252.
- [32] T. Hyun Kim and K. M. Lee, "Segmentation-free dynamic scene deblurring," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014, pp. 2766–2773.
- [33] H. Park and K. Mu Lee, "Joint estimation of camera pose, depth, deblurring, and super-resolution from a blurred image sequence," in *Proc. IEEE Int. Conf. Comp. Vis.*, Oct 2017.
- [34] S. Nayar and M. Ben-Ezra, "Motion-based motion deblurring," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 689–698, 2004.
- [35] L. Xu and J. Jia, "Depth-aware motion deblurring," in *IEEE International Conference on Computational Photography*, 2012, pp. 1–8.
- [36] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, July 2017.
- [37] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, July 2017.
- [38] T. Hyun Kim, K. Mu Lee, B. Scholkopf, and M. Hirsch, "Online video deblurring via dynamic temporal blending network," in *Proc. IEEE Int. Conf. Comp. Vis.*, Oct 2017.
- [39] T. H. Kim, S. Nah, and K. M. Lee, "Dynamic video deblurring using a locally adaptive blur model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2374–2387, Oct 2018.
- [40] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2018.
- [41] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2018.
- [42] M. Jin, G. Meishvili, and P. Favaro, "Learning to extract a video sequence from a single motion-blurred image," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2018.
- [43] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 1777–1784.
- [44] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting," in *Proc. Brit. Mach. Vis. Conf.*, vol. 2, no. 7, 2014, p. 8.
- [45] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 3395–3402.
- [46] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan 2018.
- [47] N. Sundaram, T. Brox, and K. Keutzer, "Dense point trajectories by gpu-accelerated large displacement optical flow," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2010, pp. 438–451.
- [48] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 447–456.
- [49] N. Shankar Nagaraja, F. R. Schmidt, and T. Brox, "Video segmentation with just a few strokes," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 3235–3243.

- [50] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 3899–3908.
- [51] W.-D. Jang and C.-S. Kim, "Online video object segmentation via convolutional trident network," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 5849–5858.
- [52] T. Wang, B. Han, and J. Collomosse, "Touchcut: Fast image and video segmentation using single-touch interaction," *Computer Vision and Image Understanding*, vol. 120, pp. 14–30, 2014.
- [53] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, and B. Chen, "Jumpcut: non-successive mask transfer and interpolation for video cutout," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 195–1, 2015.
- [54] Y. Yan, C. Xu, D. Cai, and J. J. Corso, "Weakly supervised actor-action segmentation via robust multi-task ranking," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 1298–1307.
- [55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 770–778.
- [56] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, July 2017.
- [57] Y.-H. Tsai, G. Zhong, and M.-H. Yang, "Semantic co-segmentation in videos," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2016, pp. 760–775.
- [58] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Deep learning markov random field for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1814–1828, 2018.
- [59] T. Tani, S. N. Sinha, and Y. Sato, "Fast multi-frame stereo scene flow with motion segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, July 2017.
- [60] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2018.
- [61] 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 Conf. Comp. Vis. Patt. Recogn.*, June 2018.
- [62] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 2758–2766.
- [63] 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. Comp. Vis. Patt. Recogn.*, July 2017.
- [64] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black, "Optical flow with semantic segmentation and localized layers," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2016.
- [65] K. Yamaguchi, D. McAllester, and R. Urtasun, "Robust monocular epipolar flow estimation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 1862–1869.
- [66] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, "Non-uniform deblurring for shaken images," *Int. J. Comp. Vis.*, vol. 98, no. 2, pp. 168–186, 2012.
- [67] S. Dai and Y. Wu, "Motion from blur," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* IEEE, 2008, pp. 1–8.
- [68] Y.-W. Tai, X. Chen, S. Kim, S. J. Kim, F. Li, J. Yang, J. Yu, Y. Matsushita, and M. S. Brown, "Nonlinear camera response functions and image deblurring: Theoretical analysis and practice," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2498–2512, 2013.
- [69] Y.-W. Tai, H. Du, M. S. Brown, and S. Lin, "Correction of spatially varying image and video motion blur using a hybrid camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1012–1028, 2010.
- [70] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 3213–3223.
- [71] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *IEEE Intelligent Vehicles Symposium (IV)*, 2011, pp. 963–968.
- [72] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-laplacian priors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1033–1041.
- [73] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568–1583, 2006.
- [74] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.
- [75] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, 2008.
- [76] D. Zhou, V. Frémont, B. Quost, Y. Dai, and H. Li, "Moving object detection and segmentation in urban environments from a moving platform," *Image and Vision Computing*, 2017.
- [77] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

Liyuan Pan is currently pursuing the Ph.D. degree in the College of Engineering and Computer Science, Australian National University, Canberra, Australia. She received her B.E degree from Northwestern Polytechnical University, Xian, China in 2014. Her interests include deblurring, flow estimation, depth completion, and event camera.

Yuchao Dai is currently a Professor with School of Electronics and Information at the Northwestern Polytechnical University (NPU). He received the B.E. degree, M.E degree and Ph.D. degree all in signal and information processing from NPU, Xian, China, in 2005, 2008 and 2012, respectively. He was an ARC DECRA Fellow with the Research School of Engineering at the Australian National University, Canberra, Australia from 2014 to 2017 and a Research Fellow with the Research School of Computer Science at the Australian National University, Canberra, Australia from 2012 to 2014. His research interests include structure from motion, multi-view geometry, low-level computer vision, deep learning, compressive sensing and optimization. He won the Best Paper Award in IEEE CVPR 2012, the DSTO Best Fundamental Contribution to Image Processing Paper Prize at DICTA 2014, the Best Algorithm Prize in NRSFM Challenge at CVPR 2017, the Best Student Paper Prize at DICTA 2017 and the Best Deep/Machine Learning Paper Prize at APSIPA ASC 2017. He served as Area Chair for WACV 2019/2020 and ACM MM 2019.

Miaomiao Liu is a Lecturer and an ARC DECRA Fellow in the Research School of Engineering, the Australian National University. She was a Research Scientist at Data61/CSIRO from 2016-2018. Prior to that she was a researcher in NICTA. She received the BEng, MEng, and PhD degrees from Yantai Normal University, Yantai, China, Nanjing University of Aeronautics and Astronautics, Nanjing, China, and the University of Hong Kong, Hong Kong SAR, China, in 2004, 2007, and 2012, respectively. Her research interests include 3D vision, 3D reconstruction and 3D scene modeling and Understanding. She is a member of the IEEE.

Fatih Porikli is an IEEE Fellow and a Professor in the Research School of Engineering, Australian National University. He is acting as the Chief Scientist at Huawei, Santa Clara. He received his Ph.D. from New York University in 2002. His research interests include computer vision, pattern recognition, manifold learning, image enhancement, robust and sparse optimization and online learning with commercial applications in video surveillance, car navigation, intelligent transportation, satellite, and medical systems.

Quan Pan is the Dean of the Automation School of Northwestern Polytechnical University (NPU). He received the Ph.D. degree from NPU, Xian, China, in 1997. He is a Member of IEEE, a Member of the International Society of Information Fusion, a Board Member of the Chinese Association of Automation, and a Member of Chinese Association of Aeronautics and Astronautics. He obtained the 6th Chinese National Youth Award for Outstanding Contribution to Science and Technology in 1998 and the Chinese National New Century Excellent Professional Talent in 2000.