Depth Map Completion by Jointly Exploiting Blurry Color Images and Sparse Depth Maps

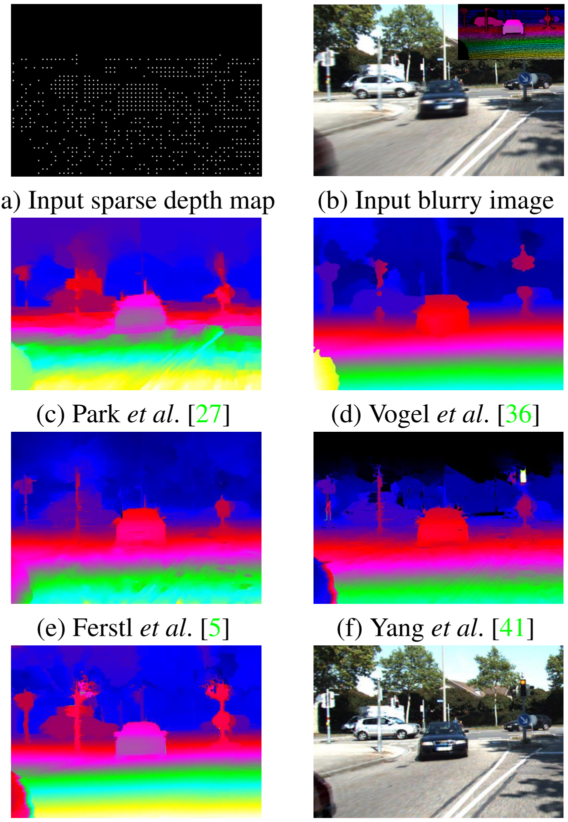
Liyuan Pan [1,2], Yuchao Dai[1,2], Miaomiao Liu[3,2] and Fatih Porikli[2]
[1] Northwestern Polytechnical University, Xi'an, China
[2] Australian National University, Canberra, Australia
[3] Data61, CSIRO, Canberra, Australia
panliyuan@mail.nwpu.edu.cn, daiyuchao@nwpu.edu.cn, {miaomiao.liu, fatih.porikli}@anu.edu.au

Abstract

We aim at predicting a complete and high-resolution depth map from incomplete, sparse and noisy depth measurements. Existing methods handle this problem either by exploiting various regularizations on the depth maps directly or resorting to learning based methods. When the corresponding color images are available, the correlation between the depth maps and the color images are used to improve the completion performance, assuming the color images are clean and sharp. However, in real world dynamic scenes, color images are often blurry due to the camera motion and the moving objects in the scene. In this paper, we propose to tackle the problem of depth map completion by jointly exploiting the blurry color image sequences and the sparse depth map measurements, and present an energy minimization based formulation to simultaneously complete the depth maps, estimate the scene flow and deblur the color images. Our experimental evaluations on both outdoor and indoor scenarios demonstrate the state-of-the-art performance of our approach. [1]

1. Introduction

High-precision and high-resolution 3D information play significant role in a variety of computer vision tasks including autonomous navigation [10, 19], 3D reconstruction and modeling [20, 35], and image deblurring [3, 13, 32, 37] just to count a few. However, the acquisition of such accurate depth maps is a challenging task. Although high-resolution depth maps can be computed from stereo images, the quality of the depth map relies on the calibration process and the apparent scene flow. Besides, stereoscopic depth estimation is problematic in low texture areas. As an alternative, active depth sensors provide depth information in a single shot. Unfortunately, measurements from the best depth sensors are still imperfect, which might be in low-resolution, noisy, and contaminated with large holes due to reflective surfaces and distant objects in the scene.



(a) Input sparse depth map (b) Input blurry image

(c) Park *et al.* [27] (d) Vogel *et al.* [36]

(e) Ferstl *et al.* [5] (f) Yang *et al.* [41]

(g) Ours: completed depth map and deblurred image

Figure 1. Qualitative comparisons on depth completion performance. (a) Input: incomplete and noisy depth map. (b) Corresponding blurry color image (with ground-truth complete depth map overlaid in the corner). (c) Estimated depth by [27] (d) Estimated depth by [36]. (e) Estimated depth by [5]. (f) Estimated depth by [41]. (g) Our depth completion and deblurring result. Compared to the stereo method (i.e. [36]) that ranks as the 1^{st} on the KITTI dataset and the remaining three state-of-the-art depth completion methods (i.e. [27, 5, 41]) shown above, our method achieves the best performance. (Best viewed on screen).

[1] Yuchao Dai is the corresponding author.

Depth super-resolution and *depth completion* techniques are designed to overcome these limitations by mainly leveraging the information from high-resolution and sharp color images [5, 41] to improve the quality of the depth map. Nevertheless, in real world settings, the quality of color images could be significantly variable due to camera vibration and relative motion of the dynamic objects in the scene.

Existing works use multiple-view blurry images to estimate the depth and deblur the image [3, 13, 32]. Even though they demonstrate that the depth estimation would also benefit image deblurring, their frameworks cannot be directly adopted to solve our problem, since they make strong assumptions that the scene is static, and the blur is only due to camera shake. In outdoors scenarios, the blur is also generated by the motion of dynamic objects and limited field of depth of the camera. Ismael *et al*. [2] recently proposed to adopt temporal information to super-resolve depth videos. Their method is constrained to particular types of 3D motion between neighboring frames such as the motion then can be decoupled into a lateral term and radial displacements. Although they attempted depth improvement, their method does not enhance the color image quality.

In Fig. 1(a), we provide a sample outdoor traffic scene image depicting camera and object motions. The undesired blur in the image causes the loss of details, which further hinders depth completion results. As indicated in Table 1, the performance of the state-of-the-art depth completion methods quickly deteriorate in the presence of blur in color images. On the other hand, the quality of the depth map has significant influence in deblurring color images. In Fig. 2, we compare the deblurring results with different resolution depth maps using our method. We observe that the deblurring performance improves with the increase of depth map quality. Therefore, we conclude that depth map completion and image deblurring are interwoven and strongly co-dependent where the solution of one benefits the other.

In this paper, we focus on handling realistic scenarios and tackling the problem of joint depth map completion and image deblurring by exploiting the spatio-temporal constraints in color images and depth map sequences. Our work is motivated by the recent progress in image deblurring and depth completion. It has been demonstrated [31] that scene flow estimation from stereo pairs can significantly improve the deblurring performance. This indicates that depth information can lead to a better deblurring in varying conditions compared to solely image-based methods. Likewise, deblurred images can support depth completion to estimate high-quality depth maps [2, 43].

To this end, we introduce a new framework for joint restoration of scene depth map and the latent clean image from given sparse depth maps and their corresponding blur color image sequences. Specifically, we use the piecewise planar assumption of the scene and represent the entire scene as a collection of 3D local planes, which significantly regularizes the problem. In this way, the joint restoration of scene depth map and latent clean image have been transformed to the estimation of the 3D geometry for each local plane, the rigid motion for each plane and the solution for the latent clean image. Our main contributions can be summarized as a comprehensive and efficient energy minimization formulation and the state-of-the-art depth completion performance using multiple images.

Table 1. Comparisons with clean/blur images on KITTI dataset.

KITTI	Flow Error(%)		Depth Error(%)	
	Clean	Blur	Clean	Blur
Vogel *et al.* [36]	2.83	13.62	4.27	8.20
Menze *et al.* [22]	3.28	14.77	4.70	6.72
Yang *et al.* [41]	/	/	3.43	4.67
D Ferstl *et al.* [5]	/	/	4.08	5.14
J Park *et al.* [27]	/	/	9.76	12.61

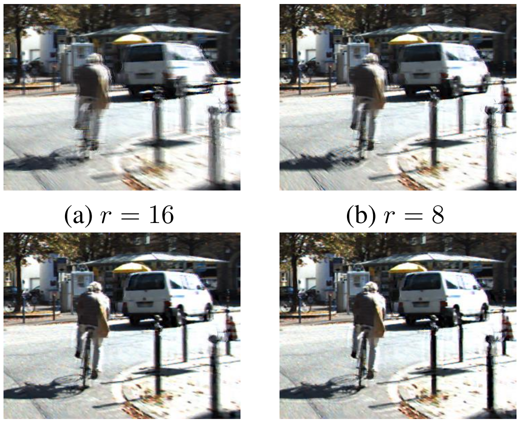

(a) $r = 16$ (b) $r = 8$
(c) $r = 4$ (d) $r = 0$

Figure 2. Performance of the image deblurring part of our method. Depth maps at different resolutions where r is the downsampling factor are shown (Best viewed on screen).

2. Related Work

Depth map completion has been widely studied in computer vision and image processing, and the research topic can be roughly divided into two categories, namely, depth map only and color image guided.

Depth map only: A common way to improve the resolution and quality of depth map is to fuse multiple depth maps into one depth map. KinectFusion [16] uses depth maps from neighboring frames to fill in the missing information during real time 3D rigid reconstruction. Newcombe *et al.* [24] took live depth data from a moving Kinect camera and created a high-quality 3D model for a static scene. Ismaeil *et al.* [2] proposed to complete low-resolution dynamic depth videos containing non-rigidly moving objects with a dynamic multi-frame super-resolution approach. This is obtained by accounting for nonrigid displacements in 3D, in addition to 2D optical flow, and simultaneously correcting the depth measurement by Kalman filtering. However, the real challenge that the research community has been facing is extending the multi-frame depth completion concept to dynamic scenes with moving objects.

Color image guided depth completion: This category of methods use additional intensity image as guidance for depth completion [7, 8, 25, 44]. Yang *et al.* [42] used bilateral filtering of a depth cost volume and a RGB image in an iterative refinement process. A more complex approach was proposed by Park *et al.* [28, 27] to use a combination of different weighting terms of a least squares optimization including segmentation, image gradients, edge saliency and non-local means for depth upsampling. Ferstl *et al.* [5] modeled the smooth term as a second order total generalized variation (TGV) regularization, and guided the depth upsampling with an anisotropic diffusion tensor calculated from a high-resolution intensity image. However, this method suffers from the blurring problems, especially areas around depth edges. Yang *et al.* [41] developed an adaptive color-guided auto-regression model for depth recovery. Aodha *et al.* [21] focused on single image upsampling as MRF labeling problem.

Deep CNN based depth completion: Recently, CNN has shown its ability in image recognition and classification task [33, 40] and has been extended to low-level vision tasks such as depth map super-resolution or depth completion. Riegler *et al.* [30] proposed a unified framework to effectively combine DCNN with total variations to generate HR depth maps. Riegler *et al.* [29] proposed to incorporate non-local variation into DCNN based framework, where the corresponding color images were also utilized. Additionally, Hui *et al.* [14] proposed a multi-scale guided convolutional network (MSG-Net). All these deep CNN based methods depend on the consistency between the training data and the testing data.

Deblurring with depth : Blur removal is an ill-posed problem, thus certain assumptions or additional constraints are required to regularize the solution space. As depth can significantly simplify the deblurring problem, depth-aware methods have been proposed to leverage the depth information. Xu *et al.* [37] inferred depth from two blurry images captured by a stereo camera and proposed a hierarchical estimation framework to remove motion blur caused by inplane translation. Hu *et al.* [13] solved it as a segment-wise depth estimation problem by assuming a discrete-layered scene where each segment corresponds to one layer. Arun *et al.* [3] proposed a geometric algorithm to estimate the camera motion from the blurry images themselves. However, they all assume that the scene to be static and the camera motion is the only source of motion blur. Recently, Sellent *et al.* [31] proposed a stereo deblurring approach, where 3D scene flow is estimated from the blurry images using a piecewise rigid 3D scene flow [36] representation. Very recently Pan *et al.* [26] proposed a single framework to jointly estimate the scene flow and deblur the images, where the motion cues from scene flow estimation and blur information could reinforce each other. Inspired by this stereo deblurring work, we aim to use a single view image sequence and its sparse and noisy depth map to complete the depth map and estimate the latent clean images.

3. Problem Formulation

Our goal is to complete the given incomplete and noisy depth maps $\tilde{\mathbf{D}}$ with the help of blurry color images $\mathbf{B}$ by exploiting the spatial-temporal constraints. Blur is caused by the motion of camera, objects, and limited depth-of-field of the camera (for large depth variations in the scene).

Towards this goal, we formulate our problem as a joint depth map completion and color image deblurring under dynamic scene settings. Since there are more variables (latent clean color images and target completed depth maps) to infer than the available measurements (blurry color images and incomplete and noisy depth measurements), we regularize this under-determined problem with the assumption that the scene can be well approximated by a collection of 3D planes [38] belonging to a finite number of objects performing rigid motions [22], *i.e.* a piecewise planar rigid motion representation. In this way, the original problem is transformed to the estimation of the geometric parameters $\mathbf{n}_i$, the local rigid motion $(\mathbf{R}_i, \mathbf{t}_i)$ for each 3D planar and the latent clean images $\mathbf{I}$.

In the following sections, we describe how to combine the geometric parameters, the motion parameters and the latent clean images together in the same objective where the solution of one variable will benefit the other variables. Our model relates to [26] and [22] in estimating the scene flow. However, our problem setting is very different where we have to exploit the sparse depth measurement constraint and blurry image constraint in a joint framework. We have introduced two new depth constraints to evaluate the consistency and the discrepancy between the sparse depth measurements and the completed depth maps.

3.1. Blur Image Formation

For complex dynamic settings such as outdoor traffic scenes, the blurry image is generated by spatially-variant per-pixel motion (optical flow). The blurry images are formed by the integration of light intensity emitted from the dynamic scene over the aperture time interval of the camera,

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

where $\mathbf{B}$ is the blurry frame, $\mathbf{x}$ denotes pixel location on image domain, $\mathbf{I}_n$ is the successive latent neighboring frames as frame m, $\mathbf{u}_n$ is the optical flow to frame n, $\mathbf{A}_m$ is the blur kernel matrix for the image m. Please refer to [15] and [31] for more details.

3.2. Formation Statement

In our setup, the incomplete and noisy depth measurements provide the depth information for each frame. Based on our piece-wise rigid planar assumption of the scene, optical flows for pixels lying on the same plane are constrained by the same 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 $i \in S$ is associated with a region $\mathcal{R}_i$ in the image and a plane variable $\mathbf{n}_{i,k} \in \mathbb{R}^3$ in 3D ($\mathbf{n}_{i,k}^T \mathbf{X} = 1$ for $\mathbf{X} \in \mathbb{R}^3$), where $k \in \{1, \cdots, |\mathcal{O}|\}$ denotes superpixel i is associated with the k-th rigid motion $\mathbf{o}_k = (\mathbf{R}_k, \mathbf{t}_k) \in \mathbb{SE}(3)$, where $\mathbf{R}_k \in \mathbb{R}^{3\times3}$ is the rotation matrix and $\mathbf{t}_k \in \mathbb{R}^3$ is the translation vector. Given the motion parameters $\mathbf{o}_k$ and geometric parameters $\mathbf{n}_{i,k}$, we can obtain the homography defined for superpixel i as

$$\mathbf{H}(\mathbf{n}_i, \mathbf{o}_k) = \mathbf{K}(\mathbf{R}_k - \mathbf{t}_k \mathbf{n}_{i,k}^T)\mathbf{K}^{-1}, \tag{2}$$

where $\mathbf{K} \in \mathbb{R}^{3\times3}$ is the intrinsic calibration matrix. $\mathbf{H}$ for each superpixel can be obtained when relates correspondences across frames $\mathbf{n}, \mathbf{o}$ are confirmed.This shows that the optical flows for pixels in a same superpixel are constrained by the corresponding homography, thus the optical flows are structured as opposed to [15].

We aim at completing the incomplete and noisy depth maps by exploiting both the spatial-temporal information in constraining the motion and the availability of corresponding blurry color images. To this end, we formulate the problem in a single framework as a discrete-continuous optimization problem to jointly complete the depth maps and deblur the color images. We explain all the constraints in the following sections.

3.3. Depth Constraint

3.3.1 Depth Consistency

The first depth term is to encourage the consistency between the sparse depth measurements and the completed depth estimation based on the piecewise planar models, which are evaluated across multiple frames. For the reference frame, the depth consistency is defined as:

$$\psi_i^1(\mathbf{n}_{i,k}) = w_1 \sum_{\mathbf{x}\in\Omega} |\tilde{\mathbf{D}}(\mathbf{x}) - \mathbf{D}(\mathbf{n}_{i,k}, \mathbf{x})|_1, \tag{3}$$

where $\tilde{\mathbf{D}}(\mathbf{x})$ denotes the sparse and noisy depth measurements from sensors such as Kinect and LiDAR, Ω denotes the image pixels with depth measurements available and $\mathbf{D}(\mathbf{n}_{i,k}, \mathbf{x})$ represents the depth estimation under the piecewise planar model.

For other frames besides the reference frame, the second depth consistency is evaluated as the discrepancy between the measured depth and the corresponding depth generated with the piecewise rigid planar motion,

$$\psi_i^2(\mathbf{n}_{i,k}, \mathbf{o}_k) = w_2 \sum_{\mathbf{x}\in\Omega} |\mathbf{D}(\mathbf{x}, \mathbf{n}_{i,k}, \mathbf{o}_k) - \tilde{\mathbf{D}}(\mathbf{H}^*\mathbf{x})|_1, \tag{4}$$

where the superscript $*$ denotes the warping direction to other color frames and the subscript of $\mathbf{H}$ is to index the corresponding homography for position $\mathbf{x}$.

3.3.2 Motion Sensitive Depth Discontinuity

Our model exploits a smoothness potential that enforcing the depth maps to be smooth and continuous, which involves both discrete and continuous variables. It is similar to the ones used in [22]. The third depth term for depth map is to enforce the motion boundaries to be co-aligned with the depth discontinuities, which is expressed as

$$\psi_{i,j}^3(\mathbf{n}_{i,k}, \mathbf{n}_{j,k'}) \\ = w_3 \begin{cases} \exp\Big(-\frac{\lambda}{|\mathcal{B}_{i,j}|} \sum\limits_{\mathbf{x}\in\mathcal{B}_{i,j}} \omega_{i,j}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{x})^2 \frac{|\mathbf{n}_i^T \mathbf{n}_j|}{\|\mathbf{n}_i\| \|\mathbf{n}_j\|}\Big) & \text{if } k \neq k', \\ 0 & \text{else.} \end{cases}$$

$\mathbf{x} \in \mathcal{B}_{i,j}$ evaluated with i-th superpixel parameter and j-th superpixel parameter, where $|\mathcal{B}_{i,j}|$ denotes the number of pixels belongs to the boundary between superpixels i and j.

3.3.3 Geometry Sensitive Depth Smoothness

The fourth depth term is to enforce the compatibility of two superpixels that share a common boundary by respecting the depth discontinuities. We define our potential function for continuous boundary as

$$\begin{aligned} \psi_{i,j}^4(\mathbf{n}_{i,k}, \mathbf{n}_{j,k'}) = &\sum_{\mathbf{x}\in\mathcal{B}_{i,j}} \rho_{\alpha_1}(d(\mathbf{n}_{i,k}, \mathbf{x}) - d(\mathbf{n}_{j,k'}, \mathbf{x})) \\ &+ \rho_{\alpha_3}\left(1 - \frac{|\mathbf{n}_{i,k}^T \mathbf{n}_{j,k'}|}{\|\mathbf{n}_{i,k}\| \, \|\mathbf{n}_{j,k'}\|}\right), \end{aligned} \tag{5}$$

$\rho_\alpha(\cdot) = \min(|\cdot|, \alpha)$ denotes the truncated penalty function.

3.4. Image Constraint

3.4.1 Brightness Consistency

Our image term involves mixed discrete and continuous variables, and are of three different kinds. The first image term encourages the corresponding pixels across the latent clean images should own similar appearance,

$$\theta_i^1(\mathbf{n}_i, \mathbf{o}, \mathbf{I}) = c_1 |\mathbf{I}(\mathbf{x}) - \mathbf{I}^*(\mathbf{H}^*\mathbf{x})|_1, \tag{6}$$

where $\mathbf{I}^* \in \mathbf{I}_n$ denotes the frame which $\mathbf{H}^*$ warping to. We use the ℓ_1 norm due to its robustness against noise and occlusions.

3.4.2 Anchor Point Constraint

While the above brightness consistency term provides dense constraint across image frames, we could also exploit the sparse and reliable feature correspondences (such as SIFT) to constrain the correspondences, which work as anchor points. Therefore our second image term is defined as

$$\theta_i^2(\mathbf{n}_i, \mathbf{o}) = \begin{cases} c_2 \rho_{\alpha_2}(||\mathbf{H}^*\mathbf{x} - \mathbf{x}^{'}||_2) & \text{if } \mathbf{x} \in \Pi \\ 0 & \text{otherwise.} \end{cases}$$

More specifically, it encodes the information that the warping of feature points $\mathbf{x} \in \Pi$ based on $\mathbf{H}^*$ should match its extracted correspondences $\mathbf{x}^{'}$ in the target view.

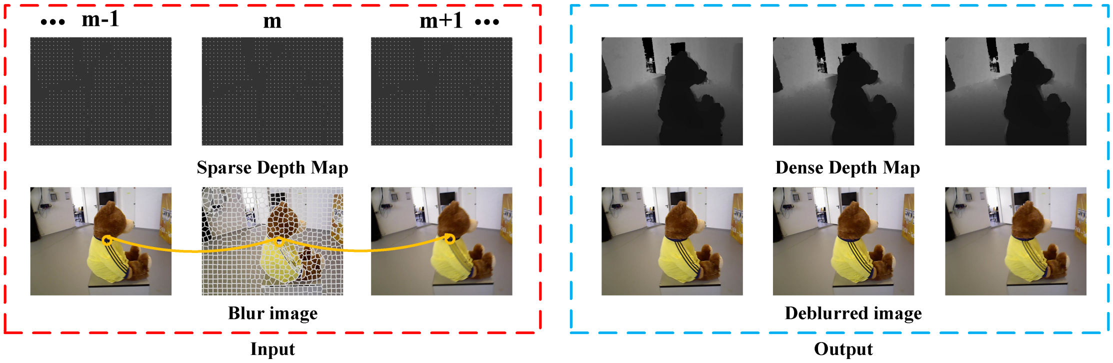


Figure 3. Illustration of our method. We simultaneously complete the depth maps and deblur the color images.

3.4.3 Deblurring Constraint

The third image term relates the observed blurry images with the latent clean images with the spatial-variant blur kernels,

$$\theta_i^3(\mathbf{n}_i, \mathbf{o}, \mathbf{I}) = c_3 \sum_m \sum_\partial \|\partial \mathbf{A}_m(\mathbf{n}_i, \mathbf{o})\mathbf{I}_m - \partial \mathbf{B}_m\|_2^2$$

where $\partial(\cdot)$ denotes the Toeplitz matrices corresponding to the horizontal, vertical derivative filters and the identity matrix. This term encourages the intensity changes and the intensity in the estimated blurry images to be close to that of the observed blurry images.

3.5. Regularization Term for Latent Clean Images

Natural images of typical real-world scenes generally obey sparse spatial gradient distributions [17, 18]. The distribution of a latent clean image can often be modeled as a generalized Laplace distribution [39], *i.e.* $P(\mathbf{I}) = \prod_{\mathbf{x}\in\mathbf{X}} \exp(-|\nabla_{\mathbf{x}}\mathbf{I}(\mathbf{x})|^p)$, where the power of p is a parameter usually within $[0.0, 1.0]$. This prior can be equivalently represented in energy minimization form, *i.e.* $\|\nabla_{\mathbf{x}}\mathbf{I}(\mathbf{X})\|^p \rightarrow \min$. We let $p = 1$ in the paper. In our model, this corresponds to a total variation term to suppress the noise in the latent image while preserving edges, and penalize spatial fluctuations.

$$\psi_m = |\nabla \mathbf{I}_m| = \|\mathbf{I}_m\|_{\text{TV}}. \tag{7}$$

3.6. Energy Minimization

Our energy minimization is defined as

$$\begin{aligned} E = & \underbrace{\sum_{i\in\mathcal{S}} \psi_i^{1,2}(\mathbf{n}_i, \mathbf{o}) + \sum_{i,j\in\mathcal{S}} \psi_{i,j}^{3,4}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{o})}_{\text{depth map}} \\ & + \underbrace{\sum_{i\in\mathcal{S}} \theta_i^{1,2,3}(\mathbf{n}_i, \mathbf{o}, \mathbf{I})}_{\text{image term}} + \underbrace{\sum_m \psi_m(\mathbf{I}_m)}_{\substack{\text{clean image}\\ \text{regularisation}}}, \end{aligned} \tag{8}$$

which consists of data terms evaluated on the color images and depth maps respectively, a smoothness term for the desired completed depth map, and a spatial regularization term for the latent clean images. Our model has been defined on three consecutive frames of RGB-D sequences. It can also allow the input with two pairs of RGB-D frames. Details are provided in section 5. In Section 4, we perform the optimization in an alternating manner to handle mixed discrete and continuous variables, thus allowing us to jointly complete the depth maps, and deblur the color images.

4. Solution of Energy Function

The optimization of our energy function defined in Eq.-(8) involves both discrete and continuous variables, which is challenging to solve. Therefore we resort to the alternative optimization manner, *i.e.*, optimizing one variable while fixing all the remaining ones. Note that our energy minimization formulation involves three set of variables, namely completed depth map $\mathbf{D}$ as indexed by $\mathbf{n}$, piecewise planar rigid motion $\mathbf{o}$ and latent clean image $\mathbf{I}$. We propose to handle the energy minimization by alternating between the following two steps,

- Fix the latent clean image $\mathbf{I}$, solve for scene geometry $\mathbf{n}$ and motion $\mathbf{o}$ (completed depth map and motion) by optimizing Eq.(9) (See Section 4.1).
- Fix the scene geometry and motion $\mathbf{n}$ and $\mathbf{o}$, solve for the latent clean image $\mathbf{I}$ by Eq.(10) (See Section 4.2).

4.1. Depth Completion and Motion Estimation

When the latent clean images are fixed as $\mathbf{I} = \tilde{\mathbf{I}}$, the joint optimization in Eq.(8) reduces to

$$\min_{\mathbf{n},\mathbf{o}} \sum_{i\in\mathcal{S}} \psi_i^{1,2}(\mathbf{n}_i, \mathbf{o}) + \sum_{i,j\in\mathcal{S}} \psi_{i,j}^{3,4}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{o}) + \theta_{i,j}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{o}, \tilde{\mathbf{I}}), \tag{9}$$

which is a discrete-continuous CRF optimization problem. We use the sequential tree-reweighted message passing (TRW-S) method [22] to find an approximate solution.

4.2. Debblurring

Given the scene geometry $\tilde{\mathbf{n}}$ and motion parameters $\tilde{\mathbf{o}}$, the blur kernel matrix $\mathbf{A}_m$ is derived based on Eq.(1). The objective function in Eq. (8) becomes convex w.r.t. $\mathbf{I}$

$$\min_{\mathbf{I}} \sum_{i \in \mathcal{S}} \theta_i^1(\tilde{\mathbf{n}}_i, \tilde{\mathbf{o}}, \mathbf{I}) + \theta_i^3(\tilde{\mathbf{n}}_i, \tilde{\mathbf{o}}, \mathbf{I}) + \sum_m \psi_m(\mathbf{I}). \tag{10}$$

In order to obtain the latent clean image $\mathbf{I}$, we adopt the conventional convex optimization method [4] and derive the primal-dual updating scheme as follows

$$\begin{cases} \mathbf{p}_{r+1} = \dfrac{\mathbf{p}_r + \gamma \nabla \mathbf{I}_r}{\mathbf{max}(1, \mathbf{abs}(\mathbf{p}_r + \gamma \nabla \mathbf{I}_r))} \\ \mathbf{q}_{r+1} = \dfrac{\mathbf{q}_r + \mu(\mathbf{I}_r - \mathbf{I}_r^*)}{\max(1, \mathbf{abs}(\mathbf{q}_r + \mu(\mathbf{I}_r - \mathbf{I}_r^*))} \\ \mathbf{I}_{r+1} = \arg\min_{\mathbf{I}} \sum_i c_3 \sum_\partial \|\partial \mathbf{A}\mathbf{I} - \partial \mathbf{B}\|_2^2 + \\ \dfrac{\left\| [\mathbf{I} - \eta((\nabla \mathbf{p}_{r+1})^T + \eta(\mathbf{q}_{r+1} - \mathbf{q}_{r+1}^*)^T)] - \mathbf{I}_r \right\|^2}{2\eta} \end{cases} \tag{11}$$

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

5. Experiments

We evaluate the performance of our method on both outdoors settings and indoors environments. For outdoors evaluation, we use the KITTI [9] autonomous driving benchmark dataset that provides monocular color images along with sparse depth maps from the LiDAR for validation. For indoors scenarios, we use the TUM RGB-D dataset [34] captured with a Kinect sensor. We present and discuss our results on both datasets in the following sections.

5.1. Experimental Setup

Initialization. Our model in Section 3 is formulated on three consecutive RGB-D images. In particular, we treat the middle frame as the reference image. We adopt the simple linear iterative clustering (SLIC) [1] to generate the superpixels, where each superpixel corresponds to a local planar in 3D. We use the penalized least squares method [6] to fast smooth the given sparse depth map for initialization. The motion hypothesis are then generated using RANSAC algorithm as suggested in [11].

Evaluations. Since our method could simultaneously complete the depth map and deblur the given images, we evaluate these two subtasks individually. We evaluate the depth completion results by counting the number of bad pixels having errors more than 3 pixels and 5% of its ground-truth. We adopt the PSNR to evaluate the deblurring performance. Thus, for each sequence, we report three performance metrics: depth errors (geometry), flow errors (motion), and PSNR (latent images) values for the reference images.

Table 2. Quantitative depth completion errors and deblur results.

	Depth Error(%)		Flow Error(%)	PSNR (dB)
	KITTI	TUM	KITTI	KITTI
Kim and Lee [15]	/	/	38.89	28.25
Sellent *et al.* [31]	8.20	/	13.62 [36]	27.75
Yang *et al.* [41]	4.67	0.43	/	/
D Ferstl *et al.* [5]	5.14	0.47	/	/
J Park *et al.* [27]	12.61	0.29	/	/
Ours(no depth term)	5.53	0.26	17.16	29.85
Ours	**3.91**	**0.22**	**13.01**	**29.83**

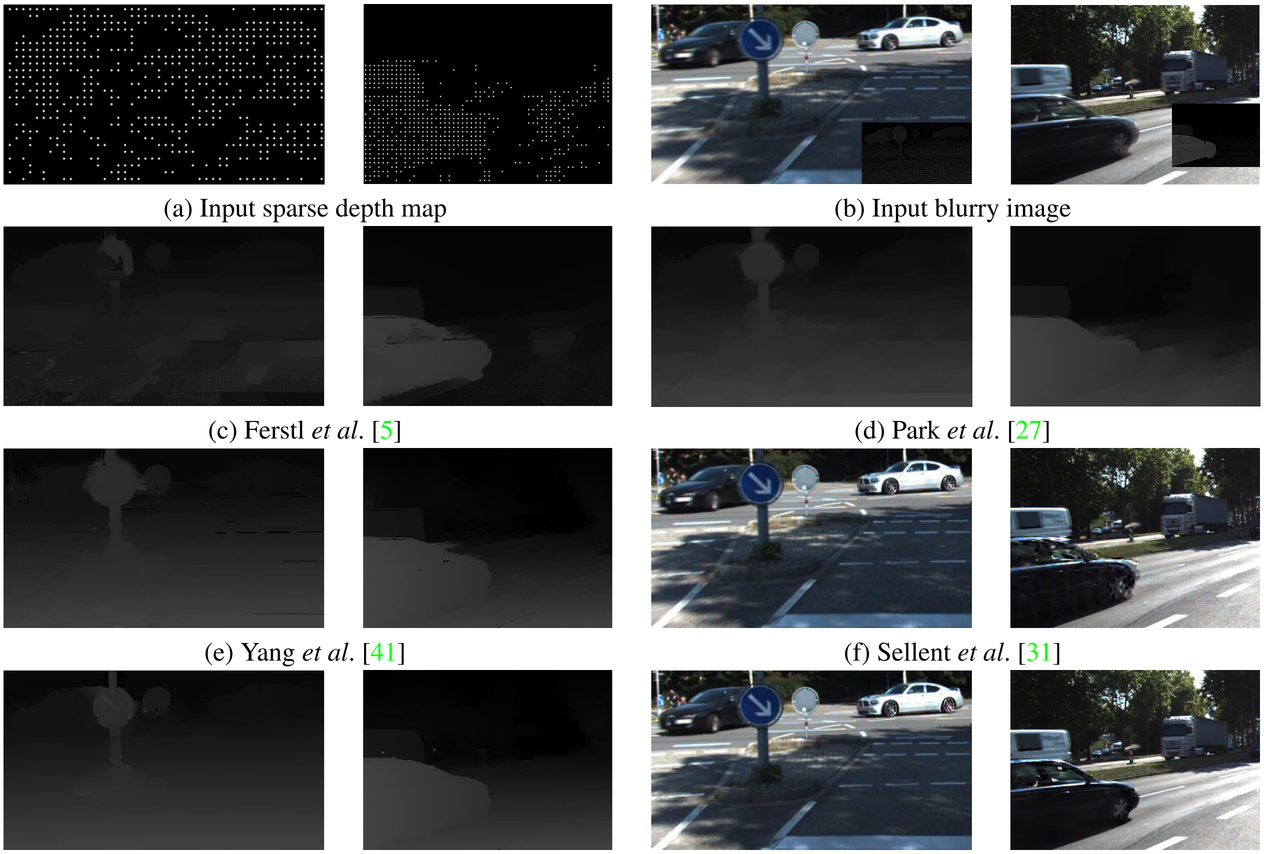
Baselines Methods. As for our depth completion results, we compare both passive and active RGB-D methods separately. For multi-view cameras system, we compare with piece-wise rigid scene flow method (PRSF) [36], which is ranked the firston the KITTI scene flow estimation benchmark and is used as the flow initialization for [31]. For active depth sensors, we compare with TGV [5], [27] and [41] which are also three applicable state-of-the-art methods. We compare our deblurring results with the state-of-the-art deblurring approach for monocular images [15], and the approach for stereo images [31]. The results are shown in Table 2.

5.2. Experimental Results

Results on KITTI. To the best of our knowledge, currently there is no realistic benchmark datasets that provide blurry images and corresponding ground-truth depth maps and the latent clean images. In this paper, we take advantage of the KITTI visual odometry dataset [9] to create a synthetic blurry image dataset on realistic scenery, where each sequence includes 6 images (375 × 1242). Then we obtained the depth sequence with a down-sample factor $r = 4$. The blurry images are generated by using the piecewise linear kernel, where frame rate is set as $\tau = 0.23$ and the number of frame is $N = 20$. Therefore, the image blur is caused by both objects motion and camera motion with occlusion and shadow. We perform block-coordinate-descent on a subset of 30 randomly selected training images to obtain the optimal model parameters $\{w, c\}$ and $\{\alpha\}$ in cross-validation.

We evaluated results on the reference image and our method consistently outperforms all baselines as illustrated in Table 2. We achieve the minimum bad pixel ratio of 3.91% for depth completion and PSNR of 29.83 for image deblurring. Fig. 4 shows qualitative depth completion results and deblurring results of our method and other competing methods on the sample KITTI sequences.

Fig. 5 shows the performance of our depth completion and image deblurring method with respect to the number of iterations, where the performance of both depth completion and image deblurring improves with the increase of iterations. While we use 6 iterations for all our experiments, the experiments indicate that 3 iterations are sufficient in most cases to reach an optimal performance for our formulation.



(a) Input sparse depth map

(b) Input blurry image

(c) Ferstl *et al*. [5]

(d) Park *et al*. [27]

(e) Yang *et al*. [41]

(f) Sellent *et al*. [31]

(g) Our completion depth map and deblurred image

Figure 4. Depth completion and image deblurring results on the KITTI dataset. (a) Input: sparse depth map. (b) Corresponding blurry color image (with ground-truth depth map in the corner). (c) Estimated depth map by [5]. (d) Estimated depth by map [27]. (e) Estimated depth map by [41]. (f) Deblurring result of [31]. (g) Our depth completion and deblurring result. Compared to the recent stereo deblurring method (i.e. [31]) and the remaining three state-of-the-art depth completion methods (i.e. [5, 27, 41]) shown above, our method achieves the best performance for both depth completion and deblurring. Best viewed on screen.

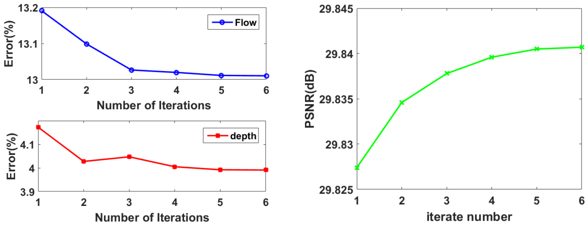


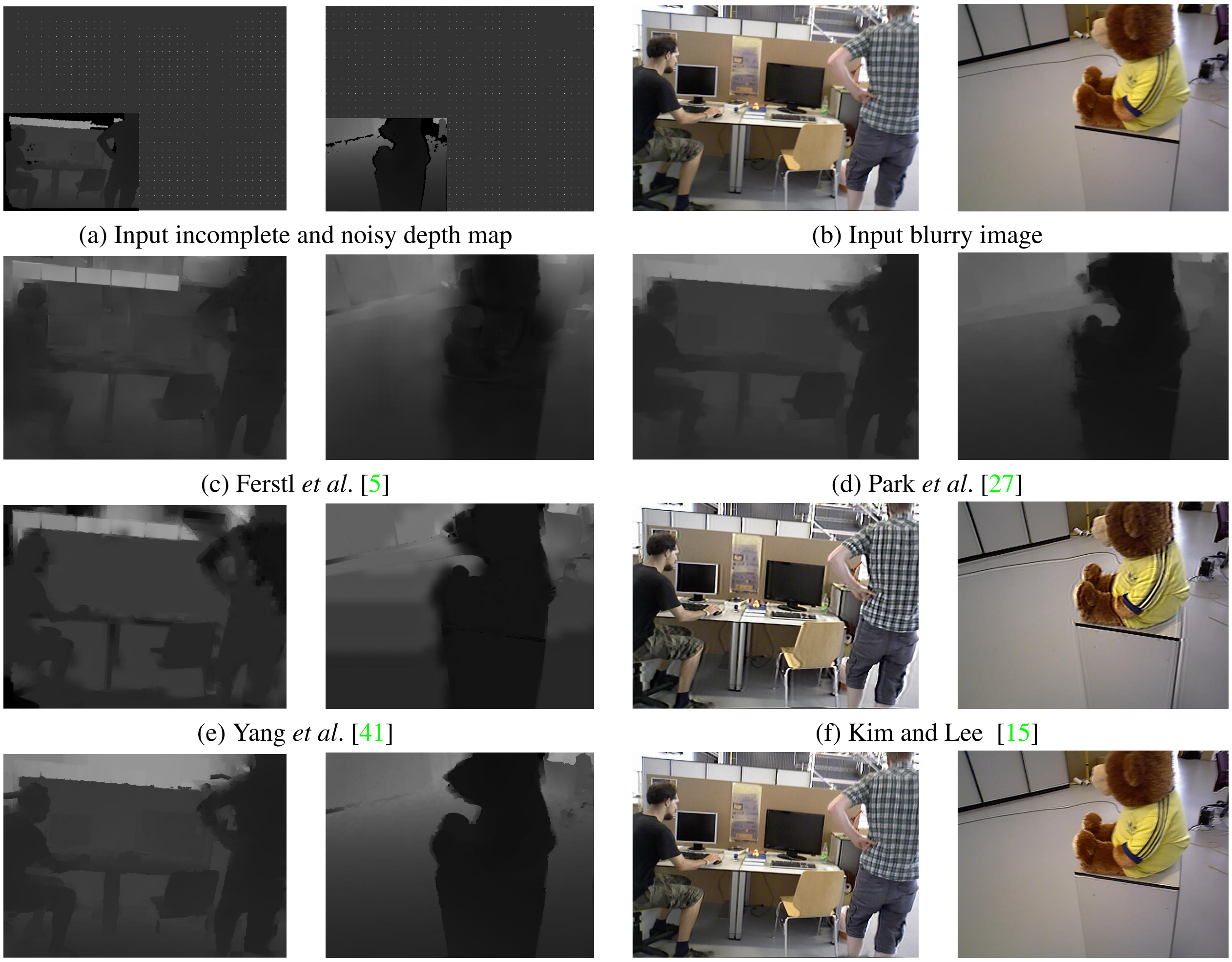
Figure 5. Performance of our method on KITTI (flow error, depth error, PSNR) with the respect to the number of iterations.

Table 3. Quantitative evaluation on the KITTI dataset where the blur images are generated by averaging three consecutive frames.

	PSNR(dB)	SSIM(%)	Depth Error(%)
Yang *et al*. [41]	/	/	6.15
D Ferstl *et al*. [5]	/	/	3.22
J Park *et al*. [27]	/	/	9.63
Kim and Lee [15]	23.21	0.781	/
Sellent *et al*. [31]	23.31	0.764	/
Ours	**23.89**	**0.786**	**2.77**

Results on TUM. In order to evaluate the performance of our method on real dataset, we use the TUM RGB-D dataset [34] which included motion blur. The captured depth maps and color images are of size 640×480. We down-sample the obtained depth maps with rate $r = 16$ to simulate sparse depth maps. We evaluated our results on the reference image and achieve the minimum bad pixel ratio of 0.22% for depth completion, consistently outperforms all baseline methods. Fig. 6 shows the visually completed depth map and deblurring results of our method comparing with other methods on sample TUM sequences.

Results on Another Blur Model. Even though the TUM dataset contains blurry images, they cannot be used for quantitative evaluation since no ground truth clean images are available. To perform such quantitative evaluation, synthetic images have been widely used [15, 23, 12]. We have



(a) Input incomplete and noisy depth map

(b) Input blurry image

(c) Ferstl *et al*. [5]

(d) Park *et al*. [27]

(e) Yang *et al*. [41]

(f) Kim and Lee [15]

(g) Our completion depth map and deblurred image

Figure 6. Depth completion and image deblurring results on the TUM dataset. (a) Input: incomplete and noisy depth map (with ground-truth depth map in the corner). (b) Corresponding blurry color image. (c) Estimated depth map by [5]. (d) Estimated depth map by [27]. (e) Estimated depth map by [41]. (f) Deblurring result of [15]. (g) Our depth completion and deblurring result. Compared to the monocular deblurring method (i.e. [15]) and the remaining three state-of-the-art depth completion methods (i.e. [5, 27, 41]) shown above, our method achieves the best performance for both depth completion and deblurring. Best viewed on screen.

evaluated our method under the spatial-variant blur generation model. Here we tested our method on another blur generation model (the blur image is simply an average of consecutive three frames). The results are shown in Table 3, where our method again achieves the best performance.

6. Conclusion

In this paper, we present a joint optimization framework to tackle the challenging task of depth map completion with the guidance of blurry color images, where depth completion and sequence images deblurring are solved in a coupled manner. Under our formulation, the motion cues from depth completion and blurry images could benefit each other, and produce superior results than conventional depth completion or deblurring methods. The performance of our method has been evaluated on both outdoor and indoor scenarios.

Acknowledgement

This work was supported in part by China Scholarship Council (201506290130), Australian Research Council (ARC) grants (DP150104645, DE140100180), and Natural Science Foundation of China (61420106007, 61473230, 61135001), State Key Laboratory of Geo-information Engineering (NO.SKLGIE2015-M-3-4), Natural Science Foundation of Shaanxi (2017JM6027, 2017JQ6005) and Aviation fund of China (2014ZC53030). We thank all reviewers for their valuable comments.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012. 6

[2] Kassem Al Ismaeil, Djamila Aouada, Thomas Solignac, Bruno Mirbach, and Bjorn Ottersten. Real-time enhancement of dynamic depth videos with non-rigid deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016. 2

[3] M Arun, AN Rajagopalan, and Gunasekaran Seetharaman. Multi-shot deblurring for 3d scenes. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 19–27, 2015. 1, 2, 3

[4] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. 6

[5] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 993–1000, 2013. 1, 2, 3, 6, 7, 8

[6] Damien Garcia. Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational statistics & data analysis*, 54(4):1167–1178, 2010. 6

[7] Frederic Garcia, Djamila Aouada, Bruno Mirbach, Thomas Solignac, and Björn Ottersten. A new multi-lateral filter for real-time depth enhancement. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 42–47, 2011. 3

[8] Frederic Garcia, Djamila Aouada, Bruno Mirbach, Thomas Solignac, and Björn Ottersten. Unified multi-lateral filter for real-time depth map enhancement. *Image and Vision Computing*, 41:26–41, 2015. 3

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, pages 1231–1237, 2013. 6

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3354–3361, 2012. 1

[11] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium*, pages 963–968, 2011. 6

[12] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton van den Hengel, and Qinfeng Shi. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. *arXiv preprint arXiv:1612.02583*, 2016. 7

[13] Zhe Hu, Li Xu, and Ming-Hsuan Yang. Joint depth estimation and camera shake removal from single blurry image. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2893–2900, 2014. 1, 2, 3

[14] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *Proc. Eur. Conf. Comp. Vis.*, pages 353–369. Springer, 2016. 3

[15] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5426–5434, 2015. 3, 4, 6, 7, 8

[16] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *ACM symposium on User interface software and technology*, pages 559–568, 2011. 2

[17] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1033–1041, 2009. 5

[18] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 233–240, 2011. 5

[19] L. Liu, H. Li, Y. Dai, and Q. Pan. Robust and efficient relative pose with a multi-camera system for autonomous driving in highly dynamic environments. *IEEE Transactions on Intelligent Transportation Systems*, PP(99):1–13, 2017. 1

[20] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[21] Oisin Mac Aodha, Neill DF Campbell, Arun Nair, and Gabriel J Brostow. Patch based synthesis for single depth image super-resolution. In *Proc. Eur. Conf. Comp. Vis.*, pages 71–84. Springer, 2012. 3

[22] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3061–3070, 2015. 2, 3, 4, 5

[23] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. *arXiv preprint arXiv:1612.02177*, 2016. 7

[24] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE international symposium on Mixed and augmented reality*, pages 127–136, 2011. 2

[25] Roy Or-El, Guy Rosman, Aaron Wetzler, Ron Kimmel, and Alfred M Bruckstein. Rgbd-fusion: Real-time high precision depth recovery. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5407–5416, 2015. 3

[26] Liyuan Pan, Yuchao Dai, Miaomiao Liu, and Fatih Porikli. Simultaneous stereo video deblurring and scene flow estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[27] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S Brown, and In So Kweon. High-quality depth map upsampling and completion for rgb-d cameras. *IEEE Trans. Image Proc.*, 23(12):5559–5572, 2014. 1, 2, 3, 6, 7, 8

[28] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S Brown, and Inso Kweon. High quality depth map upsampling for 3d-tof cameras. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1623–1630, 2011. 3

[29] Gernot Riegler, David Ferstl, Matthias Rüther, and Horst Bischof. A deep primal-dual network for guided depth super-resolution. In *Proc. Brit. Mach. Vis. Conf.*, 2016. 3

[30] Gernot Riegler, Matthias Rüther, and Horst Bischof. Atgv-net: Accurate depth super-resolution. In *Proc. Eur. Conf. Comp. Vis.*, pages 268–284. Springer, 2016. 3

[31] Anita Sellent, Carsten Rother, and Stefan Roth. Stereo video deblurring. In *Proc. Eur. Conf. Comp. Vis.*, pages 558–575. Springer, 2016. 2, 3, 6, 7

[32] Hee Seok Lee and Kuoung Mu Lee. Dense 3d reconstruction from severely blurred images using a single moving camera. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 273–280, 2013. 1, 2

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[34] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 573–580. IEEE, 2012. 6, 7

[35] Min Sun, Gary Bradski, Bing-Xin Xu, and Silvio Savarese. Depth-encoded hough voting for joint object detection and shape recovery. In *Proc. Eur. Conf. Comp. Vis.*, pages 658–671. Springer, 2010. 1

[36] Christoph Vogel, Konrad Schindler, and Stefan Roth. 3d scene flow estimation with a piecewise rigid scene model. *Int. J. Comp. Vis.*, 115(1):1–28, 2015. 1, 2, 3, 6

[37] Li Xu and Jiaya Jia. Depth-aware motion deblurring. In *IEEE International Conference on Computational Photography*, pages 1–8, 2012. 1, 3

[38] Koichiro Yamaguchi, David McAllester, and Raquel Urtasun. Robust monocular epipolar flow estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1862–1869, 2013. 3

[39] J. Yang, H. Li, Y. Dai, and R. T. Tan. Robust optical flow estimation of double-layer images under transparency or reflection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1410–1419, June 2016. 5

[40] Jingxiang Yang, Yong-Qiang Zhao, and Jonathan Cheung-Wai Chan. Learning and transferring deep joint spectral-spatial features for hyperspectral classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2017. 3

[41] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao Wang. Color-guided depth recovery from rgb-d data using an adaptive autoregressive model. *IEEE Trans. Image Proc.*, 23(8):3443–3458, 2014. 1, 2, 3, 6, 7, 8

[42] Qingxiong Yang, Ruigang Yang, James Davis, and David Nistér. Spatial-depth super resolution for range images. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1–8, 2007. 3

[43] Guofeng Zhang, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao. Consistent depth maps recovery from a video sequence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(6):974–988, 2009. 2

[44] Jiejie Zhu, Liang Wang, Ruigang Yang, James E Davis, et al. Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(7):1400–1414, 2011. 3