Joint Estimation of Camera Pose, Depth, Deblurring, and Super-Resolution from a Blurred Image Sequence

Haesol Park Kyoung Mu Lee
Department of ECE, ASRI, Seoul National University, 151-742, Seoul, Korea
haseol.park@gmail.com, kyoungmu@snu.ac.kr

Abstract

The conventional methods for estimating camera poses and scene structures from severely blurry or low resolution images often result in failure. The off-the-shelf deblurring or super-resolution methods may show visually pleasing results. However, applying each technique independently before matching is generally unprofitable because this naïve series of procedures ignores the consistency between images. In this paper, we propose a pioneering unified framework that solves four problems simultaneously, namely, dense depth reconstruction, camera pose estimation, super-resolution, and deblurring. By reflecting a physical imaging process, we formulate a cost minimization problem and solve it using an alternating optimization technique. The experimental results on both synthetic and real videos show high-quality depth maps derived from severely degraded images that contrast the failures of naïve multi-view stereo methods. Our proposed method also produces outstanding deblurred and super-resolved images unlike the independent application or combination of conventional video deblurring, super-resolution methods.

1. Introduction

Structure from motion or multi-view stereo (MVS) is a very interesting problem in computer vision that aims to determine the underlying 3D scene structure and camera configuration from multiple images. Despite the inherent difficulty of this inverse problem, contemporary algorithms show a satisfactory performance when applied on public datasets [10, 20].

Despite their encouraging achievements, some limitations prevent the aforementioned methods from being applied in highly realistic scenarios. Among these challenges are the blurs resulting from camera motion [14, 17], which becomes serious when using handheld cameras or cameras attached to moving vehicles. Blur operation acts differently on each pixel according to the scene depth and camera motion, and it breaks brightness constancy assumption among consecutive frames.

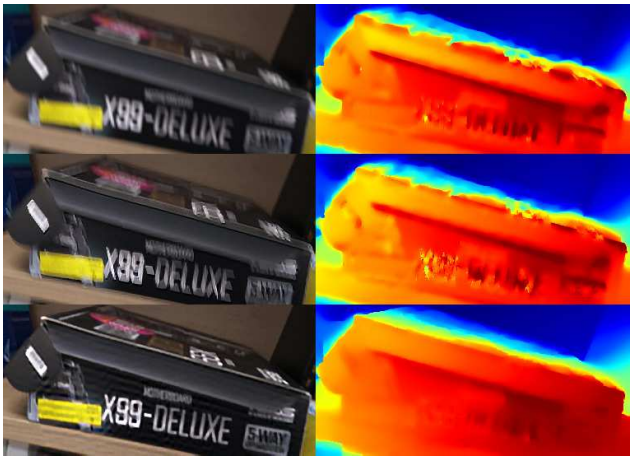


Figure 1: Comparison of depth estimation and image restoration results for blurry, low-resolution images. The left column shows the estimated latent images, while the right column shows their corresponding depth maps. The images from top to bottom are obtained via (a) a simple bicubic interpolation, (b) the independent use of deblurring [30] after applying the super-resolution algorithm [25], and (c) the proposed method, respectively. The depth maps for the first two rows are estimated via baseline variational depth estimation.

Low-resolution (LR) input images also often affect stereo matching accuracy [5, 15], because low-quality cameras are frequently used considering the limitations in cost or space for some applications. However, even in a high-resolution (HR) image, the actual scene resolution is spatially uneven and dependent on depth because of the perspective projection.

The aforementioned problem becomes especially challenging when the image frames are simultaneously corrupted by blur and low resolution. This problem can be directly addressed by applying the super-resolution [24] or

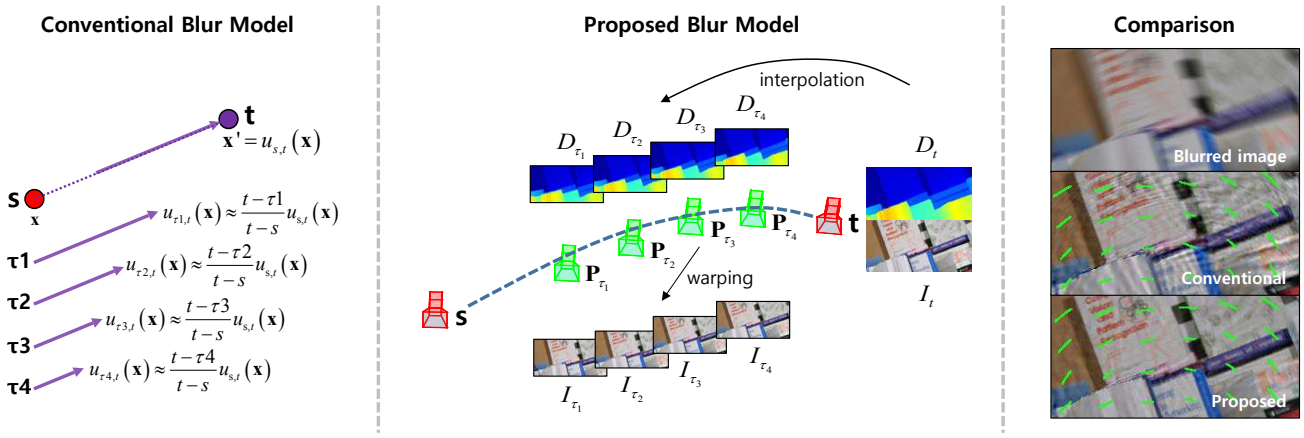


Figure 2: Comparison of the proposed blur model and the conventional blur model used in [14, 17]. Both models illustrate the blur procedures for the frame at time t, where s represents the time of the previous frame. The proposed model approximates the intermediate images I_τ's during the shutter time using the interpolated camera poses $\mathbf{P}_\tau$'s and depth maps D_τ's, while the conventional model depends on a single optical flow map from s to t, $u_{s,t}$ (e.g. $u_{\tau_1,t}$ is used to approximate I_{τ_1}). The deblurred images with the overlaid blur kernels of each model are also presented for comparison. Although both images are obtained using the ground-truth depth map and camera poses, the image obtained using the conventional blur model exhibits more artifacts because of inaccurate blur kernels.

deblurring method [30] before matching, which may produce visually pleasing images. However, the results obtained using this approach are worse than those obtained using original images in terms of matching as shown in Figures 1 (a) and (b), because single-image super-resolution or deblurring algorithms ignore and break the brightness constancy among neighboring frames.

In this study, we consider the four inter-related problems of camera pose estimation, dense depth reconstruction, deblurring, and super-resolution as a whole by formulating them as a unified energy function. To the best of our knowledge, this study is the first to solve the four aforementioned problems jointly in a single framework. Our proposed method clearly outperforms the independent use of existing techniques for each problem. By exploiting the multi-view geometry explicitly, our proposed method can handle more general blur kernels that may result from camera rotations and forward motions.

2. Related Works

Few researchers have attempted to perform image matching on blurry images. Portz *et al*. [17] proposed an optical flow method that uses a blur-aware matching procedure originally introduced in tracking methods [12, 16]. Based on the assumed commutativity of blur operations, this method blurs the input images with the kernels of one another instead of deblurring them using their own kernels.

Lee *et al*. [13, 14] extended this idea and proposed several methods for handling blurred input images in camera pose estimation [13] and dense stereo matching [14]. However, given that scene depth and camera motion can generate the exact blur kernels only when both values are correct, estimating these parameters separately would be inappropriate. Moreover, the aforementioned works [14, 17] are limited by a simple assumption that the blur kernel can be modeled by using linear optical flow vectors between consecutive frames.

By contrast, our proposed blur model (Section 3) covers more general camera motions by adopting the linear model in an Lie algebra $\mathfrak{se}(3)$space [6]. The blur kernel is explicitly approximated by interpolating the camera path and depth maps between adjacent frames. Figure 2 shows the difference between the conventional and proposed blur models.

Recent works [21, 32] have attempted to solve stereo matching and image deblurring jointly using the same blur model as the proposed one. However, both of these methods depend on additional or external data. The method proposed by Sellent *et al*. [21] can only handle stereo video sequences, in which per-frame depth cues are available. Zhen and Stevenson [32] proposed a method for single-view image sequences, but this method requires additional data, including inertial measurements and sharp noisy frames.

Some methods solve the super-resolution and MVS problems in a single framework [5, 15], which is shown to increase the accuracy of both the restored images and depth maps. However, the multi-frame super-resolution framework used in [5, 15] only works when accurate matching

information is available in sub-pixel units. Therefore, these methods cannot jointly estimate super-resolved images and depth maps for blurry input images because of the large errors in correspondences.

Some researchers proposed to solve super-resolution and deblurring jointly [23, 4]. The method proposed by Bascle *et al*. [4] relies on external tracking information to estimate the blur kernel using the trajectory of the object and establish the sub-pixel correspondence for multi-frame super-resolution. However, this method is applicable only on some objects, which should be easy to track, not on the entire image. Takeda and Milanfar [23] proposed an intriguing method to handle a spatio-temporal super-resolution and deblurring problem in a spatially invariant 3D deconvolution framework. However, this method cannot handle large blur kernels because the size of motion vectors between consecutive frames is limited by a few pixels.

3. Modeling Imaging Process

We examine an image sequence captured by a single moving camera where the target scene is assumed to be static to enable stereo matching and camera pose estimation. In this study, an image is defined as a mapping that uses a 2D pixel coordinate vector as input and a 3D color vector as output (in the case of typical RGB images). The color value of the pixel $\mathbf{x} = [x, y]^T$ of image I is given by $I(\mathbf{x})$. When the query 2D coordinate has non-integer values, the color values are interpolated using the color values of the neighboring grid points. We apply bilinear interpolation throughout this paper.

In the following, the input images are denoted by B_t's, with t representing the time when the images are captured. An acquired image B_t is assumed to be the accumulation of the sensor output from the opening (t_o) to the closing ($t_c = t$) of the camera shutters. We model this capturing process by assuming the presence of ideal clean and HR images during the shutter time. By denoting the ideal images at time τ as I_τ, a real image B_t is considered as the downsampled version of the integral of I_τ as follows:

$$B_t = \frac{1}{t_c - t_o}\left(\int_{t_o}^{t_c=t} I_\tau d\tau\right)\downarrow, \tag{1}$$

where the down arrow represents the downsampling process.

A blur is generated for the static scene because of the camera movement during the shutter time. Therefore, the images I_τ's change over time. However, the difference between I_τ and I_t, clean HR image at time t, is not large because such variation is caused by the slight camera motions that take place within a short period (less than $t_c - t_o$). Therefore, I_τ can be approximated by warping I_t, if the relative poses of the cameras and the scene structure are both known.

We denote the pose of the camera and the depth map of the image I_t by $\mathbf{P}_t$ and D_t, respectively, and then denote the time-invariant camera intrinsic matrix by $\mathbf{K}$. Using these notations, the warping process is expressed as follows:

$$I_\tau(\mathbf{x}) \approx I_t\left(W^{\tau\to t}(\mathbf{x})\right), \tag{2}$$

where the function $W^{\tau\to t}(\cdot)$ computes the warped pixel position from the camera $\mathbf{P}_\tau$ to $\mathbf{P}_t$, and can be expressed as follows:

$$\begin{aligned}\mathbf{x}_t &= W^{\tau\to t}(\mathbf{x}) \\ &= i_2\left(\mathbf{K} i_3\left((\mathbf{P}_t)^{-1}\mathbf{P}_\tau h_3\left(\frac{1}{D_\tau(\mathbf{x})}\mathbf{K}^{-1}h_2(\mathbf{x})\right)\right)\right),\end{aligned} \tag{3}$$

where functions $i_2(\cdot)$ and $i_3(\cdot)$ convert the homogeneous coordinates into inhomogeneous coordinates in the 2D and 3D spaces, respectively, while $h_2(\cdot)$ and $h_3(\cdot)$ convert the inhomogeneous coordinates into homogeneous coordinates in the same spaces.

The integral in Equation (1) is approximated using a finite sum of images. The insertion of the image warping Equation (2) generates the following:

$$B_t \approx \Psi_t \circ I_t, \tag{4}$$

$$(\Psi_t \circ I)(\mathbf{x}) = \left(\frac{1}{M}\sum_{m=1}^{M} I\left(W^{\tau_m\to t}(\mathbf{x})\right)\right)\downarrow, \tag{5}$$

where $\tau_m = (m/M)(t_c - t_o) + t_o$ and M controls the degree of discretization. We define $\Psi_t(\cdot)$ as the operator on a general image I to approximate the degradation resulting from the capturing process at time t. Figure 2 illustrates the concept of this blur operation.

In practice, the values of $D_{\tau_m}{}'s$ and $\mathbf{P}_{\tau_m}{}'s$ are approximated using D_t, $\mathbf{P}_t$, and $\mathbf{P}_s$, where s represents the time of the previous frame. $\mathbf{P}_{\tau_m}$ is sampled from the interpolated camera path between $\mathbf{P}_t$ and $\mathbf{P}_s$. The interpolation is conducted in the Lie algebra $\mathfrak{se}(3)$space [6]. Given $\Delta\mathbf{P}_{t,s} = \log\left(\mathbf{P}_t \cdot (\mathbf{P}_s)^{-1}\right)$, the interpolation is performed as follows:

$$\mathbf{P}_{\tau_m} = \exp\left(\frac{\tau_m - s}{t - s}\cdot\Delta\mathbf{P}_{t,s}\right)\cdot\mathbf{P}_s, \tag{6}$$

where $\log$ and $\exp$ denote the logarithmic and exponential maps between the Lie group $\mathbb{SE}(3)$space, where the actual camera pose matrices resides, and the Lie algebra $\mathfrak{se}(3)$space [6]. Note that the proposed method might work unreliably when the camera motion between the consecutive frames is too complex to be approximated by the simple interpolation scheme in Equation (6), for example, when the camera vibrates with a frequency much higher than the camera frame rate.

After obtaining the camera pose at time τ_m, the depth map D_{τ_m} can be approximated by warping the closest depth map D_t. The warped depth value can be computed by re-projecting D_t to the world coordinate and then projecting this map the virtual camera at $\mathbf{P}_{\tau_m}$. The projected value is actually the depth of the point from $\mathbf{P}_{\tau_m}$. The capturing operator $\Psi_t(\cdot)$ is only dependent on D_t, $\mathbf{P}_t$, and $\mathbf{P}_s$.

4. Unified Energy Formulation

This study aims to estimate the latent images I_t's with the corresponding depth maps D_t's and camera poses $\mathbf{P}_t$'s from a blurred, LR image sequence, B_t's. We assume that the intrinsic parameters $\mathbf{K}$ are previously known. Given that the target variables are interrelated, the proposed method estimates them altogether by optimizing a single unified energy function.

The total energy function E is defined by the sum of energy functions, E_t, which is defined for each single frame at time t. E_t comprises three terms, with each term having a unique physical meaning:

$$E = \sum_t E_t, \tag{7}$$

$$E_t = E_t^m + E_t^s + E_t^r, \tag{8}$$

where the matching, self-consistency, and regularization terms are presented from left to right.

4.1. Matching term

The first term relates the images from the consecutive frames based on the scene structure and camera motion. Given the static target scene, the images warped into a specific frame must coincide if the warping is based on correct depth maps and camera poses.

In the proposed matching term, we match the input blurred LR image, B_t, with the latent images of the neighboring frames, I_s's, where $s \in N(t)$ denotes the time index for the neighboring frames of t. Therefore, an additional one-way blur operation for matching is performed, where I_s's must be blurred and downsampled by the capturing operator of B_t. The matching term is defined as follows:

$$E_t^m = \sum_{s \in N(t)} \sum_{\mathbf{x} \in \Omega_{ts}} \left\| B_t(\mathbf{x}) - \Psi_t \circ I_s\left(W^{t \to s}(\mathbf{x})\right) \right\|_1. \tag{9}$$

The matching term only considers the pixels in the set Ω_{ts}, which represents the visible area of the image domain at time t in terms of the camera at s. Section 5.4 discusses how this area is determined. We use L1-norm, which generates reliable results and is highly robust to the presence of noise and occlusion [26].

In terms of MVS matching, the proposed methods try to determine the plausible scene structure and camera poses that satisfy the brightness constancy assumption by minimizing the matching term. The same matching term is also used as the evidence of super-resolution for restoring I_s's from LR observations based on the estimation of the latent images.

4.2. Self-consistency term

The self-consistency term E_t^s is derived from the imaging process in Equation (5) as follows:

$$E_t^s = \lambda_s \sum_{\mathbf{x}} \left\| B_t(\mathbf{x}) - \Psi_t \circ I_t(\mathbf{x}) \right\|_1, \tag{10}$$

which makes the solution consistent with the observation. Based on the depth maps and camera poses, the capturing operator $\Psi_t(\cdot)$ is constant and the equation is similar to the conventional data term in the extant deblurring methods. The parameter λ_s controls the strength of this constraint.

4.3. Regularization term

Although the matching and self-consistency terms can compensate each other, they both rely on possibly noisy input images. The additional term regularizes the solutions to suppress the errors. In the proposed framework, we use typical total variation (TV) priors for the depth maps and latent images. Although originally introduced for denoising signals, TV priors has been frequently used in addressing image deblurring [29], super-resolution [9], and stereo matching problems [18].

The TV priors used in the proposed method is defined as follows:

$$E_t^r = \lambda_d \sum_{\mathbf{x}} g_t(\mathbf{x}) \left\| \nabla D_t(\mathbf{x}) \right\|_2 + \lambda_i \sum_{\mathbf{x}} \left\| \nabla I_t(\mathbf{x}) \right\|_2, \tag{11}$$

where $\nabla I(\mathbf{x})$ represents the gradient value of image I at pixel $\mathbf{x}$. The weighting function $g_t(\mathbf{x})$ is used for edge-preserving smoothness with the same definition as proposed in [11]. We use the magnitude of L2-norm to make the TV priors isotropic while preserving the discontinuities in the images and depth maps. The parameters λ_d and λ_i determine the degree of regularization on the depth maps and latent images, respectively.

5. Optimization

The optimization of Equation (7) is a complex process that serves as a function of many variables (D_t's, $\mathbf{P}_t$'s, and I_t's for all frames). This process is also highly nonlinear because of the warping operations. Therefore, instead of obtaining the global optimum, we attempt to secure a favorable approximated solution by adopting several strategies. At the core of this solution is a divide-and-conquer strategy or an iterative and alternating optimization of variables. The proposed framework uses two-phase iterations in which the

Algorithm 1 The overall optimization procedure

```
% initialization
for t = 1 to T do
    Initialize D_t, P_t by minimizing Equation (15)
end for
% main loop
for iteration = 1 to max_iter do
    % first phase : update images
    for t = 1 to T do
        update I_t by minimizing Equation (14)
    end for
    % second phase : update depths and cameras
    approximate Equation (7) using Equation (12)
    update D_t's and P_t's by using IRLS
end for
```

structures (cameras and depth maps) and latent images are alternatingly updated.

Algorithm 1 presents the overall optimization procedure. T denotes the number of frames in the input image sequence, while *max_iter* denotes the number of iterations set by users. The solutions almost converge after three iterations, which is the chosen *max_iter* value of the proposed method.

5.1. Update of the depth maps and camera poses

In the first phase of each iteration, we optimize the variables on the scene structure, D_t's and $\mathbf{P}_t$'s, with the fixed latent images, I_t's. The energy function then becomes similar to that of the variational framework for optical flow [22] and we follow the optimization strategy employed in [22]. At each iteration of this iterative optimization, the functions in the L1-norm for Equations (9) and (10) are approximated using the first-order Taylor expansion at the current solution.

The linear approximation is conducted by calculating the partial derivatives of the warping equation in terms of individual depth value and camera pose as parameterized by the six-dimensional vector on $\mathfrak{se}(3)$. Suppose that the current solution of our iterative algorithm lies at a point in the solution space, D_t^0, $\mathbf{P}_t^0$, and $\mathbf{P}_s^0$. The backward image warping procedure from the frame at time s to t can be approximated as follows:

$$I_s^0(\mathbf{x}) = I_s\left(W^{t\to s}(\mathbf{x})\right)\big|_{D_t=D_t^0, P_t=P_t^0, P_s=P_s^0},$$

$$\begin{aligned} &I_s\left(W^{t\to s}(\mathbf{x})\right) \\ &= I_s^0(\mathbf{x}) + \frac{\partial I_s^0}{\partial \mathbf{u}}\left(\frac{\partial \mathbf{u}}{\partial D_t(\mathbf{x})}\Delta D_t(\mathbf{x}) + \frac{\partial \mathbf{u}}{\partial \varepsilon_t}\varepsilon_t + \frac{\partial \mathbf{u}}{\partial \varepsilon_s}\varepsilon_s\right), \end{aligned} \tag{12}$$

where $\mathbf{u}$ is the warping-generated flow that serves as a function of the depth and camera parameters. The partial derivatives are actually Jacobians [6].

$\Delta D_t(\mathbf{x})$, ϵ_t, and ϵ_s are variables that contribute to the solution as follows:

$$\begin{aligned} D_t(\mathbf{x}) &= D_t^0(\mathbf{x}) + \Delta D_t(\mathbf{x}), \\ \mathbf{P}_t &= \exp(\varepsilon_t)\,\mathbf{P}_t^0, \\ \mathbf{P}_s &= \exp(\varepsilon_s)\,\mathbf{P}_s^0. \end{aligned} \tag{13}$$

Given that all terms in the L1-norm have been linearized, these variables can be efficiently estimated using the simple iteratively reweighted least square (IRLS) method [19].

5.2. Update of the latent images

The latent images are optimized in the second phase of the outer loop. The L1-norm functions for the target image I_t in the matching term, Equation (9), provides information about the different blur and sampling of latent image I_t. The self-consistency term in Equation (10) and the smoothness imposed by the regularization term in Equation (11) are considered to provide a frame-by-frame representation of the energy function on I_t as follows:

$$\begin{aligned} &\sum_{s\in N(t)}\sum_{\mathbf{x}\in\Omega_{ts}} \left\|\Psi_s \circ I_t\left(W^{s\to t}(\mathbf{x})\right) - B_s(\mathbf{x})\right\|_1 \\ &+\lambda_s \sum_{\mathbf{x}} \left\|\Psi_t \circ I_t(\mathbf{x}) - B_t(\mathbf{x})\right\|_1 \\ &+\lambda_i \sum_{\mathbf{x}} \left\|\nabla I_t(\mathbf{x})\right\|_2, \end{aligned} \tag{14}$$

which is optimized by finding the most plausible values that satisfy these competing constraints simultaneously.

We apply bilinear interpolation to sample the color values of non-grid points in image warping, and then apply simple box filtering for downsampling in the capturing operation. This process makes the warping and capturing operations act as linear operators on the latent image after fixing the depth maps and camera poses. Consequently, the Equation (14)denotes the sum of L1-norm and L2-norm on the linear functions of I_t that can be easily optimized using IRLS [19].

5.3. Initialization

We initialize the camera poses of the first two frames using a structure from motion software [27, 28]. After determining the camera poses of the first two frames, the depth maps D_t's and remaining camera poses P_t's can be initialized by sequentially minimizing the following equation frame-by-frame in a coarse-to-fine manner [22]:

$$\begin{aligned} E_t^{init} = &\sum_{\mathbf{x}} \left\|\left(\Psi_t \circ B_s\right)\left(W^{t\to s}(\mathbf{x})\right) - \Psi_s \circ B_t(\mathbf{x})\right\|_1 \\ &+ \lambda_d \sum_{\mathbf{x}} \left\|\nabla D_t(\mathbf{x})\right\|_2, \end{aligned} \tag{15}$$

where s denotes the time of the previous frame. Given that the estimated depth maps have LR, we upsample these maps

Table 1: The performance comparison of deblurring performance for synthetic datasets. All the PSNR(dB) values are averaged for the whole frames in each sequence.

Methods	**Dolls**	**Reindeer**	**InteriorScene** [1]	**WorkDesk** [2]	**avg.**
Bicubic interpolation (Bic.)	23.52	29.54	26.82	20.45	25.08
Bic. + Lee and Lee[14]	11.17	22.52	15.19	10.88	14.94
Timofte *et al*. [24] + Lee and Lee[14]	10.60	16.71	13.29	9.74	12.59
Wang *et al*. [25] + Lee and Lee[14]	11.07	21.44	15.08	10.86	14.61
Bic. + Xu *et al*. [30]	22.47	26.88	26.43	19.77	23.89
Timofte *et al*. [24] + Xu *et al*. [30]	19.68	22.66	23.52	17.71	20.89
Wang *et al*. [25] + Xu *et al*. [30]	22.62	27.00	26.40	19.71	23.93
Bic. + Kim and Lee[11]	25.96	31.03	28.55	24.23	27.44
Timofte *et al*. [24] + Kim and Lee[11]	22.41	24.20	25.82	20.51	23.23
Wang *et al*. [25] + Kim and Lee[11]	26.11	31.56	28.65	24.18	27.63
Kim and Lee[11] + Wang *et al*. [25]	25.56	29.86	28.39	23.84	26.91
Xu *et al*. [30] + Wang *et al*. [25]	21.24	24.10	24.82	18.33	22.12
Proposed(w/o SR) + Bic.	27.33	31.11	22.48	22.17	25.77
Bic. + Proposed(w/o SR)	26.92	30.97	27.73	24.71	27.58
Proposed	**28.39**	**32.48**	**29.06**	**25.29**	**28.81**

Table 2: Depth and camera pose estimation performance of synthetic datasets. The errors are measured using PSNR and relative errors (rel.) for depth, and absolute trajectory error (e_{ate}) for pose [8]. All errors are averaged for the whole frames in each sequence.

Datasets	**Methods**	**Depth errors**		**Pose errors**
		PSNR(dB)	**rel.**	**traj.**(e_{ate})
Dolls [10]	Bic. + Lee *et al*. [13]	-	-	0.1220
	Bic. + Lee and Lee [14]	19.76	0.6700	-
	Bic. + Baseline	41.79	0.0560	0.0046
	[25] + [30] + Baseline	40.51	0.0676	0.0028
	[25] + [11] + Baseline	41.70	0.0568	0.0078
	Proposed(w/o SR) + Bic.	43.47	0.0396	**0.0005**
	Bic. + Proposed(w/o SR)	43.50	0.0375	0.0011
	Proposed	**45.37**	**0.0336**	0.0027
Reindeer [10]	Bic. + Lee *et al*. [13]	-	-	0.0107
	Bic. + Lee and Lee [14]	23.00	0.4982	-
	Bic. + Baseline	37.79	0.1084	0.0021
	[25] + [30] + Baseline	37.23	0.2026	0.0022
	[25] + [11] + Baseline	37.72	0.1099	0.0036
	Proposed(w/o SR) + Bic	36.52	0.1321	0.0005
	Bic. + Proposed(w/o SR)	37.41	0.1143	**0.0005**
	Proposed	**37.99**	**0.1055**	0.0012
InteriorScene [1]	Bic. + Lee *et al*. [13]	-	-	1.9355
	Bic. + Lee and Lee [14]	23.15	0.4641	-
	Bic. + Baseline	30.82	0.1647	0.1548
	[25] + [30] + Baseline	30.93	0.1627	0.1288
	[25] + [11] + Baseline	30.41	0.1812	0.0923
	Proposed(w/o SR) + Bic	21.26	0.5253	0.0974
	Bic. + Proposed(w/o SR)	30.19	0.1802	**0.0281**
	Proposed	**31.28**	**0.1617**	0.1461
WorkDesk [2]	Bic. + Lee *et al*. [13]	-	-	2.8334
	Bic. + Lee and Lee [14]	26.01	0.4411	-
	Bic. + Baseline	36.85	0.0949	0.1392
	[25] + [30] + Baseline	36.23	0.1057	0.1950
	[25] + [11] + Baseline	30.82	0.2479	0.4953
	Proposed(w/o SR) + Bic	36.16	0.1031	0.3481
	Bic. + Proposed(w/o SR)	**39.90**	**0.0544**	**0.0914**
	Proposed	38.13	0.0781	0.5048

to match the resolution of the target latent images, and then begin the main loop of the optimization. We adopt a simple bicubic interpolation method for the upsampling.

5.4. Occlusion Handling

Although the use of L1-norm for the matching term in Equation (9) makes the proposed method robust to existence of occlusion, modeling the visible area in Ω_{ts} can help generate precise depth values around the discontinuities. Therefore, we update the visible area Ω_{ts} whenever the depth maps and camera poses are updated. Given the updated depth maps and camera poses, we update Ω_{ts} as follows:

$$\Omega_{ts} = \left\{ \mathbf{x} \left| D_t\left(\mathbf{x}\right) > D_t\left(\mathbf{y}\right), \forall \mathbf{y} \in \Theta_{ts}\left(\mathbf{x}\right) \right. \right\}, \quad (16)$$

where Θ_{ts} represents the set of pixels in the camera at time t that fall in the same area after warping.

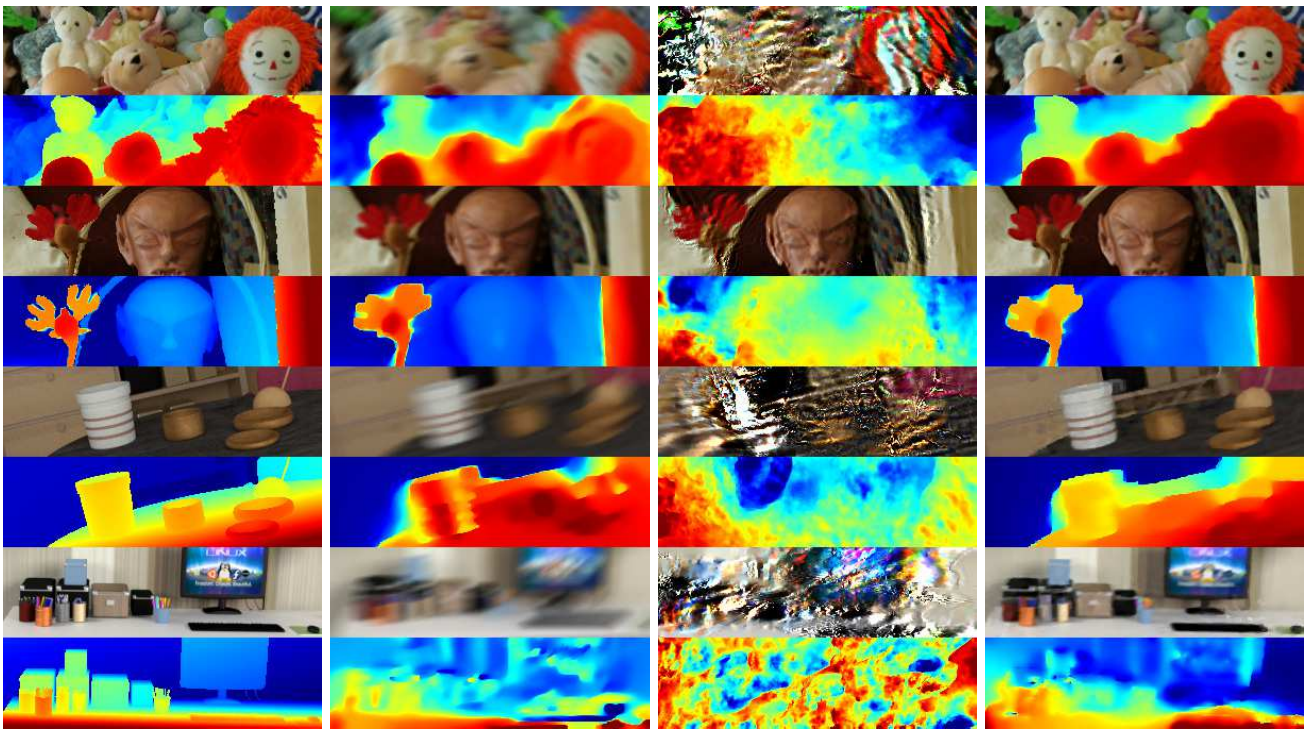
$$\Theta_{ts}\left(\mathbf{x}\right) = \left\{ \mathbf{y} \left| \left| W^{t\to s}\left(\mathbf{y}\right) - W^{t\to s}\left(\mathbf{x}\right) \right| \leq 0.5 \right. \right\}. \quad (17)$$

6. Experimental Results

We test the validity of our proposed method on synthetic and real datasets. For comparison, we use the simple variational matching method as the baseline. This method solves the same optimization problem as the proposed method, except that the capturing operations are missed in the energy terms and the images are fixed to input images.

The values of some parameters are empirically determined. The proposed algorithm converges to favorable solutions when *max_iter* is 3 and M is 50. We use a large value of λ_s (30) for all datasets to provide strong constraints on the solutions. We tune the value of λ_d between 8 to 10 and the value of λ_i between 0.3 to 0.6 based on the dataset. We set the upscale factor of the method to 2.

Our proposed framework is limited by its computational complexity. Specifically, we spend approximately five



(a) Ground truth (b) Bic. + baseline (c) Bic. + [14] (d) Proposed

Figure 3: Comparison of the depth maps and latent images of synthetic datasets. Each pair of rows shows results on **Dolls** [10], **Reindeer** [10], **InteriorScene** [1], and **WorkDesk** [2] dataset from top to bottom .

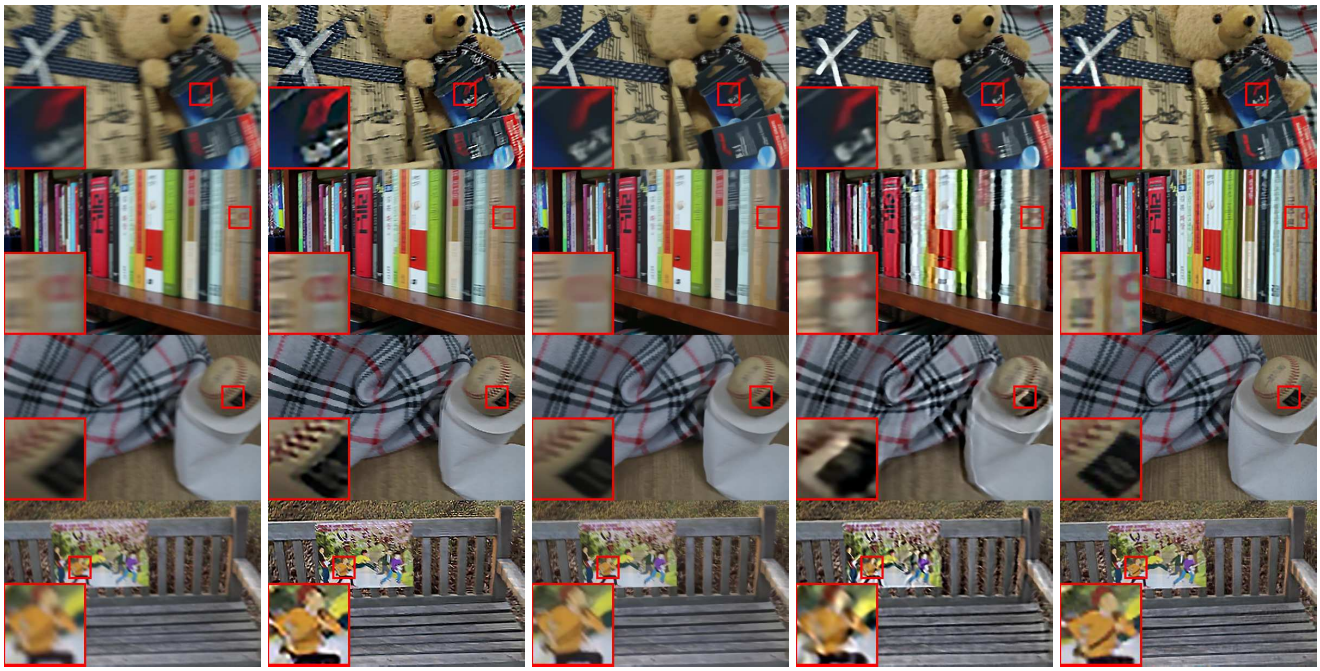
hours to process 10 frames of 320×240 images in our Matlab implementation using a quad-core 3.2GHz CPU. This complexity may be addressed by running many parts of the algorithm on GPU in parallel.

6.1. Synthetic datasets

No public datasets provide blurry LR images with corresponding ground-truth latent images, depth maps, and camera poses. The desired datasets can be generated by synthesizing a simulated blur sequence using the Middlebury stereo datasets [10]. Given two images with ground-truth depth maps, the images between these two viewpoints are interpolated by assuming an imaginary camera path between the two reference views. Afterward, with a preset imaginary shutter time, the blurry images in each frame are approximated by summing up the intermediate images while the shutters are open. Similarly, we can generate a more realistic dataset using Blender [3]. The intermediate images for these datasets are accurately rendered by using full 3D models. Figure 3 presents examples of synthesized datasets with their corresponding experimental results.

Table 2 and Table 1 present the quantitative comparison results. Table 2 shows the quantitative results of depth and camera pose estimation, while Table 1 compares the deblurring results in terms of peak signal-to-noise ratio (PSNR). The depth and image estimation errors are measured by comparing the reconstructed results with the closest intermediate sharp ground-truth ones (followed by scaling to address scale ambiguity for the case of depth maps). When we compute the depth errors for **Dolls** and **Reindeer** datasets [10], we cropped the depth maps to be the 70% of original size at image center to ignore invalid regions around image boundary introduced by warping.

The third, fourth, and fifth rows of Table 2 show that using per-frame super-resolution and deblurring independently before matching may degrade the stereo matching performance as expected. The method proposed by Lee and Lee [14] and Lee *et al*. [13] performs worse than the baseline despite employing blur-aware matching. This result may be attributed to the fact that the degree of blur in our experiments is much more severe than that in the datasets used in [13, 14] and, furthermore, the scene structures in our datasets are more complex than the nearly planar structures in [13, 14]. The pose estimation performance of the proposed method seems less impressive compared to the depth estimation performance. However, comparing the ground-truth trajectory to the restored camera trajectory itself can be problematic for blurry input images because it is ambiguous to specify a camera pose during the shutter time, especially when the the size of blur kernel is large as in **In-**



(a) Bicubic interpolation (b) [30] + [24] (c) HR images + [7] (d) [25] + [11] (e) Proposed

Figure 4: Comparison of the deblurring results on real datasets.

teriorScene [1] and **WorkDesk** [2] datasets. By contrast, the depth estimation errors are more significant because we can find the closest intermediate ground-truth depth map for a estimated depth map without ambiguity.

Table 1 shows that the proposed method outperforms the combination of super-resolution methods [24, 25] and deblurring method [14, 30, 11], which implies that jointly estimating inter-related problems is effective in terms of image restoration. The method proposed in [11] also joinlty estimates the pixel correspondences (optical flow) and latent images (deblurring) from a video sequence. Still, Table 1 shows that leveraging multi-view constraints for deblurring problem and jointly solving it with super-resolution is more profitable.

The use of super-resolution clearly improves the accuracy of image restoration and depth estimation except for the case of **WorkDesk** [2] dataset. The surfaces in this dataset are weakly-textured and repeated, making the pixel-wise matching and super-resolution results less reliable.

6.2. Real datasets

For the real datasets, we use the proposed approach in [31] for camera calibration. The shutter time and frames per second (FPS), both of which are necessary for interpolating the camera path and for simulating blurs for each frame, are obtained as metadata by taking an image sequence using commercial cameras.

Figure 4 presents the comparison results of our proposed approach with those of other image restoration methods [30, 24, 7, 25, 11]. Given that the images are blurred by real camera motions, we generate LR images by downsampling them manually to compare the super-resolution performances of these methods. The proposed method clearly outperforms the others even if the results of [7] are obtained using the original HR images. Some characters become recognizable and the textures representing the materials in the scene are well-restored in our results. Figure 4b also shows that performing the super-resolution after deblurring results in exaggeration of undesired artifacts, explaining the low PSNR values of 'super-resolution after deblurring' approaches in Table 1.

7. Conclusion

We proposed a pioneering framework for jointly solving four inter-related computer vision problems, including dense depth reconstruction, camera pose estimation, super-resolution, and deblurring. We jointly modeled these problems using an energy function that is derived by revisiting the blurry image formulation. Our model allows more general camera motions and nonlinear blur kernels than the previously proposed blur-aware matching methods. Our experiments show that the proposed method outperforms the other related methods that only address one or two target problems in terms of depth maps and latent images.

References

[1] http://www.blendswap.com/blends/view/72340/. 6, 7, 8

[2] http://www.blendswap.com/blends/view/69052/. 6, 7, 8

[3] http://www.blender.org/. 7

[4] B. Bascle, A. Blake, and A. Zisserman. Motion deblurring and super-resolution from an image sequence. *Proceedings of the European Conference on Computer Vision*, pages 571–582, 1996. 3

[5] A. Bhavsar and A. Rajagopalan. Resolution enhancement in multi-image stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1721–1728, Sept 2010. 1, 2

[6] J.-L. Blanco. A tutorial on se(3) transformation parameterizations and on-manifold optimization. Technical report, University of Malaga, Sept. 2010. 2, 3, 5

[7] S. Cho, J. Wang, and S. Lee. Video deblurring for hand-held cameras using patch-based synthesis. *ACM Transactions on Graphics (TOG)*, 31(4):64, 2012. 8

[8] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. In *arXiv:1607.02565*, July 2016. 6

[9] S. Farsiu, M. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344, Oct 2004. 4

[10] H. Hirschmuller and D. Scharstein. Evaluation of cost functions for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007. 1, 6, 7

[11] T. Hyun Kim and K. Mu Lee. Generalized video deblurring for dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5426–5434, 2015. 4, 6, 8

[12] H. Jin, P. Favaro, and R. Cipolla. Visual tracking in the presence of motion blur. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 18–25 vol. 2, June 2005. 2

[13] H. S. Lee, J. Kwon, and K. M. Lee. Simultaneous localization, mapping and deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1203–1210, Nov 2011. 2, 6, 7

[14] H. S. Lee and K. M. 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*, pages 273–280, June 2013. 1, 2, 6, 7, 8

[15] H. S. Lee and K. M. Lee. Simultaneous super-resolution of depth and images using a single camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 281–288, June 2013. 1, 2

[16] C. Mei and I. Reid. Modeling and generating complex motion blur for real-time tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008. 2

[17] T. Portz, L. Zhang, and H. Jiang. Optical flow in the presence of spatially-varying motion blur. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1752–1759, June 2012. 1, 2

[18] R. Ranftl, S. Gehrig, T. Pock, and H. Bischof. Pushing the limits of stereo using variational stereo estimation. In *IEEE Intelligent Vehicles Symposium*, pages 401–407, June 2012. 4

[19] J. A. Scales and A. Gersztenkorn. Robust methods in inverse theory. *Inverse problems*, 4(4):1071, 1988. 5

[20] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–195–I–202 vol.1, June 2003. 1

[21] A. Sellent, C. Rother, and S. Roth. Stereo video deblurring. In *Proceedings of the European Conference on Computer Vision*, pages 558–575. Springer, 2016. 2

[22] D. Sun, S. Roth, and M. Black. Secrets of optical flow estimation and their principles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2432–2439, June 2010. 5

[23] H. Takeda and P. Milanfar. Removing motion blur with space–time processing. *IEEE Transactions on Image Processing*, 20(10):2990–3000, 2011. 3

[24] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *IEEE Asian Conference on Computer Vision*, 2014. 1, 6, 8

[25] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deep networks for image super-resolution with sparse prior. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 370–378, 2015. 1, 6, 8

[26] A. Wedel, T. Pock, C. Zach, D. Cremers, and H. Bischof. An improved algorithm for TV-L1 optical flow. In *Proc. of the Dagstuhl Motion Workshop*, LNCS. Springer, September 2008. 4

[27] C. Wu. Visualsfm: A visual structure from motion system. http://ccwu.me/vsfm/, 2011. 5

[28] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3057–3064. IEEE, 2011. 5

[29] L. Xu and J. Jia. Two-phase kernel estimation for robust motion deblurring. In *Proceedings of the European Conference on Computer Vision*, ECCV'10, pages 157–170, Berlin, Heidelberg, 2010. Springer-Verlag. 4

[30] L. Xu, S. Zheng, and J. Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1107–1114, June 2013. 1, 2, 6, 8

[31] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pages 666–673 vol.1, 1999. 8

[32] R. Zhen and R. L. Stevenson. Motion deblurring and depth estimation from multiple images. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 2688–2692. IEEE, 2016. 2