# Multi-image blind super-resolution of 3D scenes

Abhijith Punnappurath*, T. M. Nimisha, A. N. Rajagopalan, *Senior Member, IEEE*

EDICS: TEC-ISR – Interpolation, Super-Resolution, and Mosaicing

***Abstract*—We address the problem of estimating the latent high-resolution (HR) image of a 3D scene from a set of non-uniformly motion blurred low-resolution (LR) images captured in the burst mode using a hand-held camera. Existing blind super-resolution (SR) techniques that account for motion blur are restricted to fronto-parallel planar scenes. We initially develop an SR motion blur model to explain the image formation process in 3D scenes. We then use this model to solve for the three unknowns – the camera trajectories, the depth map of the scene and the latent HR image. We first compute the global HR camera motion corresponding to each LR observation from patches lying on a reference depth layer in the input images. Using the estimated trajectories, we compute the latent HR image and the underlying depth map iteratively using an alternating minimization framework. Experiments on synthetic and real data reveal that our proposed method outperforms state-of-the-art techniques by a significant margin.**

***Index Terms*—Non-uniform blur, super-resolution, depth map.**

## I. Introduction

Super-resolution (SR) algorithms employ signal processing techniques to recover a high-resolution (HR) image from a set of low-resolution (LR) images. Their study is of high contemporary relevance since they offer a cheap and attractive means to retrieve high quality images from low-resolution observations without the use of additional hardware. The basic principle of multi-image SR is that downsampled (aliased) subpixel shifted LR images provide new information that can be utilized to reconstruct the HR image [1].

Traditional SR algorithms [2], [3] assume that the camera is stationary during the exposure time itself, and that the shift or motion is only *between* one LR image to the next i.e., the only blurriness in the captured images is due to downsampling, and the blurring process is known a priori. However, camera shake is a common occurrence in hand-held imaging devices such as cell phones which have now become ubiquitous. Motion of the camera *during* the exposure duration manifests as *motion blur* in the acquired image. In such situations, super-resolution makes little sense without compensating for the effect of the unknown motion blur.

The class of algorithms that estimate the unknown blur in addition to the HR image are called blind super-resolution algorithms. The critical part of such algorithms is precise estimation of the blur, which, in the context of camera shake, depends on the kind of motion the camera undergoes during exposure, and the nature of the scene being imaged. Prior works that have addressed the blind SR problem ([4], [5] for instance) assumed that the images are uniformly blurred. Sroubek et al. [4] constrained the motion of the camera to pure in-plane translations and assumed a flat constant-depth scene. This allowed them to model the blur as a convolution with an unknown kernel or point spread function (PSF), and the parameters of this unknown PSF had to be estimated. A more recent work [6] too assumes a flat scene but allows for space-varying blur due to general motion of the camera during exposure. In such cases, the convolution model with a single kernel for the entire image is no longer applicable. To tackle this challenge, a projective motion blur model based on homographies was used in [6]. However, this global homography model breaks down if there are depth variations in the scene because homographies apply only to planes. Thus, the task of 3D blind SR from non-uniformly blurred LR images is severely ill-posed because there is now an added third unknown to be solved for – the underlying depth map of the scene – in addition to the camera motion and the HR image. To the best of our knowledge, no algorithms for resolution enhancement exist if the motion blurred images are of a 3D scene, and it is this problem that we address in this work.

### A. Related works

Deblurring and super-resolution, though two extensively studied topics, have mostly been dealt with independently. We first briefly review single/multi-image blind deblurring algorithms for planar and 3D scenes. We also look at traditional multi-image SR techniques that do not consider motion blur. We would like to add that our survey is not exhaustive since there are hundreds of papers on these two topics; we mention below only some of the most influential works in these areas. Finally, we undertake a careful scrutiny of the few blind SR approaches in the literature that are most closely connected to our work.

**Deblurring:** A lot of papers exist in the literature that focus on the issue of removing motion blur due to camera shake from images. Traditionally, the blurred image resulting from camera shake has been modeled as the convolution of the latent sharp image with a blur kernel [7], [8], [9]. This model assumes that the camera undergoes only in-plane translations, and the scene is fronto-parallel planar. The seminal work of Fergus et al. [7] attempted to solve the single image blind deblurring problem by applying a sparsity prior on the PSF, and enforcing the image gradients to follow a heavy-tailed distribution. Shan et al. [8] addressed the same issue by introducing a

Abhijith Punnappurath, T. M. Nimisha, and A. N. Rajagopalan are with the Department of Electrical Engineering, Indian Institute of Technology, Chennai 600 036, India. E-mail: jithuthatswho@gmail.com; ee13d037@ee.iitm.ac.in; raju@ee.iitm.ac.in.

local smoothness prior to reduce ringing artifacts, while Xu and Jia [9] proposed a guided edge selection strategy that detects large-scale structures and subdues small edges not useful for kernel refinement. However, more recent deblurring algorithms [10], [11], [12], [13] allow for general motion of the camera since it is now well-established that tilts and rotations occur frequently in hand-held imagery [10]. These techniques typically model the motion blurred image as an average of projectively warped instances of the latent sharp image assuming a flat constant-depth scene. The deblurring schemes proposed by Hu and Yang [10], Gupta et al. [11], Hirsch et al. [12] consider the camera motion to be comprised only of in-plane translations and rotations. Hu and Yang [10] use locally estimated PSFs to constrain the possible camera poses to a low-dimensional subspace. Gupta et al. [11] model the camera motion as a motion density function, while Hirsch et al. [12] employ an efficient filter flow framework for blur removal. It is important to note that the above methods assume a fronto-parallel scene with constant depth. On the other hand, Whyte et al. [13] approach the non-uniform blind deblurring task using a depth-independent rotational model where the blurring function is represented on a 3D grid corresponding to the three directions of camera rotations. The only work, to our knowledge, to jointly estimate depth and remove non-uniform blur caused by camera motion is that of Hu et al. [14]. However, their matte-based approach requires manual intervention in the form of a scribble for each depth layer from the user.

TABLE I
Overview of related works.

| | References | Inputs | | 3D | SV |
|---|---|---|---|---|---|
| | | S | M | | |
| BD | [7], [8], [9] | ✓ | | × | × |
| | [10], [11], [12] | ✓ | | × | ✓ |
| | [13]#,[14] | ✓ | | ✓ | ✓ |
| | [15], [16], [17] | | ✓ | × | × |
| | [18], [19] | | ✓ | ✓ | ✓ |
| SR | {[20], [21], [22]}# | ✓ | | ✓ | NA |
| | [3], [23], [24], [25] | | ✓ | × | |
| | [26], [27], [28] | | ✓ | ✓ | |
| MBSR | [4], [5] | | ✓ | × | × |
| | [6] | | ✓ | × | ✓ |
| | Proposed | | ✓ | ✓ | ✓ |

The deblurring problem is ill-posed if there is only a single input observation, and is difficult to solve in a fully blind form. The above methods do not exploit the potential of the multi-image framework, where missing information about the latent image in one observation is supplemented by information in the other observations. Sroubek and Flusser [15] assume a fronto-parallel planar scene and solve the multi-image blind deblurring problem for the case of pure in-plane translational camera motion using a variational prior. Under similar assumptions of a planar constant depth scene, Delbracio and Sapiro [16] and Ito et al. [17] leverage the burst mode feature in cameras to obtain a deblurred result from multiple images. Paramanand and Rajagopalan [18] tackle the bilayer case comprising of a foreground and a background, and additionally allow for in-plane camera rotations. Lee and Lee [19] have proposed a blur-aware algorithm for reconstructing 3D scenes in which the blur kernel and the depth of each pixel are simultaneously estimated. However, their method assumes knowledge of the camera parameters. Although multi-image blind deblurring algorithms require little or no prior information about the blurs, they can hardly cope with the downsampling operator in the SR model.

**Super-resolution:** A large number of papers have addressed the classical multi-image SR problem when the images are not motion blurred. A good survey can be found in Park et al. [1]. Maximum likelihood, maximum a posteriori (MAP), the set theoretic approach using projection on convex sets, and fast Fourier techniques have all been shown to provide a solution to the SR problem. Spatial-domain SR methods are preferred over frequency-domain approaches since they can incorporate complex image priors for regularization [1]. Farsiu et al. [3] proposed a robust SR framework using $l_1$-norm minimization and bilateral filtering. Employing a variational Bayesian analysis, an algorithm for joint image registration and super-resolution has been proposed in [23]. This work was later improved in [24] using a combination of sparse and non-sparse image priors. A coordinate-descent approach for simultaneous global registration and multi-image SR has been mooted in [25]. It must be noted that these methods assume a fronto-parallel planar scene. Mudenagudi et al. [26] approach the issue of super-resolution of 3D scenes using a MAP-MRF framework, while Bhavsar and Rajagopalan [27] present an integrated strategy to estimate the HR depth and the SR image from multiple LR stereo observations. Lee and Lee [28] integrate depth map estimation and image super-resolution into a single energy minimization framework with a convex cost function. Example-based SR (also termed 'image hallucination') techniques [20], [21] that seek an HR image from a *single* LR image have also been proposed. These methods employ a database of LR and HR image pairs to learn correspondences between LR and HR image patches. When a new LR image is presented, its most likely HR version is recovered based on these learned patch correspondences. However, these techniques are known to hallucinate HR details that may not be present in the true HR image. Based on the observation that patches in a natural image tend to recur within the same image, both at the same as well as at different scales, Glasner et al. [22] sought to combine the strengths of traditional multi-image SR as well as example-based SR. It is important to note that state-of-the-art SR techniques achieve remarkable results of resolution enhancement only when there is no motion blur.

**Blind super-resolution from motion blurred LR images:** Sroubek et al. [4] take on the blind SR problem by building a regularized energy function and minimizing it alternately with respect to the original HR image and the camera motion. The method of Ma et al. [5] is based on the premise that the same region is not equally blurred across frames. They propose a temporal region selection scheme to select the least blurred pixels from each frame. Both these approaches, however, assume that the images are uniformly blurred. Zhang and Carin [6] allow for space-varying blur due to general camera motion and present a joint formulation for the tasks of alignment, deblurring and resolution enhancement. Note that all three of the above methods [4], [5], [6] assume that the scene is flat and at a constant depth from the camera. The blind super-resolution problem is extremely under-determined when there are depth variations, and there are no published algorithms that increase the image resolution if the blurred observations are of a 3D scene. A succinct overview of the works discussed thus far is given in Table I, where the following notations have been used – BD: blind deblurring, MBSR: multi-image blind super-resolution, S: single image, M: multiple images, and SV: space-varying blur. A '×' entry in the column '3D' denotes that only planar scenes can be handled, while a '✓' signifies that both 3D and planar scenes can be modeled. Likewise, a '×' entry in the column 'SV' indicates that only space-invariant blur can be accounted for, whereas a '✓' implies both space-invariant and space-varying blur can be dealt with. The 'NA' (not applicable) entry is because traditional SR algorithms do not model blur. The '#' symbol applies to methods that employ a depth independent model (i.e., do not explicitly solve for the depth map).

**Our proposed method:** The focus of this paper is on the problem of recovering the latent HR image of a 3D scene given multiple low-resolution observations that are non-uniformly motion blurred due to camera shake during exposure. The burst mode feature available in almost all modern digital cameras, including cellphones, point-and-shoots, and SLRs, allows the user to take a sequence of images in quick succession with a single click. Images captured thus will have negligible change in viewpoint since the motion is only due to incidental camera shake. In this work, we consider input observations captured using the burst mode since such images will not have large registration errors. In addition, the narrow baseline ensures that occlusion and parallax effects at depth boundaries are not very large. We initially review the super-resolution motion blur image formation model for fronto-parallel planar scenes, and then propose an elegant extension to the 3D case using a layered approach. Using this observation model, we propose an algorithm to recover the latent HR image of the scene, the underlying depth map and the associated HR camera trajectories from the input LR observations. We leverage the inter-image misalignment that results from capturing images hand-held to first coarsely segment the scene into different depth layers. This is achieved by running an optical flow (OF) algorithm on a carefully selected image pair from the set of input observations, and computing the magnitude of the flow vectors to reveal the depth map. Small patches lying on a constant depth layer in this depth map are extracted from the LR images, and the global HR camera motion is computed using only the information in these local patches. Such a patch-based approach allows us to circumvent the need to solve for all three unknowns jointly. Once the HR camera trajectories have been estimated, we iteratively solve for the latent HR image and the depth map using an alternating minimization (AM) framework. Judiciously chosen priors on the image and the depth map ensure that our AM scheme converges within just a few iterations.

To summarize, the main contributions of this paper are:

- This is the first attempt to formally address the problem of estimating the latent HR image of a 3D scene given a set of LR observations that are non-uniformly blurred due to camera shake during image capture.
- We advocate a 3D super-resolution motion blur model to explain the image formation process, and an algorithm based on this model to recover the underlying HR image.
- We propose an elegant patch-based approach to compute the global HR camera motion directly using only local information in the LR images.
- We also develop an alternating minimization framework to jointly recover the latent HR image and the depth map of the scene.

The organization of the rest of the paper is as follows: we describe our 3D super-resolution motion blur model in Section II. We initially begin by considering a planar scene and later extend it to the 3D case. In Section III, we first discuss in detail how the camera motion at HR is estimated from the LR images. Next, we elaborate on our alternating minimization scheme to jointly recover the latent HR image and the depth map of the scene. Section IV contains results of the proposed method on synthetic and real data, along with comparisons with state-of-the-art techniques. Section V concludes the paper.

## II. The 3D super-resolution motion blur model

In this section, we first discuss the super-resolution motion blur model for fronto-parallel planar scenes. Later, we generalize our framework to 3D scenes using a layered approach.

### A. Planar scene

Let us initially consider a static constant-depth planar scene imaged using a hand-held camera. When the camera motion is not restricted to pure in-plane translations, the convolution model with a single blur kernel does not hold because the apparent motion of scene points in the image will vary at different locations resulting in space-variant blur. In such a scenario, the projective motion blur model [11], [13], [29], [30] can be used to represent the blurred image resulting from camera shake as a weighted average of *warped* instances of the latent sharp image. In the context of super-resolution, this extends to modeling the blurred LR image as a *downsampled* version of the weighted average of warped instances of the latent HR image. In the discrete domain, the operation of blurring and downsampling can be represented by the following equation

$$\mathbf{g} = \mathbf{D}\left(\sum_{\mathbf{c}_l \in \mathcal{C}} \omega_{\mathbf{c}_l} \mathbf{H}_{\mathbf{c}_l} \mathbf{f}\right) + \mathbf{n}. \tag{1}$$

Here $\mathbf{g}$ denotes the blurred LR observation, while $\mathbf{f}$ represents the latent HR image of the scene as viewed by a camera placed at the origin of the world coordinate system. The vector $\mathbf{g} \in \mathbb{R}^{M_1 M_2 \times 1}$ is the lexicographically ordered version of the 2D discrete LR image $\mathbf{G} \in \mathbb{R}^{M_1 \times M_2}$, where $M_1$ and $M_2$ denote the number of LR rows and columns, respectively. Likewise $\mathbf{f} \in \mathbb{R}^{N_1 N_2 \times 1}$ is the lexicographically ordered vector version of $\mathbf{F} \in \mathbb{R}^{N_1 \times N_2}$, where $N_1$ and $N_2$ indicate the rows and columns at HR, respectively, and $N_1/M_1$ and $N_2/M_2$ are the downsampling factors along the two directions. The parameter $\omega$ depicts the camera motion i.e., for each camera pose $\mathbf{c}_l \in \mathcal{C}$, the scalar $\omega_{\mathbf{c}_l}$ denotes the fraction of the total exposure duration for which the camera stayed in the pose $\mathbf{c}_l$. The discrete camera pose space $\mathcal{C}$, on which $\boldsymbol{\omega}$ is defined, is the finite set of sampled camera poses that the camera is free to undergo i.e., $\mathcal{C} = \{\mathbf{c}_l\}_{l=1}^{|\mathcal{C}|}$, where $|\cdot|$ represents cardinality and $\boldsymbol{\omega}$ denotes the vector of weights $\omega_{\mathbf{c}_l}$, $\mathbf{c}_l \in \mathcal{C}$. The pose space is discretized in such a way that the difference in the displacements of a point light source due to two different camera poses from the discrete set $\mathcal{C}$ is at least one pixel. Akin to a PSF, $\sum_{\mathbf{c}_l \in \mathcal{C}} \omega_{\mathbf{c}_l} = 1$ and $\omega_{\mathbf{c}_l} \geq 0$. The matrix $\mathbf{H}_{\mathbf{c}_l} \in \mathbb{R}^{N_1N_2 \times N_1N_2}$ warps $\mathbf{f}$ according to the camera pose $\mathbf{c}_l$, and the vector $\mathbf{n} \in \mathbb{R}^{M_1M_2 \times 1}$ denotes the observation noise. It is to be noted that $\omega_{\mathbf{c}_l}$ in equation (1) describes the motion at HR since $\mathbf{H}_{\mathbf{c}_l}$ operates on $\mathbf{f}$, the HR image. Thus, the term inside the bracket on the RHS represents the non-uniform blurring of the latent HR image.

$\mathbf{D} \in \mathbb{R}^{M_1M_2 \times N_1N_2}$ is the downsampling operator or the decimation operator that mimics the behaviour of the digital sensors. The downsampling process consists of two stages – a convolution with the sensor PSF, followed by sampling. Sensor blur results from the finite-sized sensor integrating impinging light over its surface during exposure. The sensor has maximum sensitivity at its center while it falls off towards the edges with a Gaussian-like decay. The suitability of the Gaussian function to model the sensor PSF has been experimentally verified in [2] and, therefore, we use it here in our work. The sampling operation can be viewed as the multiplication by a sum of delta functions placed on an evenly spaced grid. In matrix form, it can be represented as the Kronecker product of a 1D sampling matrix with itself.

### *B. 3D scene*

We now extend the image formation model in equation (1) to a 3D scene using a layered approach. Let us assume that there are $R$ depth layers in the scene, and the scene depth of each layer is given by $\{d_r\}_{r=1}^{R}$. We denote one of these layers as the reference depth layer $r_{\text{ref}}$, and its corresponding depth as $d_{\text{ref}}$. We define the relative depth of each layer $\delta_r$ with respect to this reference depth as $\delta_r = \frac{d_{\text{ref}}}{d_r}$. Using the $\delta_r$ values at each pixel, we can construct the relative depth map of the scene $\boldsymbol{\chi} \in \mathbb{R}^{N_1N_2 \times 1}$. Based on the depth map, we can also split the HR image $\mathbf{f}$ into $R$ disjoint constant-depth regions as

$$\mathbf{f} = \sum_{r=1}^{R} \mathbf{f}_r. \tag{2}$$

Here the notation $\mathbf{f}_r$ indicates the $r^{\text{th}}$ depth region in the image. The intensity of a pixel in $\mathbf{f}_r$ is same as in the latent HR image $\mathbf{f}$ if it belongs to the $r^{\text{th}}$ depth region, and 0 otherwise.

Equation (1) can now be rewritten for the 3D case as

$$\mathbf{g} = \mathbf{D}\left(\sum_{\mathbf{c}_l \in \mathcal{C}} \omega_{\mathbf{c}_l}\left(\sum_{r=1}^{R} \mathbf{H}_{(\delta_r,\mathbf{c}_l)}\mathbf{f}_r\right)\right) + \mathbf{n}. \tag{3}$$

In equation (3), warped images from *all the depth layers* are subjected to a weighted averaging followed by downsampling to produce $\mathbf{g}$. Although the camera motion is the same, even for a single camera pose, the warps experienced on the image plane vary with the depth. Therefore, the warping matrix $\mathbf{H}$ is now a function of both the relative depth $\delta_r$ and the camera pose $\mathbf{c}_l$. Each warp can be described by a homography $\boldsymbol{\mathcal{P}}_{(\delta_r,\mathbf{c}_l)}$ as [13]

$$\boldsymbol{\mathcal{P}}_{(\delta_r,\mathbf{c}_l)} = \mathbf{K}_v\left([\mathbf{R}_l] + \frac{\delta_r}{d_{\text{ref}}}\mathbf{T}_l[0\ \ 0\ \ 1]\right)\mathbf{K}_v^{-1}, \tag{4}$$

where $\mathbf{T}_l = [T_{X_l}\ \ T_{Y_l}\ \ T_{Z_l}]^T$, $\mathbf{R}_l = [\theta_{X_l}\ \ \theta_{Y_l}\ \ \theta_{Z_l}]^T$ are the camera translation and rotation vectors, respectively, and $[\mathbf{R}_l]$ is the matrix exponential equivalent of $\mathbf{R}_l$ [13]. The camera intrinsic matrix $\mathbf{K}_v$ is assumed to be of the form $\mathbf{K}_v = \text{diag}(v, v, 1)$, where $v$ is the focal length. A camera that is free to undergo any general motion has six degrees of freedom, three arising from translations $\mathbf{T}$ along, and three from rotations $\mathbf{R}$ about the three axes. However, it has been pointed out [10], [11], [13] that in most practical scenarios, three degrees of freedom are sufficient to model general camera motion. While [10], [11] used in-plane translations and rotations, [13] used rotations about the three axes. In this work, we adopt the former model since it also accounts for parallax. It is assumed that the general motion of the camera can be approximated by translations along the image plane and in-plane rotations. In such a case, the homography $\boldsymbol{\mathcal{P}}_{(\delta_r,\mathbf{c}_l)}$ simplifies to

$$\boldsymbol{\mathcal{P}}_{(\delta_r,\mathbf{c}_l)} = \mathbf{K}_v \begin{bmatrix} \cos\theta_{Z_l} & \sin\theta_{Z_l} & \frac{\delta_r T_{X_l}}{d_{\text{ref}}} \\ -\sin\theta_{Z_l} & \cos\theta_{Z_l} & \frac{\delta_r T_{Y_l}}{d_{\text{ref}}} \\ 0 & 0 & 1 \end{bmatrix} \mathbf{K}_v^{-1}. \tag{5}$$

The camera pose space $\mathcal{C}$ then becomes a 3D space defined by the axes $T_X$, $T_Y$ and $\theta_Z$, and $\boldsymbol{\mathcal{P}}_{(\delta_r,\mathbf{c}_l)}$ is parameterized by $(\delta_r, T_{X_l}, T_{Y_l}, \theta_{Z_l})$.

For the camera pose $\mathbf{c}_l$, the homography $\boldsymbol{\mathcal{P}}_{(\delta_r,\mathbf{c}_l)}$ corresponds only to the depth layer $r$. However, if the translation and rotation observed on the image plane due to the camera pose $\mathbf{c}_l$ at a particular depth layer, say $r_{\text{ref}}$, are known, then, with the knowledge of $\delta_r$, we can compute the translation and rotation induced on any other layer due to $\mathbf{c}_l$ i.e., the homography $\boldsymbol{\mathcal{P}}_{(\delta_r,\mathbf{c}_l)}$ at any other layer can be estimated. Say the depth layer $r_{\text{ref}}$ at a depth $d_{\text{ref}}$ underwent the motion $(T_{X_l}, T_{Y_l}, \theta_{Z_l})$, then the motion at other depth layers can be computed as $(\delta_r T_{X_l}, \delta_r T_{Y_l}, \theta_{Z_l})$ [18]. Observe that rotation is invariant to the scene depth, while in-plane translations are scaled by the relative depth value – scene points near the camera experience more shift when compared to points that are farther away. To sum up, if the motion induced on the image plane at a reference depth $d_{\text{ref}}$ due to the global camera motion is known, then the motion at all other depths can be estimated.

Motion blur in conjunction with the downsampling operation makes the problem of super-resolution from a single image highly ill-posed. Hence, as discussed in the introduction,

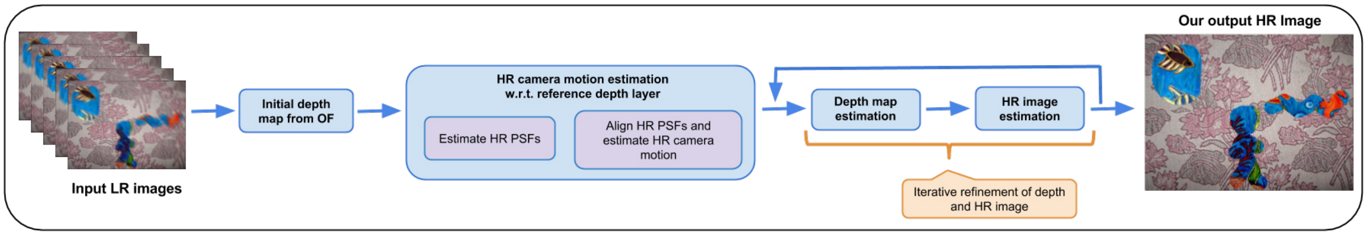

Input LR images
Initial depth map from OF
HR camera motion estimation w.r.t. reference depth layer
Estimate HR PSFs
Align HR PSFs and estimate HR camera motion
Depth map estimation
HR image estimation
Iterative refinement of depth and HR image
Our output HR Image

Fig. 1. Block diagram of the proposed framework.

we supplement the missing information in a single observation using multiple images. In this work, we assume that $K$ LR images $\mathbf{g}^k$, with $k = 1$ to $K$, are available, and $\boldsymbol{\omega}^k$ is the associated HR camera trajectory.

## III. The Proposed Method

Consider $K$ motion blurred LR observations $\{\mathbf{g}^k\}_{k=1}^K$ of a 3D scene which are related to the latent HR image $\mathbf{f}$ through the depth map $\boldsymbol{\chi}$ and the HR camera trajectories $\{\boldsymbol{\omega}^k\}_{k=1}^K$. The objective is to recover $\mathbf{f}$ given only $\{\mathbf{g}^k\}_{k=1}^K$. To this end, we first crudely segment the scene into different depth layers by applying an optical flow algorithm on the LR observations, and label the layer with the maximum area as the reference depth layer $r_{\text{ref}}$. Next, we compute the HR camera motion with respect to $r_{\text{ref}}$ using HR blur kernels estimated at a few points lying on the reference depth layer in the LR images. With the knowledge of the HR trajectories, we eventually solve for the latent HR image $\mathbf{f}$ and the depth map $\boldsymbol{\chi}$ within an alternating minimization framework. The details of these steps are explained in the following subsections. A block diagram of the proposed framework is shown in Fig. 1.

To aid explanation, we consider the synthetic example in Fig. 2. The latent HR image and the corresponding depth map are shown in the first row. By convention, a scene point that is closer to the camera has a higher intensity value in the depth map than one that is farther away. To simulate burst mode capture, we manually generated five connected camera trajectories within the motion space and initialized the weights. The parameters of the 3D camera motion space $\mathcal{C}$ were selected as follows: $\theta_Z$ ranged between $-2°$ to $2°$ in steps of $0.2°$, $T_X$ and $T_Y$ ranged between -10 to 10 pixels in increments of one pixel. The cardinality of the set $\mathcal{C}$ can be calculated as: $|\mathcal{C}|$ = (Number of translation steps along $X$-axis) $\times$ (Number of translation steps along $Y$-axis) $\times$ (Number of rotation steps about $Z$-axis) $=$ $(-10:1:10$ pixels along $X$-axis) $\times$ $(-10:1:10$ pixels along $Y$-axis) $\times$ $(-2°:0.2°:2°$ about $Z$-axis) $= 21 \times 21 \times 21 = 9261$. Five blurred LR images were then generated from the latent HR image by first performing depth-dependent blurring followed by downsampling by a factor of two (using the SR motion blur model of Section II). We then added white Gaussian noise with signal-to-noise ratio (SNR) of 30 dB, where SNR $= 10 \log\left(\frac{\sigma_{\mathbf{f}}^2}{\sigma_{\mathbf{n}}^2}\right)$, $\sigma_{\mathbf{f}}$ and $\sigma_{\mathbf{n}}$ being image and noise standard deviations, respectively. The space-variantly blurred LR images are shown in row two of Fig. 2.

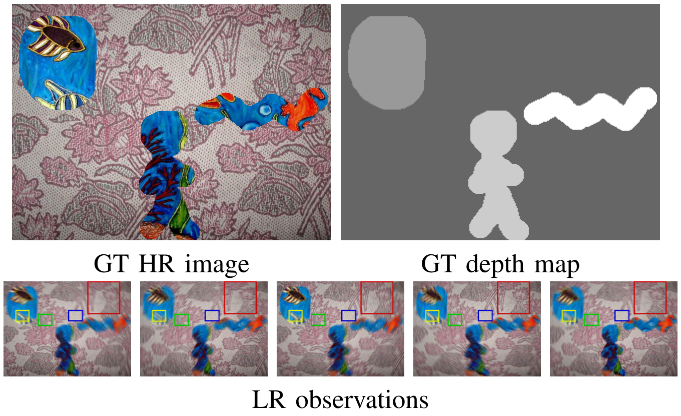
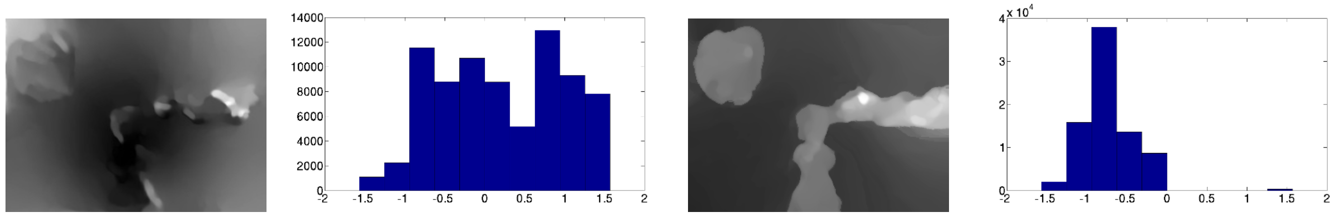


Fig. 2. Synthetic experiment of a 3D scene containing four depth layers. Row one: ground truth (GT) HR image and depth map, and row two: blurred LR observations. Note that the LR and HR images are not displayed to scale; the SR factor in this example is 2.

### A. Initial depth from optical flow

An initial rough estimate of the depth map of the scene can be obtained using optical flow. Optical flow vectors yield the displacement field between a pair of misaligned images. Since our LR observations are captured using a hand-held camera, not only can there be motion during the exposure of a single image, there is also incidental motion *between* successive captures. While intra-image motion results in blur, inter-image misalignment enables optical flow estimation. The flow vectors can be used to coarsely segment the scene into different depth layers.

If the camera undergoes pure in-plane translational motion between the capture of the two observations, then the depth is inversely proportional to the magnitude of the optical flow vector at that location. Hence, the magnitude of these vectors directly yields the relative depth map of the scene. However, this does not hold true in the case of camera rotation. Since we assume that the camera is free to undergo translations as well as rotations, our goal is to identify the pair of images from the set of $K$ LR observations with minimum camera rotation between them. The depth map recovered from such a pair using the magnitude of the flow will have minimum error. To identify this pair, we compute optical flow between all $\binom{K}{2}$ pairs of images. Next, we extract the histogram of the phase of the optical flow vectors. Ideally, if the inter-image motion is pure in-plane translation, then all the flow vectors will have the same phase even though the magnitude varies with depth. Therefore, the image pair whose histogram



Depth map: 1–2 Histogram: 1–2 Depth map: 3–5 Histogram: 3–5

Fig. 3. The depth map obtained using image pairs 1–2 and 3–5 are shown in columns one and three, respectively. The histogram of the phase of the optical flow vectors for the same two pairs are plotted in columns two and four.

has the fewest active bins[1] is the pair most suited for depth map computation. The depth value that occurs the maximum number of times in this depth map is the reference depth $d_{\text{ref}}$, and the pixels having this depth value (need not be contiguous) are flagged as belonging to the reference depth layer $r_{\text{ref}}$.

We used the optical flow algorithm of Brox et al. [31] to estimate the flow vectors. We note that their method is designed for sharp images, and there can be minor errors in the estimated flow when a blurred pair is processed. However, our experiments revealed that our proposed framework is robust to such small errors since this initial depth map is used merely to jump-start our AM scheme, and roughly localize the dominant depth layer in the scene. To demonstrate our depth map initialization step through an example, we select two pairs of images from the LR observations in row two of Fig. 2. The depth map and the histogram obtained from the two image pairs are shown in columns one to four of Fig. 3. Notice that the pair 1–2 has many active bins in its histogram because of significant in-plane rotation between these two images. Therefore, the associated depth map is also largely in error. The images 3 and 5, on the other hand, have the least rotation among all the $\binom{5}{2}$ pairs, and the corresponding histogram has all its values concentrated in just 4 bins. Therefore, we use the depth map from the 3–5 image pair to initialize our AM scheme.

### *B. Estimation of HR camera trajectories*

We first explain how we compute HR PSFs from suitably selected patches in the LR images. We then elaborate on how these locally-estimated HR PSFs reveal the global HR camera motion. We also briefly describe why HR camera trajectory estimation should be preceded by a kernel alignment step.

*1) HR PSF estimation:* Our goal is to use the LR images to estimate HR PSFs at points lying on the reference depth layer. To this end, we use the algorithm of Hu and Yang [32] to determine points with good texture and long edges belonging to $r_{\text{ref}}$ in the first LR image that are suitable for blur kernel estimation. We randomly select $S_p$ spatially-separated point locations from this set such that patches cropped around these points from the $K$ LR observations lie entirely in the reference depth layer $r_{\text{ref}}$. We denote the patches as $\{\mathbf{g}_i^1\}_{i=1}^{S_p}$, $\{\mathbf{g}_i^2\}_{i=1}^{S_p}$,..., $\{\mathbf{g}_i^K\}_{i=1}^{S_p}$. Although the blur is space-varying across the image, we assume it to be uniform within each small patch. We provide the set of patches $(\mathbf{g}_i^1, \mathbf{g}_i^2, ..., \mathbf{g}_i^K)$ as input to the blind SR technique of Sroubek et al. [4] to compute the HR blur kernels $(\mathbf{h}_i^1, \mathbf{h}_i^2, ..., \mathbf{h}_i^K)$ at the $i^{\text{th}}$ location. Fig. 4 shows the ground truth and estimated HR PSFs computed from the red patches in the LR observations in row two of Fig. 2. We repeat this HR PSF estimation step at all $S_p$ locations. We found from our experiments that the estimates of the HR PSFs returned by the method of [4] are quite accurate. See Fig. 4. It is to be noted that image and motion are the only two unknowns in this step of HR PSF estimation (depth is a constant since the image patches are extracted from a single depth layer in the scene), and the method of Sroubek et al. [4] alternately minimizes these two parameters to obtain the optimal HR PSFs.

*2) HR camera trajectories from HR PSFs:* For each LR image $\mathbf{g}^k, k = 1, 2, ..., K$, our aim is to estimate the HR camera trajectory $\boldsymbol{\omega}^k$ that concurs with the $S_p$ observed HR blur kernels $\{\mathbf{h}_i^k\}_{i=1}^{S_p}$ and their locations. Following [18], we express the blur kernel $\mathbf{h}_i^k$ as $\mathbf{h}_i^k = \mathbf{M}_i^k \boldsymbol{\omega}^k$ for $i = 1, 2, ..., S_p$. Here, $\mathbf{M}_i^k \in \mathbb{R}^{S_h \times |\mathcal{C}|}$ is a matrix whose entries are determined by the location of the blur kernel and the bilinear interpolation coefficients, and $S_h$ is the number of elements in $\mathbf{h}_i^k$ (i.e., for a blur kernel $\mathbf{h}_i^k \in \mathbb{R}^{U_1 U_2 \times 1}$, the scalar $S_h = U_1 \times U_2$). Note that the $S_p$ point locations were chosen on the LR grid and patches were cropped around these points from the LR observations. Since the camera trajectories are being estimated on an HR grid, the $S_p$ point locations should be scaled by the SR factor, and our camera motion estimation step differs from the method proposed in [18] in this important respect. By stacking all the $S_p$ blur kernels as a vector $\overline{\mathbf{h}^k}$, and suitably concatenating the matrices $\mathbf{M}_i^k$ for $i = 1, 2, ..., S_p$, the HR blur kernels can be related to the HR camera trajectory as

$$\overline{\mathbf{h}^k} = \mathbf{M}^k \boldsymbol{\omega}^k, \tag{6}$$

where the matrix $\mathbf{M}^k$ is of size $S_p S_h \times |\mathcal{C}|$. Note that $\boldsymbol{\omega}^k$ will be a sparse vector in practice because only a few camera poses $\mathbf{c}_l$ out of all the possible poses in the motion space $\mathcal{C}$ will be active during the exposure time. This allows us to impose a sparsity constraint on $\boldsymbol{\omega}^k$, and we estimate the HR camera trajectory by minimizing the following cost

$$E(\boldsymbol{\omega}^k) = ||\overline{\mathbf{h}^k} - \mathbf{M}^k \boldsymbol{\omega}^k||_2^2 + \lambda ||\boldsymbol{\omega}^k||_1 \quad \text{subject to } \boldsymbol{\omega}^k \geq \mathbf{0}. \tag{7}$$

We solve equation (7) using the *nnLeastR* function of the Lasso algorithm in [33] which considers the additional non-negativity and $l_1$-norm constraints. Here, $\lambda$ is a positive scalar

[1] We treat a bin as 'active' only if the frequency of occurrence is higher than a certain threshold.

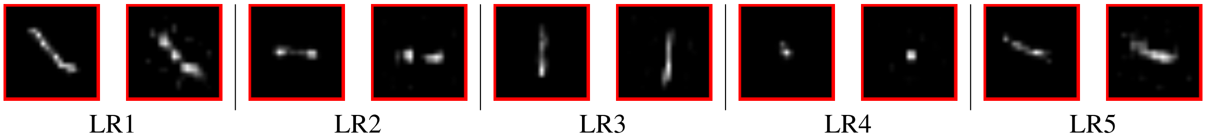


Fig. 4. The five columns correspond to the kernels at the center of the red patch in the five LR observations shown in row two of Fig. 2. In each column, the PSF on the left represents the ground truth HR kernel, while the one on the right is estimated using the algorithm of Sroubek et al. [4].

that controls the extent of sparsity of the vector $\boldsymbol{\omega}^k$. We estimate the HR trajectory $\boldsymbol{\omega}^k$ corresponding to each LR observation $\mathbf{g}^k$, $k = 1, 2, ..., K$, separately by minimizing equation (7).

The synthetically generated camera path and the estimated HR trajectory for one of the five input observations are shown in Fig. 5. It can be observed from the plots that the ground truth and recovered trajectories are very similar in shape demonstrating our algorithm's ability to compute global HR motion directly from locally estimated kernels.

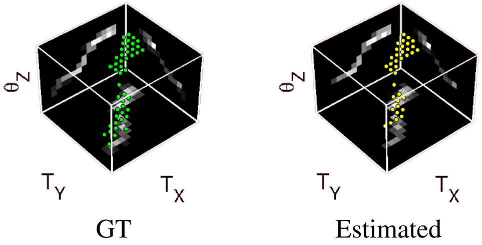


Fig. 5. Ground truth and estimated HR trajectories.

*3) Alignment of PSFs:* If $\mathbf{\Gamma}(.)$ denotes a translational shift, then a blurred LR image patch $\mathbf{g}_i^k$ which is given by $\mathbf{g}_i^k = \mathbf{D}(\mathbf{f}_i * \mathbf{h}_i^k)$ would also be equal to $\mathbf{D}(\mathbf{\Gamma}^{-1}(\mathbf{f}_i) * \mathbf{\Gamma}(\mathbf{h}_i^k))$, where $*$ denotes convolution, and $\mathbf{f}_i$ is the corresponding patch from the latent HR image. Therefore, while solving for the local HR PSFs $\mathbf{h}_i^k$, there can be incidental shifts of small magnitude in the estimated HR blur kernels with respect to the 'true' blur kernels (which are induced at a point as a result of blurring the latent HR image with the true HR camera motion). Since the blur kernels at a particular location are estimated independently with respect to kernels at other locations, the shift could vary from one location to the next. Unless the shifts in the HR PSFs are accounted for, they cannot be related to a single $\boldsymbol{\omega}^k$. Since the blurred patches $\{\mathbf{g}_i^k\}_{k=1}^K$ at a given location $i$ are related to the same latent HR image patch $\mathbf{f}_i$, if the shift of the latent image patch is $\mathbf{\Gamma}^{-1}$, then the shift for all the $K$ blur kernels will have to be $\mathbf{\Gamma}$. Hence, we need to determine the shifts of the blur kernels corresponding to only one of the $K$ observations, say $k = 1$, and these shifts remain the same for the remaining $k = 2, ..., K$ observations. The camera motion $\boldsymbol{\omega}^k$ estimated from the aligned blur kernels should have a low value of error $||\overline{\mathbf{h}^k} - \mathbf{M}^k \boldsymbol{\omega}^k||_2^2$. We consider that one of the blur kernels, say $\mathbf{h}_1^k$ does not undergo any shift and align the other blur kernels with respect to this. We need to determine two translation parameters for each of the other blur kernels. For all possible combinations of the translations, we shift the blur kernels $\mathbf{h}_2^k, \mathbf{h}_3^k, ..., \mathbf{h}_{S_p}^k$, and evaluate the solution of equation (7). Since the magnitude of the shifts is generally small, and the number of blur kernels used is typically low (around 4), finding the optimum shifts (that minimize the error $||\overline{\mathbf{h}^k} - \mathbf{M}^k \boldsymbol{\omega}^k||_2^2$) is not computationally prohibitive.

### *C. Alternating Minimization*

Once the camera motion has been computed, the next step is to iteratively estimate the depth map $\boldsymbol{\chi}$ and the latent HR image $\mathbf{f}$ using the $K$ LR observations $\{\mathbf{g}^k\}_{k=1}^K$. We propose an alternating minimization strategy to solve for the two variables wherein we fix one unknown and compute the other, in an iterative manner. The minimization sequence $(\boldsymbol{\chi}^p, \mathbf{f}^p)$, where $p$ indicates the iteration number, can be built by alternating between two minimization subproblems. Starting with an initial estimate of depth map $\boldsymbol{\chi}^0$ (the depth map from optical flow upsampled by the super-resolution factor), the two alternating steps are: step 1) estimate the latent HR image $\mathbf{f}^p$ using the previous iterate $\boldsymbol{\chi}^{p-1}$ of the depth map, step 2) use the current estimate of the image $\mathbf{f}^p$ to compute the depth map $\boldsymbol{\chi}^p$.

Since the scene is 3D, a scene point that is visible in one camera pose may be occluded by a foreground depth layer when the camera moves to a different pose. When the camera moves during the exposure duration of an image, it passes through a finite set of poses from the discretized camera pose space $\mathcal{C}$. For every observation $k$ and for each camera pose $\mathbf{c}_l$, we define a visibility function $V_{\mathbf{c}_l}^k$ on the HR grid. The binary function $V_{\mathbf{c}_l}^k(\mathbf{y})$, defined with respect to the current estimate of depth in our AM scheme, takes the value 1 if the pixel at a particular location $\mathbf{y}$ is visible[2] and has a positive camera pose weight $\omega_{\mathbf{c}_l}^k$ associated with it, and is 0 otherwise. The overall visibility $\overline{V^k}(\mathbf{y})$ of a pixel $\mathbf{y}$ in the observation $k$ is then defined as

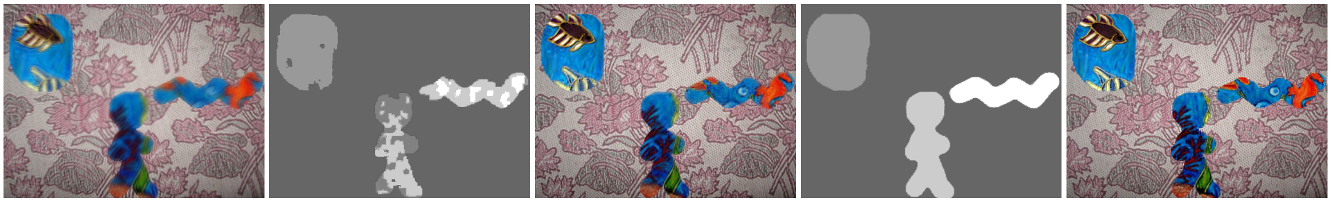
$$\overline{V^k}(\mathbf{y}) = \begin{cases} 1, & \text{if } \exists \mathbf{c}_l \in \mathcal{C} \text{ such that } V_{\mathbf{c}_l}^k(\mathbf{y}) = 1, \\ 0, & \text{if } \forall \mathbf{c}_l \in \mathcal{C}, V_{\mathbf{c}_l}^k(\mathbf{y}) = 0. \end{cases} \tag{8}$$

In other words, we label a pixel $\mathbf{y}$ in the observation $k$ as visible if it is unoccluded in at least one of the poses that the camera passes through during exposure. Note that our notion of *overall* visibility is different from the definition of visibility in Wei and Quan [34] in that ours encompasses a *set* of camera poses. Wei and Quan [34] limit their discussion to a single homography since they do not consider motion blur. We also introduce a binary modulating function on the LR grid as

$$W^k(\mathbf{x}) = \begin{cases} 1, & \text{if } \overline{V^k}(\mathbf{y}) = 1, \text{ and } \forall \underline{\mathbf{y}} \in \boldsymbol{\mathcal{Y}}, \overline{V^k}(\underline{\mathbf{y}}) = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Here $\mathbf{x}$ is the LR pixel obtained by applying the downsampling operator $\mathbf{D}$ on $\mathbf{y}$ and its neighbours $\boldsymbol{\mathcal{Y}}$ lying on the HR grid.

[2] We use the definition of visibility in Wei and Quan [34] which states that a pixel after warping is 'visible' if it is unoccluded by foreground depth layers.



Iter 1: image Iter 3: depth map Iter 3: image Iter 5: final depth map Iter 5: final image

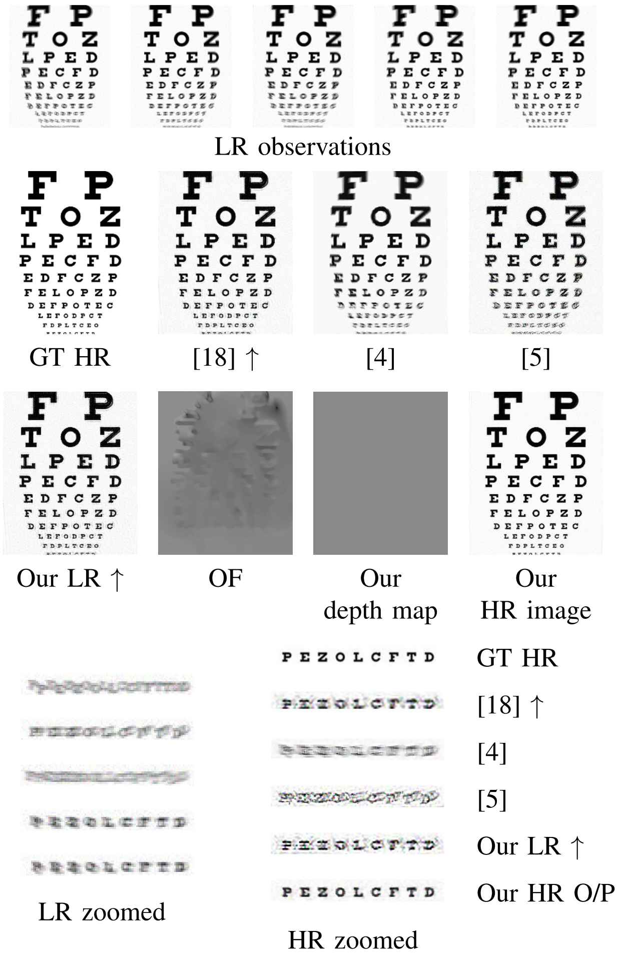
Fig. 6. The progress of our alternating minimization framework with iterations.

The function $W^k(\mathbf{x})$ takes the value 1 only when $\mathbf{y}$ and *all* its neighbours $\boldsymbol{\mathcal{Y}}$ are visible. During each iteration of our AM scheme, we modulate the data term in our cost function using $W^k$ as we shall see in the following discussion.

*1) HR image estimation:* When the HR camera motion and the depth map $\boldsymbol{\chi}^{p-1}$ are known, and $\mathbf{f}$ is to be estimated, equation (3) can be expressed in the matrix-vector notation as

$$\boldsymbol{g}^k = \mathbf{D}\boldsymbol{\mathcal{H}}^k\mathbf{f} + \mathbf{n}^k, \tag{10}$$

where $\boldsymbol{\mathcal{H}}^k \in \mathbb{R}^{N_1N_2 \times N_1N_2}$ is the matrix that performs the depth-dependent non-uniform blurring operation of the various depth layers in the scene.



LR observations

GT HR [18] ↑ [4] [5]

Our LR ↑ OF Our depth map Our HR image

LR zoomed

GT HR

[18] ↑

[4]

[5]

Our LR ↑

Our HR O/P

HR zoomed

Fig. 7. Synthetic example of a planar scene. Row one: blurred LR observations, row two: ground truth HR image, the output of Paramanand and Rajagopalan [18] super-resolved using the single image SR algorithm of [21] (super-resolution step denoted by ↑), and SR results of Sroubek et al. [4] and Ma et al. [5], row three: our LR result super-resolved using [21], depth map from optical flow, our estimated depth map, and our HR output image, and row four, column one: zoomed-in regions of the bottom line of text from the LR observations in row one, and column two: zoomed-in regions from GT HR, [18] ↑, [4], [5] in row two and our LR ↑, our HR output in row three, respectively. Note that the LR zoomed-in regions have been scaled to the same size as the HR zoomed-in regions for display.

To solve for the latent HR image $\mathbf{f}$, we formulate an energy function based on the observation error and a regularization term as

$$E\left(\mathbf{f}\right) = \sum_{k=1}^{K} ||\mathbf{W}^k(\mathbf{D}\boldsymbol{\mathcal{H}}^k\mathbf{f} - \mathbf{g}^k)||_2^2 + \alpha\mathbf{f}^T\mathbf{L}\mathbf{f}, \tag{11}$$

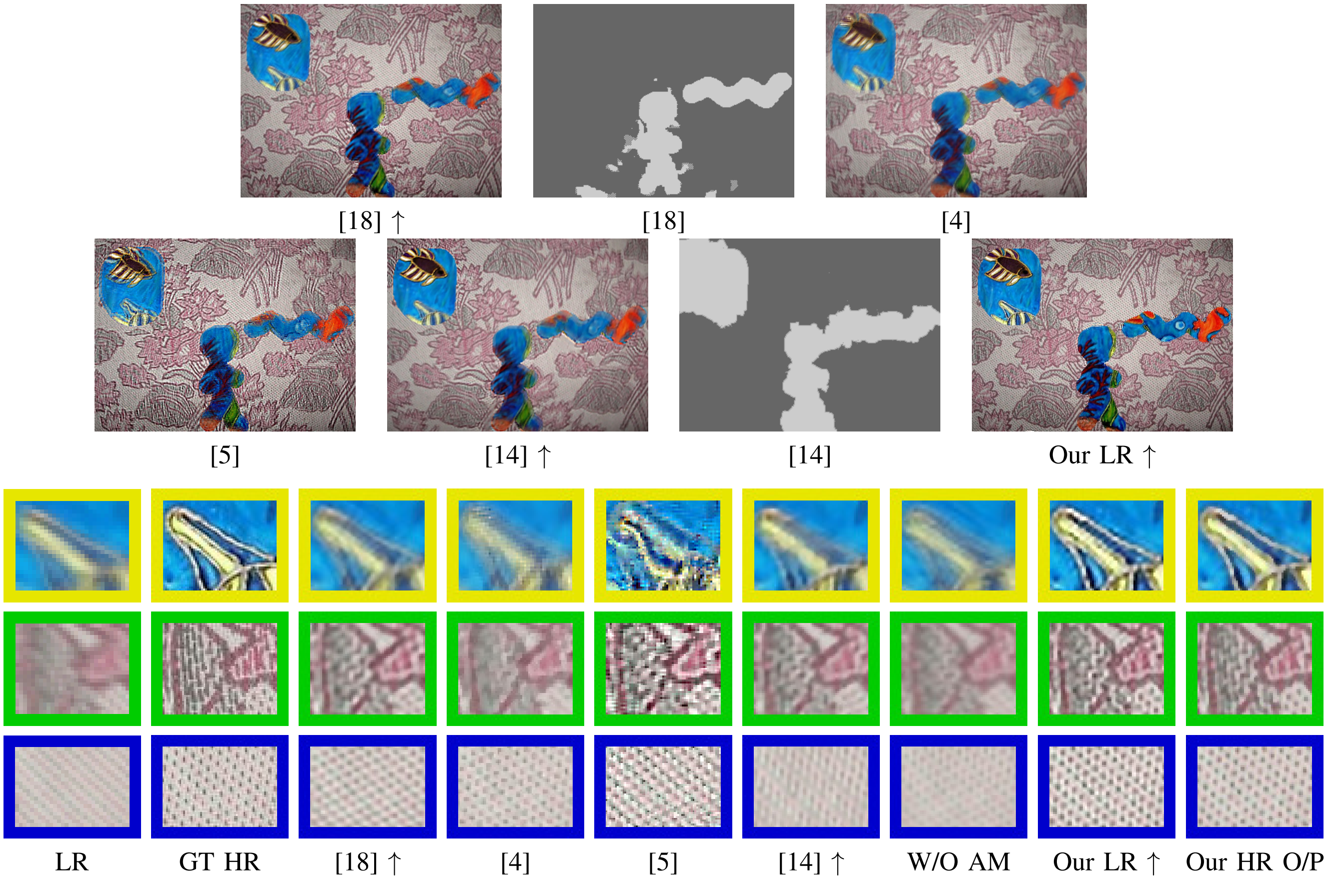
where the matrix $\mathbf{L}$ comprises of elements that depend on the gradient of $\mathbf{f}$. It is the discrete equivalent of the variational prior and is a positive semidefinite block tridiagonal matrix [4]. It exhibits isotropic behaviour in smooth areas, while also preserving edges. Here $\mathbf{W}^k \in \mathbb{R}^{M_1M_2 \times M_1M_2}$ is a diagonal matrix constructed from $W^k$ in equation (9) that decides whether or not the data cost should be enforced for a particular LR pixel in each input observation[3]. To obtain the current estimate of the HR image $\mathbf{f}^p$, we minimize equation (11)

$$\begin{aligned} \mathbf{f}^p = \underset{\mathbf{f}}{\text{argmin}}\ E(\mathbf{f}) &\Rightarrow \frac{\partial E}{\partial \mathbf{f}} = 0 \\ \Longleftrightarrow \left(\sum_{k=1}^{K} \boldsymbol{\mathcal{H}}^{k^T}\mathbf{D}^T\mathbf{W}^{k^T}\mathbf{W}^k\mathbf{D}\boldsymbol{\mathcal{H}}^k + \alpha\mathbf{L}\right) \mathbf{f} &= \\ \sum_{k=1}^{K} \boldsymbol{\mathcal{H}}^{k^T}\mathbf{D}^T\mathbf{W}^{k^T}\mathbf{W}^k\mathbf{g}^k. & \end{aligned} \tag{12}$$

We used the method of conjugate gradients to solve equation (12).

*2) Depth map estimation:* As discussed in Section II-B, if the motion induced on the image plane at a reference depth $d_{\text{ref}}$ due to the global camera motion is known, then the motion at other depths can be computed by scaling the translational parameters. The scale factor at each pixel is equal to the relative depth $\delta_r$ at that location, and our objective is to determine the scale factors at all pixels (i.e., the relative depth map $\boldsymbol{\chi}^p$) using the camera motion $\boldsymbol{\omega}^k$ estimated with respect to $r_{\text{ref}}$, and the current iterate of the image $\mathbf{f}^p$. To this end, we model the depth map as an MRF and obtain the MAP estimate using the loopy belief propagation (LBP) algorithm

[3] Since $\boldsymbol{\mathcal{H}}^k, \mathbf{W}^k$ are built using the previous iterate of the depth map $\boldsymbol{\chi}^{p-1}$, the correct notation for them in equations (10) and (11) would be $\boldsymbol{\mathcal{H}}^{k^{p-1}}, \mathbf{W}^{k^{p-1}}$. However, we drop the superscript $p-1$ for notational brevity.



[18] ↑ [18] [4]

[5] [14] ↑ [14] Our LR ↑

LR GT HR [18] ↑ [4] [5] [14] ↑ W/O AM Our LR ↑ Our HR O/P

Fig. 8. Row one: Output image of Paramanand and Rajagopalan [18] super-resolved using [21], depth map of [18], and result of Sroubek et al. [4], row two: result of Ma et al. [5], output image of Hu et al. [14] super-resolved using [21], depth map of [14], and our LR output super-resolved using [21], and row three: zoomed-in regions from the first LR observation in row two of Fig. 2, the GT HR image in column one of row one of Fig. 2, [18] ↑, [4] in row one, [5], [14] ↑ in row two, without AM in column one of Fig. 6, our LR ↑ in row two, and our HR output in column five of Fig. 6.

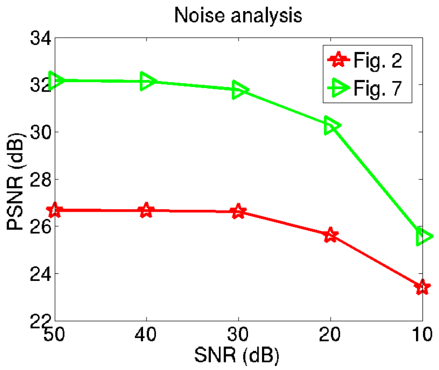


Fig. 9. Noise analysis for the examples in Figs. 2 and 7.

proposed in [35]. The algorithm is iterative, and the MAP estimate improves after each iteration until convergence is attained. The advantage of using such an approach is two-fold – (i) the MAP-MRF framework of [35] is modeled as a label assignment problem (with $\delta_r$ being the labels) which goes hand-in-hand with our discrete layered 3D model, and (ii) we can avoid the evaluation of derivatives which is quite tedious especially in the case of space-varying blur. Furthermore, we can incorporate regularization by defining a smoothness cost since the depth map is homogeneous in most regions. This regularizing term also ensures that sharp boundaries between depth layers are preserved.

The cost function for the LBP algorithm for assigning the relative depth value $\delta_r$ at a particular pixel $\mathbf{y}$ on the HR grid is given by

$$E(\delta_{r_{\mathbf{y}}}) = \sum_{k=1}^{K}\Bigg(\mathbf{g}^k(\mathbf{x}) - \mathbf{D}\Bigg(\Bigg(\sum_{\mathbf{c}_l\in\mathcal{C}} \omega^k_{\mathbf{c}_l}\mathbf{H}_{(\delta_{r_{\mathbf{y}}},\mathbf{c}_l)}\mathbf{f}^p\Bigg)(\mathbf{y}) + \sum_{\underline{\mathbf{y}}\in\boldsymbol{\mathcal{Y}}}\Bigg(\sum_{\mathbf{c}_l\in\mathcal{C}} \omega^k_{\mathbf{c}_l}\mathbf{H}_{(\delta_{r_{\underline{\mathbf{y}}}},\mathbf{c}_l)}\mathbf{f}^p\Bigg)(\underline{\mathbf{y}})\Bigg)\Bigg)^2 + \sum_{\underline{\mathbf{y}}\in\boldsymbol{\mathcal{Y}}} \mu \,\min(|\delta_{r_{\mathbf{y}}} - \delta_{r_{\underline{\mathbf{y}}}}|, \beta). \quad (13)$$

The first term in the above equation corresponding to the data cost is formulated based on the observation model – only when the HR pixels are warped and averaged according to the correct HR depth will they form a group which, when downsampled, attains the least cost when compared with their corresponding LR pixel intensity. The second term is the smoothness cost that penalizes the difference in the labels of neighboring pixels. To allow for discontinuities, this cost should take a constant value when the difference becomes large. Therefore, we adopt

**Algorithm 1** 3D blind super-resolution from a set of non-uniformly motion blurred LR images

```
Input: : Motion blurred LR images $\{\mathbf{g}^k\}_{k=1}^K$ of a 3D scene.
Output: : HR camera trajectories $\boldsymbol{\omega}^k$, depth map $\boldsymbol{\chi}$, and latent HR image $\mathbf{f}$.
Estimate initial depth map $\boldsymbol{\chi}^0$ using optical flow (Section III-A).
Choose the depth layer having the maximum area in $\boldsymbol{\chi}^0$ as the reference depth layer $r_{\text{ref}}$.
Extract $S_p$ patches lying entirely in $r_{\text{ref}}$ from all $K$ LR observations $\{\mathbf{g}_i^1\}_{i=1}^{S_p}$, $\{\mathbf{g}_i^2\}_{i=1}^{S_p}$,..., $\{\mathbf{g}_i^K\}_{i=1}^{S_p}$ (Section III-B1).
Estimate HR PSFs $(\mathbf{h}_i^1, \mathbf{h}_i^2, ..., \mathbf{h}_i^K)$ from each set of patches $(\mathbf{g}_i^1, \mathbf{g}_i^2, ..., \mathbf{g}_i^K)$ at all $i = 1, 2, ..., S_p$ locations.
Estimate the camera motion $\{\boldsymbol{\omega}^k\}_{k=1}^K$ using aligned HR PSFs (Sections III-B2 and III-B3).
Let $p = 0$, and $\mathbf{f}^0$ be initialized to all zeros.
do
    $p = p + 1$
    Estimate the latent HR image $\mathbf{f}^p$ using equation (12) (Section III-C1).
    Estimate the depth map $\boldsymbol{\chi}^p$ using loopy belief propagation by minimizing equation (13) (Section III-C2).
while RMS$(\mathbf{f}^p - \mathbf{f}^{p-1}) >$ threshold
```

the commonly used truncated linear model [35] where the threshold $\beta$ determines when the cost stops increasing, and $\mu$ is a weighting parameter.

We define a discrete search space for the relative depth $\delta_r$ as 0.1:10 with a step size of 0.2. Note that the search space extends on either side of the reference depth layer $r_{\text{ref}}$ which has a relative depth unity because the position of $r_{\text{ref}}$ with respect to the foreground and background layers is a priori unknown to our algorithm.

The two steps of the alternating minimization scheme – latent image estimation and depth refinement – are performed iteratively until convergence. The criterion for convergence is a threshold on the root mean square error between the latent image estimate of current and previous iterations. We found from our experiments that our AM scheme exhibits good convergence properties, and attains the desired solution within 5 to 6 iterations. An overview of our approach is provided in Algorithm 1. It is important to note that our method requires only the LR images as input. It demands no additional knowledge of the camera parameters or the scene for estimating the latent HR image.

The progress of the AM scheme with iterations is displayed in Fig. 6. Column one shows the estimate of the latent HR image after the first pass, computed using the initial depth map (in column three of Fig. 3) returned by optical flow. Notice that there is residual blur in the image due to errors in the depth map. Columns two to five show the depth map and the HR image after the third and the fifth iteration. The boundaries are nicely recovered in our final depth map and our output HR image is deblurred at all depth layers. This a clear indicator of the importance of refining the depth map with iterations, and the success of our alternating minimization approach. Note that the HR image in column one is the result one would have obtained without the AM framework.

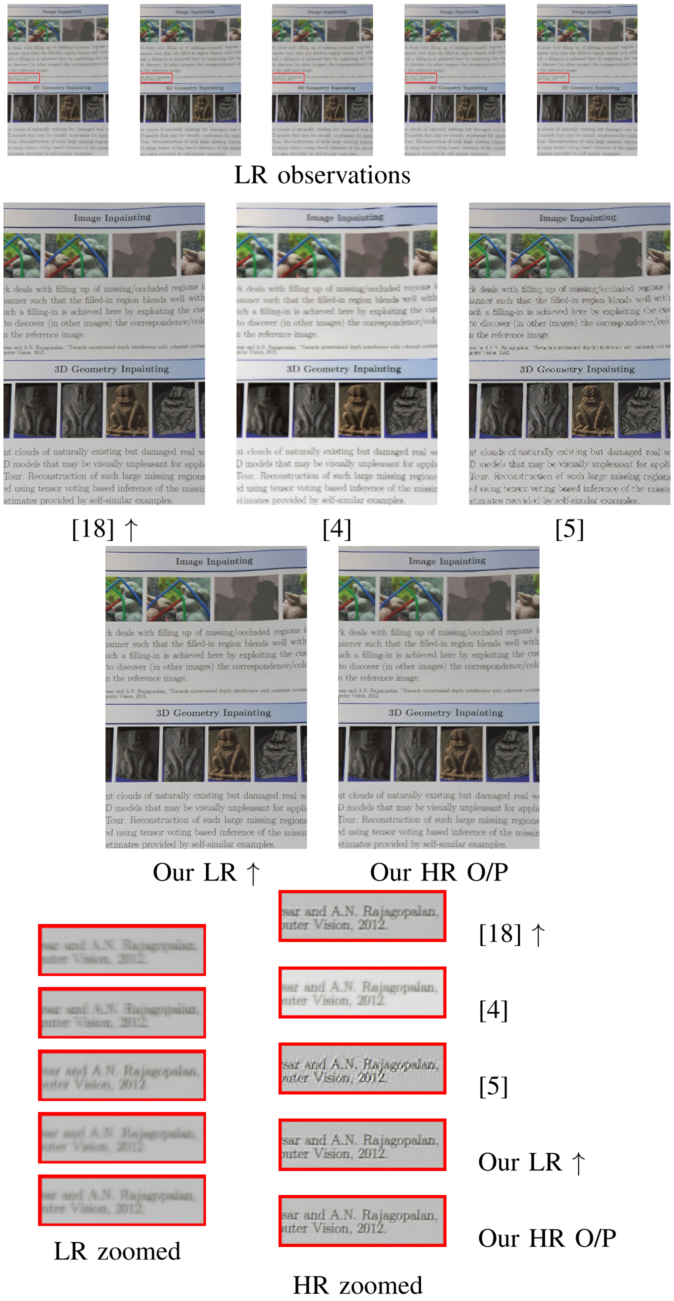


Fig. 10. Row one: blurred LR observations, row two: output image of Paramanand and Rajagopalan [18] super-resolved using [21], result of Sroubek et al. [4] and Ma et al. [5], row three: our LR output super-resolved using [21], and our HR output image, and row four, column one: zoomed-in regions from the LR observations in row one, and column two: zoomed-in regions from [18] ↑, [4], [5] in row two, and our LR ↑, our HR output in row three, respectively.

## IV. Experimental Results

We demonstrate the effectiveness of our proposed approach on synthetic as well as real images. For our synthetic experiments, we use the two metrics, PSNR (Peak Signal to Noise

Ratio) and SSIM (Structure Similarity Measure), to quantify performance.

TABLE II
QUANTITATIVE EVALUATION FOR THE EXAMPLE IN FIG. 2

| Method | [18] ↑ | [4] | [5] | [14] ↑ | W/O AM | Our LR ↑ | Our HR O/P |
|---|---|---|---|---|---|---|---|
| PSNR | 22.85 | 20.97 | 16.82 | 20.37 | 22.22 | 24.71 | **26.41** |
| SSIM | 0.811 | 0.585 | 0.497 | 0.444 | 0.698 | 0.951 | **0.956** |

TABLE III
QUANTITATIVE EVALUATION FOR THE EXAMPLE IN FIG. 7

| Method | [18] ↑ | [4] | [5] | Our LR ↑ | Our HR O/P |
|---|---|---|---|---|---|
| PSNR | 24.42 | 19.95 | 18.20 | 22.94 | **30.28** |
| SSIM | 0.813 | 0.859 | 0.649 | 0.803 | **0.892** |

We begin with a simple synthetic example of a planar scene with no depth variations. The image of an eye-chart was used as the latent HR image (see row two, column one of Fig. 7). The space-variantly blurred LR images are shown in row one of Fig. 7. They were generated in a similar manner as the LR observations in row two of Fig. 2. The parameters of the camera motion space $\mathcal{C}$ were chosen as: $T_X, T_Y = (-10 : 1 : 10$ pixels), and $\theta_Z = (-2^\circ : 0.2^\circ : 2^\circ)$. To assess the proposed method's robustness to noise, we chose a lower SNR in this case than for the example in Fig. 2; Gaussian noise with an SNR of 20 dB was added. The initial depth map obtained using optical flow is shown in row three, and it can be observed that although the OF algorithm correctly assigns the same depth value to most pixels in the image, certain small segments are in error. From the dominant depth layer in this depth map, we used the algorithm of [32] to identify pixel locations in the first LR observation that are suited for PSF estimation. Next, we randomly selected four spatially separated locations from this set. Patches around these locations were cropped from all the LR observations. We used the algorithm of [4] to determine the HR blur kernels corresponding to these patches at each of the four locations. The HR trajectories associated with each LR observation were then computed from these HR PSFs using the method described in Section III-B. The recovered depth map and the deblurred HR image obtained using our alternating minimization framework are shown in row three of Fig. 7. Our estimated depth map, in accordance with the scene, is planar i.e., has the same depth value at all pixels. For comparison, we super-resolved by a factor of two the LR output image returned by the multi-image blind deblurring method of Paramanand and Rajagopalan [18] using the single image SR algorithm of [21]. This result is shown in row two of Fig. 7. All five LR observations were given as input to the algorithm of [18]. The code of [21] is publicly available. The outputs obtained using the convolution SR model in [4], and the SR technique of [5] based on least blurred pixels are also provided in row two of Fig. 7. The SR code of [4] was made available to us on request by the authors, while the implementation of [5] is available online. We also performed a test wherein our own proposed framework was applied, but with the crucial difference that the camera motion, depth map, and image were estimated at *low resolution* (i.e., *LR* PSFs were computed from patches in Step 4 of Algorithm 1, and the camera motion in Step 5 was computed from these LR PSFs on an *LR* grid instead of HR. Subsequently, both the latent image estimation in Step 9 and the depth map estimation in Step 10 were implemented at *low resolution*). The output LR image produced by this pipeline (which, in effect, functions as a depth-aware multi-image deblurring method) was super-resolved using the SR algorithm of [21], and the result of this baseline comparison is shown in row three, column one of Fig. 7. Zoomed-in regions from the five blurred LR observations and the HR images (see caption of Fig. 7) are presented for comparison. The methods of Sroubek et al. [4] and Ma et al. [5] do not perform well since both assume space-invariant blur across the LR images. Although the deblurring scheme of Paramanand and Rajagopalan [18] allows for space-varying blur, both motion and image are computed at LR. This is also the case with the LR baseline comparison. Unlike the fragmented LR pipeline, our proposed method performs *joint* deblurring and resolution enhancement by directly computing HR motion from blurred LR frames. The improvement in deblurring quality over the methods of [18] as well as our own LR baseline can be distinctly observed from the HR zoomed-in regions. The text is crisp and clearly legible in our result.

The output image of Paramanand and Rajagopalan [18] super-resolved using the SR algorithm of [21], and the depth map returned by the method of [18], for the synthetic experiment in Fig. 2 are shown in Fig. 8. The method of [18] is tailored only for bilayer scenes (notice that their estimated depth map has only two layers), while the example we have considered has four depth layers. Thus, blur is not completely removed from all the foreground layers in their output. The result of Sroubek et al. [4] too has residual blur. The algorithm of Ma et al. [5], on the other hand, oversharpens the image. For comparison, we also provided the least blurred of the five LR observations as input to the depth-aware single image deblurring technique of Hu et al. [14]. The output image of [14] after super-resolution using [21], and the depth map estimated by [14] are shown in the second row. The method of [14] wrongly assigns all three foreground layers to a single depth value leading to poor deblurring quality. The LR baseline output has also been provided, and it can be observed that the result is inferior in quality to the HR image estimated using our proposed scheme. The yellow, green, and blue patches from the first of the five LR images in row two of Fig. 2 have been zoomed-in and displayed. HR zoomed-in regions at the corresponding locations have also been provided for qualitative assessment. We have included one more synthetic experiment in the supplementary material to demonstrate our proposed method's ability to handle even inclined planes and smoothly-varying depth values.

To quantify the performance of our algorithm we evaluated the PSNR and SSIM measures, and these values are presented in Tables II and III. The performance improvement achieved by our method over the state-of-the-art is quite evident from the results.

For a more extensive evaluation of our algorithm's noise handling capabilities, following [4], we added Gaussian noise with SNR varying from 50 dB to 10 dB, and reran the experiments in Figs. 2 and 7. We repeated the whole procedure ten times for different realizations of noise. The plot of Fig. 9

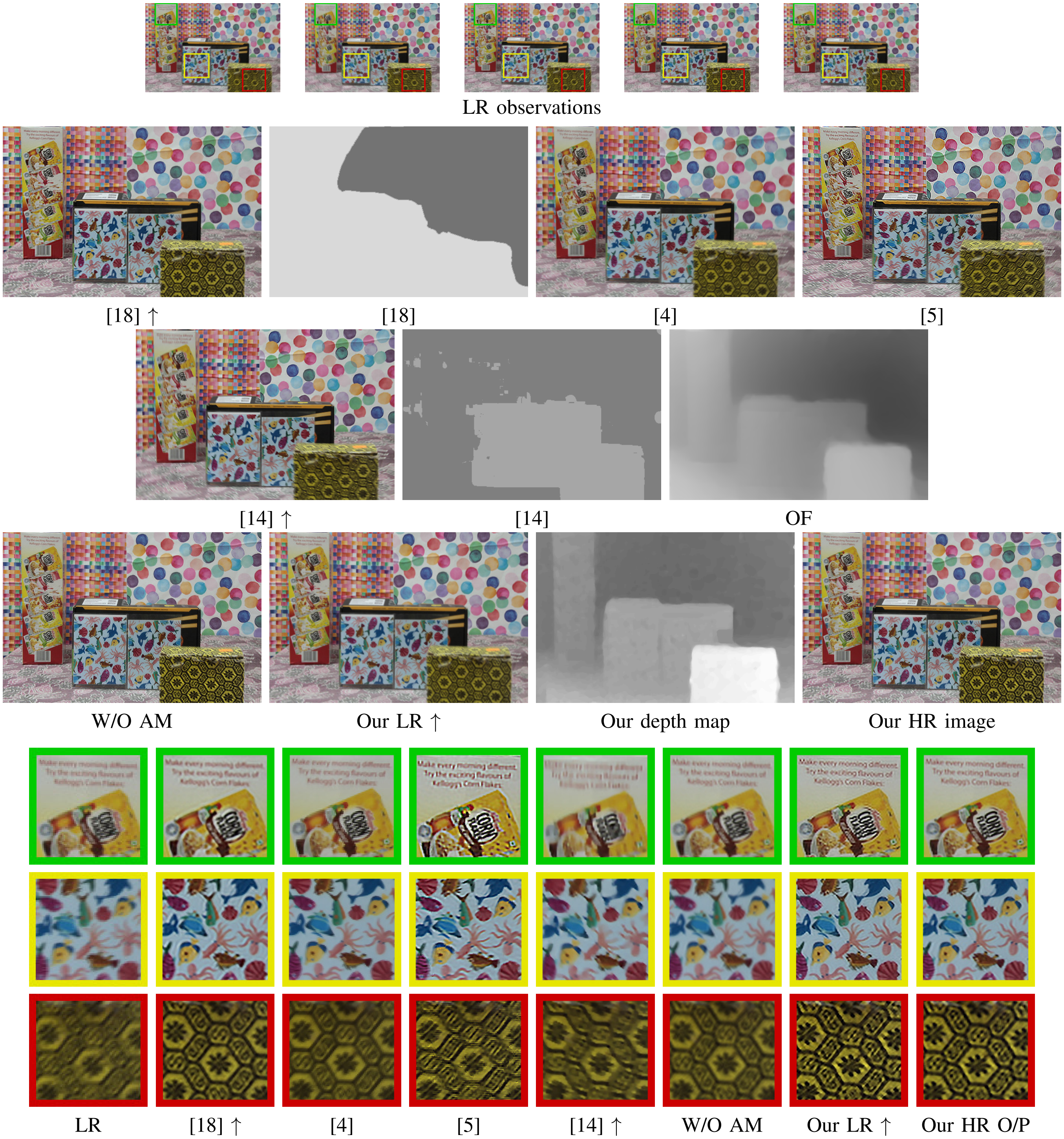
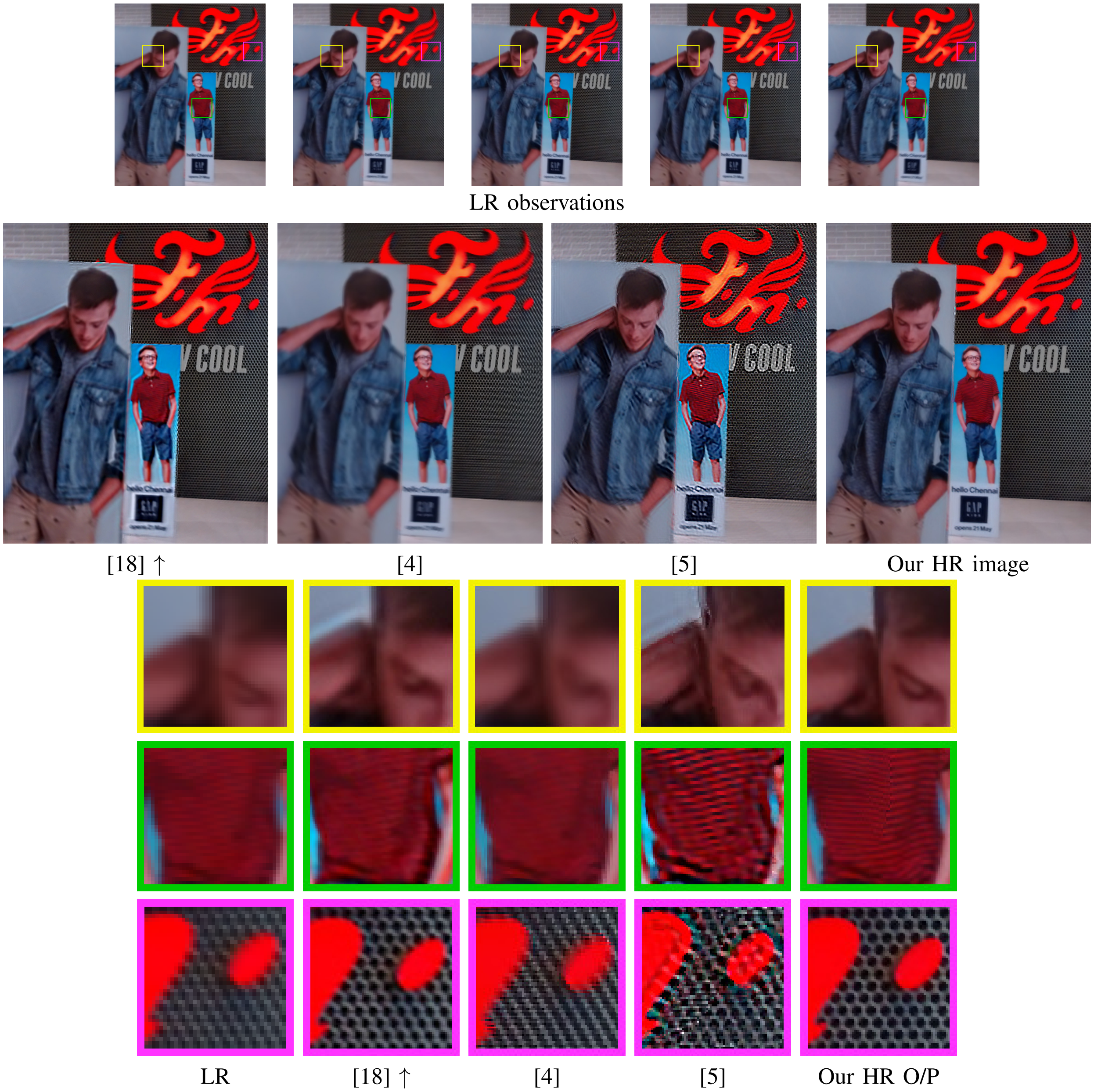


Fig. 11. Row one: blurred LR observations, row two: output image of Paramanand and Rajagopalan [18] super-resolved using [21], depth map of [18], results of Sroubek et al. [4] and Ma et al. [5], row three: output image of Hu et al. [14] super-resolved using [21], depth map of [14], depth map from optical flow, row four: result obtained without AM, our LR result super-resolved using [21], our depth map and our HR output image, and row five: zoomed-in regions from the first LR observation in row one, [18] ↑, [4], [5] in row two, [14] ↑ in row three, without AM, our LR ↑, and our HR output image in row four, respectively.

summarizes the obtained outputs in terms of average PSNR. Qualitative results have been included in the supplementary material. In practice, the level of noise depends on the amount of light during acquisition and also on the quality of the sensors. Most cameras today have SNR around 50 dB, but with decreasing illumination, it can drop to 30 dB [4]. It can be seen from the plot of Fig. 9 that our proposed algorithm maintains stable performance over this practically encountered range (50 to 30 dB). For very noisy images (20 dB and below), a drop in performance was observed. However, under normal capture conditions, such a high level of noise is uncommon.

The data for the real experiments in Figs. 10 to 14 were



LR observations

[18] ↑ [4] [5] Our HR image

LR [18] ↑ [4] [5] Our HR O/P

Fig. 12. Row one: blurred LR observations, row two: output image of Paramanand and Rajagopalan [18] super-resolved using [21], results of Sroubek et al. [4] and Ma et al. [5], and our HR output image, and row three: zoomed-in regions from the first LR observation in row one, [18] ↑, [4], [5], and our HR output in row two, respectively.

captured using a hand-held camera. The super-resolution factor was selected as two for all these examples. The first case in Fig. 10 corresponds to a planar scene of a poster. The input LR images displayed in row one have space-variant blur due to camera shake during image capture. The outputs of Paramanand and Rajagopalan [18] after super-resolving using [21], Sroubek et al. [4], Ma et al. [5], and the LR baseline are provided for comparison against the deblurred SR image obtained using our proposed scheme. Our output is sharp with clearly legible text while competing approaches either have residual blur or deblurring artifacts as can be seen from the zoomed-in regions.

We next consider a scene (see Fig. 11) similar to the real examples in the experimental section of [14]. Two textured wallpapers at a distance of approximately 2 meters from the camera formed the background, while the yellow box on the bottom-right closest to the camera was around a meter away. The translational motion of the camera was dominant in some images, and it can be observed from the LR observations that the foreground depth layers appear more blurred when compared to the layers at the back. The three zoomed-in regions shown in the last row of Fig. 11 are selected from three different depth regions. A visual examination clearly reveals that only our proposed method is able to deblur and super-resolve all three regions correctly. The deblurring technique of [18] can only handle two depth layers and introduces ringing artifacts in the middle depth layer. The SR approaches of [4] and [5] can neither cope with non-uniform blur nor depth

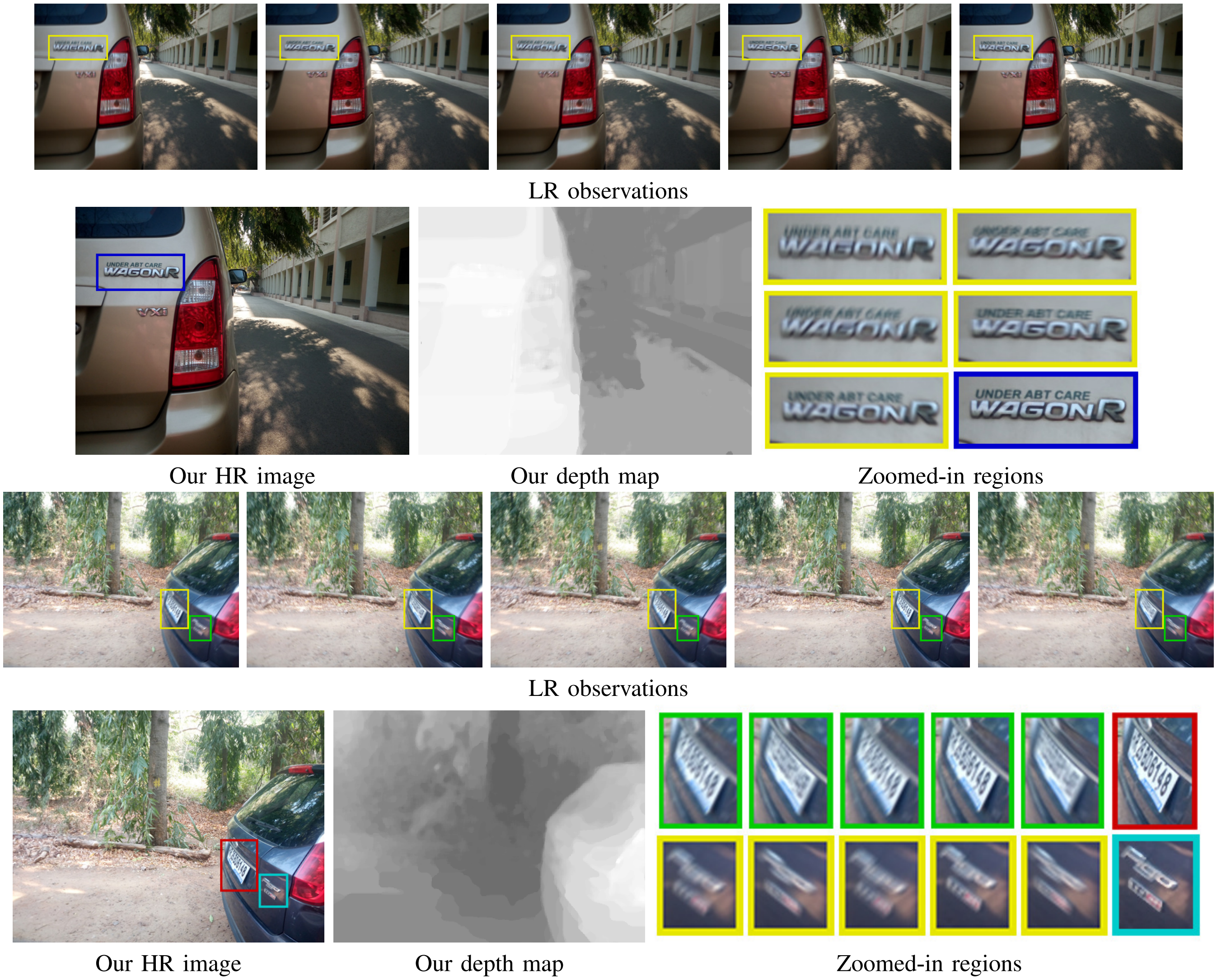


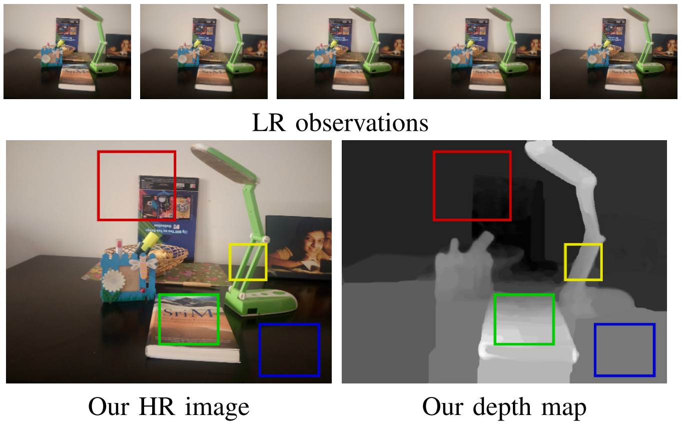
Fig. 13. Two more real examples with challenging depth variations.

variations, while the algorithm of [14] is at a disadvantage since it works with a single image. We would like to add that the first author of [14] provided us with the outputs of their method for the examples in Figs. 8 and 11. Artifacts can be observed in the green zoomed-in region of the LR baseline, whereas the text is legible in our HR output. Also notice how the depth layers corresponding to the background and the cornflakes box on the left, though merged in the initial OF depth map, have been accurately recovered in our estimated depth map.

The next example in Fig. 12 involves greater distances (of the order of 2 to 10 meters) between the camera and the scene. The scene has two advertisement boards placed at two different depths against a background, forming a piece-wise planar 3D scene. The method of [18] does deblur the background. However, the foreground layers are not properly restored (see the zoomed-in regions). The outputs of both [4] and [5] have deblurring artifacts. In contrast, our algorithm reconstructs the striped patterns on the shirt in the middle layer accurately. The black circular patterns in the background too are restored without any artifacts.

The last two real experiments in Fig. 13 are more challenging from the perspective of depth variations in the scene. The examples considered so far had depth layers that were predominantly fronto-parallel planar. The outdoor scenes in Fig. 13 are more difficult in that there are both gradual as well as sharp depth variations in the scene. Even under these demanding situations, our proposed method is quite effective at recovering the depth map, as can be observed from our results. The improvement in quality and legibility of the text post deblurring and super-resolution is evident upon examining the LR and HR zoomed-in regions.

**Limitations:** Our patch-based motion computation step allowed us to pick regions containing texture suitable for kernel estimation. Good texture is key not only to motion estimation but depth map recovery as well. As with other methods ([36]) that estimate depth using motion blur cues, our method too can yield incorrect depth maps if the intensity images have large textureless (i.e., homogeneous) regions. An example is shown in Fig. 14. Observe that although the gradual variation in the depth has been nicely captured in the green patch corresponding to the book, the depth labels in the blue patch are in error because the intensity image is mostly homogeneous in this region and contains no useful information for depth estimation. Similar is the case with the depth values in the red patch that should all ideally have been the same since the background (which is textureless) and the book are at the same depth. Another limitation is our method's inability to distinguish very fine details in the depth map i.e., objects or structures that are comparable to the blur kernel size and are only a few pixels wide (see the yellow patch). However, it should be noted that the deblurred and super-resolved image does not contain any noticeable artifacts or residual blur.


LR observations

Our HR image

Our depth map

Fig. 14. A real example illustrating the limitations of our method.

## V. Conclusions

We proposed a multi-image super-resolution technique that takes non-uniformly motion blurred LR images as input to estimate the latent HR image of the 3D scene. The underlying depth map and the associated HR camera trajectories are obtained as by-products. Global camera motion was estimated using HR PSFs computed from LR patches. While depth was refined using a loopy BP algorithm, a total variation regularizer was used to aid the image estimation step of our AM scheme. The efficacy of the proposed algorithm in advancing the state-of-the-art was amply demonstrated through challenging synthetic and real examples. Qualitative and quantitative evaluations were also provided. As future work, we plan to extend our framework to handle dynamic objects and changing illumination in the input images.

## Acknowledgements

We are grateful to Dr. Filip Šroubek and Dr. Zhe Hu for sharing the codes and results used for comparisons in this paper. The first author is deeply indebted to Dr. Paramanand Chandramouli for his insights and suggestions.

## References

[1] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *Signal Processing Magazine, IEEE*, vol. 20, pp. 21–36, May 2003.

[2] D. Capel, *Image Mosaicing and Super-resolution*. Springer, 2003.

[3] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution.," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[4] F. Sroubek, G. Cristobal, and J. Flusser, "A unified approach to super-resolution and multichannel blind deconvolution," *IEEE Transactions on Image Processing*, vol. 16, pp. 2322–2332, Sept. 2007.

[5] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu, "Handling motion blur in multi-frame super-resolution.," in *Proc. CVPR*, pp. 5224–5232, 2015.

[6] H. Zhang and L. Carin, "Multi-shot imaging: Joint alignment, deblurring, and resolution-enhancement," in *Proc. CVPR*, pp. 2925–2932, June 2014.

[7] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," in *Proc. ACM SIGGRAPH*, pp. 787–794, 2006.

[8] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *ACM Transations on Graphics*, vol. 27, pp. 73:1–73:10, Aug. 2008.

[9] L. Xu and J. Jia, "Two-phase kernel estimation for robust motion deblurring," in *Proc. ECCV*, pp. 157–170, 2010.

[10] Z. Hu and M.-H. Yang, "Fast non-uniform deblurring using constrained camera pose subspace," in *Proc. BMVC*, 2012.

[11] A. Gupta, N. Joshi, C. L. Zitnick, M. Cohen, and B. Curless, "Single image deblurring using motion density functions," in *Proc. ECCV*, pp. 171–184, 2010.

[12] M. Hirsch, C. J. Schuler, S. Harmeling, and B. Scholkopf, "Fast removal of non-uniform camera shake," *Proc. ICCV*, vol. 0, pp. 463–470, 2011.

[13] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, "Non-uniform deblurring for shaken images," *International Journal of Computer Vision*, vol. 98, pp. 168–186, June 2012.

[14] Z. Hu, L. Xu, and M.-H. Yang, "Joint depth estimation and camera shake removal from single blurry image," in *Proc. CVPR*, pp. 2893–2900, 2014.

[15] F. Sroubek and J. Flusser, "Multichannel blind deconvolution of spatially misaligned images," *IEEE Transactions on Image Processing*, vol. 14, pp. 874–883, July 2005.

[16] M. Delbracio and G. Sapiro, "Burst deblurring: Removing camera shake through fourier burst accumulation.," in *Proc. CVPR*, pp. 2385–2393, 2015.

[17] A. Ito, A. C. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk, "Blurburst: Removing blur due to camera shake using multiple images," *ACM Transactions on Graphics, Submitted*, vol. 3, no. 1, 2014.

[18] C. Paramanand and A. N. Rajagopalan, "Non-uniform motion deblurring for bilayer scenes," in *Proc. CVPR*, pp. 1115–1122, June 2013.

[19] H. S. Lee and K. M. Lee, "Dense 3D reconstruction from severely blurred images using a single moving camera," in *Proc. CVPR*, pp. 273–280, June 2013.

[20] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, pp. 2861–2873, Nov 2010.

[21] Y. Zhu, Y. Zhang, and A. Yuille, "Single image super-resolution using deformable patches," in *Proc. CVPR*, pp. 2917–2924, June 2014.

[22] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. ICCV*, pp. 349–356, Sept 2009.

[23] S. Babacan, R. Molina, and A. Katsaggelos, "Variational bayesian super resolution," *IEEE Transactions on Image Processing*, vol. 20, pp. 984–999, April 2011.

[24] S. Villena, M. Vega, D. Babacan, R. Molina, and A. Katsaggelos, "Bayesian combination of sparse and non sparse priors in image super resolution," *Digital Signal Processing*, vol. 23, no. 2, pp. 530–541, 2013.

[25] A. Snchez-Beato, "Coordinate-descent super-resolution and registration for parametric global motion models.," *Journal of Visual Communication and Image Representation*, vol. 23, no. 7, pp. 1060–1067, 2012.

[26] U. Mudenagudi, A. Gupta, L. Goel, A. Kushal, P. Kalra, and S. Banerjee, *Proc. ACCV*, ch. Super Resolution of Images of 3D Scenes, pp. 85–95. 2007.

[27] A. V. Bhavsar and A. N. Rajagopalan, "Resolution enhancement in multi-image stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1721–1728, Sept 2010.

[28] H. S. Lee and K. M. Lee, "Simultaneous super-resolution of depth and images using a single camera," in *Proc. CVPR*, pp. 281–288, June 2013.

[29] Y.-W. Tai, P. Tan, and M. Brown, "Richardson-Lucy deblurring for scenes under a projective motion path," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1603–1618, 2011.

[30] C. Paramanand and A. N. Rajagopalan, "Inferring image transformation and structure from motion-blurred images.," in *Proc. BMVC*, pp. 1–12, 2010.

[31] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” in *Proc. ECCV*, vol. 3024, pp. 25–36, May 2004.
[32] Z. Hu and M.-H. Yang, “Good regions to deblur,” in *Proc. ECCV*, pp. 59–72, 2012.
[33] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.
[34] Y. Wei and L. Quan, “Asymmetric occlusion handling using graph cut for multi-view stereo,” in *Proc. CVPR*, pp. 902–909, 2005.
[35] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient belief propagation for early vision.,” *International Journal of Computer Vision*, vol. 70, no. 1, pp. 41–54, 2006.
[36] C. Paramanand and A. N. Rajagopalan, “Shape from sharp and motion-blurred image pair,” *International Journal of Computer Vision*, vol. 107, no. 3, pp. 272–292, 2014.

**Abhijith Punnappurath** received his B.Tech. degree in Electronics and Communications Engineering in 2009. He is currently pursuing his Ph.D. degree at the Indian Institute of Technology Madras, India. His research interests lie in the areas of computer vision and image processing. He has worked on face recognition, super-resolution, dynamic object segmentation, and change detection, from moving cameras.

**T. M. Nimisha** received the B.Tech. degree in Electronics and Communications Engineering from the Amrita College of Engineering, Kollam India, in 2011, and the M.Tech. degree in Signal Processing specialization from National Institute of Technology, Calicut, India, in 2013. She is currently pursuing the Ph.D. degree at IIT Madras, Chennai, India. Her research interests lie in the areas of computer vision and image processing with the main focus on image restoration problems.

**Ambasamudram Narayanan Rajagopalan** received the Ph.D. degree in Electrical Engineering from IIT Bombay in 1998. He was with the Center for Automation Research, University of Maryland, as a Research Faculty Member, from 1998 to 2000. He joined the Department of Electrical Engineering, IIT Madras, in 2000, where he is currently a Professor. He is co-editor of the book titled Motion Deblurring: Algorithms and Systems (Cambridge University Press, 2014). He is co-author of the book titled Depth From Defocus: A Real Aperture Imaging Approach (New York: Springer-Verlag, 1999). His research interests include 3D structure from blur, motion deblurring, registration, super-resolution, video microscopy, inpainting, matting, heritage resurrection, face recognition, image forensics, and underwater imaging. He is a Fellow of the Alexander von Humboldt Foundation, Germany, the Indian National Academy of Engineering, and the Institution of Electronics and Telecommunication Engineers, India (by invitation). He was a recipient of the DAE-SRC Outstanding Investigator Award in 2012, the VASVIK Award in 2013, and the Mid-Career Research and Development Award from IIT Madras in 2014. He served as Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence from 2007 to 2011, as Associate Editor of IEEE Transactions on Image Processing from 2012 to 2016, and is currently Senior Area Editor for TIP. He was Area Chair for CVPR 2012 and ICPR 2012, and Program Co-Chair for ICVGIP 2010.