Super-Resolved Free-Viewpoint Image Synthesis Combined With Sparse-Representation-Based Super-Resolution

Ryo Nakashima*, Keita Takahashi† and Takeshi Naemura*
* Graduate School of Information Science and Technology, The University of Tokyo, Japan
E-mail: {nakashima,naemura}@nae-lab.org
† Graduate School of Engineering, Nagoya University, Japan
E-mail: keita.takahashi@ieee.org

***Abstract*—We consider super-resolved free-viewpoint image synthesis (SR-FVS), where a high-resolution (HR) image that would be observed from a virtual viewpoint is synthesized from a set of low-resolution multi-view images. In previous studies, methods for SR-FVS were proposed on the basis of reconstruction-based super-resolution (RB-SR). RB-SR uses multiple images to synthesize an HR image and thereby can naturally be applied to SR-FVS, where multi-view images are given as the input. However, the quality of the synthesized image depends on observation conditions such as the depth of the target scene, so sometimes the quality of SR-FVS can degrade severely. To mitigate such degradation, we propose integrating learning-based super-resolution (LB-SR), which uses knowledge learned from massive natural images, into the SR-FVS process. In this paper, we adopt sparse coding super-resolution (ScSR) as a LB-SR method and combine ScSR with an existing SR-FVS method.**

I. Introduction

Free-viewpoint image synthesis is the process of combining a set of multi-view images to synthesize an image from a new viewpoint where no camera is actually located. This technology has been an active area of research [1], [2], [3], [4], [5] because it can provide realistic 3-D visual experiences by enabling users to select viewpoints freely and interactively.

Conventional methods for free-viewpoint image synthesis generally consist of two steps. First, the 3-D structure of the target scene is reconstructed from input images. Then, using the reconstructed 3-D structure, the input images are registered to the coordinate system of the target image and blended to produce a new image. The blending operation in the second step can obscure the errors of 3-D reconstruction by blurring the image. However, this operation also blurs fine textures and degrades the quality of the synthesized image.

To improve the image quality, several researchers [2], [3], [4], [5] have recently replaced blending operation with reconstruction-based super-resolution (RB-SR) [6], in which a high-resolution (HR) image is reconstructed from multiple low-resolution (LR) images. We refer to this task as super-resolved free-viewpoint image synthesis (SR-FVS). Tung et al. [2] super-resolved input multi-view images to generate a complete 3-D model of a single object. For the same purpose, Goldluecke et al. [3] synthesized texture maps using RB-SR. Mudenagudi et al. [4] formulated a SR reconstruction of an entire scene as a multi-label MRF-MAP problem and solved it by graph cuts. Takahashi et al. [5] proposed a SR-FVS method that uses adaptive regularization for RB-SR to handle depth inaccuracies. It should be noted that all these methods use RB-SR for image synthesis.

In RB-SR, the SR problem is formulated as an inversion of an observation model, which describes how the LR images are generated from the underlying HR image. Therefore, the quality of the synthesized image depends on the observation model. If the observed LR images contain sufficient information to reconstruct the HR image, the quality will be high. However, if some information is lost due to occlusions or other factors[1], the quality will degrade. This degradation is undesirable for the reconstruction of the entire scene.

To mitigate such degradation, we propose integrating another class of SR approach, learning-based SR (LB-SR), into the RB-SR-FVS method [5]. LB-SR methods [7], [8], [9] synthesize an HR image by reproducing image features learned from a massive amount of natural images. Therefore, LB-SR can mitigate degradations of RB-SR due to the lack of information because such information can be compensated by the learned features. Moreover, our method can also overcome a limitation of LB-SR that it cannot recover features that are not present in training images because such features can be recovered with RB-SR as long as sufficient information is present in the input LR images. In summary, our method improves the quality of SR-FVS by utilizing RB-SR and LB-SR in a complementary manner. We adopt sparse coding super-resolution (ScSR) [8] as a LB-SR algorithm and combine ScSR with an existing RB-SR-FVS method [5].

The rest of this paper is organized as follows. In Section II, the RB-SR-FVS method [5] is introduced. In Section III, we propose our SR-FVS method that integrates ScSR [8] into the SR-FVS process. Experimental results are presented in Section IV, followed by the conclusion in Section V.

[1]For example, it is well known that when pixel shifts between images (i.e. disparity) are integers, all images contain the same information and thus RB-SR cannot improve resolution [6].

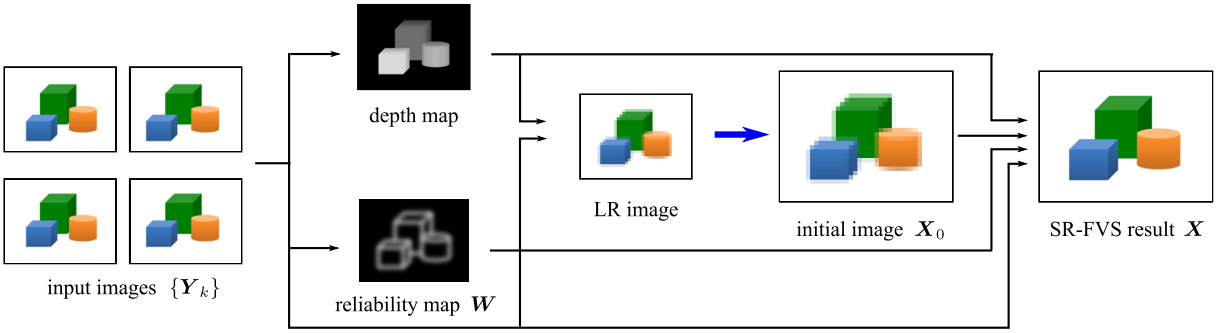


Fig. 1. Flowchart of SR-FVS. In the reconstruction-based method [5], a basic interpolation is used in the process of generating an initial image from a LR image (indicated by blue arrow). We substitute this interpolation with a LB-SR method.

II. Reconstruction-based SR-FVS method

The configuration in the SR-FVS method [5] is as follows. The input images are captured from viewpoints that are arranged on the same plane. The goal of our method is to generate an image that would be viewed from a new viewpoint, referred to as the target viewpoint. In this paper, we assume that the target viewpoint is located on the camera plane. We also assume that camera parameters are estimated in advance.

The SR-FVS method [5] consists of two steps: depth estimation and SR. In the depth estimation step, we estimate a depth map and a reliability map, which represents per-pixel reliability of the estimated depth, both from the target viewpoint. A modified version of semi-global stereo matching [10] is adopted. In the SR step, an HR image from the target viewpoint is synthesized with RB-SR using the estimated depth map and reliability map. In this paper, we leave the depth estimation step unchanged from that in [5], but newly combine ScSR [8] with the SR step. In the following text, we explain the SR step of the original method [5] as a preparation for our proposal.

A. View synthesis via RB-SR

An overview of the SR step is shown in Fig. 1. We first generate an LR image viewed from the target viewpoint by warping input images using the depth map and blending the warped images. The LR image is next upsampled to the target resolution by using a basic interpolation (e.g., bicubic interpolation in [5]) to obtain an interpolated image. We assume that the interpolated image is close to the underlying HR image, and use this image as an initial guess for SR reconstruction. Note that SR is not yet performed, and thereby, the effective resolution of the interpolated image is not increased from that of the LR image.

Then, an output HR image is synthesized via an inversion of an observation model as follows. Let K be the number of the LR images. Let $\boldsymbol{X}$, $\boldsymbol{X}_0$, and $\boldsymbol{Y}_k$ $(1 \leq k \leq K)$ be 1-D vector representations of the output HR image, the interpolated image, and the k-th input image, respectively. The relation between the k-th input image and the latent HR image $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{Y}_k = \boldsymbol{A}_k \boldsymbol{X}, \tag{1}$$

where $\boldsymbol{A}_k$ is a matrix that represents the observation model for the k-th input image. $\boldsymbol{A}_k$ can be decomposed as

$$\boldsymbol{A}_k = \boldsymbol{M}\boldsymbol{B}\boldsymbol{S}_k, \tag{2}$$

where $\boldsymbol{M}$ and $\boldsymbol{B}$ are matrices that respectively represent downsampling and blurring operations, which come from the difference of pixel sizes between the LR and HR images. $\boldsymbol{S}_k$ is a matrix that represents pixel shift (disparity) between the input and target images and is computed from the estimated depth map and camera parameters.[2]

The SR problem is formulated as a minimization of an energy function E_r given by

$$E_r(\boldsymbol{X}) = \sum_{k=1}^{K} \|\boldsymbol{Y}_k - \boldsymbol{A}_k\boldsymbol{X}\|_2^2 + \mu(\boldsymbol{X}-\boldsymbol{X}_0)^{\mathsf{T}}\boldsymbol{W}(\boldsymbol{X}-\boldsymbol{X}_0), \tag{3}$$

where μ is a positive weight and $\boldsymbol{W}$ is a diagonal matrix that represents the reliability of depth estimation. An element of $\boldsymbol{W}$ takes a small value when the estimated depth of the corresponding pixel is reliable and a large value when the estimated depth is unreliable. For details of $\boldsymbol{W}$, please refer to Appendix A.

The first term represents a reconstruction constraint, which means that the HR image $\boldsymbol{X}$ should reproduce the observed LR image $\boldsymbol{Y}_k$ after applying the observation model $\boldsymbol{A}_k$ as described in Eq. (1). The second term is a regularizer that assumes the output image $\boldsymbol{X}$ resembles the initial image $\boldsymbol{X}_0$. The matrix $\boldsymbol{W}$ determines the weight of the regularization adaptively for each pixel by using the reliability of the depth. When the estimate depth for a pixel is reliable, the weight of regularization for that pixel is set to a small value to prioritize the reconstruction constraint. Conversely, when the depth estimation for a pixel is less reliable (e.g. around occlusion boundaries), the weight is set to a larger value because the observation model (the first term) is less reliable.

The quality of the synthesized image is mainly determined by the observation model. When the observed LR images have enough information to reconstruct the HR image, fine

[2]Occlusions are considered in constructing $\boldsymbol{A}_k$. In short, if pixels of the target HR image are occluded in the k-th LR image, the corresponding columns of $\boldsymbol{A}_k$ are set to zero. The detailed procedure of occlusion handling can be found in section 4.2 of [5].

details can be recovered with the reconstruction constraint (the first term in Eq. (3)), so the quality of the resulting image becomes high. However, when such information is lost during the observation process, the reconstruction constraint cannot improve the resolution, so the quality will severely degrade. To mitigate such degradation, our method integrates a LB-SR method, ScSR, into the process of SR-FVS.

III. Proposed method

We integrate ScSR [8] into the RB-SR-FVS method [5] introduced in the previous section. In short, our method uses ScSR to generate an initial image for RB-SR ($\boldsymbol{X}_0$ in Eq. (3)). More precisely, we first synthesize an LR image from the target viewpoint, as explained in the previous section. Then we apply ScSR to the synthesized LR image to generate an initial image, instead of interpolation used in the previous section; in fact, this is the only modification to the flowchart in Fig. 1. We finally perform RB-SR by minimizing Eq. (3) to obtain the output HR image.

Our method can mitigate a drawback of RB-SR that the quality of SR-FVS depends on the observation model. This is because when the reconstruction constraint is not effective, the resulting image converges to the initial image, which is already super-resolved by ScSR. Also, note that our method is less subject to a limitation of ScSR that it cannot recover features that are not present in training images. This is because RB-SR can recover such features by using the reconstruction constraint. Therefore, our method improves the SR-FVS quality by combining RB-SR and LB-SR in a complementary manner.

In the remainder of this section, we briefly introduce ScSR and explain how ScSR can be utilized for SR-FVS.

A. Sparse coding super-resolution

ScSR is a LB-SR method designed for single image SR, and thus, it was not originally designed for SR-FVS. The key feature of ScSR is the use of a sparsity prior, which assumes that patches of natural images can be represented as a sparse linear combination of atoms of an appropriate dictionary.

ScSR consists of two steps: patch-wise SR and global reconstruction. In the patch-wise SR step, patches are extracted from an input LR image and then super-resolved via sparse representation to obtain an HR image. In the global reconstruction step, the HR image is refined with the reconstruction constraint. In fact, the latter step is equivalent to the RB-SR explained in Section II, and thus, it can be omitted. We use only the first patch-wise SR step, which will be introduced next, to generate an initial image for RB-SR.

B. Super-resolution via sparse representation

In the patch-wise SR step, we use two coupled dictionaries, $\boldsymbol{D}_H$ for HR patches and $\boldsymbol{D}_L$ for LR patches, which are trained in advance from natural image patches. Given an LR patch $\boldsymbol{y}$, we first compute a sparse representation of $\boldsymbol{y}$ with respect to the LR dictionary $\boldsymbol{D}_L$. We denote the coefficients of the sparse representation as $\boldsymbol{\alpha}$. $\boldsymbol{\alpha}$ can be computed via minimization of an energy function, E_s, given by

$$E_s(\boldsymbol{\alpha}) = \|\boldsymbol{F}\boldsymbol{y} - \boldsymbol{F}\boldsymbol{D}_L\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \tag{4}$$

where $\boldsymbol{F}$ is a feature extraction operator (e.g., derivative filter in [8]) and λ is a weight for controlling the sparseness of $\boldsymbol{\alpha}$.

Using the computed coefficient $\boldsymbol{\alpha}$ and the HR dictionary $\boldsymbol{D}_H$, the HR patch $\boldsymbol{x}$ is computed as

$$\boldsymbol{x} = \boldsymbol{D}_H\boldsymbol{\alpha}. \tag{5}$$

This procedure is applied to all LR patches to produce the initial image $\boldsymbol{X}_0$, which is used for RB-SR to obtain the output HR image.

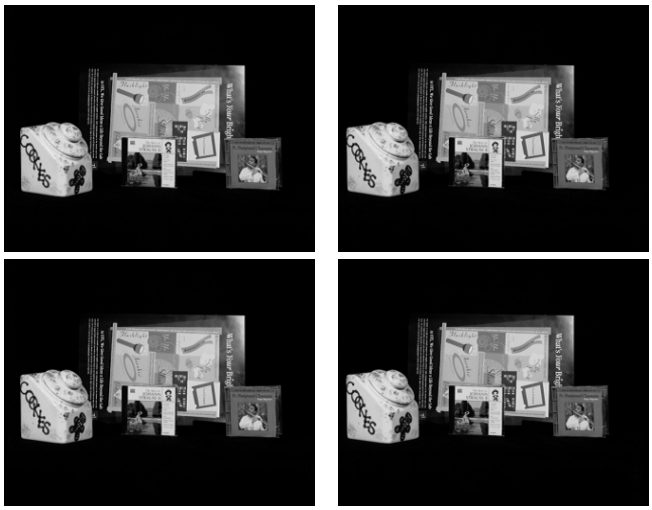


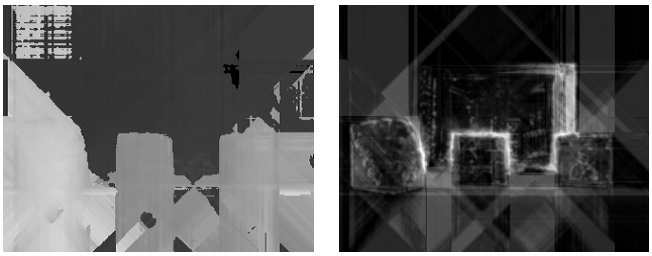
Fig. 2. Input images

IV. Experiments

A. Subjective evaluation

The four images shown in Fig. 2, which were taken from "CD cases and a poster (unoccluded)" in the Stanford light field dataset [11], were used as the input. The input viewpoints were located at the corners of a $60 \times 50\mathrm{mm}$ rectangle. In accordance with the notation of the database, the viewpoints were indexed as 30, 32, 72, and 74. The original images were 650×515 pixels in RGB color. We converted them to grayscale and then resized them to 325×258 pixels to generate the input images. To reduce the image size, we first padded the original image with an intensity of zero at the bottom edge by one pixel and then averaged 2×2 pixels of the padded image to produce a pixel of the resized image. From these LR images, we reconstructed the HR image viewed from the center of the rectangle, which was indexed as 52 in accordance with the database notation.

A depth map and a reliability map, shown in Fig. 3, were estimated by using the stereo matching method described in [5]. The estimated depth map is very noisy around the black background regions because it is inherently impossible to reliably determine stereo correspondences. However, these



(a) Depth map (b) Reliability map

Fig. 3. Depth map and reliability map. Linear structures are visible in the background due to the nature of the semi-global optimization used in [5].

errors do not affect visual quality of the resulting image because colors of these regions can be determined as black even if the estimated depths are erroneous.

We used dictionaries $\boldsymbol{D}_L$ and $\boldsymbol{D}_H$ trained with 69 images, which are available at the website of the authors of ScSR [8]. We set $\mu = 3.0 \times 10^{-19}$ in Eq. (3) and $\lambda = 0.2$ in Eq. (4).

We compared four SR-FVS methods: (a) without SR (bicubic interpolation of the LR image), (b) RB-SR only (equivalent to [5]), (c) LB-SR only, and (d) RB-SR and LB-SR (our method). The SR-FVS results are shown in Fig. 4.

We first focus on the region around the CD case (middle column of Fig. 4). The result without SR (a) was very blurry. For example, letters on the upper part of the CD case were not readable. With LB-SR only (c), these letters were still unreadable, although some textures became clear. This is because these letters were missing in the training images. In contrast, the methods with RB-SR ((b) and (d)) produced much better results, and the letters were much more clearly visible. These results confirm the effectiveness of RB-SR.

Next, let us look at the region around the illustration of a lamp (right column of Fig. 4). Similarly to the above, the result without SR (a) was blurry. With RB-SR only (b), although the result looked much better, it still contained aliasing. For example, jaggies were observed at the edges around the lamp illustration. This indicates that information to reconstruct these edges was missing in the input images, and thus, RB-SR alone could not improve the resolution. These artifacts were not present in the results from the methods with LB-SR ((c) and (d)). Therefore, LB-SR was also effective for SR-FVS.

In summary, some parts are improved by RB-SR, while other parts are improved by LB-SR, and our method received benefits from both of the approaches. Thereby, our method produced the best results among these four methods.

B. Quantitative evaluation

We also quantitatively evaluated the quality of the synthesis. We used five multi-view image datasets; the CD cases (used in the previous experiment), Humvee, Bunny, Bulldozer, and City datasets. The CD cases, Humvee, Bunny and Bulldozer datasets were taken from the Stanford light field dataset [11], and the City dataset was taken from the Multi-View Image Database of University of Tsukuba, Japan. Configurations and parameters for each dataset are summarized in Table I. We generated an HR image view from a virtual viewpoint using four input images, as described in the previous subsection. The ground-truth images for the datasets (except for the CD cases dataset) are shown in Fig. 5.

We calculated the PSNR values against the ground-truth images to evaluate the image quality. As for the CD cases, Humvee, Bunny, and Bulldozer datasets, the background regions are entirely black. Around these regions, pixel colors can be determined as black regardless of what kind of SR is used, and thus these pixels are not suitable for our evaluation. Therefore, we excluded these regions when calculating PSNR. More precisely, we segmented objects by hand and removed the pixels which were more than 10 pixels out of the object boundaries. As for the City dataset, we excluded 24 pixels from the image boundaries to avoid the effect of non-overlapping regions between the input images.

The calculated PSNR values are shown in Table II. It is obvious that (d) RB-SR and LB-SR consistently produced the best results among the four methods, because this method comprises the advantages of RB-SR and LB-SR. The improvement of the proposed method over (b) RB-SR and (c) LB-SR may seem to be slight. This is because most of the details in the image can be recovered either by RB-SR or by LB-SR, and thus only the small portion of the image receives the benefit from the proposed method that combines RB-SR and LB-SR. This improvement contributes little to the PSNR of the entire image. Yet, when focusing on the improved parts, the effectiveness of our method is apparent, as shown in the previous experiment.

Next, let us compare (b) RB-SR only and (c) LB-SR only. (b) was better for the CD cases dataset and (c) was better for the other datasets. These results can be explained by different characteristics of RB-SR and LB-SR. RB-SR uses information from input images, thus RB-SR is suitable when input images contain much information to reconstruct a latent HR image. On the other hand, LB-SR uses learned features, thereby LB-SR performs well when an HR image can be well expressed using these features. Therefore, to generate a high-quality image for both cases, both RB-SR and LB-SR are necessary. These results also support the effectiveness of our method.

V. Conclusion and Future Work

We proposed a novel SR-FVS method that utilizes RB-SR and LB-SR in a complementary manner. We integrate ScSR [8] into the existing SR-FVS method [5]. We confirmed the effectiveness of our method through experiments.

Our primary future work is to speed up computation. In our unoptimized MATLAB implementation, it took about 5 minutes to generate a single image for the CD cases dataset. Most of the computation time was taken to find sparse representation for each image patch (optimization of Eq. (4)). Such patch-wise computations can be executed in parallel using Graphics Processing Unit (GPU). Actually, some researchers [12], [13] have demonstrated that objective functions that are similar to Eq. (4) can be efficiently optimized using GPU. Therefore, we are now planning to implement our method on GPU.

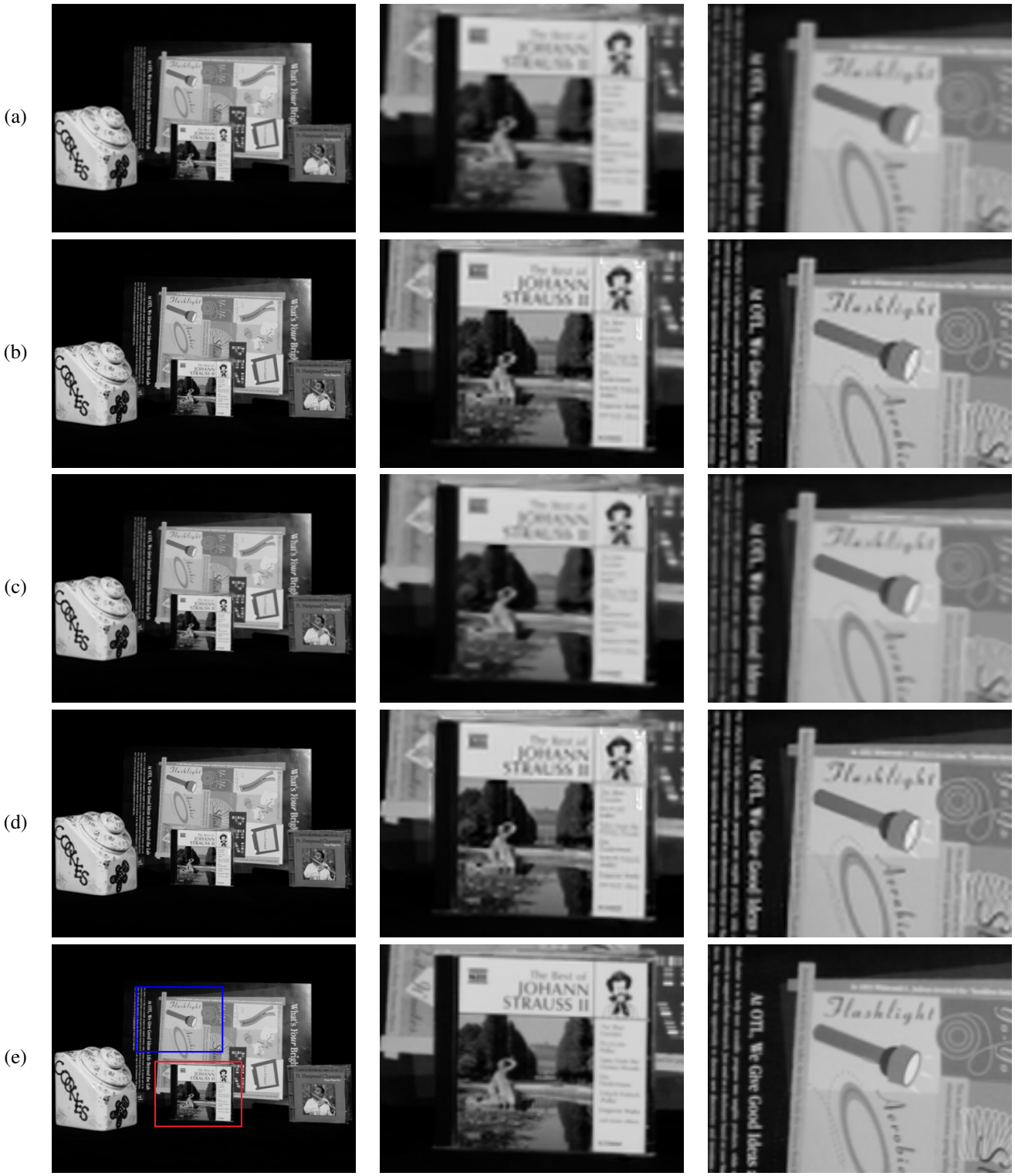


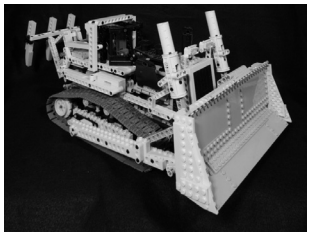
Fig. 4. SR-FVS results. (a) without SR, (b) RB-SR only [5], (c) LB-SR only, (d) RB-SR and LB-SR (proposed method), (e) ground truth. Left column: whole image, middle column: close-up around CD case (indicated by the red rectangle), right column: close-up around illustration of lamp (indicated by the blue rectangle). Best viewed on screen.

TABLE I
CONFIGURATIONS OF DATASETS. INPUT VIEWPOINTS AND TARGET VIEWPOINT ARE DISPLAYED ACCORDING TO DATABASE NOTATION.

Datasets	CD cases	Humvee	Bunny	Bulldozer	City
LR image size (pixels)	325×258		256×256	384×288	160×120
HR image size (pixels)	650×515		512×512	768×576	320×240
Input viewpoints	30, 32, 72, 74	119, 121, 151, 153	(7, 7), (7, 9), (9, 7), (9, 9)		(1, 6), (1, 8), (3, 6), (3, 8)
Target viewpoint	52	136	(8, 8)		(2, 7)
μ (in Eq. (3))	3.0×10^{-19}	5.0×10^{-13}	1.0×10^{-12}	3.0×10^{-19}	5.0×10^{-13}
λ (in Eq. (4))	0.2				


Humvee


Bunny


Bulldozer


City

Fig. 5. Datasets used for quantitative evaluation.

TABLE II
PSNR VALUES (dB) OF SYNTHESIZED IMAGES.

Dataset	CD cases	Humvee	Bunny	Bulldozer	City
(a) without SR	28.30	29.80	35.43	28.36	31.73
(b) RB-SR only	30.00	31.32	38.11	31.02	33.02
(c) LB-SR only	29.75	31.37	39.22	30.58	32.67
(d) RB-SR and LB-SR (proposed)	**30.68**	**32.19**	**40.41**	**32.18**	**33.30**

ACKNOWLEDGMENT

Ryo Nakashima is supported by JSPS Research Fellowships for Young Scientists and KAKENHI Grant Number $24 \cdot 8321$.

REFERENCES

[1] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview imaging and 3dtv," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 10–21, 2007.

[2] T. Tung, S. Nobuhara, and T. Matsuyama, "Simultaneous super-resolution and 3d video using graph-cuts," in Proceedings of *Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

[3] B. Goldluecke and D. Cremers, "Superresolution texture maps for multiview reconstruction," in Proceedings of *International Conference on Computer Vision (ICCV)*, pp. 1677–1684, 2009.

[4] U. Mudenagudi, A. Gupta, L. Goel, A. Kushal, P. Kalra, and S. Banerjee, "Super resolution of images of 3d scenecs," in Proceedings of *Asian Conference on Computer Vision (ACCV)*, pp. 85–95, 2007.

[5] K. Takahashi and T. Naemura, "Super-resolved free-viewpoint image synthesis based on view-dependent depth estimation," *IPSJ Transactions on Computer Vision and Applications*, vol. 4, pp. 134–148, 2012.

[6] S.C. Park, M.K. Park, and M.G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[7] W.T. Freeman, T.R. Jones, and E.C. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[8] J. Yang, J. Wright, T.S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[9] S. Wang, D. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in Proceedings of *Computer Vision and Pattern Recognition (CVPR)*, pp. 2216–2223, 2012.

[10] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[11] Stanford University Computer Graphics Laboratory, "The (new) stanford light field archive," `http://lightfield.stanford.edu/`.

[12] R. Raina, A. Madhavan, and A.Y. Ng, "Large-scale deep unsupervised learning using graphics processors," in Proceedings of *International Conference on Machine Learning (ICML)*, pp. 873–880, 2009.

[13] P. Sattigeri, J.J. Thiagarajan, K.N. Ramamurthy, and A. Spanias, "Implementation of a fast image coding and retrieval system using a GPU," in Proceedings of *Emerging Signal Processing Applications (ESPA)*, pp. 5–8, 2012.

APPENDIX A: RELIABILITY MAP

To evaluate reliability of the estimated depth at a pixel $\overrightarrow{p}$, the SR-FVS method [5] uses the matching cost at the estimated depth $S_{\min}(\overrightarrow{p})$. $S_{\min}(\overrightarrow{p})$ takes small values for pixels where the images are precisely aligned and the estimated depths are considered to be reliable. In contrast, $S_{\min}(\overrightarrow{p})$ takes large values around occlusion boundaries, for example, where the estimated depths are likely to be erroneous. On the basis of this idea, $\boldsymbol{W}$ is constructed as follows. Let $w(\overrightarrow{p})$ be a diagonal element of $\boldsymbol{W}$ which correspond to the pixel $\overrightarrow{p}$. $w(\overrightarrow{p})$ is defined as

$$w(\overrightarrow{p}) = \max(S_{\min}(\overrightarrow{p})^k, w_{\min}), \tag{6}$$

where k and $w_{\min}$ are chosen empirically. In our experiments, we set k to 6 for the CD cases and Bulldozer datasets, and 4 for the other datasets. $w_{\min}$ was set to 10 for all datasets.