IEICE TRANSACTIONS

on Information and Systems

VOL. E105-D NO. 1
JANUARY 2022

The usage of this PDF file must comply with the IEICE Provisions on Copyright.
The author(s) can distribute this PDF file for research and educational (nonprofit) purposes only.
Distribution by anyone other than the author(s) is prohibited.

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Image Adjustment for Multi-Exposure Images Based on Convolutional Neural Networks

Isana FUNAHASHI[†a)], ***Student Member***, **Taichi YOSHIDA**[†b)], ***Member***, **Xi ZHANG**[†c)], ***and*** **Masahiro IWAHASHI**[††d)], ***Senior Members***

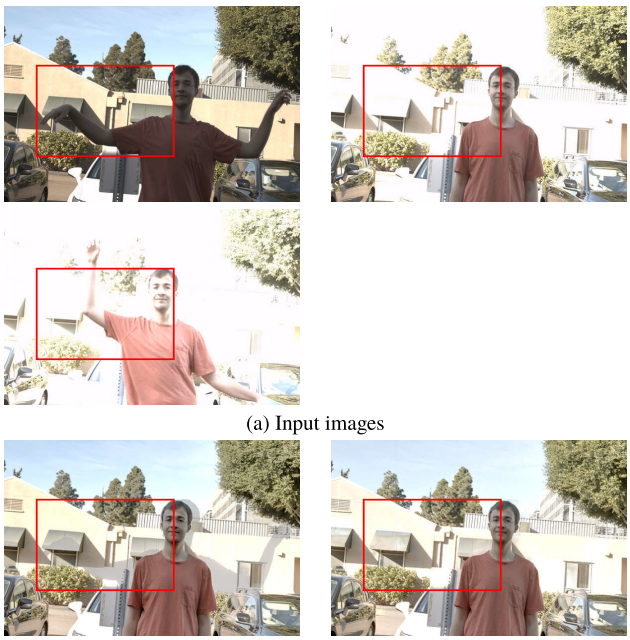
SUMMARY In this paper, we propose an image adjustment method for multi-exposure images based on convolutional neural networks (CNNs). We call image regions without information due to saturation and object moving in multi-exposure images lacking areas in this paper. Lacking areas cause the ghosting artifact in fused images from sets of multi-exposure images by conventional fusion methods, which tackle the artifact. To avoid this problem, the proposed method estimates the information of lacking areas via adaptive inpainting. The proposed CNN consists of three networks, warp and refinement, detection, and inpainting networks. The second and third networks detect lacking areas and estimate their pixel values, respectively. In the experiments, it is observed that a simple fusion method with the proposed method outperforms state-of-the-art fusion methods in the peak signal-to-noise ratio. Moreover, the proposed method is applied for various fusion methods as pre-processing, and results show obviously reducing artifacts.

key words: *high dynamic range imaging, multi-exposure image fusion, artifact removal, convolutional neural networks*

1. Introduction

A set of multi-exposure images consists of ones that capture the same scene with different exposure-settings of digital cameras, and is variously utilized in image applications. Since the images are taken by ordinary digital cameras with low dynamic range (LDR), they respectively have over- and under-exposure regions for high dynamic range (HDR) scene, where are called saturated regions in this paper. The images have the information in saturated regions of each other, and therefore the fusion of the sets produces a fine image without saturated ones. The sets are utilized in image applications such as HDR imaging, multi-exposure image fusion, and so on [1]–[15].

The ghosting artifact occurs in an image produced by fusing a set of multi-exposure images in which locations of objects and details are different. Since ordinary digital cameras cannot take multi-exposure images at exactly the same time, the difference generally occurs for natural scenes. When images with the difference are fused, the same object appears at different locations in resultant one, which is the ghosting artifact. Figure 1 shows an example of multi-exposure fusion, where (a)–(c) show an input set of multi-exposure images, an image fused by the conventional method [15], and by a simple fusion method [6] with the proposed method. From Fig. 1 (b), it is observed that the artifacts occur at the wall of the building, for example in the red frame.



(a) Input images

(b) [15] (c) Prop. + [6]

Fig. 1 Results of multi-exposure image fusion. (a) input images, (b) image fused by the conventional method [15], and (c) by the simple fusion method [6] with the proposed method, where the top-right of (a) is defined as standards for location of objects and deteils in fusion process.

For avoiding the ghosting artifact, various image fusion and correction methods have been proposed [10]–[12], [15]. The traditional methods use the patch matching [7], [8]. For a local region of an image, they effectively choose patches in other ones that have the information of its same location without saturation, and fuse them to produce a fine image. The recent methods often use convolutional neural networks (CNNs) [11], [14], [15]. They use CNN models for refining multi-exposure images, fusing them, and estimating weights for the fusion. CNN-based methods generally show superior

Manuscript received April 21, 2021.
Manuscript revised August 25, 2021.
Manuscript publicized October 21, 2021.

[†]The authors are with the Dept. of Computer and Network Engineering, Univ. of Electro-Commun., Chofu-shi, 182–8585 Japan.
[††]The author is with the Dept. of Electrical Electronics and Information Engineering, Nagaoka Univ. of Tech., Nagaoka-shi, 940–2188 Japan.
a) E-mail: i.funahashi@uec.ac.jp
b) E-mail: t-yoshida@uec.ac.jp
c) E-mail: zhangxi@uec.ac.jp
d) E-mail: iwahashi@vos.nagaokaut.ac.jp
DOI: 10.1587/transinf.2021EDP7087

Copyright © 2022 The Institute of Electronics, Information and Communication Engineers

fused images to the traditional ones.

Unfortunately, the methods still produce the artifact in areas where the set has no-information due to the saturation and the object moving, which are called lacking areas in this paper. For example, the lacking area is shown at the wall of the building in Fig. 1 (a) that are highlighted with the red frame. It is observed that these images have no-information in that area. Almost all methods use the matching of image features for avoiding the artifact. However, they induce failed matchings in lacking areas, because they cannot obtain image features of there.

In this paper, we propose a method that adjusts objects and details in multi-exposure images with adaptive inpainting based on CNN for reducing the bad influence of lacking areas. The proposed method adjusts a set of multi-exposure images and estimates the lacking areas via adaptive inpainting by three CNNs, warping and refining, detection, and inpainting networks. The first network provides pre-refinement for the set, the second one detects its lacking areas, and finally, the third one restores there. Fusing multi-exposure images adjusted by the proposed method derives a fine image without not only the ghosting artifact but also other artifacts, and thus it can be used as pre-processing for HDR imaging, multi-exposure image fusion, and so on. Figure 1 (c) shows a natural image that has no ghosting artifacts on the lacking area. In the experiments, the proposed method shows fused images compared with state-of-the-art ones [12], [15]. For general datasets of multi-exposure images, the simple fusion method [6] with the proposed one shows better scores in peak signal-to-noise ratio (PSNR). Moreover, through showing images fused by several methods [9], [12], [13] with and without the proposed one, it is recognized that the proposed one obviously reduces artifacts. The contribution of this paper is as follows: The definition of lacking areas is introduced, and it is shown that the specific estimation of their information is effective for fusion of multi-exposure images as pre-processing through experiments. Based on this knowledge shown in this paper, fusion methods are expected to reduce the artifact more efficiently.

The rest of the paper is organized as follows: Section 2 describes related work of image adjustment methods for multi-exposure images, fusion methods with artifact removal, and inpainting methods. Section 3 presents the fundamentals of inpainting and CNN. Section 4 describes details of the proposed method. Section 5 shows evaluation results with the proposed method and state-of-the-art methods for the ghosting artifact removal, and evaluation results with the proposed method as pre-processing. In Sect. 6, we conclude this paper.

2. Related Work

Image Adjustment for Multi-exposure Images

Methods of image adjustment for multi-exposure images take a reference image from a set of multi-exposure images, and then adjust objects and details of others to it [2]–[5]. The traditional methods use matching algorithms with features based on edges and key-points in images such as SIFT [2], [3]. Tomaszewska and Mantiuk have proposed a method that globally adjusts images based on SIFT features [3], and Hu et al. have also proposed one based on patch matching [16]. Unfortunately, there are no methods that tackle lacking areas.

Fusion of Multi-exposure Images with Artifact Removal

Fusion methods with artifact removal for multi-exposure images produce a fine image without artifacts from a set of multi-exposure ones [10], [11], [14], [15], [17]. Sen et al. have proposed a fusion method based on the optimization of a patch-wise objective function [8]. A method that uses dense SIFT features for measuring the spatial consistency between each exposure image has been proposed by Liu et al. [9]. Recently, several methods based on CNN have been proposed [11], [14], [15]. Kalantari and Ramamoorthi have proposed a method that uses alpha blending and estimates its weights by a CNN model [11]. A method proposed by Prabhakar et al. has a two-step network consisting of a refinement step of input images and a fusion step of resultant ones [15]. Unfortunately, they induce miss-adjustments in their adjustment and fusion processes due to lacking areas, and therefore the artifacts still occur in their fused images.

Image Inpainting

Methods of image inpainting restore missing regions in images [18]–[26]. The traditional methods use the diffusion of pixel values and the matching of image patches [20]–[22]. Criminisi et al. have proposed a method that copies patches based on residual structures and textures [20]. State-of-the-art methods of image inpainting are usually based on CNNs with supervised learning algorithms and generative adversarial networks (GANs) [18], [23]–[26]. Pathak et al. firstly used GANs to restore large missing regions [23]. Iizuka et al. have proposed a method that has two discriminator networks calculating the global and local consistency of output images for training [18]. The method has been developed by Yu et al. based on the contextual attention layer, and it can restore fine details [25]. Therefore, the proposed inpainting network is realized based on it and uses the contextual attention layer.

3. Fundamentals for Propposd Method

3.1 Dilated Convolution Layer

The dilated convolution layer is one of the famous convolutional one for increasing receptive fields of CNNs [27]. Dilated convolution layers apply filters that have fixed spaces between each filter elements for input signals. The operation of dilated convolution layers is written as

$$\boldsymbol{y}_{i,j} = \sum_{u=-k'_h}^{k'_h} \sum_{v=-k'_w}^{k'_w} \boldsymbol{W}_{k'_h+u,k'_w+v} \boldsymbol{x}_{i+lu,j+lv}, \tag{1}$$

where y, x, W, and l denote output and input signals, values

of filter elements, and a factor of fixed spaces,

$$k'_h = \frac{k_h - 1}{2}, \tag{2}$$

$$k'_w = \frac{k_w - 1}{2}, \tag{3}$$

and k_h and k_w denote the size of filter kernels along two directions, respectively [18]. Note that l is called a dilation factor in this paper. Dilated convolution layers with $l = 1$ are normal convolution ones. Dilated convolutional layers with an exponential increase of dilation factors exponentially grow the size of their receptive fields. Therefore, CNNs with the layers can refer to image features in a large region.

3.2 Contextual Attention Layer

The contextual attention layer reconstructs detailed features of the output image from feature maps of input images [25]. The layer uses two feature maps of images, where one image is called a target and another is a source. Firstly, the layer extracts 3×3 patches from a feature map of the source image. The layer calculates the normalized inner product of the target patch $t_{i',j'}$ and the source patch $s_{i,j}$ with the softmax operation as

$$p_{i,j,i',j'} = \mathrm{softmax}\left(\left\langle \frac{\boldsymbol{s}_{i,j}}{\|\boldsymbol{s}_{i,j}\|}, \frac{\boldsymbol{t}_{i',j'}}{\|\boldsymbol{t}_{i',j'}\|} \right\rangle\right), \tag{4}$$

where $p_{i,j,i',j'}$ denotes the attention score of each pixel. The operations are easily implemented by convolutional layers and channel-wise softmax operations. Finally, the layer applies the transposed convolution that uses the source patches as filters to the feature map of the target image. Thanks to the above operations, features of the source image that is most similar to the target patch of the target image are chosen, and are copied to the target image.

3.3 Residual Block

The residual block is one of CNN architectures that enable CNN methods to have a great depth. Residual blocks are constructed with two convolutional layers, activation functions, batch normalizations [28], and a skip connection [29]. It has been reported in [30] that the batch normalization reduces the flexibility of CNNs by its normalizing process. To avoid this problem, a residual block without batch normalizations has been proposed [31]. Therefore, residual blocks in the proposed CNN architectures also avoid batch normalizations.

4. Proposed Method

4.1 Overview

In this paper, we propose a CNN-based adjustment method for a set of multi-exposure images. The proposed method adjusts locations of objects and details in a source image denoted as I_{src} to ones in a reference image denoted as I_{ref}. For more than two images, I_{ref} is firstly defined, then the proposed method is applied for each source image, one by one. For example, for three images, I_1, I_2, and I_3, if I_1 is defined as I_{ref}, the proposed method is applied for I_2, and then for I_3. In Fig. 2, a fusion process of the proposed method is shown with multi-exposure images, in which the middle exposure image is determined as I_{ref}. Thanks to the proposed method as pre-processing of multi-exposure image fusion, fused images avoid artifacts.

The proposed method is shown in Fig. 3 and consists of three steps as follows; (1) warping and refining I_{src} based on the conventional method [15], (2) detecting areas for inpainting, and (3) inpainting the detected areas. In the proposed method, we warp I_{src} with its calculated optical flows and apply refinement for I_{src} based on CNN [15]. Then lacking areas are detected via the detection network, and the detected results are used in the inpainting network as a mask of inpainting.

4.2 Detection of Lacking Areas with CNN

Inputs of the proposed detection network are I_{src}, I_{ref}, and estimated results of occlusion regions O_{src}. They are concatenated along the color channel. O_{src} indicates locations

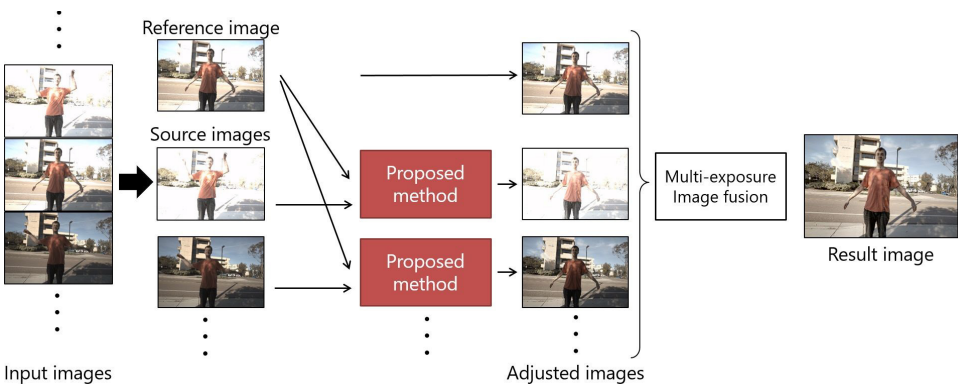


Fig. 2 Fusion process with the proposed method for more than two of images. (Reference image: middle exposure one)

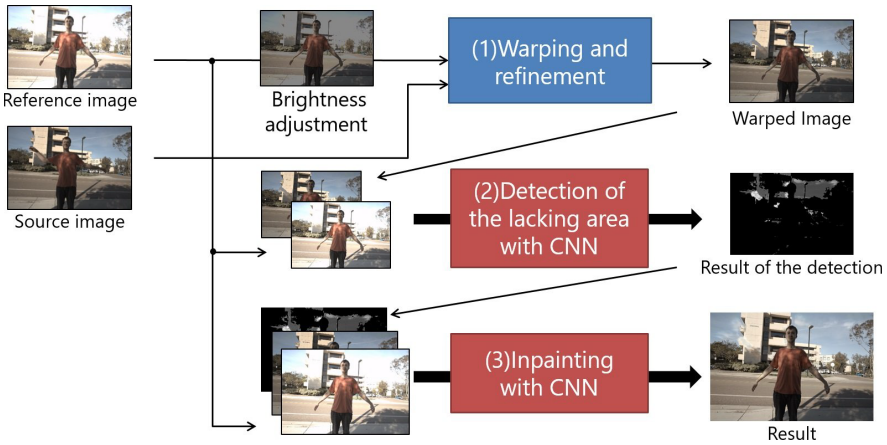


Fig. 3 Overview of the proposed method.

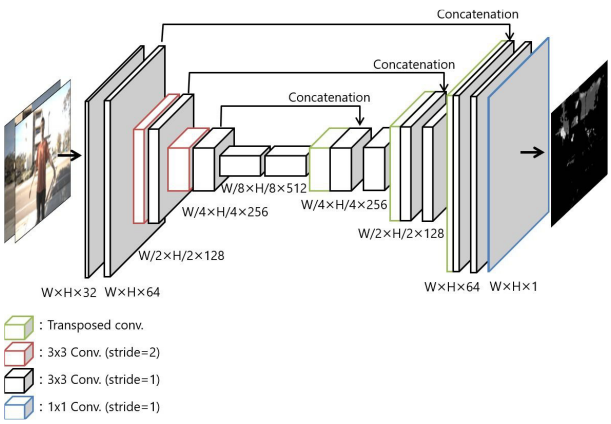


Fig. 4 Architecture of the proposed detection network.

of occlusion regions and is calculated from optical flows of input images [32]. O_{src} is a binary map whose size is equal to input images. If a pixel of I_{src} are in occlusion regions, the corresponding element of O_{src} are 1, and otherwise, 0. The d-th element of O_{src} is 1 when the following condition is satisfied:

$$\begin{aligned}&\left|f_{\mathrm{fw}}(\mathrm{d}) + f_{\mathrm{bw}}(\mathrm{d} + f_{\mathrm{fw}}(\mathrm{d}))\right|^2 \\ &< 0.01 \times \left(\left|f_{\mathrm{fw}}(\mathrm{d})\right|^2 + \left|f_{\mathrm{bw}}(\mathrm{d} + f_{\mathrm{fw}}(\mathrm{d}))\right|^2\right) + 0.5,\end{aligned} \tag{5}$$

where d is an index of pixels, and f_{fw} and f_{bw} denote optical flows of I_{src} estimated by I_{ref} and I_{ref} by I_{src}, respectively. These optical flows are estimated in the warping process (1) shown in Fig. 3. The proposed detection network is represented as follows:

$$W = F_{\mathrm{dtn}}(I_{\mathrm{ref}}, I'_{\mathrm{src}}, O_{\mathrm{src}}), \tag{6}$$

where F_{dtn} and W denote the proposed detection network and the map for I'_{src}, respectively.

Figure 4 shows an architecture of the proposed detection network, which is based on the U-Net [33], and its parameters are shown in Table 1, where "Output ch." and "conv." means the number of output channels and the convolution, respectively. The network has three downsampling and concatenation processes. Different from the U-Net, we use the convolution with stride 2 for the downsampling. Thanks to the process, the proposed detection network uses multi-resolution information of images. The ReLU [34] is used as activation functions in the network except the last layer , and the sigmoid function is applied after the last layer to normalize into [0, 1].

Table 1 Parameters of the proposed detection network.

Layer type	Filter size	Stride	Padding	Output ch.
Conv.	5×5	1	2	32
Conv.	3×3	1	1	64
Conv.	3×3	2	1	128
Conv.	3×3	1	1	128
Conv.	3×3	2	1	256
Conv.	3×3	1	1	256
Conv.	3×3	2	1	512
Conv.	3×3	1	1	512
Transposed conv.	4×4	2	1	256
Concatenation	-	-	-	512
Conv.	3×3	1	1	256
Transposed conv.	4×4	2	1	128
Concatenation	-	-	-	256
Conv.	3×3	1	1	128
Transposed conv.	4×4	2	1	64
Concatenation	-	-	-	128
Conv.	3×3	1	1	64
Conv.	1×1	1	0	1

4.3 Inpainting for Lacking Areas

The proposed inpainting network F_{inp} restores lacking areas as follows:

$$\hat{I}_{\mathrm{src}} = F_{\mathrm{inp}}(I_{\mathrm{ref}}, I'_{\mathrm{src}}, W), \tag{7}$$

Table 2 Parameters of the proposed inpainting network.

	Layer type		Filter size	Stride	Padding	Dilation	Output ch.
.5	Conv.		5×5	1	2	-	32
(a)	Conv.		3×3	2	1	-	64
	Conv.		3×3	1	1	-	64
	Conv.		3×3	2	1	-	128
.175	Contextual attention		-	-	-	-	128
(b)	Residual block	Conv.	3×3	1	1	-	128
		Conv.	3×3	1	1	-	128
.4	Conv.		5×5	1	2	-	32
	Conv.		3×3	2	1	-	64
	Conv.		3×3	1	1	-	64
(c)	Conv.		3×3	2	1	-	128
	Residual block	Dilated conv.	3×3	1	2	2	128
		Dilated conv.	3×3	1	4	4	128
	Residual block	Dilated conv.	3×3	1	8	8	128
		Dilated conv.	3×3	1	16	16	128
.4	Concatination		-	-	-	-	256
	Residual block	Dilated conv.	3×3	1	2	2	256
		Dilated conv.	3×3	1	4	4	256
(d)	Transposed conv.		4×4	2	1	-	128
	Conv.		3×3	1	1	-	128
	Transposed conv.		4×4	2	1	-	64
	Conv.		3×3	1	1	-	32
	Conv.		1×1	1	1	-	3

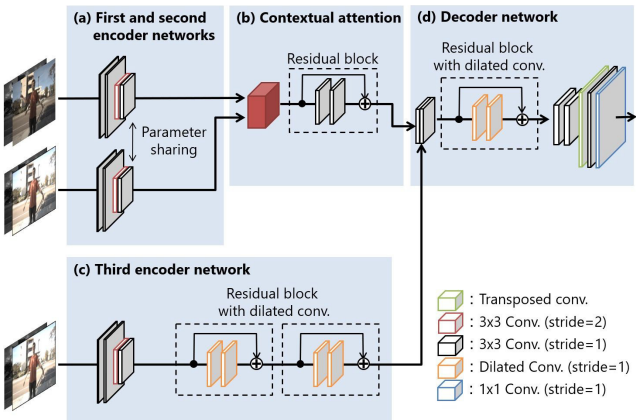


Fig. 5 Architecture of the proposed inpainting network.

where $\hat{I}_{\text{src}}$ and I'_{src} are restored and warped source images, respectively. For inputs of the network, W is concatenated to I_{ref} and I'_{src} along the color channel. Finally, the adjusted image $\hat{Y}$ is calculated by

$$\hat{Y} = I'_{\text{src}} \odot (1 - W) + \hat{I}_{\text{src}} \odot W, \tag{8}$$

where $\odot$ denotes the pixel-wise multiplication.

The proposed inpainting network is constructed with contextual attention layers [25] and dilated convolution layers [27] mentioned in Sect. 3, which is shown in Fig. 5, and Table 2 shows its parameters. First, the network extracts feature maps by three parallel encoder networks. The first and second encoder networks extract image features from I_{ref} and I'_{src} for the contextual attention layer. Since the contextual attention layer uses two features for calculating their similarity, the encoder networks have same parameters of convolutional layers. The third encoder network extracts global features of images using the residual block with the dilated convolution. The resultant features are concatenated and merged by the decoder network to produce the inpainting result. The leaky ReLU [35] is used as activation functions in the network expect the last layer and its parameter is set 0.2. A sigmoid function is also applied after the last layer to normalize into $[0, 1]$.

4.4 Training for Proposed Networks

The proposed networks are trained through two steps that are the pre-training of the proposed inpainting network and the main training of the detection and inpainting network. Unfortunately, there are no datasets for the pre-training of the detection network, and therefore the two-step training is used. Note that pixel values of datasets are linearly normalized into $[0, 1]$ in this training.

First, the proposed inpainting network is trained on a large dataset of image inpainting. The input image I, which randomly has lost regions, is produced by

$$I = x \odot m, \tag{9}$$

where x is the raw image of I without the lost regions and m is a binary mask that indicates their locations, respectively.

x is also used as a ground truth image for the pre-training. The resultant image $\hat{x}$ is produced by

$$\hat{x} = I \odot (1 - m) + F_{\text{inp}}(I, I, m) \odot m, \tag{10}$$

where I is used twice as input instead of I_{ref} and I'_{src} in the pre-training. For the loss function of the pre-training, the pixel-wise mean absolute error (MAE) is used as

$$L_{\text{pre}}(\hat{x}, x) = \frac{1}{3wh} \sum_{i=1}^{w} \sum_{j=1}^{h} \|\hat{\boldsymbol{x}}_{i,j} - \boldsymbol{x}_{i,j}\|_1, \tag{11}$$

where w and h denote the width and height of images, and $x_{i,j}$ and $\hat{x}_{i,j}$ are vectors of RGB values at (i, j) pixel of x and $\hat{x}$ respectively. The ADADELTA algorithm is used in the optimization [36].

Next, the proposed networks are trained on a dataset of multi-exposure images. The source image is randomly chosen from several exposure images with under- and over-exposures, and hence the proposed model covers images of several exposures for image adjustment. (10) shows that the proposed method only uses differentiable operations for combining results of the proposed networks. Therefore, the computational graph of the detection and inpainting networks is connected, and they are trained at a time. For the loss function, MAE is also used as

$$L_{\text{main}}(\hat{Y}, Y) = \frac{1}{3wh} \sum_{i=1}^{w} \sum_{j=1}^{h} \|\hat{\boldsymbol{Y}}_{i,j} - \boldsymbol{Y}_{i,j}\|_1, \tag{12}$$

where $\hat{Y}_{i,j}$ and $Y_{i,j}$ denote vectors of RGB values at (i, j) pixel of the output image $\hat{Y}$ and the ground truth image Y, respectively. The ADADELTA algorithm is also used in the optimization [36].

5. Experiment

5.1 Training Details

Dataset

We used the Places2 dataset [37] for the pre-training shown in Sect. 4. We randomly chose 10000 images from the dataset, and separated them into training and validation sets, which have 9000 and 1000 images, respectively. We used images of resolution 256×256 with a 128×128 hole and their masks as inputs, and used raw images without the hole as ground truth.

We used the Kalantari dataset [11] for the main training shown in Sect. 4. The dataset contains 74 sets of three multi-exposure images and their HDR images. We determined middle exposure images in the sets as the reference one. As ground truth, we produced sets of multi-exposure images from the HDR images via several camera response functions and appropriate exposure values [38]. Training images were resized into 384×256 and clipped to 256×256 patches with a 64-pixel step, and we applied the rotation and the horizontal flipping for the patches. Finally, the training dataset has 4440 sets of I_{src}, I_{ref}, and Y.

Implementation

We implemented the proposed network on Chainer v7, CUDNN v7, CUDA v10.0, and trained it on Intel(R) Xeon(R) E5-2620 v4 and GeForce(R) GTX1080ti. We used a batch-size of 12 and 225000 iterations for the pre-training, and a batch-size of 10 and 177600 ones for the main training. The method proposed by Mertens et al., which is simple and ignores reducing artifacts, was used for the fusion [6].

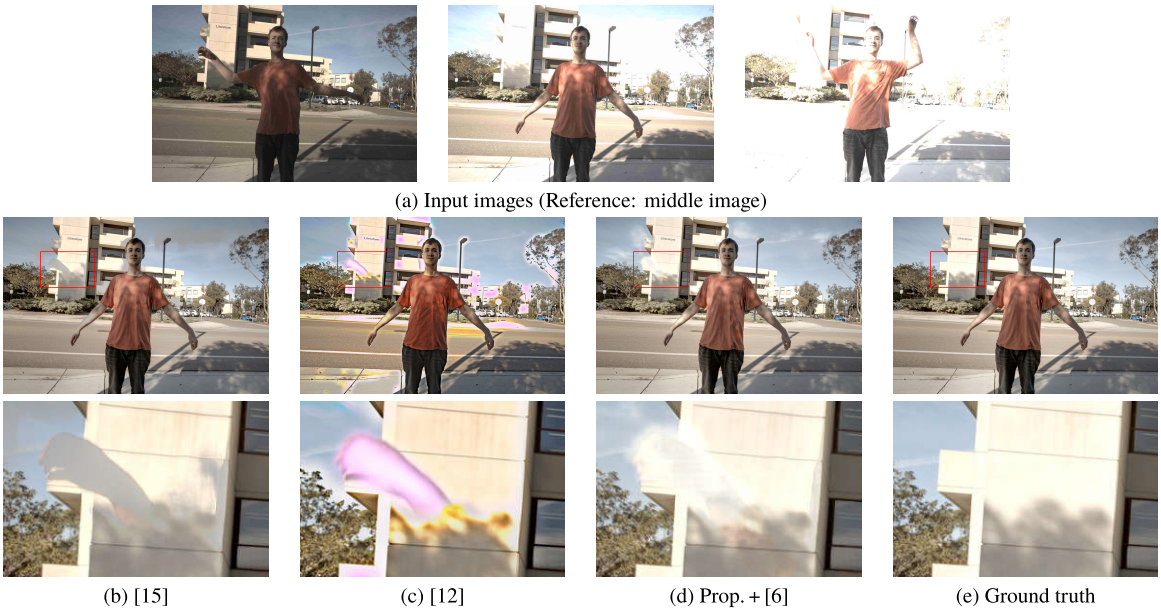
5.2 Quantitative Evaluation

Quantitative evaluation for the proposed method is shown here via the test set of the Kalantari dataset. In this section, the proposed method is compared with state-of-the-art ones of multi-exposure image fusion, proposed by Prabhakar et al. [15] and Ma et al. [12]. The dataset has natural sets of multi-exposure images with objects moving and their HDR images. Therefore, for a fair comparison, LDR images are produced as ground-truth by applying the global tone-mapping for the HDR images, where exposure settings are fit to ones of reference images and the Agfacolor Future 100DC, which is modeled in [38], is used as the camera response curve. Table 3 shows their PSNR scores, and Fig. 6 (a), (b)–(d), and (e) show the input of multi-exposure images, resultant images of each method, and the fused ground truth for Image 3 in Table 3, respectively. Images in the bottom row of Fig. 6 show enlarged images of each resultant image. From Table 3, the proposed method achieves the highest score in the average. Unfortunately, Prop. + [6] for Image 1 shows a lower PSNR score compared with the conventional one, but it is observed from Fig. 6 that the proposed method clearly reduces artifacts and hence produces perceptually outperformed images. Thanks to the adaptive inpainting, the proposed method reduces the artifacts and produces better results.

Unfortunately, the proposed method sometimes produces inaccurate restorations for large lacking areas, shown in enlarged images of Fig. 6 (d) and (e), and we will tackle this problem in future work since it is caused by the inpainting network. Generally, the restoration of large areas is the challenging task in inpainting [18]. Moreover, the training dataset [11] includes a few sets of images having large lacking areas, because it is constructed without considering lacking areas. In future work, we will try to improve the architecture based on state-of-the-art methods of inpainting that

Table 3 Results of quantitative evaluation on the Kalantari dataset.

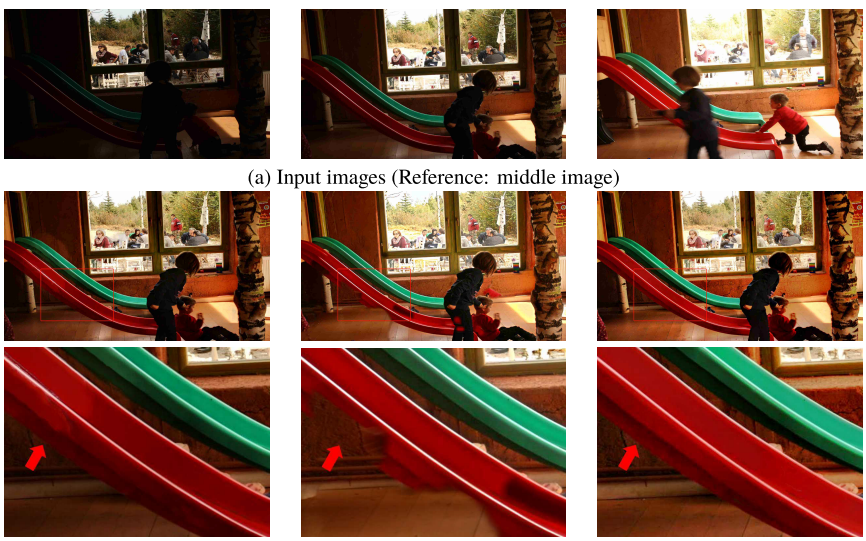
	PSNR [dB]		
	[15]	[12]	Prop. + [6]
Image 1	**27.29**	21.40	26.88
Image 2	24.18	20.53	**27.86**
Image 3	24.32	22.07	**27.56**
Image 4	29.21	25.96	**31.14**
Image 5	30.06	26.03	**32.07**
Average on the Kalantari dataset (15 sets)	26.92	22.87	**27.34**



(a) Input images (Reference: middle image)

(b) [15] (c) [12] (d) Prop. + [6] (e) Ground truth

Fig. 6 Results of multi-exposure image fusion for Image 3. (a) Input images, (b)–(d) fused results by [12], [15], and Prop.+ [6], and (e) ground truth, respectively, where in (b)–(e), the bottom row shows enlarged images at red frames in the top row.



(a) Input images (Reference: middle image)

(b) [15] (c) [12] (d) Prop. + [6]

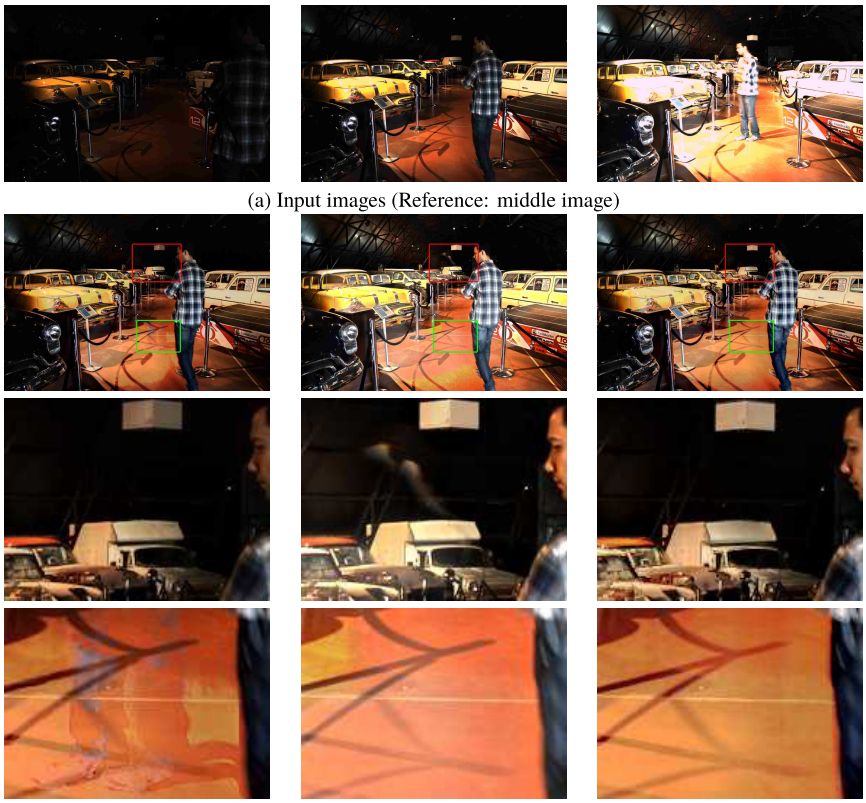
Fig. 7 Results of multi-exposure image fusion for a set of the Karaduzovic-Hadziabdic dataset. (a) Input images, and (b)–(d) fused results by [12], [15], and Prop.+ [6], respectively, where in (b)–(d), the bottom row shows enlarged images at red frames in the top row, and particular artifacts are highlighted by red arrows.

are effective for various sizes of restored areas and gather enough training sets that have lacking areas with various sizes.

5.3 Evaluation in Natural Scenes

We show resultant images of the methods for other datasets that have multi-exposure images without ground-truth fused images. The Karaduzovic-Hadziabdic dataset [39] and the Tursun dataset [40], which are taken in natural scenes and widely used for the evaluation of artifact removal, are used. Unfortunately, since they only have input images, this experiment skips quantitative evaluation.

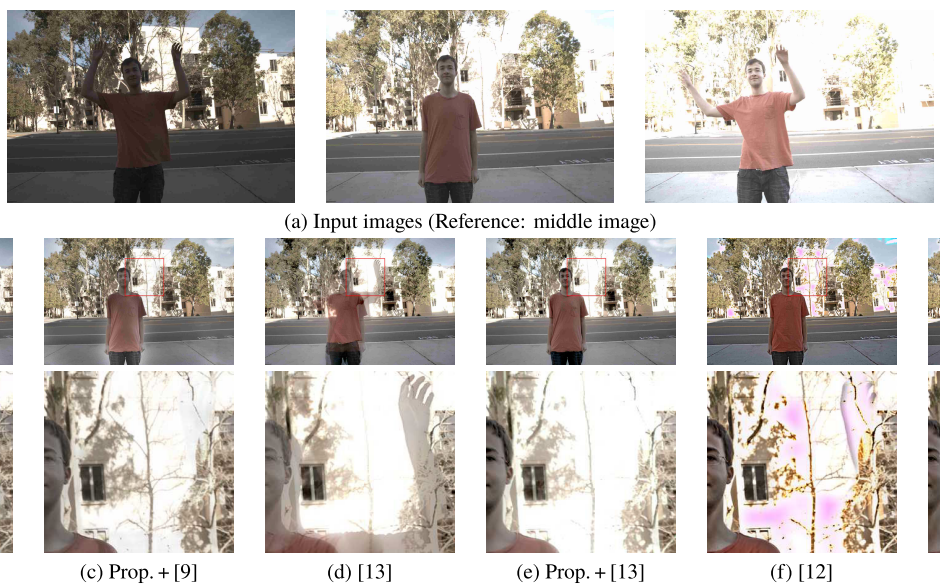
Figure 7 and Fig. 8 show results for the Karaduzovic-Hadziabdic and the Tursun dataset, respectively, where (a) and (b)–(d) show input images and ones fused by each method. In Fig. 7, the bottom row shows enlarged images of each resultant image. Since the Karaduzovic-Hadziabdic dataset has very complex motions, unfortunately, Fig. 7 (b) shows the visual artifact that looks like a shadow of the slide



(a) Input images (Reference: middle image)

(b) [15] (c) [12] (d) Prop. + [6]

Fig. 8 Results of multi-exposure image fusion for a set of the Tursun dataset. (a) Input images, and (b)–(d) fused results by [12], [15], and Prop.+ [6], respectively, where in (b)–(d), the middle and bottom row shows enlarged images at red and green frames in the top row, respectively.

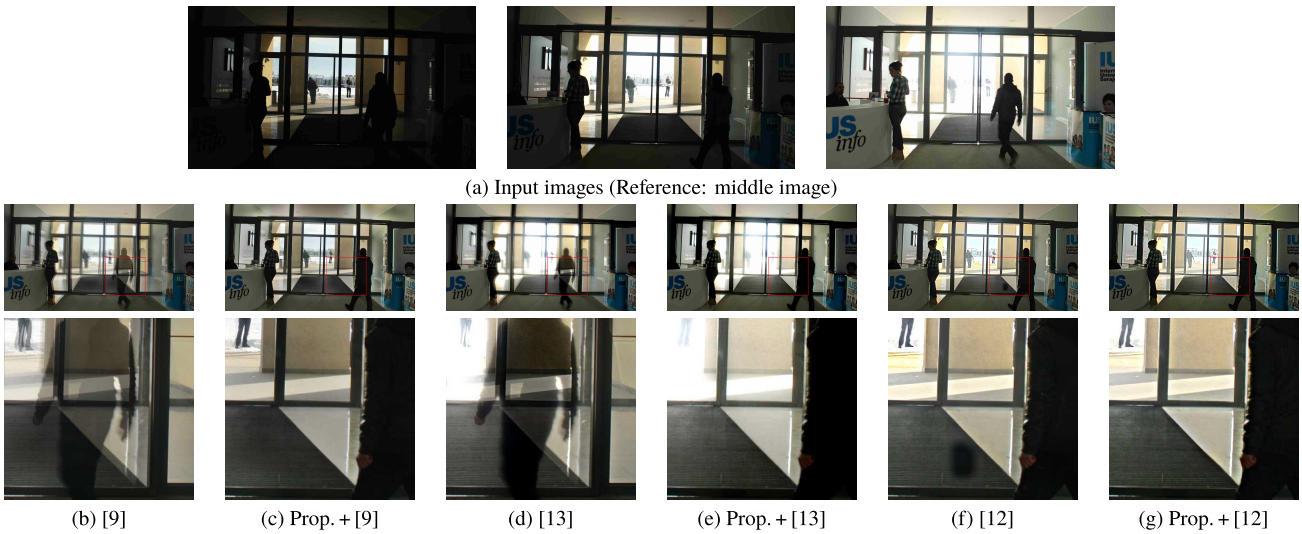


(a) Input images (Reference: middle image)

(b) [9] (c) Prop. + [9] (d) [13] (e) Prop. + [13] (f) [12] (g) Prop. + [12]

Fig. 9 Results of multi-exposure image fusion for a set of the Kalantari dataset. (a) Input images, and (b)–(g) fused results by [9], Prop. + [9], [13], Prop. + [12], [13], and Prop. + [12], respectively, where in (b)–(g), the bottom row shows enlarged images at red frames in the top row.

and Fig. 7 (c) also shows several artifacts on the slide. However, Fig. 7 (d) shows no visual artifacts on that region. In Fig. 8, the middle and lower rows are enlarged resultant images. Several artifacts are observed in Fig. 8 (b) and (c), but (d) shows reducing them. It is observed from Figs. 7 and 8 that the proposed method contributes to visually reduce



(a) Input images (Reference: middle image)

(b) [9] (c) Prop. + [9] (d) [13] (e) Prop. + [13] (f) [12] (g) Prop. + [12]

Fig. 10 Results of multi-exposure image fusion for a set of the Karaduzovic-Hadziabdic dataset. (a) Input images, and (b)–(g) fused results by [9], Prop. + [9], [13], Prop. + [12], [13], and Prop. + [12], respectively, where in (b)–(g), the bottom row shows enlarged images at red frames in the top row.

artifacts, compare with state-of-the-art methods.

5.4 Evaluation as Pre-Processing for Fusion Methods

This evaluation shows the efficacy of the proposed method as pre-processing for conventional fusion methods. Conventional methods used in this section are proposed by Liu and Wang [9], and Li et al. [13], and Ma et al. [12]. They are typical fusion methods, and some of them tackle the ghosting artifact. The Kalantari dataset and the Karaduzovic-Hadziabdic dataset [11], [39] are used.

Figure 9 and Fig. 10 show results for the Kalantari dataset and the Karaduzovic-Hadziabdic dataset, respectively, where (a) shows input images, (b)–(g) shows resultant fused ones by each method without and with the proposed method, and the lower row shows enlarged resultant ones. Figure 9 (b), (d), and (f) show ghosting artifacts that look a translucent hand on the building, and they are visually reduced in Fig. 9 (b), (d), and (f) respectively. Figure 10 (b) and (d) show ghosting artifacts that look a translucent person, and (f) shows artifacts on the floor. It is observed from Figs. 9 and 10 that the proposed method has the efficacy to reduce artifacts as pre-processing for several fusion methods.

6. Conclusion

In this paper, we introduced the lacking areas and proposed a method that adjusts objects and details in multi-exposure images with adaptive inpainting based on CNN for tackling them. The proposed method consists of three CNNs, warping and refining, detection, and inpainting networks. The second and third networks detect and restore the lacking areas, respectively, and the proposed method provides a set of multi-exposure images without object moving and missing information. It is shown through experiment that a simple fusion method with the proposed method objectively outperforms state-of-the-art methods, which tackle the ghosting artifact, and the proposed one is effective as pre-processing in fusion of multi-exposure images. Thanks to the detection and inpainting network, the proposed method estimates information of lacking areas, and reduces artifacts on the area.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 18K11260.

References

[1] P.E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," Proc. the 24th Annual Conf. Comput. Graphics and Interactive Techniques, pp.369–378, 1997.
[2] G. Ward, "Fast, Robust Image Registration for Compositing High Dynamic Range Photographs from Hand-Held Exposures," Journal of Graphics Tools, vol.8, no.2, pp.17–30, 2012.
[3] A. Tomaszewska and R. Mantiuk, "Image Registration for Multi-exposure High Dynamic Range Image Acquisition," The 15-th Intl. Conf. in Central Europe on Comput. Graphics, Visualization and Comput. Vis., pp.49–56, 2007.
[4] T. Kartalov, Z. Ivanovski, and L. Panovski, "A real time global motion compensation for multi-exposure imaging algorithms," Proc. IEEE EUROCON - Intl. Conf. Comput. as a Tool, pp.1–4, 2011.
[5] K.R. Prabhakar and R.V. Babu, "Ghosting-free multi-exposure image fusion in gradient domain," Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Process., pp.1766–1770, 2016.
[6] T. Mertens, J. Kautz, and F.V. Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," Computer Graphics Forum, vol.28, no.1, pp.161–171, 2009.
[7] O. Gallo, N. Gelfandz, W.-C. Chen, M. Tico, and K. Pulli, "Artifact-free High Dynamic Range imaging," Proc. IEEE Intl. Conf. Computational Photography, pp.1–7, 2009.
[8] P. Sen, N.K. Kalantari, M. Yaesoubi, S. Darabi, D.B. Goldman, and E. Shechtman, "Robust patch-based hdr reconstruction of dynamic

scenes," ACM Trans. Graphics, vol.31, no.6, pp.1–11, 2012.
[9] Y. Liu and Z. Wang, "Dense SIFT for ghost-free multi-exposure fusion," J. Vis. Commun. Image Represent., vol.31, pp.208–224, 2015.
[10] T.-H. Oh, J.-Y. Lee, Y.-W. Tai, and I. Kweon, "Robust high dynamic range imaging by rank minimization," IEEE Trans. Patt. Anal. Machine Intell., vol.37, no.6, pp.1219–1232, 2015.
[11] N.K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," ACM Trans. Graphics, vol.36, no.4, 2017.
[12] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust Multi-Exposure Image Fusion: A Structural Patch Decomposition Approach," IEEE Trans. Image Process., vol.26, no.5, pp.2519–2532, 2017.
[13] Z. Li, Z. Wei, C. Wen, and J. Zheng, "Detail-Enhanced Multi-Scale Exposure Fusion," IEEE Trans. Image Process., vol.26, no.3, pp.1243–1252, 2017.
[14] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," Proc. the European Conf. Comput. Vis., pp.120–135, 2018.
[15] K.R. Prabhakar, R. Arora, A. Swaminathan, K.P. Singh, and R.V. Babu, "A Fast, Scalable, and Reliable Deghosting Method for Extreme Exposure Fusion," Proc. IEEE Intl. Conf. Computational Photography, pp.1–8, 2019.
[16] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR Deghosting: How to Deal with Saturation?," Proc. IEEE Conf. Comput. Vis. Patt. Recognit., pp.1163–1170, 2013.
[17] C. Lee, Y. Li, and V. Monga, "Ghost-Free High Dynamic Range Imaging via Rank Minimization," IEEE Signal Process. Letters, vol.21, no.9, pp.1045–1049, 2014.
[18] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," ACM Trans. Graphics, vol.36, no.4, pp.107–121, 2017.
[19] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," Proc. the 27th Annual Conf. Comput. Graphics and Interactive Techniques, pp.417–424, 2000.
[20] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," IEEE Trans. Image Process., vol.13, no.9, pp.1200–1212, 2004.
[21] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," IEEE Trans. Image Process., vol.12, no.8, pp.882–889, 2003.
[22] C. Barnes, E. Shechtman, A. Finkelstein, and D.B. Goldman, "PatchMatch: a randomized correspondence algorithm for structural image editing," ACM Trans. Graphics, vol.28, no.3, pp.1–11, 2009.
[23] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A.A. Efros, "Context Encoders: Feature Learning by Inpainting," Proc. IEEE Conf. Comput. Vis. Patt. Recognit., pp.2536–2544, 2016.
[24] R.A. Yeh, C. Chen, T.Y. Lim, A.G. Schwing, M. Hasegawa-Johnson, and M.N. Do, "Semantic Image Inpainting with Deep Generative Models," Proc. IEEE Conf. Comput. Vis. Patt. Recognit., pp.6882–6890, 2017.
[25] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T.S. Huang, "Generative Image Inpainting with Contextual Attention," Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recognit., pp.5505–5514, 2018.
[26] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-Form Image Inpainting With Gated Convolution," Proc. IEEE/CVF Intl. Conf. Comput. Vis., pp.4470–4479, 2019.
[27] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," Proc. Intl. Conf. Learning Represent., 2016.
[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Proc. Intl. Conf. Machine Learning, pp.448–456, 2015.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2016.
[30] S. Nah, T.H. Kim, and K.M. Lee, "Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring," Proc. IEEE Conf. Comput. Vis. Patt. Recognit., pp.257–265, 2017.
[31] B. Lim, S. Son, H. Kim, S. Nah, and K.M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," Proc. IEEE Conf. Comput. Vis. Patt. Recognit. Workshops, pp.1132–1140, 2017.
[32] S. Meister, J. Hur, and S. Roth, "UnFlow: Unsupervised learning of optical flow with a bidirectional census loss," Proc. AAAI Conf. Artificial Intell., 2018.
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Proc. Medical Image Computing and Computer-Assisted Intervention, pp.234–241, 2015.
[34] V. Nair and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines," Proc. Intl. Conf. Machine Learning, pp.807–814, 2010.
[35] A.L. Maas, A.Y. Hannun, and A.Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," Proc. Intl. Conf. Machine Learning, 2013.
[36] M.D. Zeiler, "ADADELTA: an adaptive learning rate method," arXiv:1212.5701, 2012.
[37] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," IEEE Trans. Patt. Analysis and Machine Intell., vol.40, no.6, pp.1452–1464, 2018.
[38] M.D. Grossberg and S.K. Nayar, "Modeling the space of camera response functions," IEEE Trans. Patt. Analysis and Machine Intell., vol.26, no.10, pp.1272–1282, 2004.
[39] K. Karaduzovic-Hadziabdic, J.H. Telalovic, and R. Mantiuk, "Expert evaluation of deghosting algorithms for multi-exposure high dynamic range imaging," HDRi2014-Second Intl. Conf. and SME Workshop HDR Imag., 2014.
[40] O.T. Tursun, A.O. Akyüz, A. Erdem, and E. Erdem, "An objective deghosting quality metric for hdr images," Proc. the 37th Annual Conf. of the European Association for Comput. Graphics, vol.35, no.2, pp.139–152, 2016.

Isana Funahashi received the B.Eng. and M.Eng. degrees from the Nagaoka University of Technology, Nagaoka, Japan, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo, Japan. His research interests include image processing and computer vision.

Taichi Yoshida received the B.Eng., M.Eng., and Ph.D. degrees in engineering from Keio University, Yokohama, Japan, in 2006, 2008, and 2013, respectively. In 2014, he joined the Nagaoka University of Technology. In 2018, he joined the University of Electro-Communications, where he is currently an Assistant Professor with the Department of Computer and Network Engineering. His research interests include filter bank design and image coding applications.

Xi Zhang received the B.E. degree in communication engineering from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 1984, and M.E. and Ph.D. degrees in electronic and information engineering from University of Electro-Communications (UEC), Tokyo, Japan, in 1990 and 1993, respectively. He was with Department of Electronic Engineering at NUAA from 1984 to 1987, and with Department of Electronic and Information Engineering at UEC from 1993 to 1996, all as an Assistant Professor. He was with Department of Electrical Engineering at Nagaoka University of Technology (NUT), Niigata, Japan, as an Associate Professor, from 1996 to 2004. Currently, he is with Department of Computer and Network Engineering at UEC, as an Professor. He was a Visiting Scientist of the MEXT of Japan with Massachusetts Institute of Technology (MIT), Cambridge, from 2000 to 2001. His research interests are in the areas of digital signal processing, graph signal processing, filter design theory, filter banks and wavelets, and its applications to image and video processing. Dr. Zhang is a senior member of the IEEE. He received the third prize of the Science and Technology Progress Award of China in 1987, and the challenge prize of Fourth LSI IP Design Award of Japan in 2002. He served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS from 2002 to 2004.

Masahiro Iwahashi received the B. Eng., M. Eng., and D. Eng. degrees in Electrical Engineering from Tokyo Metropolitan University in 1988, 1990, and 1996, respectively. In 1990, he joined Nippon Steel Co., Ltd. From 1991 to 1992, he was seconded to Graphics Communication Technology Co., Ltd. In 1993, he joined Nagaoka University of Technology, where he is currently a Professor in the Department of Electrical Engineering, Faculty of Technology. From 1995 to 2001, he was also a lecturer at Nagaoka Technical College. From 1998 to 2001, he relocated to Thammasat University, Thailand, and to the Electronic Engineering Polytechnic Institute of Surabaya, Indonesia, as a JICA expert. His research interests are digital signal processing, multi-rate systems, and image compression.