Convolutional Neural Network-Based Infrared Image Super Resolution Under Low Light Environment

Tae Young Han, Yong Jun Kim, Byung Cheol Song
Department of Electronic Engineering
Inha University
Incheon, Republic of Korea
ty_han@inha.edu, yongjk@inha.edu, bcsong@inha.ac.kr

***Abstract*—Convolutional neural networks (CNN) have been successfully applied to visible image super-resolution (SR) methods. In this paper, for up-scaling near-infrared (NIR) image under low light environment, we propose a CNN-based SR algorithm using corresponding visible image. Our algorithm firstly extracts high-frequency (HF) components from low-resolution (LR) NIR image and its corresponding high-resolution (HR) visible image, and then takes them as the multiple inputs of the CNN. Next, the CNN outputs HR HF component of the input NIR image. Finally, HR NIR image is synthesized by adding the HR HF component to the up-scaled LR NIR image. Simulation results show that the proposed algorithm outperforms the state-of-the-art methods in terms of qualitative as well as quantitative metrics.**

Keywords—Near-infrared and visible images; super-resolution; convolutional neural networks; low light images

I. INTRODUCTION

With the development of infrared (IR) sensor technology, the field of application of IR images has widened. IR imaging is most commonly used in the military and security sectors, which use IR imaging to monitor enemies and detect and remove hidden explosives early on [1]. In addition, there is a famous Microsoft's Kinect [2], which is a typical example of utilizing IR sensor. Kinect provides depth information and skeleton tracking by analyzing the characteristics of point patterns by projecting specific IR dot patterns onto the object. Recently, the importance of IR images is increasing in autonomous vehicles, which may be a big market in the automobile industry. Unfortunately, it is difficult to recognize an object with only visible (VIS) image in the nighttime with low illumination. Therefore, IR image-based object recognition is preferred for driver assistance in the nighttime [3].

In spite of increasing necessity of IR technology, the resolution of IR image is normally lower than that of VIS image due to the limited nature of IR sensor, and blur phenomenon often occurs in the edge area of IR images. So, many algorithms for improving the visual quality of IR images, e.g., super-resolution have been developed.

Note that IR image is more effective in low-light environment than bright environment. But, acquisition of high-resolution (HR) IR image requires high cost. As a result, effective SR technique is required to generate HR images from low resolution (LR) IR images. For example, Zhao et al. presented a reconstruction method for super-resolving IR images based on sparse representation [4]. Still, there is a limit to improve the resolution only by the IR image.

Meanwhile, various approaches for acquiring IR images and corresponding VIS images together and fusing them to generate a desired IR image have been proposed [5-8]. Recently, Ma et al. proposed an IR/VIS fusion method based on gradient transfer and total variation (TV) minimization so that it can keep both the thermal radiation and the appearance information in the source images [7]. Bavirisetti and Dhuli utilized anisotropic diffusion to decompose the source images into approximation and detail layers, and computed final detail and approximation layers with the help of Karhunen-Loeve transform (KLT) [8]. They produced a fused image from the linear combination of final detail and approximation layers. Such methods can generate fused images with enhanced contrast, but they seldom improve the resolution of the IR image itself in nighttime or low light environments.

A remaining problem is that IR images obtained in low light environments are usually bright, but suffer from blur phenomenon and low spatial resolution. On the other hand, in a low light environment, VIS images are generally noisy and dark, while their spatial resolution and definition are relatively better than those of IR images.

In this paper, we propose a CNN-based SR algorithm to improve the resolution of near-IR (NIR) images by using LR IR images and HR VIS images simultaneously acquired in low light environments. First, the HF components are extracted from the input LR NIR image and the corresponding HR VIS image. The extracted heterogeneous HF images are input together into CNN. The two input images are concatenated as soon as entering into the network, and pass through the learned convolutional layer to synthesize HR HF image(s). Finally, the synthesized HR HF NIR image is added to the LR NIR image to generate a HR NIR image. The experimental results show that the proposed algorithm provides a higher PSNR of 0.94 dB than the state-of-the-art [7] in low light environments.

ISBN 978-0-9928626-7-1

II. Related Work

This section briefly reviews recently published CNN-based SR techniques [9-11]. Although conventional CNN-based SR techniques have been developed for VIS images only, they are meaningful because they can be directly applied to NIR images.

Dong et al. [9] applied CNN technique to SR for the first time in the world. Dong et al.'s method, i.e., SRCNN directly learned an end-to-end mapping between the LR and HR images which is represented as a deep CNN that takes the LR image as the input and outputs the HR one.

As an extension of SRCNN, Kim et al. [10] introduced a very deep CNN-based SR (VDSR) with deeper network structure by employing visual geometry group (VGG) network. They used residual-learning and extremely high learning rates to optimize a very deep network fast. Also they adopted gradient clipping to ensure the training stability. As a result, they have demonstrated that VDSR outperforms SRCNN on various benchmarked images.

Kappeler et al. proposed a CNN that is trained on both the spatial and the temporal dimensions of videos to enhance their spatial resolution [11]. Consecutive frames were motion compensated and they were input to the CNN that provides super-resolved video frames as output. This multiple-image-based SR called VSRnet is meaningful in that it is the first example of applying adjacent frames in a video together with CNN.

The proposed algorithm differs from conventional CNN-based SR schemes in the following aspects.

- It focuses on SR of NIR image, not VIS image in low light environment.
- It utilizes VIS image obtained at the same time as auxiliary information.
- It is based on a CNN structure that simultaneously receives HF information of NIR image and VIS image.

III. Proposed Method

Fig. 1 shows the overall structure of the proposed algorithm. The proposed algorithm consists of three steps: extraction of HF components from input NIR and VIS images, and CNN step, and image generation step. The CNN step of producing HR HF information is composed of concatenate layer and convolutional layer(s). It combines VIS image and NIR image to synthesize HR HF component corresponding to LR NIR image. Finally, the HR NIR image is reconstructed by adding the output of the CNN step and the LR NIR image. It is assumed that the NIR LR image has already been up-scaled by a bi-cubic filter so that it has the same spatial resolution as the NIR HR image.

A. High-Frequency Extraction

This section describes the HF extraction step, which is the first step of the proposed algorithm (see Fig. 2). First, when NIR LR image or VIS HR image is input, it is down-scaled through bicubic filter (D) and then up-scaled through bicubic interpolator (U) to generate low-frequency (LF) component image(s). Here, D and U have a scale factor of 2. Next, by subtracting the LF component image from the input image, an HF component image is obtained. In this way, the HF components are extracted from the NIR LR image and the VIS HR image, respectively. Finally, the extracted HF component images are input to the following CNN module.

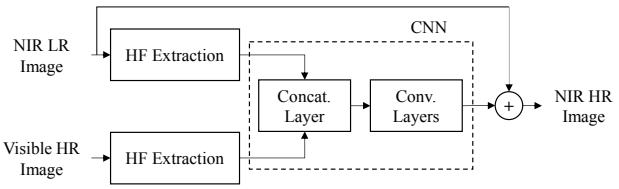

Fig. 1. The block diagram of the proposed method. Our method consists of three parts which is HF extraction part, CNN part and image generation part.

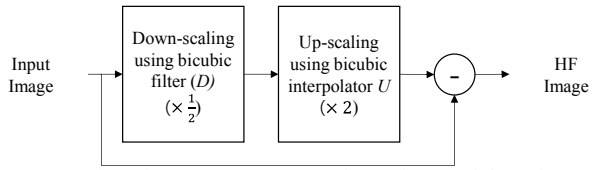

Fig. 2. Extraction of HF component. Input image is passed through *D* and *U*, and the result is subtracted from the input image.

B. CNN Architecture

Inspired by VSRnet [11], we have designed a network architecture where both NIR and corresponding VIS images are input as shown in Fig. 3. The network architecture (architecture A) corresponds to the dotted line in Fig. 1. First, the HF components of the VIS/NIR images which were extracted from Section III.A are input to this CNN module. Each passes through the first convolutional layer and is concatenated before passing through the second convolutional layer. Each concatenated feature merges into one as shown in the dotted line in Fig. 3, and the feature depth increases. For example, when the size of the image is M× N and the number of filters of the n-th convolutional layer is C_n, the size of the NIR data and the VIS data passing through the first convolutional layer are $M \times N \times C_1$. Therefore, the input size of the second convolutional layer after the concatenate layer becomes $2 \times C_1 \times M \times N$.

After the concatenate layer, the features of VIS and NIR images are extracted to the 19th convolutional layer. Here, the number of layers in the proposed algorithm is assumed to be 20. The final convolutional layer is the NIR HR HF reconstruction process by fusing the HF components of VIS HR and NIR LR images.

On the other hand, the structure of the proposed algorithm can be changed according to the location of the concatenate layer after the convolutional layer as shown in Fig. 4. For example, if we place a concatenate layer after the n-th convolutional layer, the input size of the (n + 1) th convolutional layer is $2 \times C_n \times M \times N$. And, the output size of the concatenate layer of architectures B and C becomes $2 \times C_{10} \times M \times N$ and $2 \times C_{19} \times M \times N$, respectively. The

ISBN 978-0-9928626-7-1 © EURASIP 2017

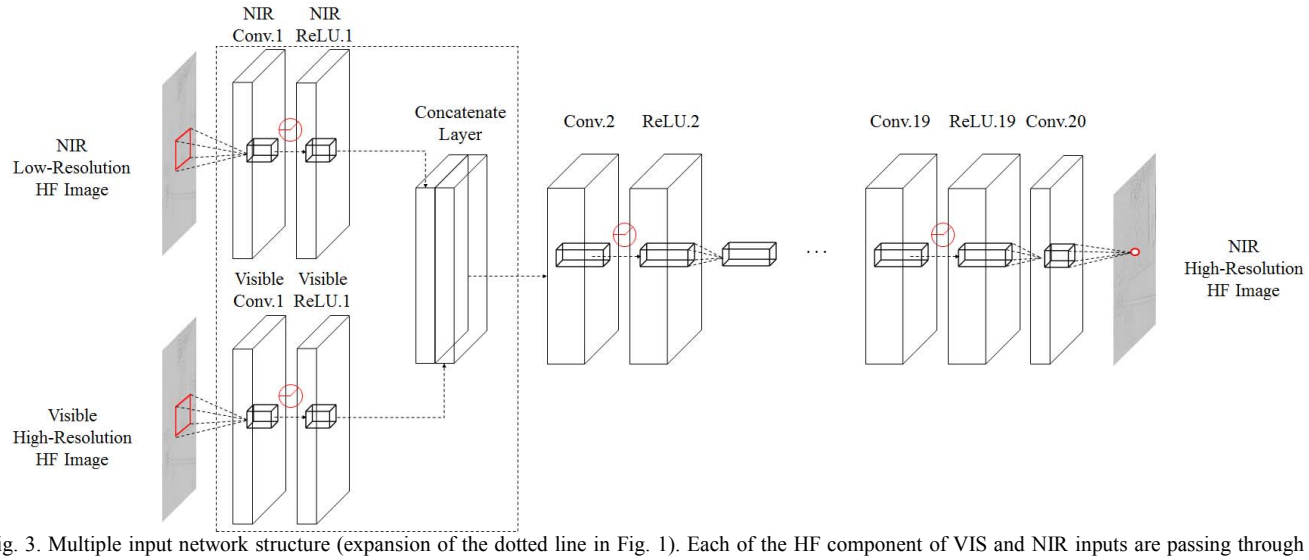

Fig. 3. Multiple input network structure (expansion of the dotted line in Fig. 1). Each of the HF component of VIS and NIR inputs are passing through the first convolutional layer separately, and concatenated after the first layer as the part shown in dotted line. We cascade a pair of convolutional layer and ReLU layer repeatedly after the concatenate layer to fuse visible information with NIR information. We denote this structure as Architecture A.

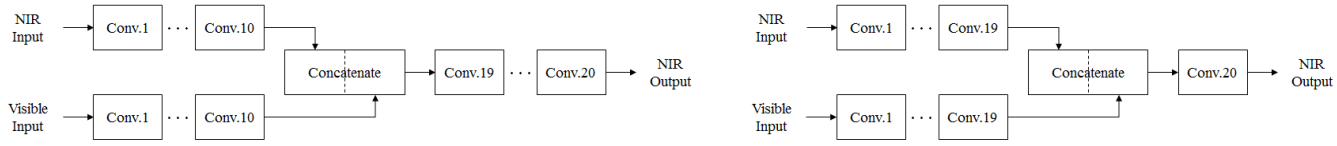

Fig. 4. Examples of network architectures. Multiple inputs are concatenated after the 10th convolutional layer in the Architecture B. Architecture C concatenates both data after the 19th convolutional layer.

performance change of the proposed algorithm according to the CNN structure will be described in Section IV.C.

We arranged a pair of cascaded pairs of Rectified Linear Unit Layers (ReLU) [12] after the convolutional layer, and used a total of 20 convolutional layers. The feature depth of the convolutional layer prior to the concatenate layer was 64 as in [10]. Since CNN has two inputs as described above, the output depth of the convolutional layer after the concatenate layer is doubled, so that the feature depth is 128.

C. CNN Learning Stage

In the CNN learning stage, the original images, i.e., the HR NIR images are used as the label images. As shown in Fig. 5, NIR LR images are generated from NIR HR images, and those pairs are used during learning stage. Here, D and U with a scale factor of 2 were used.

Finally, the HF component of the NIR HR image of Fig. 3 is generated by the process of Fig. 2. For the learning, input and label data are mapped on 41×41 patch basis.

IV. Experimental Results

A. Experimental Condition

For the experiment, PC with Intel Core i7-6700K CPU@4 GHz and 64 GHz RAM and GeForce GTX Titan X graphics card was used. As CNN module, Caffe library [13] was adopted. All the dataset images used for learning and testing were the VIS and NIR image pairs captured by the RGB-NIR camera [14-17]. The training images are 20 sets of 1024 × 682 VIS / NIR images [14] taken indoors, and the images used in the test are some of the images used in [15-17] as shown in Fig. 6. In addition, test 1 to 4 images in Fig. 6 are VIS and NIR image pairs in a bright environment, and test 5 to 8 are VIS and NIR image pairs taken in a low light environment. NIR and VIS images have the same spatial resolution. Test 1 to 4 are all 1024 × 682 resolution images, and the resolution of test 5 to 8 is 1442 × 1006, 1147 × 800, 800 × 600, and 800 × 600, respectively. Note that the NIR LR image used as the test

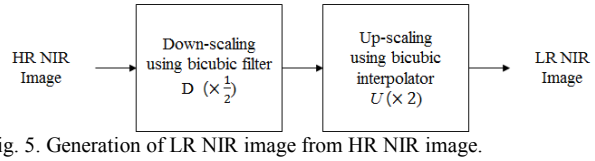

Fig. 5. Generation of LR NIR image from HR NIR image.

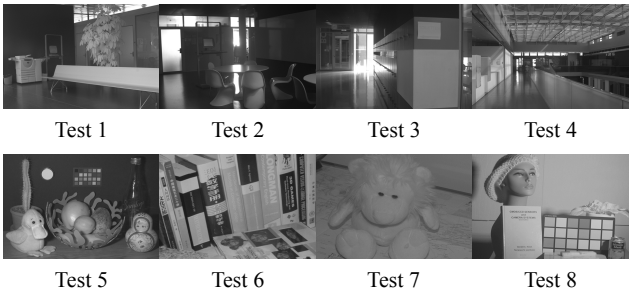

Fig. 6. Thumbnails of NIR test set. Test 1-4 (1024×682) were taken at a bright environment, and test 5 (1442×1006), 6 (1146×800), test 7-8 (800×600) were taken at a low light environment.

ISBN 978-0-9928626-7-1 © EURASIP 2017

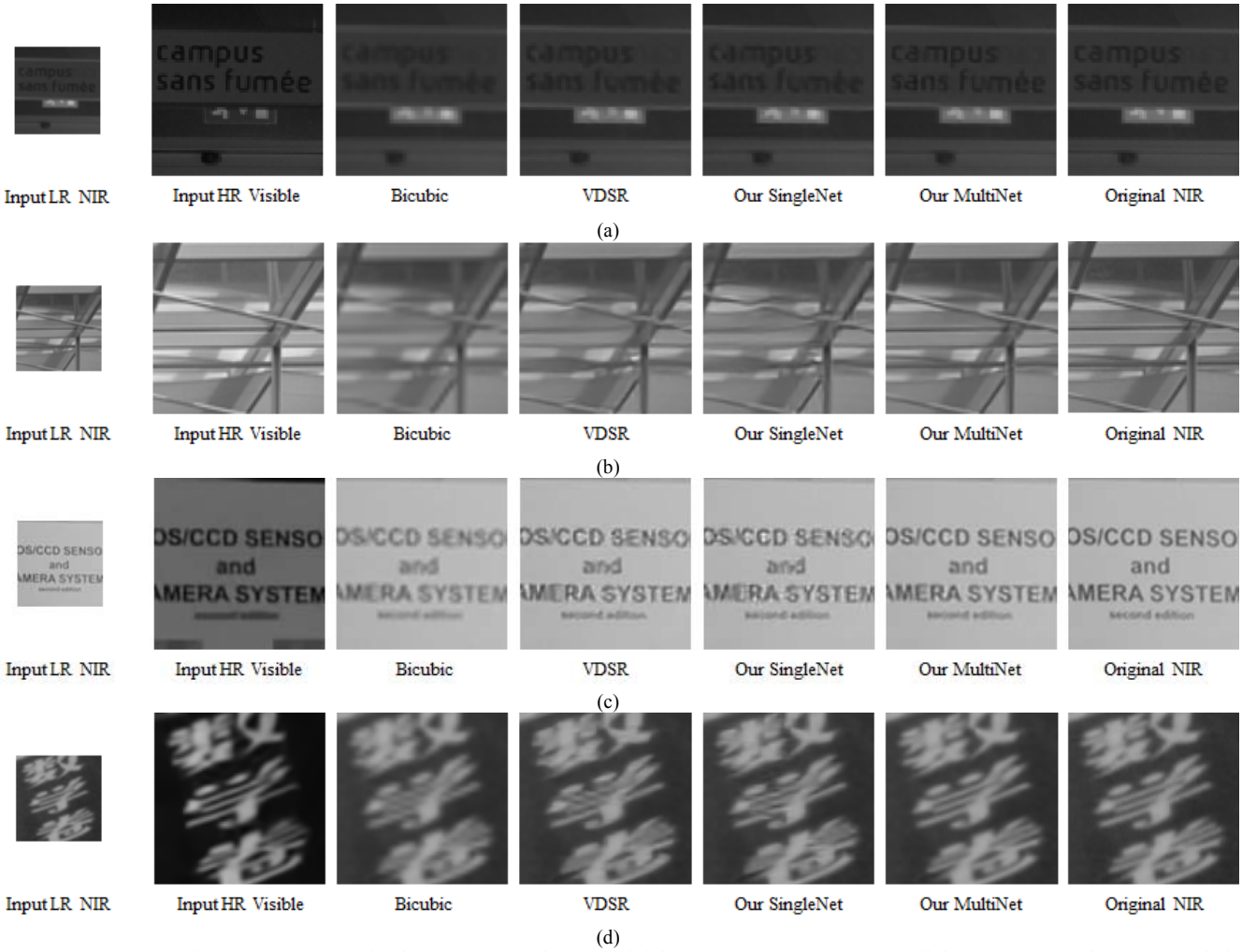

Fig. 7. Comparison of result images. (a) Test 3 (bright environment) (b) Test 4 (bright environment) (c) Test 6 (low light environment) (d) Test 8 (low light environment).

image is generated from the NIR HR image according to Fig. 5. VIS HR images were used in the original resolution. As a result, the proposed scheme up-scales the input NIR LR image with an up-scaling ratio of × 2.

Bi-cubic and the latest CNN-based SR algorithm, VDSR [10] were chosen for comparison with the proposed algorithm. In order to evaluate the performance of CNN itself used in the proposed scheme, the version that receives only NIR which is called SingleNet is also compared with them together. The proposed algorithm that receives both NIR and VIS inputs is called MultiNet relatively. Those algorithms are evaluated in terms of quantitative metric, i.e., PSNR as well as subjective visual quality.

B. Evaluation Results

Fig. 7(a), (b) partly shows the results for 'test 3' and 'test 4' images acquired in a bright environment. We can find that VDSR provides better image quality than bicubic. However, we can still see it blurred. SingleNet also has similar image quality to VDSR. On the other hand, MultiNet, which is the authentic proposed technique, improved the resolution of NIR image clearly like VIS image. Similarly, Fig. 7(c), (d) compares the results for 'test 6', 'test 8' images obtained in low light condition. We can observe that the image quality of the proposed algorithm is better than bicubic and VDSR. Note that the image quality very close to the original NIR HR image is obtained.

TABLE I. TABLE PSNR COMPARISON WITH THE STATE-OF-THE-ART AT BRIGHT ENVIRONMENT (TEST 1-4) AND LOW LIGHT ENVIRONMENT (TEST 5-8). THE BOLDFACED TYPE INDICATES THE BEST PERFORMANCE.

	Bicubic	VDSR	SingleNet	MultiNet
Test 1	34.64	35.94	36.07	**39.21**
Test 2	38.44	44.04	43.53	**45.27**
Test 3	39.02	42.41	41.97	**44.62**
Test 4	32.47	35.49	34.67	**38.45**
Average	36.14	39.47	39.06	**41.89**
Test 5	40.46	42.06	42.07	**42.96**
Test 6	42.93	44.43	44.51	**45.76**
Test 7	44.01	45.31	45.40	**45.67**
Test 8	35.84	39.65	38.65	**40.80**
Average	40.81	42.86	42.66	**43.80**

ISBN 978-0-9928626-7-1 © EURASIP 2017

TABLE II. PSNR PERFORMANCE ACCORDING TO THE LOCATION OF CONCATENATE LAYER. THE BOLDFACED TYPE INDICATES THE BEST PERFORMANCE.

	Architecture A	Architecture B	Architecture C
Test 5	**42.96**	42.93	42.66
Test 6	**45.76**	45.73	45.42
Test 7	**45.67**	45.64	45.63
Test 8	**40.80**	40.58	40.53
Average	**43.80**	43.72	43.56

Table 1 shows the results in terms of PSNR. For a test set obtained in a bright environment, MultiNet has a PSNR of 2.42 dB higher than VDSR on average. Although the disparity is somewhat reduced for test set obtained in low-light environment, MultiNet still has an average PSNR of 0.94 dB higher than VDSR for test 5-8 images.

The proposed algorithm shows clear SR result because it replaces the HF component, which is fundamentally insufficient for NIR image, with the HF component of VIS image. Also, the proposed algorithm adopted CNN to improve the resolution while suppressing possible artifacts.

C. *Evaluation of Network Architecture*

In Section III.B, we mentioned that the performance of the proposed algorithm depends on the position of the concatenate layer. This is because the features of the input image passing through each convolutional layer are extracted differently depending on the position of the concatenate layer. So we performed the verification of architectures A, B, and C in Fig. 3 and 4 to investigate the performance of the proposed algorithm based on concatenate layer location.

Table 2 shows the comparison results for three architectures. As shown in Table 2, the performance of architecture A is the best, and the performance of architecture B and C are degraded as the position of the concatenate layer approaches the final convolutional layer. The reason is that architecture A can pass through more convolutional layers after concatenate layer than architectures B and C, hence it can extract and utilize further information of VIS and NIR images.

V. CONCLUSION

In this paper, we proposed a CNN-based NIR image SR technique using VIS image in a low light environment. The proposed algorithm fuses the HF components of NIR image and VIS image based on a CNN structure. As a result, the missing HF component of the NIR image was effectively reconstructed by the HF component of the corresponding VIS image. In the low light environment, PSNR of the proposed algorithm is improved by 0.94dB on average in comparison with a state-of-the-art SR, i.e., VDSR.

ACKNOWLEDGEMENT

This work was supported by the Industrial Strategic Technology Development Program (10073154, Development of human-friendly human-robot interaction technologies using human internal emotional states recognition) funded By the Ministry of Trade, industry & Energy (MI, Korea)

REFERENCES

[1] K. H. Ghazali, and M. S. Jadin, "Detection Improvised Explosive Device (IED) emplacement using infrared image," International Conference on Computer Modelling and Simulation, 2014.

[2] Z. Zhang, "Microsoft Kinect sensor and its effect," IEEE Multimedia Magazine, vol. 19, mo. 2, pp. 4-10, February 2012.

[3] T. Y. Han, and B. C. Song, "Night vision pedestrian detection based on adaptive preprocessing using near infrared camera," IEEE International Conference on Consumer Electronics-Asia, 2016.

[4] Y. Zhao et al., "A novel infrared image super-resolution method based on sparse representation," Infrared Physics & Technology, vol. 71, pp. 506-513, July 2015.

[5] X. Li and S. Y. Qin, "Efficient fusion for infrared and visible images based on compressive sensing principle," IET Image Processing, vol. 5, no. 2, pp. 141-147, 2011.

[6] A. Gyaourova, G. Bebis, and I. Pavlidis, "Fusion of infrared and visible images for face recognition," European Conference on Computer Vision (ECCV), Berlin Heidelberg, 2004.

[7] J. Ma et al., "Infrared and visible image fusion via gradient transfer and total variation minimization," Information Fusion, vol. 31, pp. 100-109, 2016.

[8] D. P. Bavirisetti and R. Dhuli, "Fusion of infrared and visible sensor images based on anisotropic diffusion and Karhunen-Loeve transform," IEEE Sensors Journal, vol. 16, no. 1, pp. 203-209, 2016.

[9] C. Dong, C. C. Loy, K. He and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295-307, 2016.

[10] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[11] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," IEEE Transactions on Computational Imaging, vol. 2, no. 2, pp. 109-122, June 2016.

[12] A. L. Mass, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," International Conference on Machine Learning, 2013.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," arXiv preprint arXiv: 1408.5093, 2014.

[14] M. Brown, and S. Susstrunk, "Multi-spectral SIFT for scene category recognition," IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[15] X. Shen, L. Xu, Q. Zhang, and J. Jia, "Multi-modal and multi-spectral registration for natural images," European Conference on Computer Vision, 2014.

[16] D. Krishnan and R. Fergus, "Dark flash photography," ACM Transactions on Graphics, vol. 28, no. 96, August 2009.

[17] Q. Yan, X. Shen, L. Xu, S. Zhuo, X. Zhang, L. Shen, and J. Jia, "Cross-field joint image restoration via scale map," International Conference on Computer Vision, 2013.