3D VIDEO SUPER-RESOLUTION USING FULLY CONVOLUTIONAL NEURAL NETWORKS

Yanchun Xie[1], *Jimin Xiao*[1], *Tammam Tillo*[1], *Yunchao Wei*[2], *Yao Zhao*[2]

[1]Xi'an Jiaotong - Liverpool University, [2]Beijing Jiaotong University

{jimin.xiao,tammam.tillo}@xjtlu.edu.cn

ABSTRACT

Large amount of redundant information and huge data size have been a serious problem for multiview video systems. To address this problem, one popular solution is mixed-resolution, where only few viewpoints are kept with full resolution and other views are kept with lower resolution. In this paper, we propose a super-resolution (SR) method, where the low-resolution viewpoints in the 3D video are up-sampled using a fully convolutional neural network. By simply projecting the neighboring high resolution image to the position of the low resolution image, we learn the relationship of high and low resolution patches, and reconstruct the low resolution images into high resolution ones using the projected image information. We propose to use a fully convolutional neural network to establish a mapping between those images. The network is barely trained on 17 pairs of multiview images, and tested on other multiview images and video sequences. It is observed that our proposed method outperforms existing methods objectively and subjectively, with more than 1 dB average gain achieved. Meanwhile, our network training procedure is efficient , with less than 3 hours using one Titan X GPU.

Index Terms— super resolution, mix-resolution, virtual view, depth map, convolutional neural network , training

1. INTRODUCTION

Multiview video with mixed-resolution has proven to be an effective method for multiview video compression in recent years [1, 2, 3]. For such multiview video, only few view points are kept with full resolution and other views are down-sampled. At the decoder and display side, the down-sampled viewpoint could be super-resolved to full resolution before display. In mix-resolution stereoscopic video, one of the two views is represented with a lower resolution compared to the other one, while, according to the binocular suppression theory, it is assumed that the Human Visual System (HVS) fuses the two images such that the perceived quality is close to that of the higher quality view [4], [5].

For multiview videos with mix-resolution textures and depth maps, it is feasible to improve the low resolution texture video's quality by using up-sampling methods. Neighboring views share a same but shifted image which are captured from different perspectives, so that we could extract information from full resolution (FR) images to recover and improve the quality of the low resolution (LR) images. One approach to enhance LR images was proposed in [6]. This method obtains high frequency information from the adjacent FR images and the corresponding depth maps. Such information is exploited during the LR image up-sampling procedure. Some studies [7], [3] suggested that with the help of FR images in mix-resolution multiview video, LR image could be recovered well by setting appropriate threshold when choosing patch from LR images and projected view. These methods need a lot of manual work to determine global parameters to meet the requirements of various cases. Instead of establishing a model by hand, using a well-designed learning method would be a better solution.

Besides traditional signal processing-based methods for image super-resolution, convolutional neural networks (CNN) has been proven to be a powerful tool to solve this problem. Dong *et al.* [8] introduced a deep learning method to learn an end-to-end mapping between low/high resolution images. It lead to better results than dictionary learning and sparse coding SR methods [9],[10].

Inspired by the brave attempt [8], we consider that a convolution neural network would be helpful in combining related information in the mix-resolution multiview video system. In this paper, we introduce a method that directly takes two inputs (projected view and interpolated image from LR view) into the 3-layer CNN, and only 17 image pairs are used to train the network. We test the network's performance on other images and video sequences. It is observed that on the test image/video sequences, our approach outperform both methods with a significant gain.

The following parts of paper is organized as follows. We introduce the closely related works in the next section, and our proposed method is given in section 3, including the training details along with the feature analyses, and experimental results are presented in section 4.

2. RELATED WORK

2.1. SR with convolutional neural network

Before (CNN) was used for the task of SR, [11, 12, 13, 14] tried to learn the example-based mapping between image patches or internal similarities of images. All these methods require a large amount of examples to generate a dictionary of low/high resolution, which is difficult and inefficient. CNN-based SR learning method [8] takes advantage of flexible learning strategies, powerful training tools (GPUs) and abundant data. Thus, it leads to better performance compared with traditional learning methods. In [8], three layers of convolutional neural network is deployed for three operations (1.Patch extraction and representation; 2.Non-linear mapping; 3.Reconstruction.), and it also shows that the sparse-coding-based SR method [10] can be also viewed as a convolutional neural network.

2.2. Multiview image SR using depth information

Diogo *et al.* [6] presented a super-resolution method, in which low resolution views are enhanced with the aid of high frequency content from neighboring full resolution views. More specifically, they utilized the Depth-Image-Based-Rendering (DIBR) technique to generate a virtual view from the full resolution viewpoint. The high frequency components of the virtual view are extracted, and added to the interpolated low resolution view, with a consistency check step. Motivated by [6], Jin *et al.* [7] put forward a SR method for mix-resolution multiview videos. The core idea of the method is to selectively pick pixels from either the interpolated image or the full resolution virtual view. Inter-view similarity check is used during the selection process, with many manually set parameters. In both [6] [7] the final reconstructed image is generated using 2 types of distorted images, i.e., interpolated image and the full resolution virtual view.

3. 3D VIDEO SUPER-RESOLUTION USING CNN

Our goal is to recover the LR images with help of neighboring FR images and depth maps. Assuming that the depth maps of the 3D multiview video are available, we can use the FR image and its depth map to construct a 3D image and thus to generate a virtual view in the position of the LR image. Due to the change of viewpoint, for the generated virtual view, we will miss some parts which are hidden by objects in front of them. These parts are usually called occluded regions or occlusions. Meanwhile, the interpolated image is blurred due to the loss of high frequency information, especially for the object edges and rich texture regions. Generating a high quality image from a virtual view and interpolated image is a non-trivial task. Different from previous studies [3][6][7], in this paper, we do not set a threshold or perform a consistency check operation, all the information fusion work is done by the training of a fully convolutional neural network.

3.1. View projection – virtual view generation

According to the DIBR technique, a mapping of reference image and virtual view could be established through 3D-warping.

$$X_{warp} = X_{origin} + b * \frac{f}{Z}. \tag{1}$$

where the horizontal coordinates of the reference point and the virtual viewpoint are denoted as X_{origin} and X_{warp}; b represents the baseline distance, f is the focal length of the camera, and Z is the depth value of the reference image pixel.

Notice that if the Z value difference of neighboring pixels is big enough, occlusions in virtual view become inevitable. For traditional methods [6] [7], after establishing correspondences between reference view and virtual view, consistency check and pixel interpolation are performed.

3.2. Image fusion – LR view reconstruction

Besides the view projection step, another pre-processing step in our approach is batch cropping. We cropped the images into small sub-images (i.e., 33 × 33), which is also called patched, and feed them into the CNN input. Using small patches instead of the whole image will speed up the training process. We group two same position sub-images into a pair, one is the interpolated LR view, another is a projected virtual view but with high resolution. Denote the interpolated LR image and virtual image as L and V respectively, it is possible to use traditional methods to remove the noisy pixels in V, and utilize the rest information to repair L. With the same purpose, we train the neural network and learn a mapping between the two images and the ground truth image G.

3.2.1. Image fusion and feature extraction

The image merging and feature extraction step is the most important one. In this step, we use one convolutional layer to fuse L and V, and meanwhile extract features from them. Using different filters (9 × 9) for the convolution operation, we obtain a set of feature vectors from the 2 input patches (L and V) , we also call them feature maps. In this operation, we could separate the occluded regions from V. Thus, in the next convolution operation, the coefficients of these feature map regions are set with small values. Therefore, this convolution layer has the denoising effects and directly improves the overall performance. We also add a bias term to the convolutional result, and all the filters and biases are obtained by optimizing the output of the network.

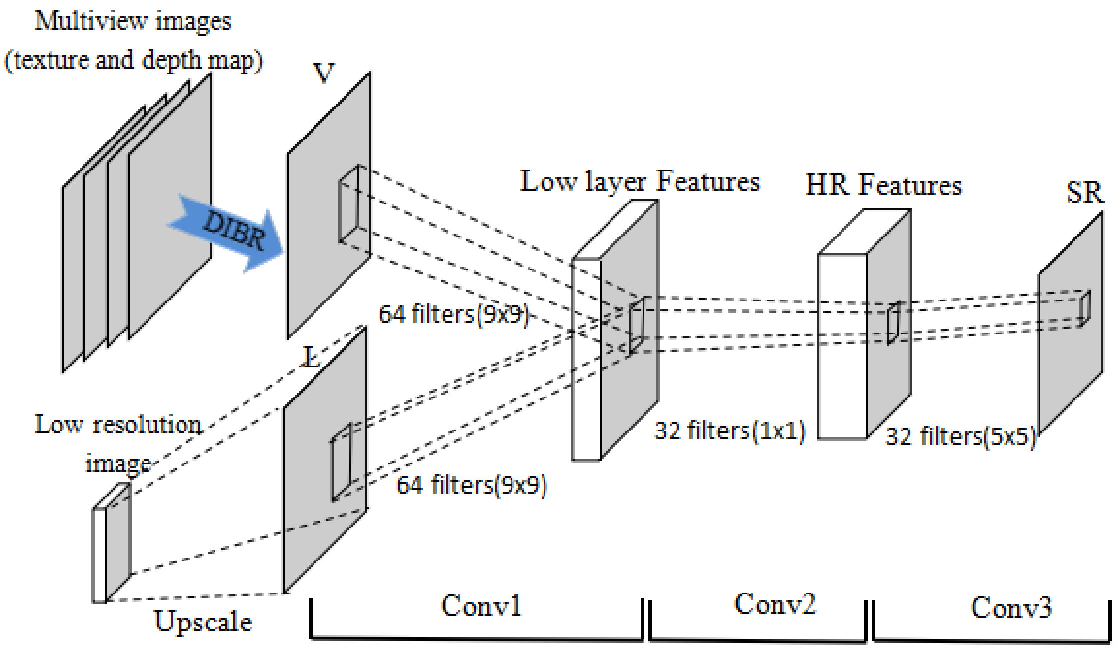


Fig. 1. Our network structure. An interpolated low resolution image (LR) and a projected virtual view (VR) go through 3 convolutional layers. In the first layer, two images are combined as 2 input channels, we obtain a set of feature maps. In the second layer, a mapping of the low level features and high resolution features is established, and then these features pass through the last convolutional layer to get the output – super-resolution image.

The convolution operation in this layer is represented as follows:

$$F_i = W_{1,i} \cdot L + W_{2,i} \cdot V + bias_i, \qquad (2)$$

where $\mathrm{W}_{1,i}$ and $\mathrm{W}_{2,i}$ represent the ith filter in the layer 1 of CNN, and F_i is the ith feature map generated after the convolutional layer 1.

We also apply the Rectified Linear Unit [15] (ReLU) on every neuron's output. The ReLU function is $max(0, F_i)$.

3.2.2. *Mapping and reconstruction*

The mapping and reconstruction step includes the second and third convolutional layers. In this step, we are going to build a nonlinear mapping between the previously generated feature maps and the final full resolution images. Notice that in the neural network, the computational complexity grows exponentially with the increase number of hidden layers and the neural cells. Hence, we use only 1 convolutional layer as hidden layer to establish a nonlinear map of feature maps with different dimensions. In the end, another convolutional layer is used to construct a high resolution image.

The above two steps are both achieved using convolution operations. Our three-layer network structure is shown in Fig.1.

3.2.3. *Training*

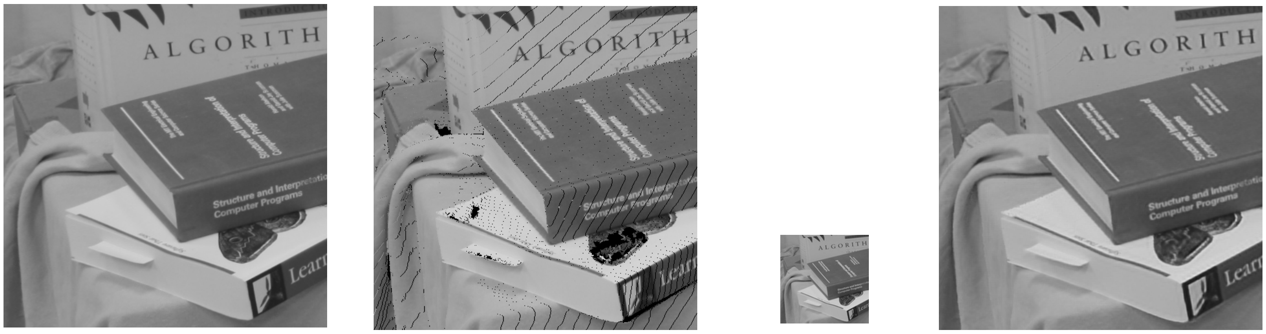
Given a training dataset $\{ x^{(i)}, y^{(i)}, l^{(i)} \}_{i=1}^{N}$, our goal is to train a model f of 3 layers that predicts value $\hat{z} = f(x, y)$, where input data x and y are the interpolated LR image and the projected virtual view, l is the ground truth full resolution image (label), and $\hat{z}$ is the estimated high resolution image. In order to learning a mapping between $x^{(i)}$, $y^{(i)}$ and $\hat{z}^{(i)}$ we use Mean Squared Error (MSE) as the loss function:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \| f(x^{(i)}, y^{(i)}, \theta) - l^{(i)} \|^2 \qquad (3)$$

where θ is the parameters of the network, including both filters and biases. Using MSE as loss function conduce to a high PSNR. However, in our framework, it is flexible for the network to adapt to a new evaluation index by modifying the loss function.

Stochastic gradient decent (SGD) is used to minimize the loss function with the standard backpropagation. At the beginning, we initialize the weights of the network with values of random gaussian distribution with zero mean. The learning rate is 10^{-3} for the first 2 layers, and 10^{-4} in the last layer for the first 10^6 backpropagations, and then we reduce the learning rate by a factor of 10.

In order to speed up the training process in a efficient way, small sub-image (patch) pairs are generated from the interpolated images and virtual views. This is because for large training images, CNN will take much more time than the small ones to complete the convolution operations. Meanwhile, more data for training comes with better results. Therefore, using small sub-image leads to a good trade-off between accuracy and speed.



(a) Orginal view
Label

(b) Virtual view
PSNR = 21.32 dB

(c) Low resolution
PSNR = 33.68 dB

(d) Super resolution
PSNR = 37.41 dB

Fig. 2. Example of our SR result. In this example, the scale is 4, and the PSNR of the Low resolution image is computed after bicubic interpolation.

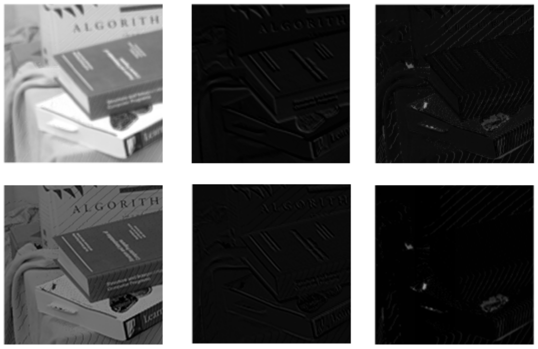


Fig. 3. Feature map visualization in conv1 of CNN

3.3. Feature visualization and analysis

Taking Fig.2 as an example, we want to find out how the filters in convolutional layers remove the noise pattern in V and how much V and L contribute to the final SR image.

Visualization of the feature maps is partly shown in Fig.3. In convolutional layer 1, feature maps can be divided into 3 categories, one is combined images without or with noise (the first 2 images); the second category is the edge features which contain high frequency information of the original image (the 3rd and 4th images) ; the third category feature maps contain some hole (occluded) regions which are from projected virtual view (the 5th and 6th images).

The first convolution layer has the denoising effect as expected, and the following 2 convolution layers will further remove the noisy patterns and build a clearer final image. Notice that in this experiment, 64×2 filters are used in convolution layer 1, 64 for both L and V, so that we can get 64 feature maps after convolution layer 1. By increasing the number of the filters, more feature map will be generated, and usually the accuracy of the final image is improved.

4. EXPERIMENTS

4.1. Training and testing data

Training datasets In order to improve the network's performance, CNN usually requires large amount of training data. SRCNN [8] uses a very large ImageNet dataset and 91 images from Yang *et al.* [10] to train the network separately. For low-level vision task as super-resolution, it turns out that the effect of big data is not as impressive as that shown in classification or detection problems. Therefore, we use 17 image pairs in Middlebury Stereo Dataset [16] for training. Each image pair consists of 2 views (left and right textures and depth maps) taken under several different illuminations and exposures. 429184 sub-image (33×33) pairs are generated from the 17 pairs of interpolated images and virtual views.

Testing datasets To test our network's performance and robustness, not only the rest image pairs in Middlebury Stereo Datasets, but also different kinds of multiview video sequences from MPEG are used. Test sequences "sticks", "storage", "sword1", "sword2", "vintage" are selected from Middlebury Stereo Datasets (2014), and the "Doorflower", "Newspaper", "Laptop" are from MPEG video sequences whose depth maps are less accurate.

Our CNN uses 3 convolution layers, $2\times$ 64 filters (9×9) in conv1, 64×32 filters (1×1) in conv2 and 32×1 filters (5×5) in conv3. For all layers, the convolution stride is 1.

4.2. Performance evaluation

We compare the proposed method with different types of existing approaches, including the depth information-based SR method [7] and the CNN-based approach SRCNN [8]. It can

Table 1. Comparison with simplified [7]

Dataset	Scale	simplified [7](dB)	Proposed (dB)
Sticks	2	38.15	40.07
Storage	2	47.26	49.00
Sword1	2	40.42	42.93
Sword2	2	47.36	50.88
Vintage	2	39.74	43.16
Doorflower	2	36.58	39.19
Newspaper	2	39.16	41.79
Laptop	2	37.14	39.17
Average	2	40.73	43.27

Table 2. Comparison with SRCNN [8], with down-sampling scale 3

Dataset	Scale	Interpolated (dB)	SRCNN[8] (dB)	Proposed (dB)
Sticks	3	34.17	35.45	36.19
Storage	3	42.75	45.05	43.14
Sword1	3	36.00	37.48	38.16
Sword2	3	42.19	43.44	44.53
Vintage	3	35.65	36.94	39.67
Doorflower	3	30.62	32.07	36.15
Newspaper	3	34.28	36.18	36.21
Laptop	3	31.32	32.61	35.71
Average	3	35.87	37.34	38.72

Table 3. Comparison with SRCNN [8], with down-sampling scale 4

Dataset	Scale	Interpolated (dB)	SRCNN[8] (dB)	Proposed (dB)
Sticks	4	32.05	33.12	35.47
Storage	4	40.27	42.02	41.47
Sword1	4	33.44	34.54	37.15
Sword2	4	39.08	42.31	42.90
Vintage	4	33.13	34.65	39.81
Doorflower	4	30.62	32.07	36.15
Newspaper	4	31.56	33.38	34.57
Laptop	4	31.32	32.61	35.71
Average	4	33.93	35.57	37.90

be found that in Table 1, 2 and 3, our method outperforms both [7] and [8].

SR method using depth information has achieved a certain degree of gain for compressed video sequences [7]. However, SRCNN uses uncompressed images for training and testing, in order to keep consistency, we reproduced the work [7] with the same algorithm but using uncompressed images. Therefore, the dataset in Table 1, 2 and 3 are the same images. Meanwhile, we simplified the work [7] by keeping the zero filling algorithm, and skipping the post image-smooth processing, which leads to far less than 1 dB gain in the final results according to the paper. The performance is shown in Table 1, and it is obvious that our well-trained network performs better than [7], where complex-calculated thresholds are used to fuse pixels from 2 images.

Table 2 and 3 show that our method achieves a significant gain, 1.38 dB and 2.33 dB higher than the SRCNN for down-sampling scale 3 and 4, respectively. It is worth mentioning that testing dataset includes depth maps of various quality levels. Notice that our method does not achieve the best result for test sequence "Storage", which is due to 2 reasons. One is that large disparity value for "Storage" leads to huge hole regions in the projected view, and finally decrease the quality of the fused image. This problem can be avoided because the disparity of neighboring view is usually small in multi-view video. Another reason is that we do not provide enough samples for the training, thus this problem can be solved by increasing the size of training data.

Particularly, one zoom-in example is given in Fig.4. After down-sampling using a large scale 4, it is difficult to recover a satisfactory image for SRCNN. Such information loss in multiview can be compensated with the assistance of neighboring images and depth maps.

5. CONCLUSION

In this work, we have presented a super-resolution method for multiview video using fully CNN. We have demonstrated that our method outperforms the existing methods on the testing image pairs and multiview video sequences. We believe our approach could achieve better result using a deeper network with more training data.

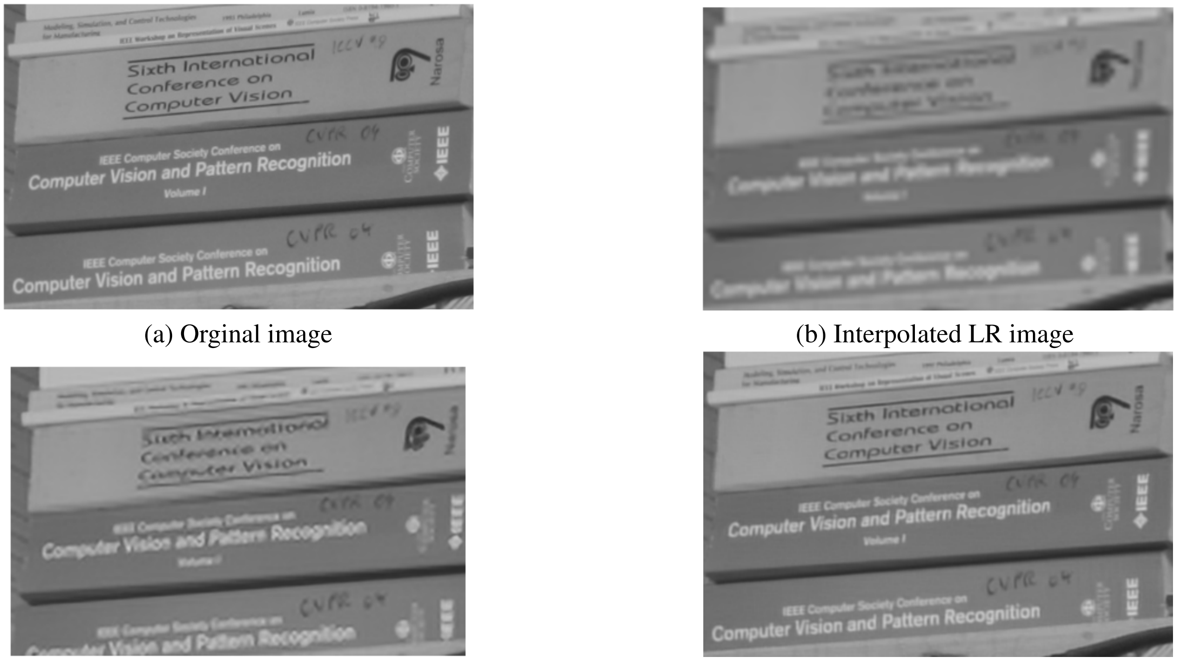
Acknowledgment This work was supported by the National Natural Science Foundation of China (NO. 61501379 and NO. 61210006) and the Jiangsu Science and Technology Programme (BK20150375)

6. REFERENCES

[1] Ankit K. Jain and Truong Q. Nguyen, "Video super-resolution for mixed resolution stereo," in *ICIP*, 2013.

[2] Zhizhong Fu, Zening Li, Lan Ding, and Truong Nguyen, "Translation invariance-based super resolution method for mixed resolution multiview video," in *ICIP*, 2014.

[3] Michal Joachimiak, Payman Aflaki, Miska M. Hannuksela, and Moncef Gabbouj, "Evaluation of depth-based super resolution on compressed mixed resolution 3d video," in *ACCV*, 2014.



(a) Orginal image

(b) Interpolated LR image

(c) SRCNN result

(d) Proposed SR result

Fig. 4. Image quality comparison

[4] P. Aflaki, M. M. Hannuksela, J. Hakkinen, P. Lindroos, and M. Gabbouj, "Impact of downsampling ratio in mixed-resolution stereoscopic video," in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*, 2010, pp. 1–4.

[5] Lew Stelmach, Wa James Tam, Dan Meegan, and Andre Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 188–193, 2000.

[6] Edson M. Hung, Camilo C. Dorea, Diogo C. Garcia, and Ricardo L. de Queiroz, "Transform-domain super-resolution for multiview images using depth information," in *EUSIPCO*, 2010.

[7] Zhi Jin, Tammam Tillo, Chao Yao, Jimin Xiao, and Yao Zhao, "Virtual view assisted video super-resolution and enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. pp, no. 99, pp. 1, 2015.

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014.

[9] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *CVPR*, 2012.

[10] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, pp. 2861–2873, 2010.

[11] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang, "Single-image super-resolution: A benchmark," in *ECCV*, 2014.

[12] Gilad Freedman and Raanan Fattal, "Image and video upscaling from local self-examples," *ACM Trans. Graph.*, vol. 30, pp. 12, 2011.

[13] Zhen Cui, Hong Chang, Shiguang Shan, Bineng Zhong, and Xilin Chen, "Deep network cascade for image super-resolution," in *ECCV*, 2014.

[14] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, "Single image super-resolution from transformed self-exemplars," in *CVPR*, 2015.

[15] Vinod Nair and Geoffrey E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[16] Daniel Scharstein, Heiko Hirschmller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *DAGM*, 2014.