applied sciences

MDPI

Article

Stereoscopic Image Super-Resolution Method with View Incorporation and Convolutional Neural Networks

Zhiyong Pan [1], Gangyi Jiang [1,*], Hao Jiang [2], Mei Yu [1,*], Fen Chen [1] and Qingbo Zhang [2]

[1] Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China; zhiyong_pan@126.com (Z.P.); chenfen@nbu.edu.cn (F.C.)

[2] Intelligent Household Appliances Engineering Center, Zhejiang Business Technology Institute, Ningbo 315012, China; jd_jh@zjbti.net.cn (H.J.); zqbb0805@126.com (Q.Z.)

* Correspondence: jianggangyi@126.com (G.J.); yumei@nbu.edu.cn (M.Y.); Tel.: +86-574-8760-0411 (G.J.)

Academic Editor: Martin Richardson

Received: 6 March 2017; Accepted: 12 May 2017; Published: 26 May 2017

Abstract: Super-resolution (SR) plays an important role in the processing and display of mixed-resolution (MR) stereoscopic images. Therefore, a stereoscopic image SR method based on view incorporation and convolutional neural networks (CNN) is proposed. For a given MR stereoscopic image, the left view of which is observed in full resolution, while the right view is viewed in low resolution, the SR method is implemented in two stages. In the first stage, a view difference image is defined to represent the correlation between views. It is estimated by using the full-resolution left view and the interpolated right view as input to the modified CNN. Accordingly, a high-precision view difference image is obtained. In the second stage, to incorporate the estimated right view in the first stage, a global reconstruction constraint is presented to make the estimated right view consistent with the low-resolution right view in terms of the MR stereoscopic image observation model. Experimental results demonstrated that, compared with the SR convolutional neural network (SRCNN) method and depth map based SR method, the proposed method improved the reconstructed right view quality by 0.54 dB and 1.14 dB, respectively, in the Peak Signal to Noise Ratio (PSNR), and subjective evaluation also implied that the proposed method produced better reconstructed stereoscopic images.

Keywords: stereoscopic imaging and coding; mixed-resolution stereoscopic image; super-resolution; view difference; convolutional neural networks

1. Introduction

With advancements in imaging, processing, and display technologies in recent years, stereoscopic video entertainment and communication have emerged as promising services of novel visual user experiences such as three-dimensional (3D) television [1], free-viewpoint video [2], and video conferencing [3]. Compared with monocular images, stereoscopic images provide depth perception and engender an immersive user experience [4]. Meanwhile, the immense amount of data generated by stereoscopic imaging requires large storage and transmission capabilities and thus must be efficiently encoded and processed. On the basis of binocular suppression theory [5], higher quality views will be received as the perceived quality of stereo vision by the human visual system (HVS). Thus, mixed resolution (MR) stereoscopic image processing techniques are motivated by binocular perception theory. Specifically, one view of the MR stereoscopic image is provided with full resolution (FR), whereas the other view is degraded by the MR stereoscopic image observation model. To decrease the amount of data while preserving the high definition and stereo vision experience, the low-resolution

(LR) view must be super-resolved to a high resolution (HR) at the decoder and display side. In recent years, MR stereoscopic imaging and processing techniques have proven to be effective approaches for stereoscopic imaging and compression [6].

Existing super-resolution (SR) methods are used to reconstruct the FR image from its LR version. These methods are mainly divided into three types; interpolation [7,8], reconstruction [9,10], and learning [11–15]. Among them, the learning-based method has become widely used owing to its outstanding performance. Its basic idea is to establish a mapping relation between the LR and HR image patches and then to find the optimal solution from the LR image. Thus, to study the common prior knowledge between the image patches, most renowned methods adopt the learning-based strategy [16]. Chang et al. [11], for example, adopted the concept of local linear embedding to propose an SR reconstruction method based on neighborhood embedding. Furthermore, the above-mentioned neighborhood embedding is deduced to a more complex sparse coding formulation in Yang et al.'s work [12,13]. They determined that a linear combination of atoms from an over-complete dictionary can well represent natural image patches. Therefore, after the training process on HR and LR image patches, HR and LR dictionaries are jointly obtained. Then, according to the observed LR patch and LR dictionary, the sparse coefficients are achieved and applied to the HR dictionary to produce the final HR patches.

To improve the SR speed while maintaining SR accuracy, Timofte et al. [14] used several smaller complete dictionaries to replace the single large over-complete dictionary, thereby greatly reducing the computational cost. In recent years, with the development of deep learning, an increasing number of researchers have employed deep learning for image processing such as image classification [17], object detection [18], and image denoising [19]. Additionally, some researchers have begun establishing a depth model for SR reconstruction. Cui et al. [20] adopted stacked auto-encoders which combined the internal example-based approach to gradually upsample LR images layer by layer. Moreover, Dong et al. [15] combined dictionary learning and neural networks to establish a model of the SR convolutional neural network (SRCNN). This model showed better performance than traditional methods, such as dictionary learning and sparse coding. Furthermore, Liu et al. [21] emphasized the importance of traditional sparse representation. They integrated it into deep learning to further improve the SR results. Although the LR view of the MR stereoscopic image can be directly upsampled by these single-view SR methods, these methods do not take advantage of the correspondence between views for stereoscopic image SR.

The observed FR image in the neighboring view of the stereoscopic image can provide richly detailed information of the scene. Thus, the relativity between views has been utilized to strengthen the particular LR views. Garcia et al. [22] proposed an SR method that exploits depth information for the MR multi-view video. On the basis of the available depth maps, their approach enhanced the observed LR image by extracting the high-frequency content from the neighboring FR view. However, the acquisition of the depth maps was not discussed in their work as these are not easy to accurately estimate. In addition, Brust et al. [23] employed the estimated depth map calculated in advance from the original FR stereoscopic pairs to render the LR view from other HR views. Again, the original FR stereoscopic pairs cannot be obtained at the decoder side. Therefore, all of the above methods are not fully consistent with the MR stereoscopic imaging and processing techniques.

Unlike existing SR methods, which require depth maps or depth estimation, we combine the correlation between the views of the MR stereoscopic image without estimating the depth map. We propose a stereoscopic image SR method based on view incorporation and convolutional neural network (CNN). The proposed stereoscopic image SR method can be implemented in two stages. In the first stage, for establishing links between views, a view difference image is defined, and the modified CNN is created to estimate a high-precision view difference image. Then, the estimated right view image is obtained by subtracting the estimated view difference image from the observed FR left view image. In the second stage, we consider that the estimated right view image should be retained with the LR right view image with regard to the MR stereoscopic image observation

model. Accordingly, we model the global reconstruction constraint for incorporating the right view by projecting the estimated right view image obtained in the first stage onto the solution space of the image observation model. The solution can be computed by iterative back projection [24]. The SR results demonstrate that the proposed SR method retained more details and well reduced ringing.

In short, the contributions of this paper are outlined as follows:

- We combine the correlation between the views of the MR stereoscopic image without estimating the depth map;
- We use the view difference image containing the image texture information, as well as the depth information of the stereoscopic pairs, as the input to the modified CNN, whose pooling layers and fully connected layers are removed according to the SR task;
- We combine the high-precision view difference image estimated by the modified CNN with the global reconstruction constraint to further improve the performance of the MR stereoscopic image SR.

The remainder of this paper is organized as follows. Section 2 describes an MR stereoscopic image observation model. Then, the proposed stereoscopic image SR method is illustrated in Section 3. Experiments are given and discussed in Section 4. Section 5 presents the conclusions.

2. MR Stereoscopic Image Observation Model

As shown in Figure 1, the observed FR left view and downsampled right view constitute the MR stereoscopic video sequence. Accordingly, the MR stereoscopic video coding model can enable a large amount of data to be compressed for storage and transmission. This is because directly encoding the FR stereoscopic videos leads to doubling the required storage and bandwidth. Actually, owing to the restricted storage space and network bandwidth, this MR stereoscopic video coding model is crucial to providing clear bitrate reduction [6]. Furthermore, for ensuring stereo vision comfort, the SR of the MR stereoscopic video is needed at the decoder. Therefore, to contribute to the MR stereoscopic video coding model, we adopt an MR stereoscopic image observation model for stereoscopic image SR.

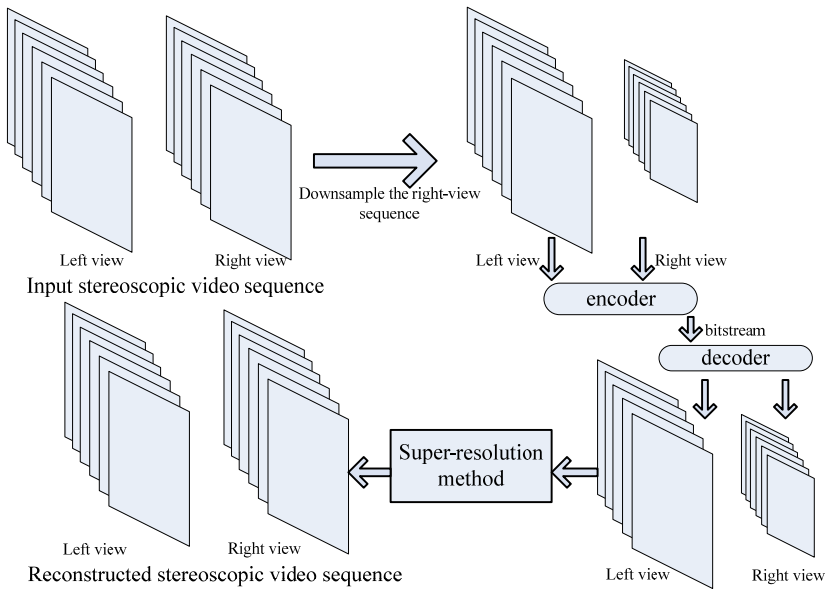


Figure 1. Mixed resolution (MR) stereoscopic video coding model.

As shown in Figure 2, a stereoscopic imaging system obtains the original FR stereoscopic image pairs (with the size of $N_1 \times N_2$ for each view). In general, we assume that there exists the observed FR

left view and observed LR right view (with the size of $M_1 \times M_2$), which is degraded on account of the blurring and downsampling operation. The degradation model is expressed by

$$Y = DBX \tag{1}$$

where $\boldsymbol{Y}$ and $\boldsymbol{X}$ denote the observed LR right view image and original FR right view image (that is the unknown FR right view image), respectively. Moreover, $\boldsymbol{Y}$ and $\boldsymbol{X}$ are both in vector format with the sizes of $M_1M_2 \times 1$ and $N_1N_2 \times 1$, respectively. $\boldsymbol{D}$ is the downsampling matrix with the size of $M_1M_2 \times N_1N_2$, and $\boldsymbol{B}$ is the blurring matrix with the size of $N_1N_2 \times N_1N_2$. In addition, let $\boldsymbol{Z}$ represent the observed FR left view image, the purpose of the proposed method is to acquire an FR right view image, $\boldsymbol{X}$, by making full use of the abundant information of the observed LR right view image $\boldsymbol{Y}$ and observed FR left view image $\boldsymbol{Z}$.

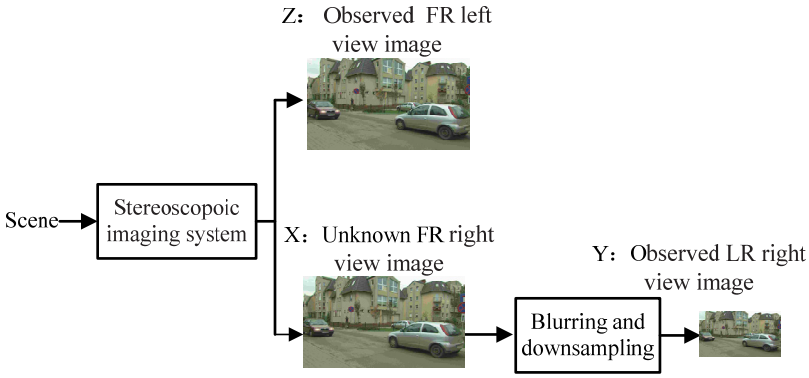


Figure 2. MR stereoscopic image observation model.

3. Proposed Stereoscopic Image SR Method with View Incorporation and CNN

In this paper, we focus on the estimation of the view difference image and global reconstruction constraint to solve the SR task of MR stereoscopic images, as depicted in Figure 3. The proposed method includes two main stages. The first stage is the estimation of a view difference image for employing the correlation between the left and right views by using a modified CNN. Firstly, the observed LR right view image is interpolated into FR by using a bicubic interpolation filter. Secondly, the view difference image between the observed FR left view image and interpolated right view image is defined and employed as the input to the modified CNN. Hence, the high-precision view difference image is provided as the output. Furthermore, the estimated right view image is obtained by combining the observed FR left view image with the high-precision view difference image. The second stage is the global reconstruction constraint process for incorporating the right view. By making the estimated right view image obtained in the first stage align with the LR right view according to the MR stereoscopic image observation model, we model the global reconstruction constraint by using the iterative back projection method. Finally, after the above two stages, the SR of the LR right view is obtained.

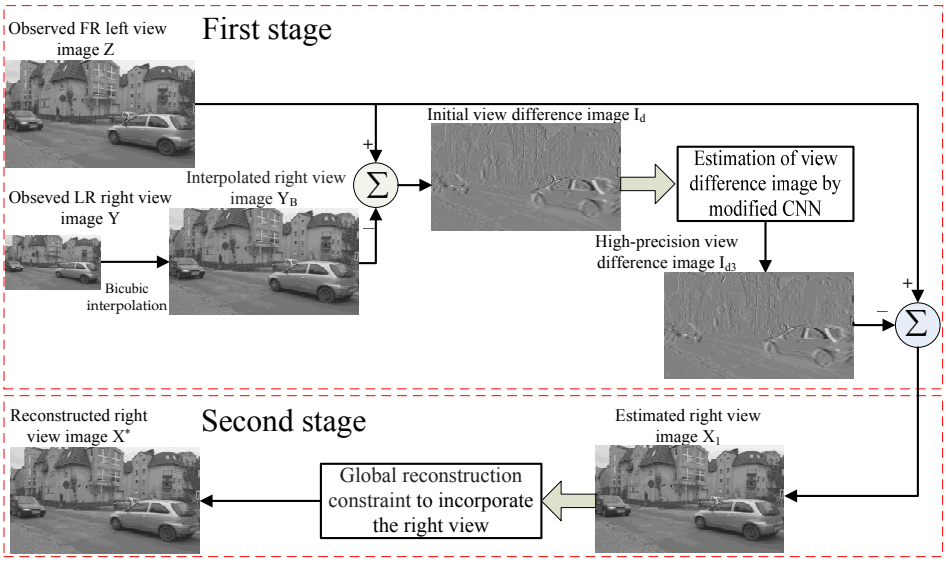


Figure 3. Overall procedure of the proposed method.

3.1. Estimation of View Difference Image with Modified CNN

The key aspect of the SR process of MR stereoscopic images is to use the correlation between views as far as possible to enhance the resolution of the LR right view. As mentioned above, the view difference image is very important for representing the correlation between views because both the image texture information and depth information of the stereoscopic pairs [25] are included in the view difference image. Therefore, we establish a modified CNN, the pooling layers and fully connected layers of which are removed according to the SR task to construct the high-precision view difference image, as shown in Figure 4. In addition to the input layer, the modified CNN training framework consists of three layers, in which the hidden layers [26] are the first two convolution layers, and the output layer is the third convolution layer. Given an initial view difference sub-image obtained by a FR left view training sub-image and an interpolated right view training sub-image, the first convolution layer of the modified CNN extracts a number of feature maps. Then the second convolution layer maps these feature maps to high-precision feature vectors. Finally, the third convolution layer produces the high-precision view difference sub-image according to these high-precision feature vectors.

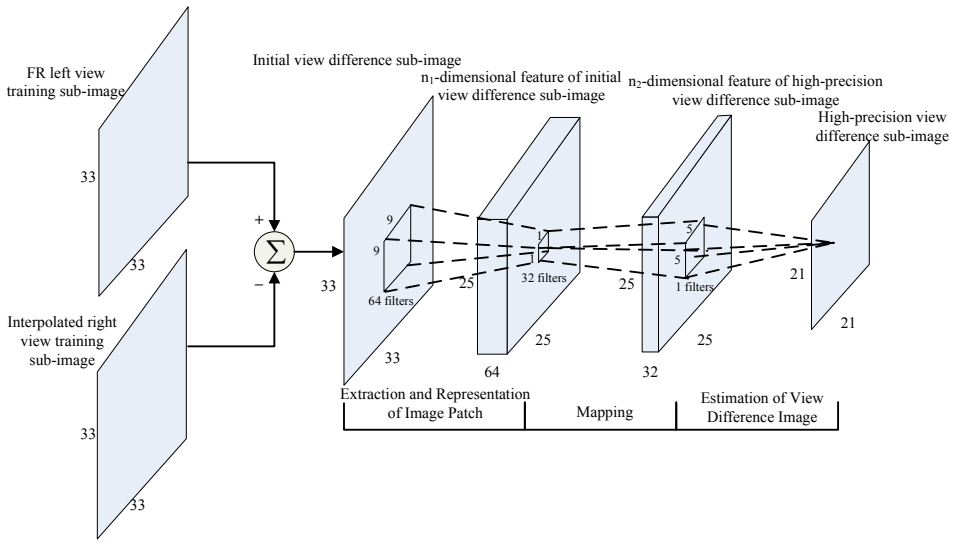


Figure 4. Modified convolutional neural networks (CNN) training framework.

3.1.1. Image Patch Extraction and Representation

The image patch extraction and representation operation extracts image patches from the view difference image and represents them as high-dimensional vectors by a number of bases. In our formulation, this is the same process as convolving the view difference image by a number of filters. These vectors obtained by convolution contain a number of feature maps that can be further mapped to finer feature vectors.

Generally, the view difference image is defined by:

$$I_d(x,y) = Z(x,y) - Y_B(x + d_x(x,y), y + d_y(x,y)) \tag{2}$$

where $\boldsymbol{Z}$ and $\boldsymbol{Y_B}$ denote the observed FR left view image and the interpolated right view image, which is the FR; that is, respectively, $\boldsymbol{Z} = \{Z(x, y)\}$ and $\boldsymbol{Y}_B = \{Y_B(x, y)\}$. In addition, $d_x(x, y)$ and $d_y(x, y)$ are the horizontal and vertical components of the disparity at position (x, y), and $\boldsymbol{I}_d = \{I_d(x, y)\}$ is the defined view difference image on the basis of the disparity information. Actually, for SR of the MR stereoscopic image, we cannot obtain the original FR stereoscopic image pair at the decoder side. Thus, the disparity map cannot be accurately identified. Similar to [25], the view difference image is directly calculated from the stereoscopic image pairs as:

$$I_d(x,y) = Z(x,y) - Y_B(x,y) \tag{3}$$

We take the initial view difference image, $\boldsymbol{I}_d$, as the input of the modified CNN, and the convolutional operation in the first layer in the CNN is represented as:

$$I_{d1} = W_1 * I_d + B_1 \tag{4}$$

where * denotes the convolutional operation. Then, we apply the rectified linear unit (RELU) [27] to alleviate the overfitting problem after the convolutional operation:

$$I_{d1} = \max(0, W_1 * I_d + B_1) \tag{5}$$

where W_1 and B_1 represent n_1 filters of the support $f_1 \times f_1 \times c_1$ and biases, which comprise an n_1-dimensional vector, respectively. Here, f_1 represents the filter size, and c_1 denotes the number of channels in the input image. Additionally, output $\boldsymbol{I}_{d1}$ is composed of n_1 feature maps of the input image.

3.1.2. Mapping and Estimation of View Difference Image and Right View Image

After extracting an n_1-dimensional feature map for each image patch in the first CNN layer, these n_1-dimensional vectors are mapped into n_2-dimensional vectors in the second CNN layer. The high-precision view difference image is estimated in the third CNN layer. Finally, we estimate the right view image after the three layer operation.

Similar to the first layer, the second layer is built through the convolution and the RELU, as follows:

$$I_{d2} = \max(0, W_2 * I_{d1} + B_2) \tag{6}$$

where W_2 and B_2 represent n_2 filters of the support $f_2 \times f_2 \times c_2$ and the biases, which comprise an n_2-dimensional vector, respectively. Output $\boldsymbol{I}_{d2}$ is composed of n_2 feature maps, which can be conceptually used as representations of high-precision view difference image patches that will constitute a full high precision view difference image.

According to the representations of high-precision view difference image patches in the second layer, the convolutional operation in the third CNN layer is defined to produce the high-precision view difference image, $\boldsymbol{I}_{d3}$, as:

$$I_{d3} = W_3 * I_{d2} + B_3 \tag{7}$$

where W_3 and B_3 represent n_3 filters of the support $f_3 \times f_3 \times c_3$ and the biases, which comprise an n_3-dimensional vector.

After the three layer operation, the right view image is estimated as:

$$X_1 = Z - I_{d3} \tag{8}$$

where $\boldsymbol{X}_1$ is the estimated right view image which is the output of the first stage.

3.1.3. Modified CNN and Training

In this paper, the modified CNN is created to estimate the high-precision view difference image. Our objective is to train network f with three layers when given a training dataset, $\{Z^{(i)}, X^{(i)}, {Y_B}^{(i)}\}_{i=1}^{N}$, so that the high precision view difference image, ${I_{d3}}^{(i)} = f(Z^{(i)} - {Y_B}^{(i)})$, is estimated, where $Z^{(i)}$ and $X^{(i)}$ denote the ground-truth left view image and ground-truth right view image, respectively. Furthermore, ${Y_B}^{(i)}$ is the interpolated right view image, and N is the number of training samples. Then, a labeled ground-truth difference image in the used CNN is defined as ${I_{d-label}}^{(i)} = Z^{(i)} - X^{(i)}$. To produce the high-precision view difference image, we use the sum of the absolute difference (SAD) as the distortion function:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left| f(Z^{(i)} - {Y_B}^{(i)}) - {I_{d-label}}^{(i)} \right| \tag{9}$$

According to Equation (9), the distortion between the estimated high-precision view difference image I_{d3} and ground-truth difference image $I_{d-label}$ is minimized. The parameters $\theta = \{W_1, W_2, W_3, B_1, B_2, B_3\}$ are hence obtained.

Stochastic gradient descent [28] is used to minimize the distortion by standard back-propagation. Then, the parameters' matrices are updated as

$$\Delta_{i+1} = M \cdot \Delta_i + r \cdot \frac{\partial L}{\partial \theta_i^l}, \theta_{i+1}^l = \theta_i^l + \Delta_{i+1} \tag{10}$$

where $l \in \{1, 2, 3\}$ represents the indices of layers, i denotes iterations, M is the momentum parameter with a value 0.9, r is the learning rate, and $\frac{\partial L}{\partial \theta_i^l}$ is the derivative of the network parameters. Initially, all filters are initialized with a random Gaussian distribution with a zero mean and standard deviation of 0.001. Meanwhile, the biases of each layer are initialized by zero. The learning rate is 2.5×10^{-4} for the first two layers and 2.5×10^{-5} for the last layer.

3.2. Global Reconstruction Constraint to Incorporate the Right View

For a given LR right view, $\boldsymbol{Y}$, there may be many HR right views $\boldsymbol{X}$ owing to the extremely ill posed character of SR. We consider that the reconstructed HR right view $\boldsymbol{X}$ should be retained with the LR right view $\boldsymbol{Y}$ in terms of the MR stereoscopic image observation model. However, the estimated right view image $\boldsymbol{X}_1$ obtained in the first stage may not satisfy this condition. Therefore, the global reconstruction constraint is enforced for incorporating the right view by projecting $\boldsymbol{X}_1$ onto the solution space of $\boldsymbol{Y} = \boldsymbol{DBX}$. It is computed as:

$$X^* = \underset{X}{\operatorname{argmin}} \|DBX - Y\|_2^2 + \|X - X_1\|_2^2 \tag{11}$$

Thus, the solution can be computed by iterative back projection. The updated equation is:

$$X_{t+1} = X_t + v\left[B^T D^T (Y - DBX_t) + (X - X_1)\right] \tag{12}$$

where X_t denotes the reconstructed right view image after the $(t-1)$-th iteration and v denotes the step size with the value of one.

We use $\boldsymbol{X}^*$ from the aforementioned optimization as the ultimately reconstructed right view image. On one hand, image $\boldsymbol{X}^*$ is as close as possible to the estimated right view image $\boldsymbol{X}_1$, obtained by estimation of the view difference image in the first stage. On the other hand, it satisfies the global reconstruction constraint in the second stage.

4. Experimental Results and Discussion

The training and testing data used in our experiments are described in this section. We explore our modified CNN and provide the convergence analysis to ensure the method is efficient. Then, the SR results of the proposed method are compared with those of recent state-of-the art methods. Next, the running time of all these methods is compared to evaluate the computational complexity. At last, subjective evaluation is adopted to analyze the perceived quality of stereoscopic images.

4.1. Training and Testing Data

Considering that the appropriate amount of training data can increase the CNN performance [29], we used 20 stereoscopic images from the Middlebury Stereo Dataset [30] as the training dataset. Each image pair was comprised of two views. Figure 5 shows the right view of each pair. To provide sufficient information for training the modified CNN, while reducing the complexity of the modified CNN, we extracted the view difference image between the 20 pairs of the FR left view image and the interpolated right view image, and we randomly cropped sub-images with the size of 33×33 from the training images. A total of 552,729 sub-images (that is N, the number of training samples) for training were generated. To avoid the boundary effect in the training process, all the convolutional layers have no padding. Moreover, although we used a changeless image size during training, the modified CNN could be employed on images of variable sizes in the testing process.

Figure 5. Right view of each stereoscopic pair in the training dataset (from the Middlebury Stereo Dataset [30]).

To test the performance and robustness of the modified CNN, we employed the first frame of various multi-view video sequences, including Champagne_tower, Dog, Pantomime, Newspaper, Pozan Street [31], Ballet, and Breakdancers [32], as well as the remaining image pairs in the Middlebury

Stereo Dataset; Sword 2, Umbrella, and Vintage. Figure 6 shows the right view of each pair in the testing dataset. For Champagne_tower, the FR left view is view 38 and view 39 is selected as the LR right view. For Dog and Pantomime, the FR left view is view 40 and view 39 is selected as the LR right view. For Newspaper, the FR left view is view 2 and view 3 is selected as the LR right view. For Poznan Street, the FR left view is view 5 and view 4 is selected as the LR right view. For Ballet and Breakdancers, the FR left view is view 2 and view 1 is selected as the LR right view.

Figure 6. Right view of each stereoscopic pair in the testing dataset.

In this paper, the specific CNN parameters are established according to practical experience, as shown in Figure 4. The MATLAB toolbox MatConvNet [33] is applied to actualize our CNN. Each parameter of the convolutional filters is set to $[f_l, f_l, c_l, n_l]_{l=1}^{3} = [9,9,1,64]_{l=1}, [1,1,64,32]_{l=2}, [5,5,32,1]_{l=3}$. To strengthen the correlation between image patches, the convolution stride is set to one for all layers. For a three-channel color image, we followed the methods of [12–15] in the experiments. Additionally, our stereoscopic image SR is used to contribute to the MR stereoscopic video coding model, whose video format is YCbCr (That is a color space. Y is the brightness (luma) component of the color space, while Cb and Cr are the blue and red concentration offset components). Thus, the color image is converted to the YCbCr color space, and only the Y component of the image is reconstructed by our SR reconstruction method, while both the Cb and Cr components are upsampled by the bicubic interpolation method [34]. Furthermore, although our network can be easily extended to multi-channel image processing, Dong's experiments performed on the YCbCr color space [15] demonstrated that the Cb and Cr channels scarcely improved the performance. Therefore, the following objective evaluation indices are calculated only in the Y channel in Sections 4.2 and 4.3.

4.2. Convergence Analysis

To verify the efficiency of the proposed method, a convergence analysis of the proposed method was conducted. Figure 7 shows the training performance on the Middleburry Stereo Dataset [30] and the testing of the performance on the image Dog. For evaluating the convergence of the modified CNN, the above-mentioned SAD was computed as a training error and Peak Signal to Noise Ratio (PSNR) was used as a testing error in each epoch. Here, an epoch denotes the training times.

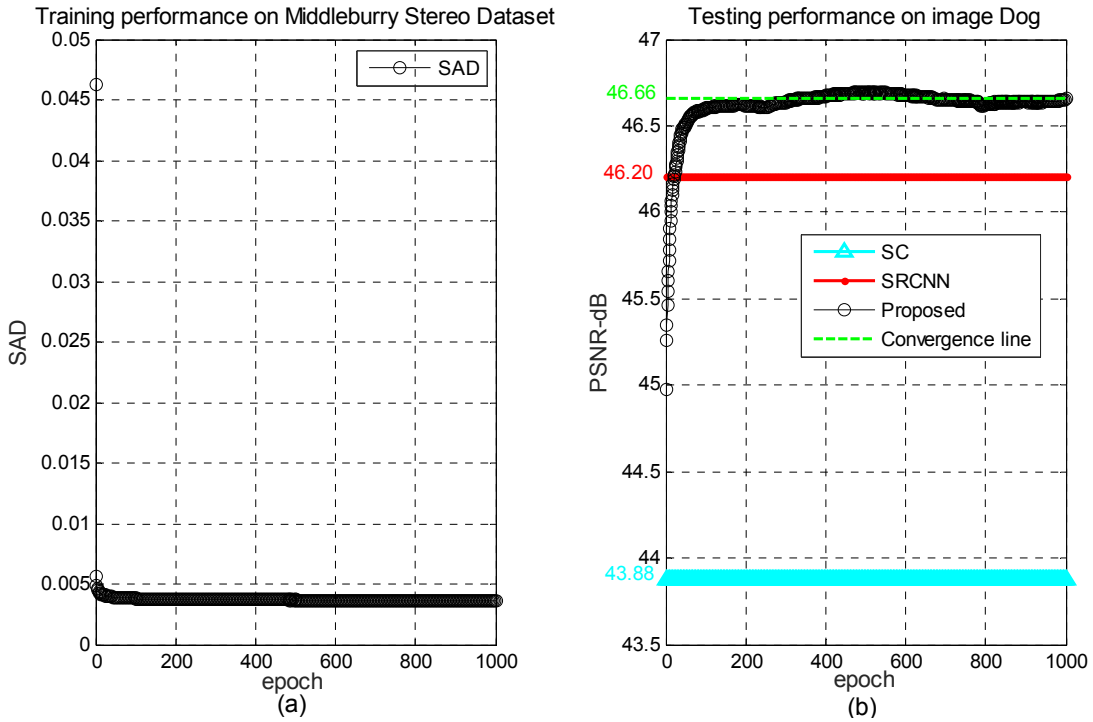


Figure 7. Convergence curve. (**a**) Training performance on Middleburry Stereo Dataset [30]; (**b**) Testing the performance on the image Dog.

It is evident in the figure that, with the increase in the epoch, SAD in training gradually decreases, whereas PSNR in testing progressively increases. Furthermore, the gradual convergence tendency with the increase in the epoch can be predicted. These results show that the reconstruction results are better with more training epochs until the modified CNN reaches stability. Furthermore, the performance of the proposed method surpasses that of the sparse coding (SC) method [13] baseline with a few training epochs, and it outperforms the SRCNN [15] with proper training epochs. Finally, it converges to the PSNR value of 46.66 dB on the image Dog.

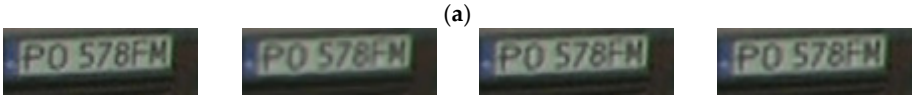
4.3. Stereoscopic Image SR Results

4.3.1. Contrast with Single-View Methods

To demonstrate that the proposed method is more effective than single-view methods, comparative experiments were carried out between the SC method [13], SRCNN method [15], and proposed method. In addition to the common PSNR index, we used two other objective evaluation indices, namely the structural similarity index (SSIM) [35] and the blind/referenceless image spatial quality evaluator (BRISQUE) [36]. Since the PSNR and SSIM indices are full-reference image quality assessments and the BRISQUE index is a no-reference image quality assessment, we considered PSNR and SSIM together for facilitating the analysis. Figure 8 depicts the reconstruction results of these three methods for the first frame of the Pozan Street sequence. It is seen that, according to the details from the local amplification region, SC [13] produced a relatively vague reconstruction. Furthermore, as shown by the edge of the letter P and the number 0, the proposed method was more effective in ringing reduction compared with SRCNN [15]. Figure 9 shows the reconstruction results of the three methods for the Umbrella image. From the magnified details of the umbrella stick and umbrella skeleton, the proposed method produced sharper edges that more closely approximate the real HR images. Figure 10 additionally shows that the details of the pantomimist clothes in the first frame of the Pantomime sequence obtained by the proposed method are clearer. Table 1 presents the PSNR and SSIM values in the Y channel obtained by SC [13], SRCNN [15], and the proposed method. As shown

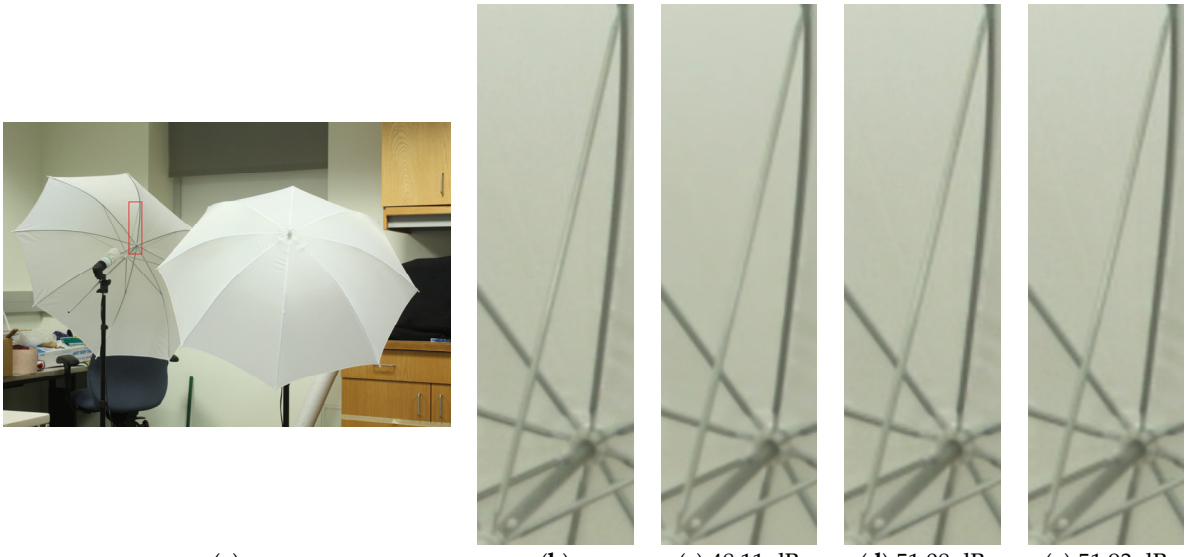
in the table, compared with SC [13] and SRCNN [15], the average PSNR value of the proposed method is increased by 2.39 dB and 0.54 dB, respectively, and the average SSIM value of the proposed method is slightly increased by 0.01 and 0.0002, respectively. Furthermore, the BRISQUE value in Table 2, which typically has a value between 0 and 100 (0 represents the best quality, 100 the worst), demonstrates the significant performance of the proposed method. The average BRISQUE value of the proposed method is decreased by 21.66 and 2.03 compared with SC [13] and SRCNN [15]. Although the CNN architecture of the proposed method is similar to that of SRCNN [15], the proposed method achieves better performance benefiting from the combination of the high-precision view difference image estimated by the modified CNN and the global reconstruction constraint. Overall, according to the tables and the reconstructed images, the proposed method achieves a better SR result than SC [13] and SRCNN [15].

(**a**)



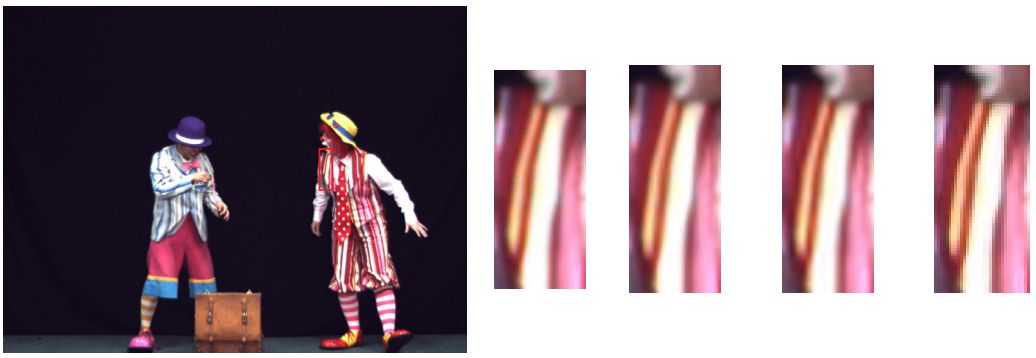
(**b**) (**c**) 40.77 dB (**d**) 42.49 dB (**e**) 42.69 dB

Figure 8. Super-resolution (SR) results and Peak Signal to Noise Ratio (PSNR) values in the Y channel for the first frame of the Pozan Street sequence. (**a**) Ground truth; (**b**) local amplification region of the ground truth; (**c**) local amplification region of sparse coding (SC) [13]; (**d**) local amplification region of SRCNN [15]; and (**e**) local amplification region of the proposed method.



(**a**) (**b**) (**c**) 48.11 dB (**d**) 51.08 dB (**e**) 51.93 dB

Figure 9. SR results and PSNR values in the Y channel for the Umbrella image. (**a**) Ground truth; (**b**) local amplification region of the ground truth; (**c**) local amplification region of SC [13]; (**d**) local amplification region of SRCNN [15]; and (**e**) local amplification region of the proposed method.



(**a**) (**b**) (**c**) 45.71 dB (**d**) 46.41 dB (**e**) 47.91 dB

Figure 10. SR results and PSNR values in the Y channel for the first frame of the Pantomime sequence. (**a**) Ground truth; (**b**) local amplification region of the ground truth; (**c**) local amplification region of SC [13]; (**d**) local amplification region of SRCNN [15]; and (**e**) local amplification region of the proposed method.

Table 1. Comparison with SC [13] and SRCNN [15] for PSNR (dB)/ structural similarity index (SSIM).

Images	Resolution	Scale	SC [13]	SRCNN [15]	Proposed
Sword2	2856 × 2000	2	46.76/0.9857	51.07/0.9951	51.35/0.9951
Umbrella	2960 × 2016	2	48.11/0.9897	51.08/0.9954	51.93/0.9955
Vintage	2912 × 1924	2	40.84/0.9855	42.11/0.9903	42.23/0.9894
Champagne_tower	1280 × 960	2	43.76/0.9856	44.83/0.9917	46.42/0.9921
Dog	1280 × 960	2	43.88/0.9799	46.20/0.9901	46.66/0.9903
Pantomime	1280 × 960	2	45.71/0.9906	46.41/0.9942	47.91/0.9948
Newspaper	1024 × 768	2	41.21/0.9720	42.60/0.9834	42.54/0.9832
Pozan Street	1920 × 1088	2	40.77/0.9600	42.49/0.9787	42.69/0.9788
Ballet	1024 × 768	2	41.01/0.9590	42.40/0.9699	42.82/0.9714
Breakdancers	1024 × 768	2	41.08/0.9402	42.44/0.9573	42.42/0.9573
Average	-	2	43.31/0.9748	45.16/0.9846	45.70/0.9848

Table 2. Comparison with SC [13] and SRCNN [15] for blind/referenceless image spatial quality evaluator (BRISQUE).

Images	Resolution	Scale	SC [13]	SRCNN [15]	Proposed
Sword2	2856 × 2000	2	67.94	48.05	46.34
Umbrella	2960 × 2016	2	77.60	41.75	40.20
Vintage	2912 × 1924	2	62.65	42.70	37.65
Champagne_tower	1280 × 960	2	62.23	39.89	36.81
Dog	1280 × 960	2	42.29	40.31	38.39
Pantomime	1280 × 960	2	79.83	40.77	40.39
Newspaper	1024 × 768	2	46.70	36.86	34.21
Pozan Street	1920 × 1088	2	47.65	41.25	38.79
Ballet	1024 × 768	2	63.22	41.77	39.26
Breakdancers	1024 × 768	2	52.09	32.56	30.53
Average	-	2	60.22	40.59	38.56

4.3.2. Contrast with Depth-Based Methods

To show the advantages of the proposed method, which fully leverages the correspondence between views without estimating the depth map, comparative experiments were carried out between the depth-based method presented by Garcia and the proposed method. As mentioned in Garcia's work [22], for fully considering the high-frequency information of original images, the test sequences, including Dog, Pantomime, Pozan Street, Ballet, and Breakdancers, were resized to 640 × 480,

640 × 480, 960 × 544, 512 × 384, and 256 × 192, respectively, by using a six-tap Lanczos interpolation filter. To ensure consistency, we employed these downsampled images for testing the original images under the same conditions.

As depicted in Table 3, the average PSNR value of the proposed method increased by 1.14 dB compared with Garcia's method [22]. Two significant characteristics of the proposed method contributed to the improvement; (1) the estimation of the high-precision view difference image between views owing to modified CNN in the first stage and (2) the gradual improvement of the reconstructed right view by enforcing the global reconstruction constraint to incorporate the right view.

Table 3. Comparison with the depth-based Garcia [22] for PSNR (dB). The fourth column data from Garcia [22]

Images	Resolution	Scale	Garcia [22]	Proposed
Dog	640 × 480	2	36.54	38.33
Pantomime	640 × 480	2	38.59	39.63
Pozan Street	960 × 544	2	35.78	35.90
Ballet	512 × 384	2	36.34	36.83
Breakdancers	256 × 192	2	39.09	41.38
Average	-	2	37.27	38.41

4.4. Running Time

To evaluate the computational complexity, the running times of all methods were compared for the ten testing stereoscopic images listed in Table 1, as shown in Table 4. All results were acquired from the corresponding authors' MATLAB code, and ours were likewise obtained on MATLAB software. All the algorithms were run on the same machine with an Intel 2.30-GHz CPU and 16 GB of RAM. It was apparent that the proposed method had a faster processing speed than the SC method [13]. This is because the proposed method does not need to solve a complex optimization problem as the SC method does [13]; thus, the running time of the proposed method was less than that of the SC method. Moreover, the proposed method's speed was close to that of SRCNN [15]. As soon as the training of our modified CNN was complete, the SR results were quickly obtained by the proposed feed-forward method.

Table 4. Running time of the ten testing stereoscopic images listed in Table 1 (unit: s).

Methods	SC [13]	SRCNN [15]	Proposed
Average	2064.80	95.05	97.54

4.5. Subjective Evaluation

4.5.1. Generation of Testing Stereoscopic Images

For evaluating the quality of the stereoscopic images generated by different SR methods, a subjective experiment was implemented, and the procedure followed the International Telecommunications Union-Radio Communications Sector (ITU-R) Recommendation BT.500 [37] so that subjective quality of the reconstructed stereoscopic images relative to the original stereoscopic images was obtained. Specifically, this experiment also used the ten aforementioned testing stereoscopic images and adopted a Double-Stimulus Continuous Quality-Scale method [37], which is equivalent to simultaneously scoring two stimuli corresponding to the reconstructed stereoscopic image and the original stereoscopic image. Thus, for ten testing stereoscopic images and three different SR methods, there are a total of 10 × 3 = 30 comparison clips produced.

4.5.2. Experimental Environment and Participants

In the experiment, the stereoscopic projection system was adopted, and the participants needed to wear polarized glasses, which separate the left and right views to the appropriate eyes. The system consisted of two projectors (BenQ PB8250 DLP), DELL real-time 3D graphics workstations, a polarized light bracket, and a metal screen (150 inches). The experiment was conducted in a specific laboratory, in which the illumination, temperature, and other experimental conditions followed ITU-R Recommendation BT.500.

A total of 20 participants were involved in the study, with an average age of 23 years, and all of the participants underwent the color vision test and met a 20:30 visual acuity test and a stereoscopic visual acuity test at 40 s-arc. They were non-experts whose professional backgrounds are not directly related to image quality. Each participant, needed to score 30 pairs of stereoscopic images. Each pair cost 40 s on display, 10 s on scoring, and 10 s on resting. According to ITU Recommendation BT.500, the display order of the 30 pairs was random. The distance between the participants and the screen was 3 m; that is, three times the height of the screen. In addition, to make the participants be familiar with the scoring process, four other pairs of stereoscopic images were displayed to them before the official scoring.

4.5.3. Ranking and Raw Data Processing

After the stereoscopic images were scored by the participants, the Difference Mean Opinion Scores (DMOS) between each pair of stereoscopic images, which include the original stereoscopic image and the reconstructed stereoscopic image, were calculated. According to ITU-R BT.500, the value range of DMOS is from 0 to 100. The formula is as follows:

$$DMOS_j = \frac{1}{N_j}\sum_{i=1}^{N_j} d_{ij} \tag{13}$$

$$d_{ij} = r_{i\,ref(j)} - r_{ij} \tag{14}$$

where r_{ij} denotes the raw quality score of the j-th reconstructed stereoscopic image evaluated by the i-th participant, $r_{iref(j)}$ is the raw quality score assigned by the i-th participant to the j-th original stereoscopic image, and N_j denotes the number of participants involved in assessment of the j-th stereoscopic image. Then $DMOS_j$ is the mean of the raw difference scores d_{ij}.

Before calculating the final DMOS of each of the reconstructed stereoscopic images, the data of the participants with poor score stability should be removed. Specially, if the value of d_{ij} is beyond the 95% confidence interval of $DMOS_j$, then d_{ij} is called an outlier.

4.5.4. Results and Analysis

Figure 11 shows the DMOS of the ten testing stereoscopic images reconstructed with the three different SR methods. It is obvious that the SC method [13] was poorer than the other two SR methods. Most of the participants thought that the stereoscopic images reconstructed by the SC method were more obscure. Benefiting from the correlation between views, the proposed method also outperformed the SRCNN [15], demonstrating its better stereo visual quality.

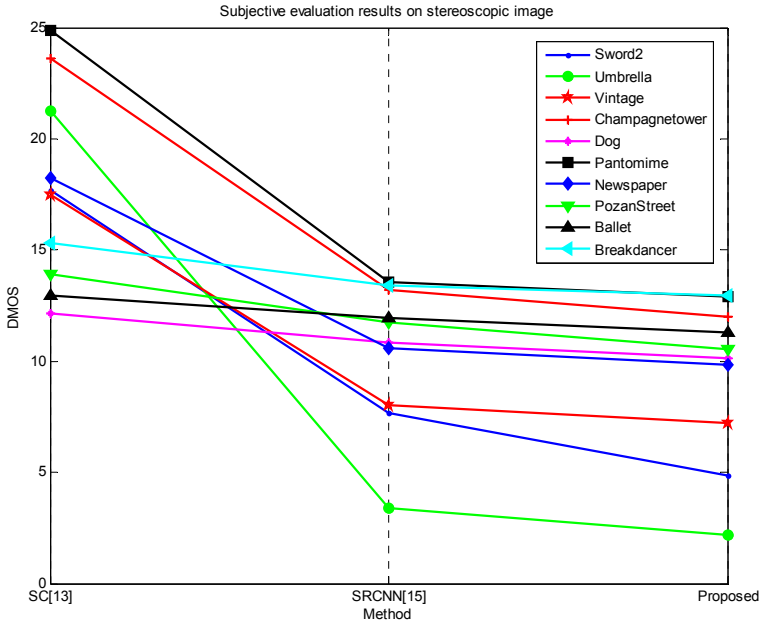


Figure 11. Difference Mean Opinion Scores (DMOS) of ten testing stereoscopic images reconstructed with three different SR methods.

5. Conclusions

To ensure the perceived quality of stereo vision, we proposed in this paper a method using CNNs and incorporating views to reconstruct a FR stereoscopic image. Compared with single-view and depth-based methods, the proposed method combines the correlation between left and right views without estimating the depth image. Firstly, a deep learning tool is used to estimate the view difference image comprising the image texture information and the depth information of the stereoscopic pairs. Then, the estimated right view image is projected onto the solution space of the MR stereoscopic image observation model. Finally, the HR reconstructed right view image is obtained. The experimental results indicated that the performance of the proposed method is superior to those of the existing methods in terms of both reconstruction effect and speed. In the future, we will focus on accelerating the CNN convergence and further consider exploiting the temporal correlation of video to research MR stereoscopic video SR.

Acknowledgments: This work was supported by the Natural Science Foundation of China under Grant Nos. U1301257, 61671258, and 61620106012, the National High-tech R&D Program of China under Grant No. 2015AA015901, and the Natural Science Foundation of Zhejiang Province under Grant Nos. LY15F010005 and LY16F010002. It was also sponsored by the K.C. Wong Magna Fund of Ningbo University.

Author Contributions: Z.P. and G.J. designed the algorithm and wrote the source code. They together wrote the manuscript. H.J. and M.Y. provided suggestions on the algorithm and revised the entire manuscript. F.C. and Q.Z. provided suggestions on the experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, J.; Chen, C.; Liu, Y.; Chen, X. Small-world brain functional network altered by watching 2D/3DTV. *J. Vis. Commun. Image Represent.* **2016**, *38*, 433–439. [CrossRef]
2. Domański, M.; Bartkowiak, M.; Dziembowski, A.; Grajek, T.; Grzelka, A.; Łuczak, A.; Mieloch, D.; Samelak, J.; Stankiewicz, O.; Stankowski, J.; et al. New results in free-viewpoint television systems for horizontal virtual navigation. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.

3. Nikolic, S.; Lee, M.J.W. Special session: Exploring learning opportunities in engineering education using 2D, 3D and immersive video augmented online technologies. In Proceedings of the IEEE Frontiers in Education Conference (FIE), Erie, PA, USA, 12–15 October 2016; pp. 1–2.
4. Passig, D.; Tzuriel, D.; Eshel-Kedmi, G. Improving children's cognitive modifiability by dynamic assessment in 3D Immersive Virtual Reality environments. *Comput. Educ.* **2016**, *95*, 296–308. [CrossRef]
5. Julesz, B. *Foundations of Cyclopean Perception*; The University of Chicago Press: Oxford, UK, 1971; p. 406.
6. Chung, K.-L.; Huang, Y.-H. Efficient multiple-example based super-resolution for symmetric mixed resolution stereoscopic video coding. *J. Vis. Commun. Image Represent.* **2016**, *39*, 65–81. [CrossRef]
7. Zhang, X.; Wu, X. Image Interpolation by Adaptive 2-D Autoregressive Modeling and Soft-Decision Estimation. *IEEE Trans. Image Process.* **2008**, *17*, 887–896. [CrossRef] [PubMed]
8. Zhu, S.; Zeng, B.; Zeng, L.; Gabbouj, M. Image Interpolation Based on Non-local Geometric Similarities and Directional Gradients. *IEEE Trans. Multimed.* **2016**, *18*, 1707–1719. [CrossRef]
9. Kim, K.I.; Kwon, Y. Single-Image Super-Resolution Using Sparse Regression and Natural Image Prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1127–1133. [PubMed]
10. Mourabit, I.E.; Rhabi, M.E.; Hakim, A.; Laghrib, A.; Moreau, E. A new denoising model for multi-frame super-resolution image reconstruction. *Signal Process.* **2017**, *132*, 51–65. [CrossRef]
11. Hong, C.; Dit-Yan, Y.; Yimin, X. Super-resolution through neighbor embedding. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004.
12. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image super-resolution as sparse representation of raw image patches. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
13. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image Super-Resolution Via Sparse Representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [CrossRef] [PubMed]
14. Timofte, R.; De Smet, V.; Van Gool, L. A+: Adjusted Anchored Neighborhood Regression for Fast Super-Resolution. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 111–126.
15. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef] [PubMed]
16. Yang, C.Y.; Ma, C.; Yang, M.H. Single-Image Super-Resolution: A Benchmark. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 372–386.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–10 December 2015; pp. 91–99.
19. Agostinelli, F.; Anderson, M.R.; Lee, H. Adaptive multi-column deep neural networks with application to robust image denoising. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 1493–1501.
20. Cui, Z.; Chang, H.; Shan, S.; Zhong, B.; Chen, X. Deep Network Cascade for Image Super-resolution. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 49–64.
21. Liu, D.; Wang, Z.; Wen, B.; Yang, J.; Han, W.; Huang, T.S. Robust Single Image Super-Resolution via Deep Networks With Sparse Prior. *IEEE Trans. Image Process.* **2016**, *25*, 3194–3207. [CrossRef] [PubMed]
22. Garcia, D.C.; Dorea, C.; de Queiroz, R.L. Super Resolution for Multiview Images Using Depth Information. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1249–1256. [CrossRef]
23. Brust, H.; Tech, G.; Mller, K. *Report on Generation of Mixed Spatial Resolution Stereo Data Base*; Technical Report Project No. 216503; MOBILE 3DTV: Tampere, Finland, 2009.
24. Irani, M.; Peleg, S. Improving resolution by image registration. *CVGIP Graph. Models Image Process.* **1991**, *53*, 231–239. [CrossRef]
25. Ma, L.; Wang, X.; Liu, Q.; Ngan, K.N. Reorganized DCT-based image representation for reduced reference stereoscopic image quality assessment. *Neurocomputing* **2016**, *215*, 21–31. [CrossRef]

26. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
27. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
28. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
29. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in neural information processing systems, Lake Tahoe, Nevada, USA, 3–6 December 2012; pp. 1097–1105.
30. Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; Westling, P. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In Proceedings of the 36th German Conference on Pattern Recognition, Münster, Germany, 2–5 September 2014; pp. 31–42.
31. Mobile 3DTV. Available online: http://sp.cs.tut.fi/mobile3dtv/stereo-video/ (accessed on 4 February 2017).
32. Zitnick, C.L.; Kang, S.B.; Uyttendaele, M.; Winder, S.; Szeliski, R. High-quality video view interpolation using a layered representation. In Proceedings of the 31st international conference on computer graphics and interactive techniques, Los Angeles, California, USA, 8–12 August 2004; pp. 600–608.
33. Vedaldi, A.; Lenc, K. Matconvnet: Convolutional neural networks for matlab. In Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 689–692.
34. Bicubic Interpolation. Available online: https://en.wikipedia.org/wiki/Bicubic_interpolation (accessed on 11 February 2017).
35. Zhou, W.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Process.* **2004**, *13*, 600–612.
36. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [CrossRef] [PubMed]
37. Methodology for the Subjective Assessment of the Quality of Television Pictures. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-11-200206-S!!PDF-E.pdf (accessed on 4 February 2017).