Enhancing the Spatial Resolution of Stereo Images using a Parallax Prior

Daniel S. Jeon Seung-Hwan Baek Inchang Choi Min H. Kim*
Korea Advanced Institute of Science and Technology (KAIST)
{sjjeon,shwbaek,inchangchoi,minhkim}@vclab.kaist.ac.kr

Abstract

We present a novel method that can enhance the spatial resolution of stereo images using a parallax prior. While traditional stereo imaging has focused on estimating depth from stereo images, our method utilizes stereo images to enhance spatial resolution instead of estimating disparity. The critical challenge for enhancing spatial resolution from stereo images: how to register corresponding pixels with subpixel accuracy. Since disparity in traditional stereo imaging is calculated per pixel, it is directly inappropriate for enhancing spatial resolution. We, therefore, learn a parallax prior from stereo image datasets by jointly training two-stage networks. The first network learns how to enhance the spatial resolution of stereo images in luminance, and the second network learns how to reconstruct a high-resolution color image from high-resolution luminance and chrominance of the input image. Our two-stage joint network enhances the spatial resolution of stereo images significantly more than single-image super-resolution methods. The proposed method is directly applicable to any stereo depth imaging methods, enabling us to enhance the spatial resolution of stereo images.

1. Introduction

With recent advances in mobile phones, a dual camera is more commonly used to estimate depth information, allowing for 3D imaging and augmented reality applications. While traditional stereo imaging has focused on depth estimation, other applications from stereo have rarely been discussed. In this work, we present a novel method that allows us to enhance the spatial resolution of stereo images. To enhance spatial resolution, multiple sampling with subpixel offsets is necessary [42]. Since we have two images in a stereo pair, the disparity exists between these images and is much larger than a pixel. While the disparity in traditional stereo imaging allows us to estimate depth, per-pixel registration using disparity is insufficient to enhance the spatial resolution of stereo images. We, therefore, propose a method that learns a subpixel parallax prior to enhance the spatial resolution of stereo images.

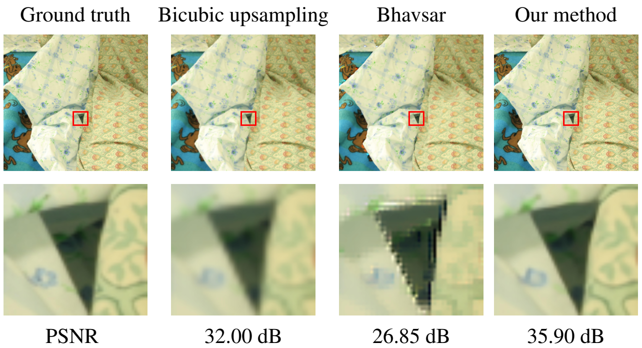

Figure 1: Compared to a naïve approach of bicubic upsampling, Bhavsar et al. [5] enhance the spatial resolution but suffer from jaggy aliasing artifacts. The proposed method can enhance the spatial resolution significantly by taking advantage of stereo input.

Traditional approaches to enhance spatial resolution utilize multiple source images of low resolution with jittered subpixel offsets. Multiple shots or sequentially-moving video frames have been used as input [42]. Recent advances in enhancing spatial resolution have recovered a high-resolution image from a single source image itself by finding similar small patches [20, 52, 24, 53, 36, 43], joint learning [15, 8], or convolutional neural networks [13, 53, 9]. These data-driven approaches are a great success indeed in enhancing the resolution of a single image. However, no work looks into stereo images as input for enhancing spatial resolution, since stereo images are different due to parallax. In this work, we propose a novel method that takes stereo images as input to enhance spatial resolution.

A simple solution for enhancing the spatial resolution of stereo images is to utilize disparity obtained from a depth-from-stereo algorithm to register pixels in a pair of stereo images [4, 5, 41]. However, the disparity accuracy obtained from stereo imaging is significantly lower than the subpixel precision required for super-resolution reconstruction [42], in addition to correspondence errors of stereo matching. Therefore, the reconstructed resolution of prior stereo en-

[*] Corresponding author's email: minhkim@kaist.ac.kr

hancements has been significantly lower than that of single image-based approaches. See Figure 1 for an example.

We, therefore, endeavor to build a deep convolutional network that directly learns an end-to-end mapping between continuous parallax shifts and a high-resolution image. Our deep network includes two subnetworks, which are trained jointly in the end-to-end training manner. Since we decompose an input pair of stereo to left/right pairs of luminance and chrominance, the first network directly learns a luminance mapping from a pair of stereo luminance images to a high-resolution luminance image. The second network learns chrominance mapping from both original chrominance and high-resolution luminance images to a final high-resolution color image. The proposed method outperforms both state-of-the-art image-enhance methods and presents effectiveness in particular for removing aliasing artifacts on slanted edges.

2. Related Work

Super-resolution imaging has been researched extensively in recent decades. For the sake of brevity, we refer readers to [42] for the foundation of this subject. This section reviews only state-of-the-art methods.

Multi-Frame Super-Resolution The classical super-resolution approach is to reconstruct a high-resolution image from multiple image inputs from a single camera jittered with subpixel offsets [42]. However, since multiple images from the same viewpoint cannot be captured simultaneously and also precise registration of fast moving objects is challenging, the applicability of multi-frame methods is restricted to static scenes [10, 11, 45, 3].

Single-Image Super-Resolution To overcome the drawbacks of the multi-frame approach, a single image-based approach has been studied more extensively like other example-based vision applications [35] in the past decade, aiming at recovering a high-resolution image from a single image input. However, it is fundamentally an ill-posed problem that has multiple solutions. Recent research of single-image approaches can be categorized into two groups: example-based and deep network-based methods.

Example-based methods exploit similarities of small patches within an image [24] or learn dictionary priors from external datasets of pairs of low- and high-resolution images [13, 53, 50, 51, 24]. For instance, Yang et al. [53] proposed a sparse representation-based method that learns a joint dictionary from pairs of low- and high-resolution training datasets. Huang et al. [24] exploit a statistical natural image prior by searching similar patches explicitly from localized planes in the source image.

Recently, the example-based approach has been extended using deep convolutional neural networks (CNN) [9, 27, 28, 48]. Dong et al. [9] extend an example-based method [53] to a super-resolution convolutional neural network (SRCNN), where a layer consists of patch extraction, non-linear mapping, and reconstruction, analogous to super-resolution sparse coding [53]. Kim et al. [27, 28] enhance the performance of SRCNN by applying a very-deep network architecture [49] and residual learning [23]. Shi et al. [48] proposed an efficient method using subpixel CNN to extract features from a low-resolution space. Note that single-image super-resolution methods do not rely on subpixel registration, but infer subpixel similarities using convolutional networks.

Light-Field Image Enhancement Since light-field imaging extends the angular resolution of light-field by sacrificing the spatial resolution, light-field super-resolution has been proposed for enhancing the reduced spatial resolution. Bishop et al. [6] estimate a point spread function to defocus light-field images. Georgiev et al. [18] reconstruct a high-resolution light field image directly from a Bayer-patterned input image by avoiding interpolation. Yoon et al. [54] train end-to-end convolutional networks that synthesize both angular and spatial light field images. Kalantari et al. [26] proposed a two-stage network architecture: one for disparity and the other for reconstruction. Flynn et al. [12] proposed a CNN-based view synthesis from plane-sweep panorama input like light-field. These light field-based methods rely on short baseline characteristics in light-field for training these reconstruction networks. However, they are inapplicable directly to an ordinary stereo setup with a larger baseline.

Stereo Image Enhancement Since estimating disparity from the stereo is also an ill-posed problem, the resolution enhancement by stereo images has often been limited by the inaccuracy of disparity. Komatsu et al. [33] preliminarily combine stereo input images on a single depth plane to reconstruct a high-resolution image. However, this method is inapplicable to real 3D scenes as they do not account for parallax. Gao et al. [14] proposed a refractive stereo method that can measure depth and enhance the image resolution. Multiple images need to be captured while moving the glass orientation and reconstruct a high-resolution image similar to multi-frame SR; therefore, it is inapplicable to the snapshot-based approach, which is our objective. Bhavsar and Rajagopalan [4, 5] and Park et al. [41] utilize stereo block matching to search pixel correspondences. The correspondences are then used to register two stereo input images. Garcia et al. [16] and Jain et al. [25] similarly use a stereo pair of low- and high-resolution video frames. They transfer the high-resolution frame to the low-resolution video frame through block matching to obtain a high-resolution image. While subpixel multi-sampling is required to reconstruct a high-resolution image [42], these state-of-the-art stereo methods have attempted to utilize the per-pixel disparity from stereo matching, which is the discrete estimation of parallax.

3. Learning a Parallax Prior

Our objective is to enhance the spatial resolution of stereo images. Since multiple sampling with *subpixel shifts* is necessary as input to enhance image resolution [42], we are motivated to avoid using *discrete disparity* from stereo matching. Instead, we directly learn an *end-to-end* mapping from a stereo pair of low-resolution images to a high-resolution output image. As shown in Figure 2, parallax occurs through *perspective projection* between the stereo images. By the geometric difference between parallax in perspective projection and affinity in stereo image transformation [22, 2, 1, 34], continuous shifts by parallax exist at a subpixel resolution in the horizontal and vertical displacement of corresponding stereo pairs. Our deep network enhances image resolution by directly utilizing these subpixel shifts caused by parallax to reconstruct a high-resolution image without estimating depth or disparity.

3.1. Network Architecture Overview

To make convolutional networks learn stereo correspondences, we create an image stack with shift intervals to feed it to train networks. Shifted pixels are designed to infer correspondence cues through the networks, which interpret them in a nonlinear fashion to yield a high-resolution image. To do so, we devise a two-network architecture, inspired by the traditional architecture of image compression, which reserves two different bandwidths for luminance and chrominance respectively, accounting for human perception [29, 30, 31]. The first network focuses on learning the high-resolution luminance mapping. The second network learns a color transformation mapping from both high-resolution luminance and low-resolution chrominance to a high-resolution color image. See Figure 3 for an overview.

3.2. Network Formulation

Instead of using general color channels, we convert three color channels of red, green and blue (RGB) into YC_bC_r coefficients of luminance Y and chrominance C_bC_r as a 3D tensor of size $H \times W \times C \in \mathbb{R}^3$, where H is the height, W is the width, and C is the number of channels in an input image. We formulate the first network that learns a luminance mapping between a stereo image stack with shift intervals and a high-resolution image. Suppose a training data set $\left\{\mathbf{X}^i, \mathbf{y}^i\right\}_{i=1}^{N}$ is given from N number of low-resolution stereo image pairs $\mathbf{X}^i = \left\{\mathbf{x}_1^i, \mathbf{x}_2^i\right\}$ and ground-truth high-resolution images $\mathbf{y}^i \in \mathbb{R}^3$. $\mathbf{x}_1^i \in \mathbb{R}^3$ denotes the left reference color image and $\mathbf{x}_2^i \in \mathbb{R}^3$ means the right image in stereo. Note that we upsample the resolution of the source low-resolution stereo images by using bicubic interpolation to match the resolution of the high-resolution images. The main objective of our network is to learn a model F that can predict a high-resolution image $\hat{\mathbf{y}} = F(\mathbf{X})$ from given input $\mathbf{X}$. Our model consists of two deep networks: a luminance network and a chrominance network. Note that we jointly train both networks as one to learn an end-to-end mapping.

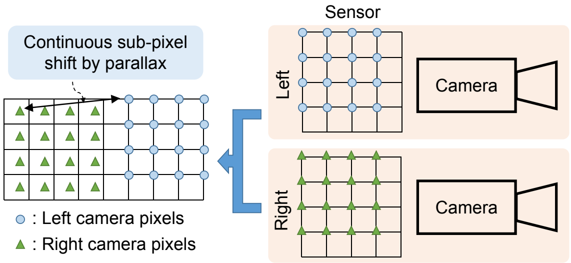

Figure 2: Registering stereo images with parallax. Affine transformation that aligns two images contains *subpixel shifts*, due to the geometric difference between perspective projection and affinity of the stereo planes. We learn a mapping from subpixel shifts to a high-resolution image.

Luminance Network Our luminance network detects similar patches in the input stereo tensor with shift using deep convolutional networks, rather than using traditional block matching that determines the discrete disparity. It means that our luminance network detects patch similarities in stereo channels that contain continuous parallax offsets. The closest patch in the source image can be used directly for reconstruction through networks, enabling subpixel precision multi-sampling for image enhancement. Finding similar patches is more effective way to enhance patch resolution regardless of patch correspondences for the disparity. Note that no disparity map is required for our enhancement of spatial resolution.

Inspired by the modern architecture of very deep networks using residual learning [27], we define a residual image of luminance $\mathbf{r}_L = \mathbf{y}_L - \mathbf{x}_{1,L} \in \mathbb{R}^2$, where $\mathbf{y}_L$ is a high-resolution luminance image and $\mathbf{x}_{1,L}$ is one of the low-resolution luminance stereo images. Since we have stereo input from a dual camera, we utilize stereo images as input to learn residuals to infer subpixel shifts by parallax through the networks. Different from [27], we make use of a stack of stereo input to reconstruct a high-resolution luminance image and design two-network architecture by separating colors into luminance and chrominance.

The first network learns the residuals $\mathbf{r}_L^i$ between a high-resolution luminance $\mathbf{y}_L^i$ and a low-resolution luminance stereo image $\mathbf{x}_{1,L}^i$ over training datasets. To account for subpixel shift by parallax, we repack the left and right images to yield a combined stereo image tensor $\tilde{\mathbf{X}}_L^i \in \mathbb{R}^3$ as follows:

$$\begin{aligned} \tilde{\mathbf{X}}_{L,j}^i(x,y) &= \mathbf{x}_{2,L}^i(x-\phi(j),y) && \text{for } j \in \{1\ldots M\}, \\ \tilde{\mathbf{X}}_{L,j}^i(x,y) &= \mathbf{x}_{1,L}^i(x,y) && \text{for } j = M+1, \end{aligned} \tag{1}$$

where $\phi(j)$ is a shifted offset of the j-th layer, and M is the number of shifts. We then minimize the mean squared errors of $\left\{\frac{1}{2}\|\mathbf{r}_L^i - f(\tilde{\mathbf{X}}_L^i)\|^2\right\}_i$, where f predicts the resid-

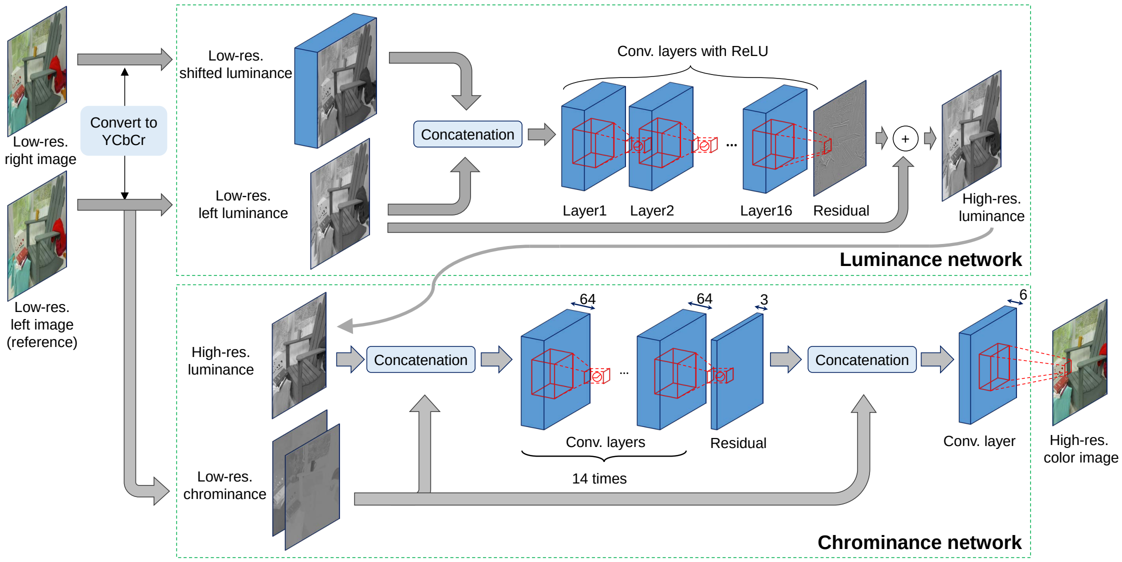

Figure 3: Schematic diagram of our stereo-enhancement network architecture. Our network comprises two subnetworks for enhancing luminance resolution and reconstructing colors, respectively. For the luminance network, we concatenate the left image and 64 of the right images with shifting as input. The number of feature maps for the input layer is 65. The last convolutional layer applies no ReLU for computing residuals following [7, 37]. For the chrominance network, we concatenate the reconstructed high-resolution luminance image and upsampled chrominance components from the low-resolution input. In this network, instead of directly adding a residual prediction to the low-resolution image, we employ the last convolutional layer after concatenating the residuals and the original image. This semi-residual learning approach allows us to increase the accuracy. Finally, we train them jointly for an end-to-end mapping.

uals of the high-resolution luminance $\hat{\mathbf{r}}_L$ with respect to the low-resolution luminance image $\mathbf{x}_{1,L}$. Note that f estimates $\hat{\mathbf{r}}_L \in \mathbb{R}^2$ from given $\tilde{\mathbf{X}}_L \in \mathbb{R}^3$, reducing the input dimension through the networks.

Once we learn the residual prediction model f, we can reconstruct a high-resolution luminance image $\hat{\mathbf{y}}_L \in \mathbb{R}^2$ by summing the left low-resolution luminance input and the predicted residual:

$$\hat{\mathbf{y}}_L = \mathbf{x}_{1,L} + f(\tilde{\mathbf{X}}_L). \tag{2}$$

The function f comprises multiple subsets of convolutional layers f_k. A k-th convolutional layer $f_k\left(\mathbf{Z}\right)$ can be represented as

$$f_k\left(\mathbf{Z}\right) = \max\left(0, \Omega_k * \mathbf{Z} + \beta_k\right), \quad \mathbf{Z} \in \mathbb{R}^3, \tag{3}$$

where $\Omega_k \in \mathbb{R}^4$ and $\beta_k \in \mathbb{R}^3$ represent the convolution filters and biases of the k-th layer, and the operator $*$ denotes convolution. When the depth of the k-th layer is n_k, the size of kernel Ω_k is $n_{k-1}\times3\times3\times n_k$, and the size of bias β_k is n_k. $\max(0,\cdot)$ represents a rectified linear unit (ReLU) [40].

Chrominance Network State-of-the-art super-resolution algorithms [9, 43, 27] reconstruct high-resolution colors by either applying the super-resolution algorithm to three color channels independently or converting the RGB input into the YC_bC_r color space and applying the super-resolution method to the Y luminance only. Different from these simple approaches, we introduce an additional network to enhance the chrominance upsampling stage instead of using bicubic interpolation. Although the luminance component is dominant for color image resolution, upsampling of low-resolution chrominance using bicubic interpolation generates a color bleeding problem over edges. See the middle column image of Figure 4 for an example. We attempt to overcome the problem by introducing the chrominance network. Figure 4 demonstrates that the chrominance upsampling artifacts of color bleeding over edges are effectively removed by the chrominance network.

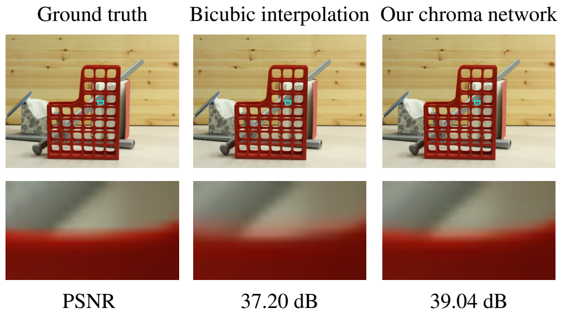

Figure 4: Comparison between the bicubic interpolation and our chrominance network for reconstructing colors. Our method shows clearer color edges.

The input of our chrominance network is the reconstructed high-resolution luminance image $\hat{\mathbf{y}}_L^i$ and the upsampled chrominance from C_b and C_r channels in a low-resolution image in stereo $\mathbf{x}^i_{1,c_b}$ and $\mathbf{x}^i_{1,c_r}$. In order to build a training dataset $\left\{\tilde{\mathbf{x}}^i, \mathbf{y}^i\right\}^N_{i=1}$, we concatenate three input channels into $\tilde{\mathbf{x}}^i$ as

$$\begin{aligned} \tilde{\mathbf{x}}^i_c &= \hat{\mathbf{y}}^i_L \quad \text{for } c = 1 \\ \tilde{\mathbf{x}}^i_c &= \mathbf{x}^i_{1,c} \quad \text{for } c \in \{2, 3\}, \end{aligned} \tag{4}$$

where $\mathbf{x}^i_{1,2}$ and $\mathbf{x}^i_{1,3}$ denote $\mathbf{x}^i_{1,c_b}$ and $\mathbf{x}^i_{1,c_r}$, respectively. The main objective of the chroma network is to reconstruct a final high-resolution color image.

Our chrominance network implicitly learns the residuals between the high-resolution color and the half-way-through low-resolution color image in stereo. To do so, we define a residual image of chrominance $\mathbf{r} = \mathbf{y} - \tilde{\mathbf{x}} \in \mathbb{R}^3$. We then minimize the mean squared errors of $\left\{\frac{1}{2}\|\mathbf{r}^i - g(\tilde{\mathbf{x}}^i)\|^2\right\}_i$, where function g predicts residuals of high-resolution colors from the combined input image $\tilde{\mathbf{x}}^i$ of low-resolution chrominance and high-resolution luminance.

Different from traditional residual learning, we follow a recent semi-residual learning approach with an additional convolutional layer, inspired by Gharbi et al. [19]. Instead of using the simple summation approach in traditional residual learning [23], we combine the predicted residuals $\mathbf{r}$ with the fast-forwarded identity $\tilde{\mathbf{x}}$ in the form of a convolutional layer. To calculate their convolution, we concatenate the predicted residual and the input colors as

$$\begin{aligned} \tilde{\mathbf{y}}_c &= \tilde{\mathbf{x}}_c \qquad \text{for } c \in \{1 \dots 3\} \\ \tilde{\mathbf{y}}_c &= g(\tilde{\mathbf{x}})_{c'} \quad \text{for } c \in \{4 \dots 6\}, \end{aligned} \tag{5}$$

where $c' = c - 3$. The last sub-layer h takes $\tilde{\mathbf{y}}$ as input without the ReLU activation function to reconstruct the final color image $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = h\left(\tilde{\mathbf{y}}\right) = \Omega_h * \tilde{\mathbf{y}} + \beta_h, \tag{6}$$

where the depth of the last layer is n_h, the size of kernel Ω_h is $3{\times}3{\times}3{\times}n_h$, and the size of bias β_h is $H{\times}W{\times}n_h$. It allows us to enhance high-frequency information in the final output through the chroma upsampling network (see Figure 8). In summary, we learn a model F that can predict a high-resolution image $\hat{\mathbf{y}} = F(\mathbf{X}) = h(g(f(\mathbf{X}), \mathbf{X}), \mathbf{X})$ from a given input $\mathbf{X}$.

3.3. Network Parameters

3.3.1 Luminance Network

Architecture Since the human eye's resolving power of contrast mainly depends on luminance, we exclusively use the luminance channel of input images to enhance spatial resolution. The first layer contains a total of 65 images, where with an image from the left-view and 64 images from the right-view with parallax shifting. The filter size of the first convolutional layer is 65×3×3×64. We use 16 layers with the same kernel size of 64×3×3×64. Each convolutional layer is followed by a ReLU. The last layer is used for the residual image reconstruction with a filter size of 64×3×3×1.

Receptive Field The size of a receptive field varies depending on the size of filters and the number of layers. 3×3 filters with 16 layers operate the receptive field size of 33×33. Inserting more layers increases not only the size of the receptive field but also its computational cost. We found that the increased receptive field does not guarantee increased performance regarding accuracy. Recently, Mayer et al. [39] found that deep CNN cannot handle relatively large disparities even with an enlarged receptive field. We also found that the convolutional networks could be inefficient in searching correspondences as CNN's test matching costs in 2D, rather than in 1D along the epipolar line.

Maximum Disparity We build a stereo image tensor with 64 shifted images with one-pixel intervals, in which the maximum disparity is assumed to be the summation of 64 shifted pixels plus 33 pixels of the width of the receptive field, approximately ∼100 pixels in total.

We use the Middlebury dataset for training our networks. Since the resolution of the original Middlebury dataset [47, 46] is too higher than others (Tsukuba and KITTI), we downsampled the Middlebury dataset in half both horizontally and vertically to make its resolution similar to others for testing. More than about 98% of disparities in the dataset are within this range except for very close objects. Figure 5 compares the impact of the number of input stereo images for learning the luminance network.

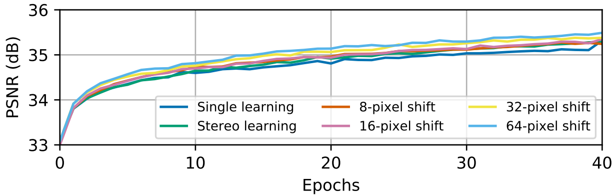

Figure 5: Impact of the size of the stereo image stack for learning the luminance network. A larger stereo tensor input increases the reconstruction accuracy.

Number of Layers Figure 6 compares the impact of convolutional layers in the luminance network. While a large number of layers improves the quality of a result image, it also inflates the computational time. Also, more than 16 layers made no significant improvement so that we chose 16 layers at the end.

Activation Functions The last convolutional layer of the luminance network reconstructs residuals for a high-resolution image. The activation function in the last layer handles the contrast variance of the residual image. We compare the performance differences by using three different types of activation functions: ReLU, sigmoid, and

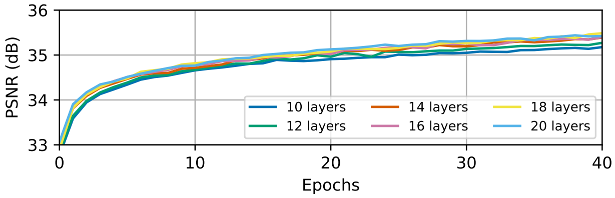

Figure 6: Impact of the number of layers in the luminance network.

an identity function (no activation), as shown in Figure 7. The case of no activation function in the last layer shows the highest peak signal-to-noise ratio (PSNR) performance. The ReLU and the sigmoid function clamp residual values between [0, ∞] and [-1, +1], respectively. We found that they cause contrast compression. As a result, we employ no activation function at the last layer to reproduce high-frequency features of the target image.

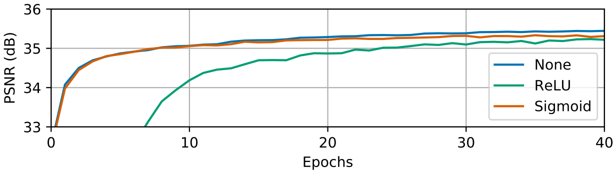

Figure 7: Impact of the activation function in the last convolutional layer in the luminance network.

3.3.2 Chrominance Network

Architecture Our chrominance network follows the typical residual learning architecture [23]. We use 15 convolutional layers for learning residual images, and an additional layer to produce the final image from low-resolution chromaticity channels and the reconstructed high-resolution luminance image. The kernel size of the first convolutional layer is 3×3×3×64. Internal convolutional layers apply 64×3×3×64 for each convolution operation. Then we concatenate residuals with the reference image and apply the last convolutional layer with a kernel size of 6×3×3×3. The last convolutional layer applies no activation function when computing final results.

Semi-Residual Learning To preserve high-frequency details from the reconstructed high-resolution luminance, we take a semi-residual learning approach, which includes a convolutional layer at the end of the network instead of the sum of the fast-forwarded identity. While details in C_b and C_r channels tend to disappear through the convolution steps, the semi-residual approach allows us to increase high-frequency details, as shown in Figure 8.

3.4. Training Networks

The number of layers of each subnetwork is similar to very deep neural networks [27]. We are therefore motivated to train this network in two stages. For initial estimation of subnetworks, we trained each subnetwork individually. We first trained the luminance network and then trained the chrominance network using high-resolution luminance reconstructed by the previous network (see Figure 3). However, from this separately learning process, the chrominance network learns the second residual for the luminance channel which has already been conducted by the previous network. To avoid the second residual issue, we additionally trained both networks *jointly* to refine the end-to-end model through the second stage of optimization.

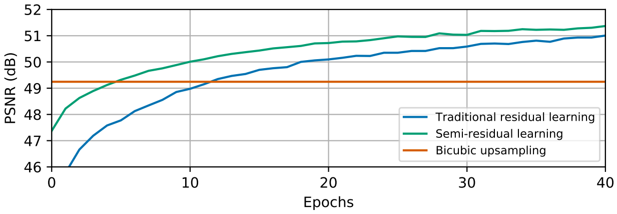

Figure 8: Impact of the additional convolutional layer at the end of the chrominance network.

4. Results

We employed the TensorFlow deep-learning framework to implement our stereo super-resolution networks. For computing and applying gradients to weights, we select the Adam optimizer [32] with an initial learning rate of 0.001. The Adam optimizer utilizes two momentum variables to compute the adaptive learning rates. We set the exponential decay rate for the first momentum as $\beta_1 = 0.9$ and the second momentum as $\beta_2 = 0.999$. The initial values of the kernels are calculated using Xavier's algorithm [21]. The training of the luminance super-resolution network takes about 3 hours on a machine with a 4.0 GHz Intel i7-6700K CPU and Titan X Pascal GPU with a batch size of 128 for 40 epochs (50,000 iterations) of 26,825 training patches with augmentation (flipping and rotation). We found that learning more than 40 epochs causes an overfitting problem, which results in performance degradation. Table 1 compares averaged computational time for reconstructing a 320×240 image by five different methods. Note that the computational cost of our method increases linearly proportional to the size of input as our computation consists of convolutions only.

Datasets To create the training datasets of low- and high-resolution stereo image pairs, we use stereo images from the Middlebury dataset [47, 46], the KITTI stereo dataset [17, 39], and the Tsukuba dataset [44]. As mentioned, the Middlebury datasets are downsampled to match the resolutions of the other datasets. For the training data, we use 60 Middlebury images dividing into 33×33 patches

Methods	Single image input			Stereo image input	
	SRCNN	VDSR	PSyCo	Bhavsar	Ours
Time	9.17s	2.34s	2.54s	57.78s	3.23s

Table 1: Average computational time for reconstructing a 640×480 high-resolution image from a 320×240 low-resolution image from the Middlebury dataset.

Dataset	Scale	Bicubic		SRCNN		VDSR		PSyCo		Ours	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Middlebury (5 images)	×2	29.64	0.9228	31.48	0.9505	31.62	0.9543	32.03	0.9542	33.05	0.9545
	×3	27.20	0.8737	28.76	0.9136	29.30	0.9240	29.40	0.9231	29.59	0.8974
	×4	25.79	0.8344	27.11	0.8814	27.23	0.8916	27.58	0.8935	26.80	0.8495
Tsukuba (16 images)	×2	36.69	0.9833	40.05	0.9846	40.50	0.9840	41.08	0.9846	43.30	0.9968
	×3	32.98	0.9659	35.89	0.9611	36.70	0.9666	36.74	0.9657	37.32	0.9754
	×4	30.89	0.9453	33.10	0.9343	33.40	0.9463	33.83	0.9418	34.40	0.9541
KITTI 2012 (16 images)	×2	28.08	0.9200	29.27	0.9162	29.48	0.9137	29.82	0.9202	30.30	0.9452
	×3	25.72	0.8701	26.98	0.8533	27.24	0.8552	27.46	0.8637	27.96	0.8889
	×4	24.33	0.8345	25.34	0.8002	25.44	0.8043	25.73	0.8142	25.75	0.8382
KITTI 2015 (100 images)	×2	27.14	0.9176	28.50	0.9193	28.60	0.9156	28.75	0.9188	29.36	0.9456
	×3	24.74	0.8597	26.19	0.8530	26.37	0.8539	26.37	0.8548	26.96	0.8879
	×4	23.34	0.8130	24.44	0.7951	24.50	0.7999	24.64	0.7998	24.83	0.8338

Table 2: Quantitative evaluation of our method with state-of-the-art SR algorithms for ×2, ×3 and ×4 magnification ratios on the Middlebury stereo dataset, Tsukuba dataset, and KITTI stereo dataset. Red color indicates the highest accuracy, and blue color presents the second highest accuracy regarding PSNR [dB] and SSIM.

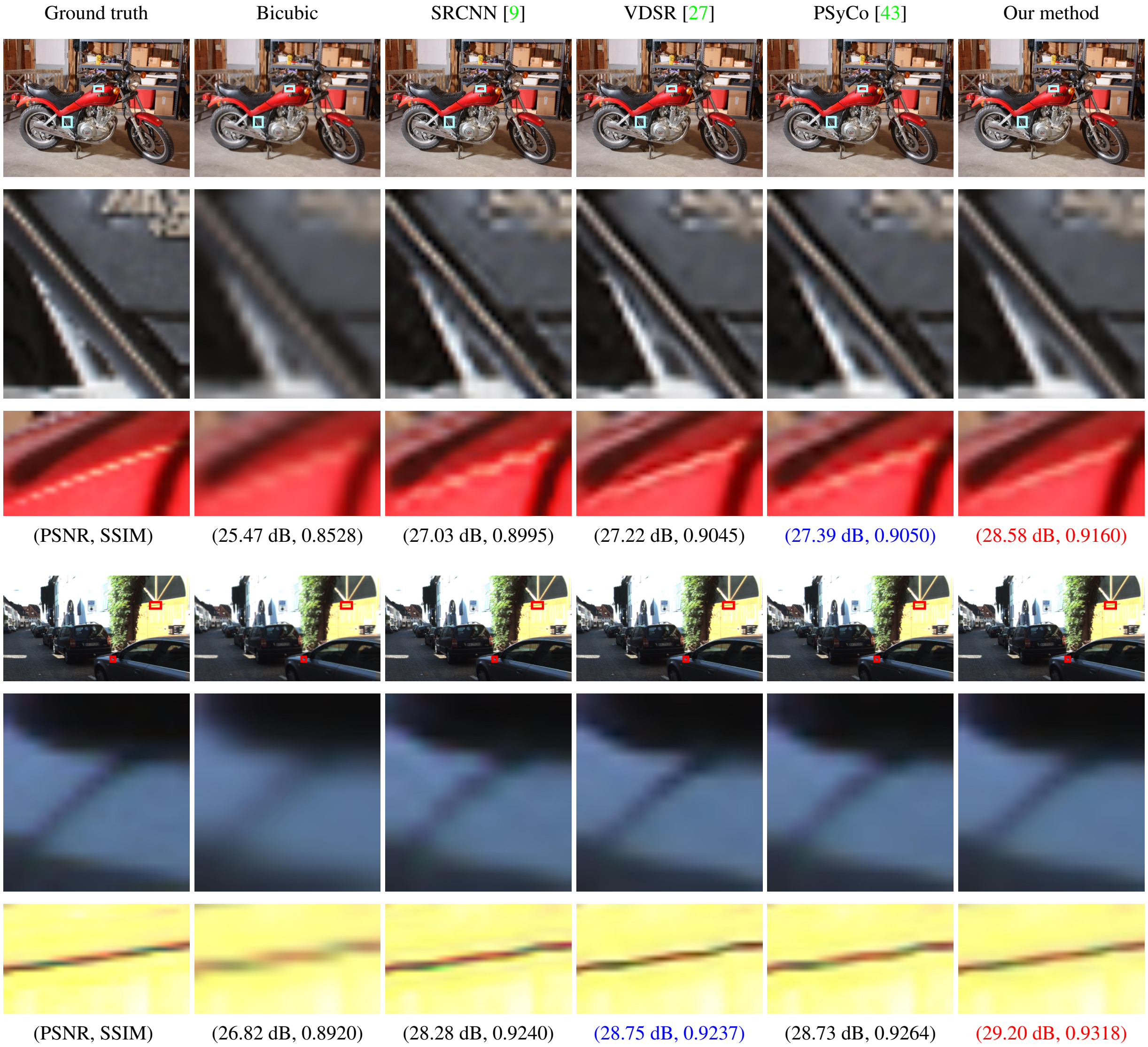

Figure 9: Results of super-resolution with the ×2 magnification factor. PSNR and SSIM values are shown under each result. The red color indicates the highest PSNR and SSIM values, and the blue color represents the second highest values.

with stride 24 and apply data augmentation to create a large number of data patches. For the test data set, we use 5 images from Middlebury, 16 images from Tsukuba, 16 images from KITTI2012, and 100 images from KITTI2015. Refer to the supplemental materials for more images.

We use learning data in the Middlebury dataset, while we use test images from other datasets (5 Middlebury images, 16 Tsukuba images, and 116 KITTI images) that are not used in the training process. To test the impact of the training dataset on accuracy, we attempted to train the original SRCNN method with the same training dataset that we used. The accuracy evaluation of the newly trained SRCNN network shows 31.39 for 5 Middlebury test images and 40.11 for 16 Tsukuba images. This accuracy is not significantly different from that of the original SRCNN model trained by the authors (the average PSNR of 31.48 for 5 Middlebury test images and 40.05 for 16 Tsukuba images, as shown in Table 2.) It validates that our results are not affected by the training dataset.

Stereo Image Input We evaluate our method with a state-of-the-art stereo super-resolution method proposed by Bhavsar and Rajagopalan [5]. Figure 10 compares our reconstructed image with the previous stereo super-resolution method. The previous method severely suffers from artifacts on edges, resulting in even lower resolution than that of bicubic upsampling.

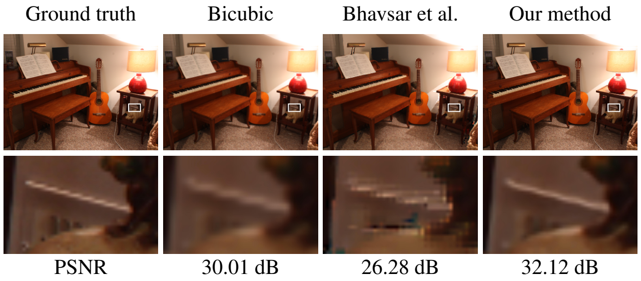

Figure 10: Comparison of our method with a state-of-the-art stereo super-resolution method by Bhavsar and Rajagopalan [5]. While their method suffers from aliasing artifacts on edges, our result is close to the ground truth.

Single Image Input We evaluate the results of our method with quantitative and qualitative comparisons. For comparison, we choose state-of-the-art single-image super-resolution methods: super-resolution convolutional neural network (SRCNN) [9], very deep super-resolution (VDSR) [27], and patch symmetry collapse (PSyCo) [43]. SRCNN and VDSR are deep learning-based approaches as our method, while PSyCo is an optimization-based solution. We use the trained models and best parameters directly provided by the authors. (9-5-5 ImageNet model for SRCNN, a network of 20 layers for VDSR, and 1024 atoms for PsyCo).

To quantitatively evaluate results, we first created test datasets by downsampling them by ×2, ×3 and ×4. Then we upscale them by the magnification ratios of ×2, ×3 and ×4. Also there is missing information around the image boundary due to parallax, so we cropped out the missing regions in the pair. Figure 9 shows super-resolution results of the magnification ratio of ×2 on the Middlebury test dataset. Our method outperforms other state-of-the-art single-image methods. Table 2 provides the average PSNRs and SSIMs on each benchmark dataset. Our method achieves the highest PSNR and SSIM values in most cases when compared with the state-of-the-art methods. For more results, refer to the supplemental materials.

Natural Objects To help judge the naturalness of our reconstruction method, we experimented with natural objects: a human face (Figure 11). We compare our result with the best prior performer, PSyCo, in our experiment. The PSNR of our result is still higher than PSyCo, while our result can provide highly natural appearance.

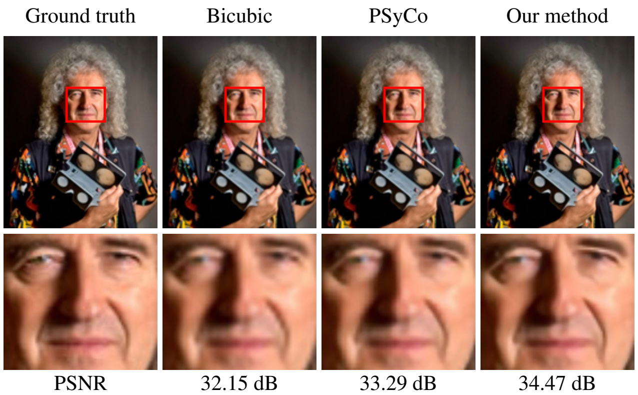

Figure 11: Our method shows not only a higher PSNR value but also the reconstructed image appears more plausible in terms of naturalness. Image courtesy of Brian May [38].

5. Conclusion

We have described a method that can enhance the spatial resolution of stereo images, which comprises two subnetworks for luminance and chrominance, respectively. Even though our method does not calculate disparity directly, it utilizes a parallax prior in stereo that can reconstruct a high-resolution image with subpixel accuracy in registration. Our method can outperform both current state-of-the-art methods in enhancing the spatial resolution of stereo images. It can be used with any other stereo imaging methods additionally to enhance spatial resolution.

Acknowledgements

Min H. Kim acknowledges Korea NRF grants (2016R1-A2B2013031, 2013M3A6A6073718) and additional support by SK-Hynix, KOCCA in MCST of Korea, Giga KOREA Project (GK17P0200), and an ICT R&D program of MSIT/IITP of Korea (2017-0-00072, 2016-0-00018).

References

[1] S.-H. Baek, I. Choi, and M. H. Kim. Multiview image completion with space structure propagation. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR 2016)*, pages 2727–2735, Las Vegas, USA, 2016. IEEE. 3

[2] S.-H. Baek and M. H. Kim. Stereo fusion: Combining refractive and binocular disparity. *Computer Vision and Image Understanding*, 146:52–66, 2016. 3

[3] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002. 2

[4] A. V. Bhavsar and A. N. Rajagopalan. Resolution enhancement for binocular stereo. In *ICPR*, pages 1–4, 2008. 1, 2

[5] A. V. Bhavsar and A. N. Rajagopalan. Resolution enhancement in multi-image stereo. *IEEE Trans. Pattern Anal. Mach. Intell*, 32(9):1721–1728, 2010. 1, 2, 8

[6] T. E. Bishop and P. Favaro. The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE Trans. Pattern Anal. Mach. Intell*, 34(5):972–986, 2012. 2

[7] I. Choi, D. S. Jeon, G. Nam, D. Gutierrez, and M. H. Kim. High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2017)*, 36(6):218:1–13, 2017. 4

[8] D. Dai, R. Timofte, and L. J. V. Gool. Jointly optimized regressors for image super-resolution. *Comput. Graph. Forum*, 34(2):95–104, 2015. 1

[9] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 184–199. Springer, 2014. 1, 2, 4, 7, 8

[10] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Trans. Image Processing*, 6(12):1646–1658, Dec. 1997. 2

[11] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Trans. Image Processing*, 13(10):1327–1344, Oct. 2004. 2

[12] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[13] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, 2002. 1, 2

[14] C. Gao and N. Ahuja. A refractive camera for acquiring stereo and super-resolution images. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2316–2323. IEEE, 2006. 2

[15] X. Gao, K. Zhang, D. Tao, and X. Li. Joint learning for single-image super-resolution via a coupled constraint. *IEEE Trans. Image Processing*, 21(2):469–480, 2012. 1

[16] D. C. Garcia, C. C. Dorea, and R. L. de Queiroz. Super resolution for multiview images using depth information. *IEEE Trans. Circuits Syst. Video Techn*, 22(9):1249–1256, 2012. 2

[17] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 6

[18] T. Georgiev, G. Chunev, and A. Lumsdaine. Superresolution with the focused plenoptic camera. In *Proc. Computational Imaging*, volume 7873 of *SPIE Proceedings*, page 78730X. IS&T/SPIE, 2011. 2

[19] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand. Deep joint demosaicking and denoising. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2016)*, December 2016. 5

[20] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *ICCV*, 2009. 1

[21] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010. 6

[22] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, June 2004. 3

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 2, 5, 6

[24] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206. IEEE, 2015. 1, 2

[25] A. K. Jain and T. Q. Nguyen. Video super-resolution for mixed resolution stereo. In *ICIP*, pages 962–966. IEEE, 2013. 2

[26] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2016)*, 35(6), 2016. 2

[27] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 3, 4, 6, 7, 8

[28] J. Kim, J. Kwon Lee, and K. Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[29] M. H. Kim. *High-Fidelity Colour Reproduction for High-Dynamic-Range Imaging*. Ph.D. Thesis, University College London, 2010. 3

[30] M. H. Kim, T. Ritschel, and J. Kautz. Edge-aware color appearance. *ACM Transactions on Graphics (Presented at SIGGRAPH 2011)*, 30(2):13:1–9, 2011. 3

[31] M. H. Kim, T. Weyrich, and J. Kautz. Modeling human color perception under extended luminance levels. *ACM Transactions on Graphics (Proc. SIGGRAPH 2009)*, 28(3):27:1–9, 2009. 3

[32] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[33] T. Komatsu, K. Aizawa, T. Igarashi, and T. Saito. Signal-processing based method for acquiring very high resolution images with multiple cameras and its theoretical analysis. *IEE Proceedings I-Communications, Speech and Vision*, 140(1):19–24, 1993. 2

[34] J. H. Lee, S.-H. Baek, and M. H. Kim. Multiview image completion with space structure propagation. In *Proc. British*

Machine Vision Conference (BMVC 2017), pages 1–11, London, England, 2017. 3

[35] J. H. Lee, I. Choi, and M. H. Kim. Laplacian patch-based image synthesis. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, USA, 2016. IEEE. 2

[36] W. Liu and S. Li. Multi-morphology image super-resolution via sparse representation. *Neurocomputing*, 120:645–654, 2013. 1

[37] J. Marco, Q. Hernandez, A. Muñoz, Y. Dong, A. Jarabo, M. H. Kim, X. Tong, and D. Gutierrez. Deeptof: Off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Trans. Graph.*, 36(6):219:1–219:12, Nov. 2017. 4

[38] B. May. Dr. Bri stereo image photographed by Kyle Cassidy. https://brianmay.com/brian/brianssb/brianssbjul15b.html, 2017. 8

[39] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6

[40] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Fürnkranz and T. Joachims, editors, *ICML*, pages 807–814. Omnipress, 2010. 4

[41] H. Park, K. M. Lee, and S. U. Lee. Combining multi-view stereo and super resolution in a unified framework. In *APSIPA*, pages 1–4. IEEE, 2012. 1, 2

[42] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *Signal Processing Magazine, IEEE*, 20(3):21–36, May 2003. 1, 2, 3

[43] E. Pérez-Pellitero, J. Salvador, J. Ruiz-Hidalgo, and B. Rosenhahn. Psyco: Manifold span reduction for super resolution. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 4, 7, 8

[44] M. Peris, S. Martull, A. Maki, Y. Ohkawa, and K. Fukui. Towards a simulation driven stereo vision system. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1038–1042. IEEE, 2012. 6

[45] M. Protter, M. Elad, H. Takeda, and P. Milanfar. Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Trans. Image Processing*, 18(1):36–51, Jan. 2009. 2

[46] D. Scharstein, H. Hirschmller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, volume 8753, pages 31–42. Springer, 2014. 5, 6

[47] D. Scharstein and R. S. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, Apr. 2002. 5, 6

[48] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[49] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2

[50] R. Timofte, V. De Smet, and L. Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013. 2

[51] R. Timofte, R. Rothe, and L. Van Gool. Seven ways to improve example-based single image super resolution. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[52] J. Yang, Z. Lin, and S. Cohen. Fast image super-resolution based on in-place example regression. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1059–1066, 2013. 1

[53] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Trans. Image Processing*, 19(11):2861–2873, 2010. 1, 2

[54] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon. Learning a deep convolutional network for light-field image super-resolution. In *ICCV Workshops*, pages 57–65. IEEE, 2015. 2