Stereoscopic Image Super-Resolution with Stereo Consistent Feature

Wonil Song, Sungil Choi, Somi Jeong, Kwanghoon Sohn*
School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
{swonil92, csi6570, somijeong, khsohn}@yonsei.ac.kr

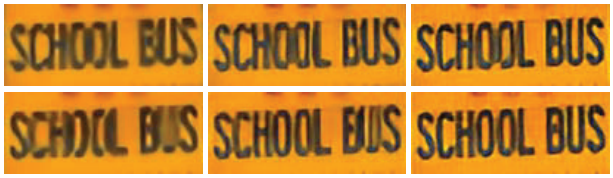
Abstract

We present a first attempt for stereoscopic image super-resolution (SR) for recovering high-resolution details while preserving stereo-consistency between stereoscopic image pair. The most challenging issue in the stereoscopic SR is that the texture details should be consistent for corresponding pixels in stereoscopic SR image pair. However, existing stereo SR methods cannot maintain the stereo-consistency, thus causing 3D fatigue to the viewers. To address this issue, in this paper, we propose a self and parallax attention mechanism (SPAM) to aggregate the information from its own image and the counterpart stereo image simultaneously, thus reconstructing high-quality stereoscopic SR image pairs. Moreover, we design an efficient network architecture and effective loss functions to enforce stereo-consistency constraint. Finally, experimental results demonstrate the superiority of our method over state-of-the-art SR methods in terms of both quantitative metrics and qualitative visual quality while maintaining stereo-consistency between stereoscopic image pair.

Introduction

With the increasing attention and popularity of stereoscopic 3D industry, stereoscopic image/video processing techniques have been spotlighted in a wide range of fields such as image inpainting (Wang et al. 2008), video stabilization (Guo et al. 2016), and style transfer (Chen et al. 2018). These 3D contents are presented via 3D displays such as AR/VR devices and 3D televisions by creating the illusion of depth from the stereo images. As the 3D devices have advanced, they demand high-resolution stereoscopic images, thus requiring stereoscopic super-resolution (SR) technique. The most important point for the stereoscopic SR is preserving the consistency between super-resolved stereo image pair to provide the illusion of depth. Otherwise, the inconsistent stereo images would cause 3D fatigue to the viewers.

Super-resolution (SR) is a fundamental problem in low-level vision tasks aiming to enhance the spatial resolution of low-resolution (LR) image by reconstructing high-resolution (HR) image. Recently, following the seminal work of Dong et al. (2015), numerous single image SR methods have leveraged convolutional neural networks (CNNs). They adopt advanced CNN techniques such as residual architecture (Kim, Lee, and Lee 2016; Zhang et al. 2018c) and perceptual loss (Johnson, Alahi, and Fei-Fei 2016; Ledig et al. 2017) to achieve the enhanced performance, but due to significant information loss in LR image, it is still extremely challenging.



VDSR PASSRnet Ours

Figure 1: The result of $4\times$ super-resolved stereo left (top) and right (bottom) image using VDSR (Kim, Lee, and Lee 2016), PASSRnet (Wang et al. 2019a), and our proposed method. These results are achieved on "Validation 57" of Flickr 1024 dataset (Wang et al. 2019b). Our result shows consistent texture details different with others.

To alleviate this limitation, some SR methods have been proposed using multiple LR images such as stereo image pairs (Jeon et al. 2018; Wang et al. 2019a) and video frames (Liao et al. 2015; Caballero et al. 2017; Tao et al. 2017). By taking advantage of additional information from multiple LR images, they show the superior SR performance. However, the disparity (motion) exists between stereo images (video frames), so conducting disparity (motion) compensation is a essential process to integrate them. However, since they focus on generating the target HR image by incorporating additional LR images, they do not consider the consistent property. As shown in Figure 1, the SR images reconstructed by conventional single and stereo SR methods (Kim, Lee, and Lee 2016; Wang et al. 2019a) show inconsistent texture details across stereo images, thus providing a visual discomfort from blurry details.

In this paper, we present a novel stereoscopic super-resolution method for overcoming the aforementioned lim-

*Corresponding author

itations of current SR methods. We propose a *self and parallax attention mechanism* (SPAM), which captures self attention map and parallax-attention map simultaneously. The self-attention map considers similarity within its own image, which helps to supplement information from similar points within itself. The parallax-attention map estimates the correspondence between stereo image pair, and is utilized to integrate information from LR stereo images. The proposed SPAM generates features implying abundant and comprehensive clues for reconstructing the HR images. In addition, disparity map derived from parallax-attention is applied on both proposed architecture and our objective functions to force the SR stereo images be consistent. To evaluate our method, we conduct extensive experiments on Middlebury (Scharstein and Pal 2007), Flickr 1024 (Wang et al. 2019b), KITTI 2012 (Geiger, Lenz, and Urtasun 2012), and KITTI 2015 (Menze and Geiger 2015) datasets compared to the state-of-the-art SR methods. We also conduct an ablation study to analyze the contribution of our attention mechanism. We demonstrate that our method achieves for reconstructing competitive SR quality while maintaining the stereo-consistency through qualitative and quantitative evaluations.

Overall, our contributions are three-fold: First, we propose the first stereoscopic super-resolution method by imposing stereo-consistency constraint on integrating stereo information and training the network simultaneously. Secondly, we propose self and parallax attention mechanism to aggregate the information for reconstructing high-quality stereoscopic SR image. Lastly, the proposed method shows the state-of-the-art SR performance compared to the conventional stereo SR method while preserving stereo-consistency on various datasets.

Related Work

Single Image Super-Resolution. The seminal work of Dong et al. (2015) was proposed super-resolution convolutional neural network (SRCNN), and achieved the superior performance with simple three convolutional layers. Kim et al. (2016) proposed a very deep network for SR (VDSR) with twenty convolutional layers, and introduced residuals for facilitating training them. By increasing the depth, VDSR achieved the better performance than SRCNN.

Later on, some methods have been proposed using various network architectures, such as Laplacian pyramid structure (Lai et al. 2017), residual blocks (Wang et al. 2018), attention module (Zhang et al. 2018b), and dense block (Zhang et al. 2018c). Since applying pixel-wise loss (*e.g.* MSE) results in over-smoothed and less high-frequency textures, some methods have been introduced diverse perceptual loss, which makes the distance between HR and SR in a feature space minimized, such as combined pixel-wise and perceptual loss (Ledig et al. 2017) and adversarial loss (Johnson, Alahi, and Fei-Fei 2016). Even though the state-of-the-art single image SR methods achieve dramatic improvements, in terms of utilizing information from single LR image only to recover HR image, they still have shown limited performance.

Multiple Image Super-Resolution. Since multiple images (*e.g.,* video frames, stereo images) can provide plenty of information from additional images, the SR performance can be improved. However, it is very challenging to incorporate them due to temporal or spatial discrepancy, so the alignment between them should be handled.

Liao et al. (2015) proposed the first work of the video frame SR. By compensating motions using conventional optical flow algorithms (Brox et al. 2004; Xu, Jia, and Matsushita 2011), they generates the SR image from combined adjacent video frames. Tao et al. (2017) proposed jointly learning frameworks for estimating motion and recovering SR images, and enhanced the video SR performance.

Jeon et al. (2018) first introduced stereo image SR (StereoSR), which estimates a parallax prior from stereo images through networks, and then generates the SR image by taking the concatenated image of left and sifted right image by its parallax as input. Because it cannot deal with large disparity variations, Wang et al. (2019a) proposed a parallax-attention stereo super-resolution network (PASSRnet). By introducing a parallax-attention module, which estimates global stereo correspondence along the epipolar line, it removes the limitation of disparity range.

Those multiple image SR methods utilize additional LR images to enrich the information for recovering HR details only. As a result, they leave out consideration of the consistent property that the correspondence points between SR images should possess consistent texture details, so they are not proper for enhancing the spatial resolution for the stereoscopic images.

Attention Mechanism. To perform the computer vision tasks, such as image classifications, segmentations, and generations, many deep generative models rely on the stack of convolution layers, which has the limited local receptive field. Since convolution is conducted within only local region, it cannot consider whole image at once, thus showing the limited performance. To alleviate this limitation, self-attention module (Zhang et al. 2018a) and non-local neural (Wang et al. 2018) network were proposed by modeling long-range dependency. By calculating the correlation of intermediate feature responses in all position and taking weighted sum of them, the long-range dependency is modeled, which presents the relations of any two positions within the whole images. Based on its performance improvement, several methods in various low-level vision applications has been adopted the long-range dependency, such as de-raning (Li et al. 2018) and image SR (Liu et al. 2018; Dai et al. 2019).

Proposed Method

The objective of our method is to estimate a stereoscopic super-resolution image pair $(\mathbf{I}^l_{SR}, \mathbf{I}^r_{SR})$ from a given stereoscopic low-resolution image pair $(\mathbf{I}^l_{LR}, \mathbf{I}^r_{LR})$, recovering high-resolution details while preserving the stereo-consistency. For achieving the high-quality SR result for each LR image using stereoscopic images together, our method establishes stereo correspondence pair $(\mathcal{D}^l, \mathcal{D}^r)$ and valid mask pair $(\mathcal{V}^l, \mathcal{V}^r)$ using novel attention mechanisms,

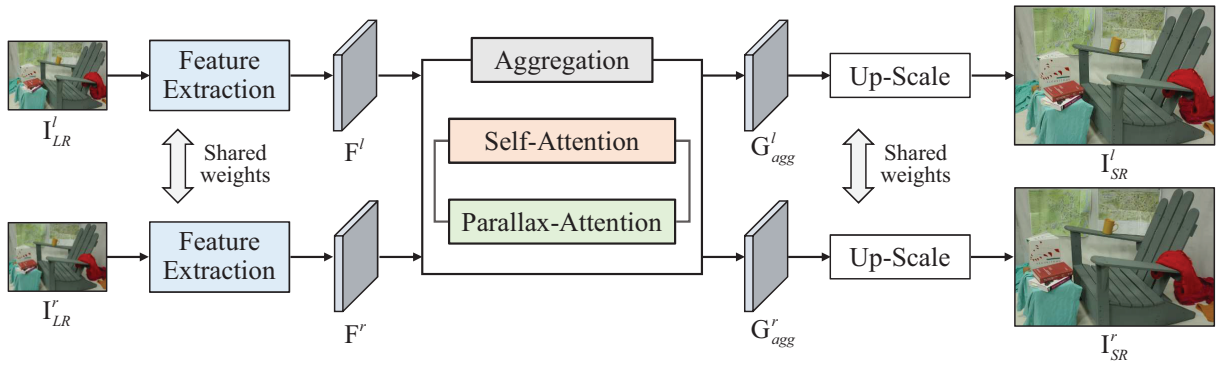


Figure 2: The architecture of proposed method. Stereo images are fed into the feature extraction part for generating highly discriminative features to find stereo correspondence. Then, the feature aggregation part produces view-symmetric and rich informative features for stereoscopic SR using a self and parallax attention mechanism (SPAM). The up-scale part outputs the SR image of each view.

and gives stereo-consistent constraints using them on not only intermediate features directly but also loss functions. As a result, the proposed method yields the stereoscopic SR image pair $(\mathbf{I}^l_{SR}, \mathbf{I}^r_{SR})$, which possesses the consistent HR details between the corresponding points in image pair, thus enhancing the illusion of depth.

Network Architecture

Our proposed method consists of three stage as shown in Figure 2: *feature extraction*, *feature aggregation*, and *up-scale part*.

Given $(\mathbf{I}^l_{LR}, \mathbf{I}^r_{LR})$, we extract a discriminative feature pair $(\mathbf{F}^l, \mathbf{F}^r)$, which is utilized for establishing the reliable $(\mathcal{D}^l, \mathcal{D}^r)$ and $(\mathcal{V}^l, \mathcal{V}^r)$, denoted as

$$\mathbf{F}^l = H_{feat}(\mathbf{I}^l_{LR}), \quad \mathbf{F}^r = H_{feat}(\mathbf{I}^r_{LR}), \tag{1}$$

where H_{feat} stands for the feature extraction part, which is composed of 3 residual blocks and 2 residual Atrous Spatial Pyramid Pooling (ASPP) blocks (Wang et al. 2019a). Note that the configuration of our feature extraction module is same as PASSRnet (Wang et al. 2019a) to compare the SR performance with respect to the attention module fairly.

The features $(\mathbf{F}^l, \mathbf{F}^r)$ are used for estimating the stereo correspondence pair $(\mathcal{D}^l, \mathcal{D}^r)$ and the valid mask pair $(\mathcal{V}^l, \mathcal{V}^r)$ in the feature aggregation part. Note that $\mathcal{D}^l$ is denoted as stereo correspondence fields from $\mathbf{I}^l_{LR}$ to $\mathbf{I}^r_{LR}$ and $\mathcal{V}^l$ is denoted as the overlapping (non-occluded) regions in $\mathbf{I}^l_{LR}$ with $\mathbf{I}^r_{LR}$, and vice versa. To obtain them, we propose a self and parallax attention mechanism (SPAM). By extracting self and parallax attention maps via SPAM, the feature aggregation part generates highly informative and stereo-symmetric feature pair $(\mathbf{G}^l_{agg}, \mathbf{G}^r_{agg})$ for stereoscopic SR. We will explain in details how those attention maps are utilized to produce the rich context stereo-symmetric features in the following section.

Finally, the stereo-symmetric features $(\mathbf{G}^l_{agg}, \mathbf{G}^r_{agg})$ are up-scaled to generate the stereoscopic SR image pair. To recover the HR details, the up-scale part consists of sub-pixel convolution layer, which generates the SR stereoscopic images as follows:

$$\mathbf{I}^l_{SR} = H_{up}(\mathbf{G}^l_{agg}), \quad \mathbf{I}^r_{SR} = H_{up}(\mathbf{G}^r_{agg}), \tag{2}$$

where H_{up} stands for the up-scale part.

Feature Aggregation Part

In order to create features, which is abundant for SR, and symmetric for imposing stereo-consistency for stereoscopic SR, we propose a self and parallax attention mechanism in the feature aggregation part. It consists of three mechanisms: parallax-attention, self-attention, and fusion mechanism. The framework is illustrated in Figure 3.

Since the main purpose of the feature extraction part is to extract the features $(\mathbf{F}^l, \mathbf{F}^r)$, which are the optimal for estimating the stereo correspondence, we construct an additional transition residual block to obtain the intermediate features, which are the optimal features for reconstructing SR images, denoted as

$$\mathbf{F}^{l*} = H_{Res-a}(\mathbf{F}^l), \quad \mathbf{F}^{r*} = H_{Res-a}(\mathbf{F}^r). \tag{3}$$

Using $(\mathbf{F}^l, \mathbf{F}^r)$ and $(\mathbf{F}^{l*}, \mathbf{F}^{r*})$, we will explain more details how these asymmetric features are translated to the informative symmetric features.

Parallax-Attention Mechanism. Inspired by PASSRnet (Wang et al. 2019a), we modified the parallax-attention mechanism (PAM), which is more suitable for stereoscopic SR. Through the PAM, the global correspondence in stereo image is captured to integrate the rich context information from $(\mathbf{I}^l_{LR}, \mathbf{I}^r_{LR})$. For simplification, we abbreviate the target view as u and the opposite view as v, *e.g.*, when u is left view, v is right view and vice versa. The features $\mathbf{F}^u$ and $\mathbf{F}^v$ are fed into different 1×1 convolution layers α and β respectively. Then, the correlation between row vector of $\alpha(\mathbf{F}^u)$ and $\beta(\mathbf{F}^v)$ are computed, and then the softmax-normalized correlation plays a role as parallax-attention $\mathbf{P}^{u\rightarrow v}$ for target to opposite view.

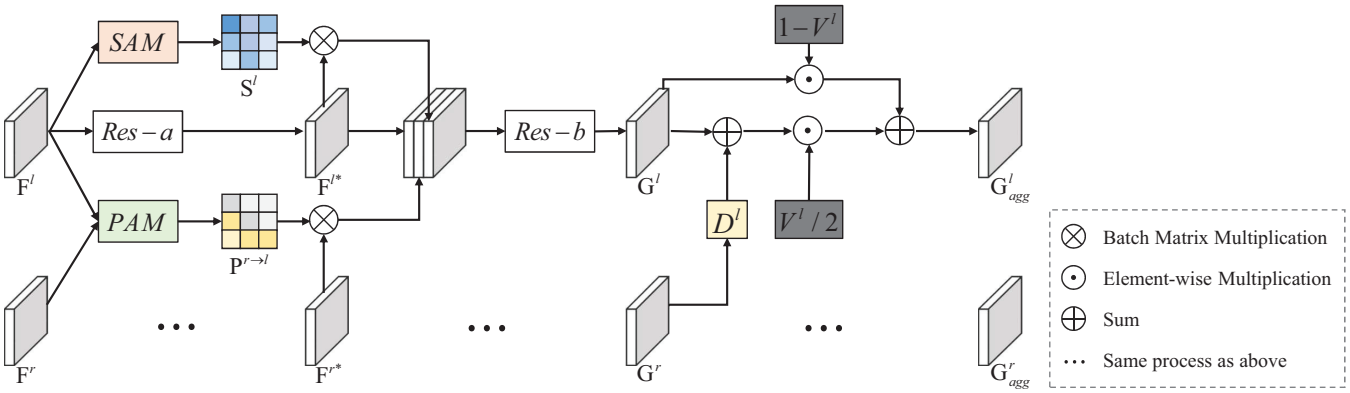


Figure 3: Illustration of a self and parallax attention mechanism (SPAM). The discriminative features $(\mathbf{F}^l, \mathbf{F}^r)$ enter the SAM and PAM for finding self-attention and parallax-attention. The optimal features for SR are generated by additional residual block. With the self and parallax attention, more abundant but asymmetric features are generated. By warping the features of one view to the other view with the disparity maps $(\mathcal{D}^l, \mathcal{D}^r)$ derived from the parallax-attention maps, stereo-symmetric features $(\mathbf{G}^l_{agg}, \mathbf{G}^r_{agg})$ are produced.

In order to enable the feature to possess the plenty of information from its own $\mathbf{F}^{u*}$ and the opposite $\mathbf{F}^{v*}$ simultaneously, we conduct forward waping of $\mathbf{F}^{v*}$ using the parallax-attention map $\mathbf{P}^{v\to u}$ as follows:

$$\mathbf{F}^u_{PAM} = W_{PAM}(\mathbf{P}^{v\to u}, \mathbf{F}^{v*}), \tag{4}$$

where W_{PAM} is the forward warping operation, conducted by matrix multiplication of $\mathbf{P}^{v\to u}$ and $\mathbf{F}^{v*}$ for the each row.

Self-Attention Mechanism. In addition, we adopt the self-attention mechanism to create abundant features by exploiting the self-similarities. However, calculating the self-similarity over whole pixels causes the expensive computational burden, which is impractical for implementation. Therefore, we conduct self-similarity measurement at the region-level (Tao et al. 2017). The feature map is divided into several $k \times k$ regions. Then, the self-similarity map $\mathbf{S}^u$ is obtained by softmax operation on the self-correlation map over the 1x1 convolution output $\gamma(\mathbf{F}^v)$ and $\delta(\mathbf{F}^v)$ in the region-level. The additional features enriching SR using SAM are as follows:

$$\mathbf{F}^u_{SAM} = W_{SAM}(\mathbf{S}^u, \mathbf{F}^{u*}), \tag{5}$$

wherw W_{SAM} is the warping operation, conducted by matrix multiplication of $\mathbf{S}^u$ and $\mathbf{F}^{u*}$ at the region-level.

Fusion Mechanism. To produce the aggregated feature $\mathbf{G}^u$, which contains the rich context features from its own image and its counterpart image, we let the concatenated features $(\mathbf{F}^{u*}, \mathbf{F}^u_{PAM}, \mathbf{F}^u_{SAM})$ pass through the several residual blocks H_{Res-b}, denoted as

$$\mathbf{G}^u = H_{Res-b}(concat(\mathbf{F}^{u*}, \mathbf{F}^u_{PAM}, \mathbf{F}^u_{SAM})). \tag{6}$$

Even though these aggregated features $(\mathbf{G}^l, \mathbf{G}^r)$ are abundant for SR, they are not symmetric according to the inherent disparity in stereoscopic images, which results in the stereo-inconsistent SR images.

To alliviate the asymmetric problem, we aggregate the features using the disparity obtained by the parallax-attention maps $(\mathbf{P}^{u\to v}, \mathbf{P}^{v\to u})$. First, we find the disparity $\mathcal{D}^u$ using an argmax operation on $\mathbf{P}^{u\to v}$ as

$$\mathcal{D}^u(i,j) = j - \arg\max_k \mathbf{P}^{u\to v}(i,j,k). \tag{7}$$

Here, $\mathbf{P}^{u\to v}(i,j,k)$ represents the attention of pixel (i,j) of view u to pixel (i,k) of view v. Also, we obtain the valid mask $\mathcal{V}^u$ by checking the left-right consistency between the disparity maps $(\mathcal{D}^u, \mathcal{D}^v)$ as follows

$$\mathcal{V}^u(i,j) = \begin{cases} 1, & if\ \|\mathcal{D}^u(i,j) - \mathcal{D}^v(i, j-\mathcal{D}^u(i,j))\|_1 < \tau \\ 0, & otherwise, \end{cases} \tag{8}$$

where τ is a threshold, which determines how much errors between left and right disparities able to be accepted. Here, we set the τ to 2.

With these disparity maps $(\mathcal{D}^u, \mathcal{D}^v)$ and valid masks $(\mathcal{V}^u, \mathcal{V}^v)$, we make the stereo-symmetric features as

$$\mathbf{G}^u_{agg} = \frac{\mathbf{G}^u + W_{disp}(\mathcal{D}^v, \mathbf{G}^v)}{2} \odot \mathcal{V}^u + \mathbf{G}^u \odot (1 - \mathcal{V}^u), \tag{9}$$

where W_{disp} is the forward warping operation based on the disparity. As a result, the stereo-symmetric features $(\mathbf{G}^l_{agg}, \mathbf{G}^r_{agg})$ are fed into the reconstruct module to generate stereoscopic SR images.

Training Loss

In this section, we introduce three loss functions that are applied for training our network: *reconstruction loss*, *parallax-attention loss* and *stereo-consistency loss*.

Reconstruction Loss. Similar with other conventional CNN-based SR methods, we use mean square error (MSE) to produce the reconstructed SR image identical to the groundtruth HR image pixel-wisely, thus achieving higher PSNR.

$$\mathcal{L}_{rec} = \|\mathbf{I}^u_{SR} - \mathbf{I}^u_{HR}\|_2^2. \tag{10}$$

Parallax-Attention Loss. Since our method does not use the groundtruth disparity to train the SPAM, we obtain the stereo correspondence pair $(\mathcal{D}^l, \mathcal{D}^r)$, which encodes the disparity information of stereo images, in an unsupervised manner. Following the unsupervised learning methods of depth estimation (Godard, Mac Aodha, and Brostow 2017; Wang et al. 2019a; Joung et al. 2019), we formulate the parallax-attention loss consisting of photometric, smoothness and cycle term for training the PAM, such as $\mathcal{L}_{pa} = (\mathcal{L}_{photo} + \mathcal{L}_{smooth} + \mathcal{L}_{cyc})$

First, we apply the photometric term to give the pixel-wise photometric consistent constraints. In the ideal case, the left and the warped right image to the left view should be identical. Based on the estimated parallax-attention map $\mathbf{P}^{v\rightarrow u}$, it enforces the photometric consistency between input stereo images, defined as

$$\mathcal{L}_{photo} = \|\mathbf{I}^u_{LR} - W_{PAM}(\mathbf{P}^{v\rightarrow u}, \mathbf{I}^v_{LR})\|_1, \quad (11)$$

where $(\mathbf{I}^u_{LR}, \mathbf{I}^v_{LR})$ are the input LR stereo image.

To alleviate the unexpected noise value in the parallax-attention maps, we adopt the smoothness term to make it locally smooth as follows

$$\mathcal{L}_{smooth} = \sum_{i,j,k} \|\mathbf{P}^{v\rightarrow u}(i,j,k) - \mathbf{P}^{v\rightarrow u}(i+1,j,k)\|_1 + \|\mathbf{P}^{v\rightarrow u}(i,j,k) - \mathbf{P}^{v\rightarrow u}(i,j+1,k+1)\|_1. \quad (12)$$

Lastly, we use the cycle-consistency term, which is the widely used concept in diverse tasks (Zhou et al. 2016; Zhu et al. 2017; Joung et al. 2019). The point in $\mathbf{I}^u$ is warped to the v view via $\mathbf{P}^{v\rightarrow u}$, and then the warped point in the v view is warpped again to the original view via $\mathbf{P}^{u\rightarrow v}$. Then, it is natural that this point should be located in the start point. Thus, we design the cycle term to regularize the PAM training as follows

$$\mathcal{L}_{cycle} = \|\mathbf{P}^{v\rightarrow u\rightarrow v} - \mathcal{I}\|_1, \quad (13)$$

where $\mathcal{I}$ is the identical matrix.

Stereo-consistency Loss. To enforce the stereo consistency into a pair of stereo SR outputs, we propose a novel stereo consistency loss. Preliminarily, Chen et al. (2018) defined a disparity loss to obtain the stereo-consistent style transfer results. It enforces the stylized result at one view to be close to the warped result from the other view explicitly. The proposed stereo-consistency loss has the same goal to the disparity loss in that the left and right SR results have to be consistent in the corresponding regions. To do so, we give the constraint that the warped SR image using the obtained disparity map should be identical to the opposite HR image, defined as

$$\mathcal{L}_{stereo} = \sum_v \|\mathcal{V}^u_{HR} \odot (W_{disp}(\mathcal{D}^v_{HR}, \mathbf{I}^v_{SR}) - \mathbf{I}^u_{HR})\|_2^2 \quad (14)$$

where $\odot$ is the element-wise multiplication. Note that in order to consider only valid region and exclude the rest region, we apply the valid mask. In addition, $\mathcal{D}^v_{HR}$ is excluded for testing the network.

Overall Loss Function. In summary, we optimize our model in a unified and end-to-end manner. The total loss function consists of three losses, expressed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{pa}\mathcal{L}_{pa} + \lambda_{stereo}\mathcal{L}_{stereo}, \quad (15)$$

where λ_{pa} and λ_{stereo} control the relative weights between them.

Experiments

Implementation Details

We used the Middlebury (Scharstein and Pal 2007), Flickr 1024 (Wang et al. 2019b), and KITTI 2012 (Geiger, Lenz, and Urtasun 2012) and KITTI 2015 (Menze and Geiger 2015) dataset to train and evaluate our method. To be specific, we divide the 60 Middlebury datasets into 30 pairs for training, 10 pairs for validation, and 20 pairs for evaluation. Following StereoSR (Jeon et al. 2018), we downsampled Middlebury by a factor of 2 to generate HR images. As provided the Flickr 1024 dataset, we used 800 pairs for trainining, 112 pairs for validation, and 112 pairs for evaluation. We select 40 pairs from KITTI 2012 and 2015 datasets, and they are used for only test. To make training patches, we first downsampled HR images using bicubic interpolation. Then we cropped 30×90 patches in the $4\times$downsampled images to take them as LR inputs, and cropped correponding patches in HR images to use it as HR groundtruth.

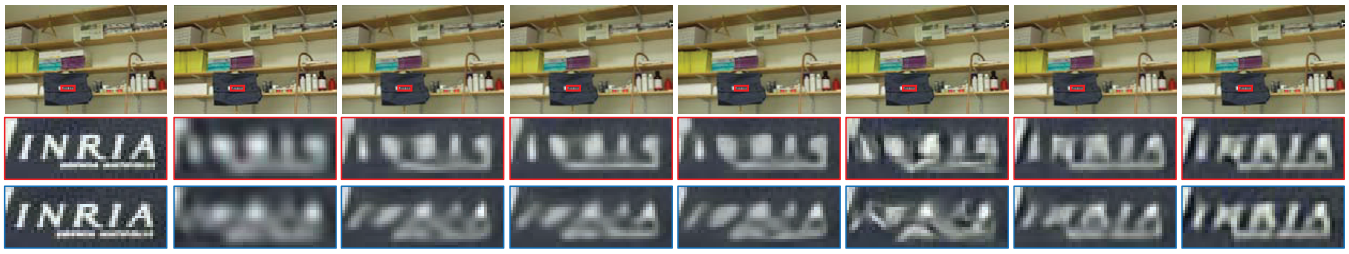
Our network was implemented using PyTorch and trained on NVIDIA GeForce GTX TitanX GPU. The weights of networks are initialized by a Gaussian distribution with mean 0 and standard deviation 0.01, and the Adam optimizer (Kingma and Ba 2014) was employed for optimization, where $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Additionally, for region-level SAM, we set $k = 4$. The initial learning rate is 10^{-4} and halved at every 30 epochs, and the batch size is 2. We set the parameters of the loss functions, such as $\lambda_{pa} = 0.005, \lambda_{stereo} = 1$.

Comparison with State-of-the-Art Methods

We evaluated our method with the state-of-the-art single image SR methods, SRCNN (Dong et al. 2015), VDSR (Kim, Lee, and Lee 2016), DRRN (Tai, Yang, and Liu 2017), and LapSRN (Lai et al. 2017), and stereo image SR methods, StereoSR (Jeon et al. 2018), and PASSRnet (Wang et al. 2019a) both quantitatively and qualitatively.

Quantitative Evaluation. We measure the SR performance using three error metrics, Peak Signal-to-Noise Rate (PSNR), Structural Similarity (SSIM), and Warping error. The quality of SR image can be analyzed using the PSNR and SSIM, where higher is better. Also, since the stereo image SR methods (Jeon et al. 2018; Wang et al. 2019a) are trained to infer SR left image only, we measured PSNR and SSIM on only SR left images for the fair comparison. The stereo consistency property can be analyzed using the warping error, which measures the mean square error between SR left and warped SR right images using the groundtruth dense disparity, so it can be measured only in Middlebury dataset.

The quantitative comparisons were shown in Table 1. In terms of PSNR and SSIM, compared to the single image



HR	SRCNN	VDSR	DRRN	LapSRN	StereoSR	PASSRnet	**Ours**
PSNR / SSIM	27.23 / 0.885	29.03 / 0.918	29.15 / 0.920	29.06 / 0.923	28.15 / 0.911	29.52 / 0.931	**29.71 / 0.932**
PSNR / SSIM	27.33 / 0.887	29.06 / 0.919	29.17 / 0.920	29.17 / 0.924	28.17 / 0.912	29.63 / 0.932	**29.77 / 0.933**

Figure 4: Qualitative results ($\times 4$) on image "Shelves" from Middlebury 2014 dataset.

HR	SRCNN	VDSR	DRRN	LapSRN	StereoSR	PASSRnet	**Ours**
PSNR / SSIM	23.69 / 0.787	24.90 / 0.834	25.266 / 0.844	24.98 / 0.840	24.04 / 0.817	25.96 / 0.870	**26.50 / 0.883**
PSNR / SSIM	23.09 / 0.752	24.12 / 0.799	24.83 / 0.816	24.14 / 0.803	23.304 / 0.781	24.96 / 0.833	**25.24 / 0.846**

Figure 5: Qualitative results ($\times 4$) on image "Validation 19" from Flickr 1024 dataset.

SR methods, it is natural that the stereo image SR methods outperform because they can utilize additional information. Among the stereo image SR methods, our method has shown the state-of-the-art performance in PSNR and SSIM. It demonstrates that our method can formulate the information well to recover the HR details well. Moreover, in terms of the warping error, our method shows the lowest score, which means that the SR stereo images obtained by other methods contain the inconsistent textures

Qualitative Evaluation. We qualitatively evaluate our method in Figure 4 for Middlebury (Scharstein and Pal 2007) dataset and Figure 5 for Flickr 1024 (Wang et al. 2019b) dataset. In Figure 4 and 5, we show the SR results of left image at the first row. In order to compare the quality of stereo-consistency, we represent magnified red regions of left SR results at second row, and corresponding blue regions of right SR results at third row in Figure 4 and 5. Super-resolved outputs of stereoSR and PASSRnet on right view are obtained by inverting an order of left and right inputs. From Figure 4 and 5, we can see that most of other SR models cannot generate consistent fine details of stereo pairs. In contrast, our model shows more consistent results and further recovers more accurate details closer to HR images. Specially, even if scale of texture in red and blue regions in Figure 4 is so small that it is hard to recover the texture, our model shows consistent results of left and right SR image. Moreover, in Figure 6, it can be observed that our model outperforms other methods on non-static scene like KITTI dataset (Geiger, Lenz, and Urtasun 2012) as well.

Ablation Study

We perform an ablation study to analyze how the component of SPAM works for the stereoscopic SR. From baseline method, which excludes SPAM, we include the parallax-attention map (+P), self-attention map (+P+S), and stereo consistency loss (+P+S+C) step by step. As shown in Table 2, by enriching the information using the parallax-attention (+P), we can obtain more high-quality results because the information from its counterpart can be added. However, by enforcing the parallax-attention in the feature-level naively, it is difficult to preserve the stereo-consistent property in SR stereo images. Also, adding the self-attention (+S) helps to improve the SR performance, but still it cannot produce the stereo-consistent results. Finally, we fig-

Table 1: Quantitative results (4×) with single image SR methods and stereo image SR methods on various datasets.

Error Metric	Dataset	Single Image SR				Stereo Image SR		
		SRCNN	VDSR	DRRN	LapSRN	StereoSR	PASSRnet	**Ours**
PSNR/SSIM	Middlebury	27.34 / 0.870	29.35 / 0.905	29.41 / 0.906	29.39 / 0.910	28.47 / 0.897	29.89 / 0.917	**30.04 / 0.921**
	Flickr 1024	22.01 / 0.715	22.73 / 0.752	22.751 / 0.755	22.71 / 0.756	22.53 / 0.723	23.21 / 0.779	**23.35 / 0.787**
	KITTI 2012	24.82 / 0.835	25.93 / 0.858	25.911 / 0.860	25.99 / 0.862	25.72 / 0.848	26.29 / 0.869	**26.34 / 0.873**
	KITTI 2015	23.39 / 0.818	24.36 / 0.846	24.291 / 0.848	24.40 / 0.850	24.01 / 0.824	24.77 / 0.861	**24.86 / 0.865**
Warping Error	Middlebury	0.000469	0.000433	0.000318	0.000321	0.000419	0.000291	**0.000259**

*Notes : The warping error is measured in only Middlebury dataset because it has the dense disparity groundtruth.
The higher PSNR and SSIM is better, and the lower warping error is better.

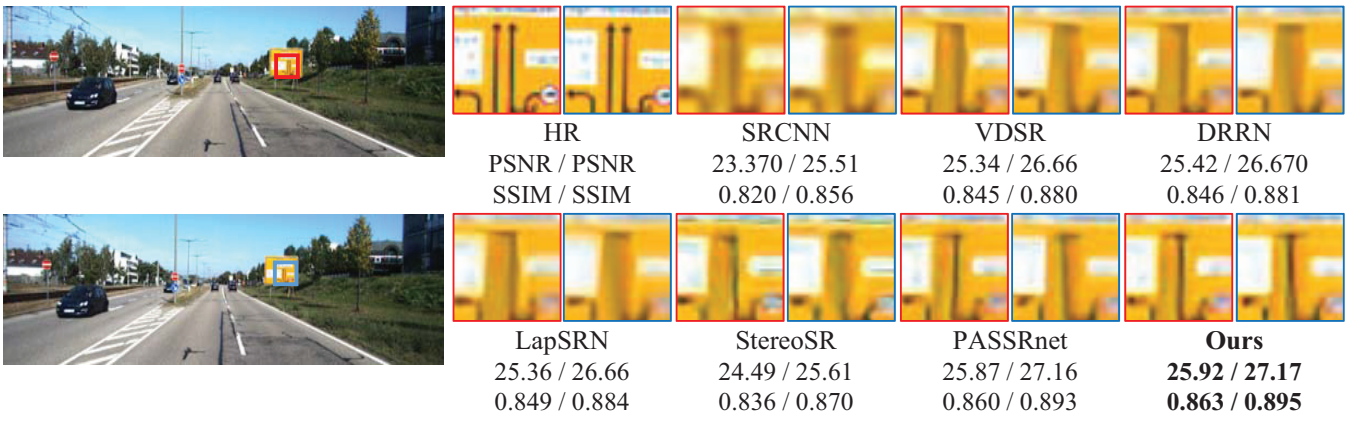


Figure 6: Qualitative results (×4) on image "testing 19" from KITTI 2015 dataset.

Table 2: Results (4×) of ablation study of SPAM on Middlebury dataset.

	Baseline	+P	+P+S	+P+S+C
PSNR	29.85	29.98	30.08	**30.11**
SSIM	0.901	0.919	0.924	**0.925**
Warping Error	0.000361	0.000328	0.000327	**0.000259**

ure out that applying our SPAM and stereo-consistency loss (+P+S+C) shows not only the best SR performance but also stereo-consistent SR images. It demonstrates that our SPAM and loss function are well organized mechanism for the stereoscopic SR tasks qualitatively. In Figure 7, we can find that the results of SPAM (+P+S+C) presents the well-reconstructed and stereo-consistent SR images.

Figure 7: Visual results of ablation studty. These results are achieved on "Midd1" of Middlebury 2006 dataset.

Conclusion

We present a novel stereoscopic super-resolution method by imposing stereo-consistency constraint on feature aggregation and training loss functions. Specifically, the proposed self and parallax attention mechanism (SPAM) in the feature aggregation part enables not only generating rich informative features for SR, but also imposing view-symmetric consistency on features for stereo-consistent SR. In addition, our method enforces the stereo-consistency in the loss function using the disparity driven by the parallax-attention mechanism in SPAM to produce the stereoscopic SR. We evaluated our proposed method on various datasets, which demonstrates that our method definitely outperforms other CNN-based image super-resolution methods.

Acknowledgments

This research was supported by R&D program for Advanced Integrated-intelligence for Identification (AIID) through the National Research Foundation of KOREA(NRF) funded by Ministry of Science and ICT (NRF-2018M3E3A1057289).

References

Brox, T.; Bruhn, A.; Papenberg, N.; and Weickert, J. 2004. High accuracy optical flow estimation based on a theory for

warping. In *ECCV*.

Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*.

Chen, D.; Yuan, L.; Liao, J.; Yu, N.; and Hua, G. 2018. Stereoscopic neural style transfer. In *CVPR*.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *CVPR*.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on Pattern Analysis and Machine Intelligence* 38(2):295–307.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*.

Godard, C.; Mac Aodha, O.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*.

Guo, H.; Liu, S.; Zhu, S.; and Zeng, B. 2016. Joint bundled camera paths for stereoscopic video stabilization. In *ICIP*.

Jeon, D. S.; Baek, S.-H.; Choi, I.; and Kim, M. H. 2018. Enhancing the spatial resolution of stereo images using a parallax prior. In *CVPR*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.

Joung, S.; Kim, S.; Park, K.; and Sohn, K. 2019. Unsupervised stereo matching using confidential correspondence consistency. *IEEE Transactions on Intelligent Transportation Systems*.

Kim, J.; Lee, J.; and Lee, K. 2016. Accurate image super-resolution using very deep convolutional networks. In *CVPR*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*.

Li, G.; He, X.; Zhang, W.; Chang, H.; Dong, L.; and Lin, L. 2018. Non-locally enhanced encoder-decoder network for single image de-raining. *arXiv preprint arXiv:1808.01491*.

Liao, R.; Tao, X.; Li, R.; Ma, Z.; and Jia, J. 2015. Video super-resolution via deep draft-ensemble learning. In *CVPR*.

Liu, D.; Wen, B.; Fan, Y.; Loy, C. C.; and Huang, T. S. 2018. Non-local recurrent network for image restoration. In *NeurIPS*.

Menze, M., and Geiger, A. 2015. Object scene flow for autonomous vehicles. In *CVPR*.

PyTorch. https://github.com/pytorch/pytorch.

Scharstein, D., and Pal, C. 2007. Learning conditional random fields for stereo. In *CVPR*.

Tai, Y.; Yang, J.; and Liu, X. 2017. Image super-resolution via deep recursive residual network. In *CVPR*.

Tao, X.; Gao, H.; Liao, R.; Wang, J.; and Jia, J. 2017. Detail-revealing deep video super-resolution. In *ICCV*.

Wang, L.; Jin, H.; Yang, R.; and Gong, M. 2008. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *CVPR*.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV*.

Wang, L.; Wang, Y.; Liang, Z.; Lin, Z.; Yang, J.; An, W.; and Guo, Y. 2019a. Learning parallax attention for stereo image super-resolution. In *CVPR*.

Wang, Y.; Wang, L.; Yang, J.; An, W.; and Guo, Y. 2019b. Flickr1024: A dataset for stereo image super-resolution. *arXiv preprint arXiv:1903.06332*.

Xu, L.; Jia, J.; and Matsushita, Y. 2011. Motion detail preserving optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(9):1744–1757.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2018a. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image super-resolution using very deep residual channel attention networks. In *ECCV*.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2018c. Residual dense network for image super-resolution. In *CVPR*.

Zhou, T.; Krahenbuhl, P.; Aubry, M.; Huang, Q.; and Efros, A. A. 2016. Learning dense correspondence via 3d-guided cycle consistency. In *CVPR*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*.