Disparity-Aware Domain Adaptation in Stereo Image Restoration

Bo Yan *, Chenxi Ma, Bahetiyaer Bare, Weimin Tan
Shanghai Key Laboratory
of Intelligent Information Processing,
School of Computer Science, Fudan University
{byan, cxma17, bahtiyarbari16, wmtan14}@fudan.edu.cn

Steven Hoi
Salesforce Research Asia
Singapore Management University
shoi@salesforce.com

Abstract

Under stereo settings, the problems of disparity estimation, stereo magnification and stereo-view synthesis have gathered wide attention. However, the limited image quality brings non-negligible difficulties in developing related applications and becomes the main bottleneck of stereo images. To the best of our knowledge, stereo image restoration is rarely studied. Towards this end, this paper analyses how to effectively explore disparity information, and proposes a unified stereo image restoration framework. The proposed framework explicitly learn the inherent pixel correspondence between stereo views and restores stereo image with the cross-view information at image and feature level. A Feature Modulation Dense Block (FMDB) is introduced to adaptively insert disparity prior throughout the whole network. The experiments in terms of efficiency, objective and perceptual quality, and the accuracy of depth estimation demonstrates the superiority of the proposed framework on various stereo image restoration tasks.

1. Introduction

With the rising interest in virtual and augmentation reality, stereo images are widely investigated in multiple computer vision fields from stereo magnification, stereo matching to depth estimation. In practice, the stereo images always suffer from various degradations. Unlike the active situation in other stereo image-related studies, the researches devoted to enhance the quality and practicality of stereo images are rarely mentioned. Hence, stereo image restoration is a promising study for its ability to release the inherent quality limitation of the degraded stereo images in research and applications.

An alternative solution to enhance stereo images is using single image restoration methods, which only exerts spatial statistics inside a degraded view and ignores the pixel registration between different views. So, restoring each image independently limits the performance, especially when predicting some details which are lost in one view but may exist in another view. Another way is multi-frame or video restoration methods, which extend the time dimension and utilize the supplementary information between adjacent frames. However, different from video captured at one viewpoint and different time points, the stereo image corresponds to different viewpoints at the same time. The pixel offsets in video and stereo images are caused by movement and parallax respectively, which makes the correlation in video differ from that in a stereo pair. The reference-based image restoration method is also not suitable to stereo images, *e.g.* RefSR [22] super-resolves a low-resolution (L-R) image with the help of high-resolution (HR) references. However the HR reference is difficult to obtain. As shown in Figure 1, the above methods are restricted to a planar scene and not fully applicable to stereo scene.

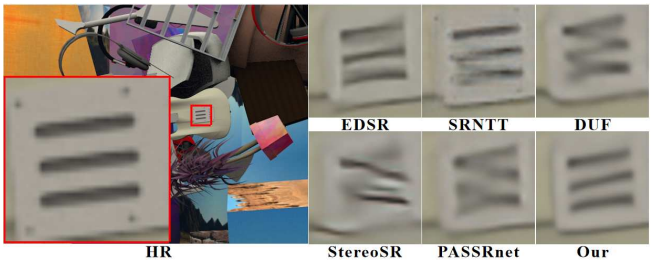

Figure 1: The $\times 4$ SR results on image "C0007" from FlyingThings3D. Compared to the state-of-the-art SISR (EDSR [11]), reference based SR (SRNTT [22]), video SR (DUF [6]), and stereo SR (StereoSR [5], PASSRnet [18]) works, the proposed approach synthesizes finer texture and restores accurate details without bringing in distortions.

From the stereo imaging process illustrated in Figure 2, these two views contain similar contents and serve as a reference to each other. The parallax refers to the inherent corresponding relationships between two views and provides sub-pixel offsets information, which is relative to pixel-wise

*This work was supported by NSFC (Grant No.: 61772137).

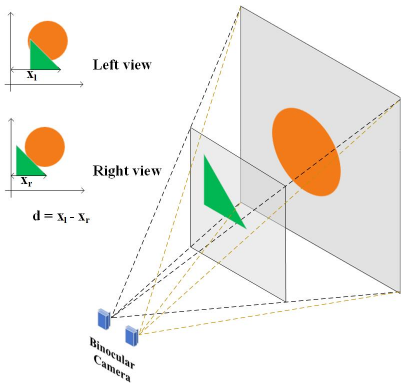

Figure 2: The illustration of stereo imaging process and visual representation of parallax.

information. When reconstructing one view, the disparity serves as a prior knowledge to make one view provide accurate reference to other view.

Recently, two deep learning based stereo image super-resolution (SR) methods are proposed to utilize the parallax. StereoSR [5] aligns the left and right views by horizontally offsetting all pixels in right view for a fixed number 64, without considering the parallax variation between different stereo images. PASSRnet [18] learns a parallax attention map to transfer the information from right to left view. However, these methods do not consider the role of disparity. Zhou *et al*. [23] proposed a stereo image deblurring network DAVNet, which estimates the disparity to align the features of two views. Though great breakthrough has been achieved, DAVNet does not fully combine the disparity prior into the whole pipeline. The influence of disparity information is only considered at one stage, that is the disparity is only utilized to warp the features of two views. Besides, these above methods all concentrate on improving the reconstruction performance at image level and neglect the more practical value of the stereo image, like the precision of disparity estimation.

Above analysis drives us to develop an end-to-end trainable stereo image restoration network (StereoIRN), which restores stereo images by fully exploring the disparity information and can be seamlessly integrated into the CNN of different stereo image tasks. The StereoIRN, composed of monocular network, disparity flow network and binocular network, captures the complicated dependency between two views and deploys the disparity prior into image restoration. Specifically, the monocular network restores each image by independently exploiting the spatial information of current view and transfers the information at image and feature levels to the binocular network. The disparity flow network utilizes multi-task learning to generate more suitable disparity prior via sharing the feature extraction layers with that of the monocular network, and registers the sub-pixel information to increase the correspondence between two views. Under the accurate guidance, the binocular network incorporates the output images and features to reconstruct stereo image details. The disparity prior is further inserted into the binocular network to refine the feature accuracy by the proposed Feature Modulation Dense Block (FMDB), which generates affine transformation parameters for spatial-wise feature modulation.

The main contributions are as follows:

- We propose a unified stereo image restoration framework (StereoIRN), guided by the feature continuity and disparity prior, to perceive the spatial and cross-view information simultaneously.
- We analyze the properties of disparity for stereo image restoration, explore how to exploit the stereo imaging nature, and propose a Feature Modulation Dense Block to refine the spatial feature by adaptively incorporating the information in disparity domain.
- We introduce two disparity attention losses, which encourage the solutions to improve the accuracy of disparity estimation.
- We are the first to evaluate the stereo image with the disparity estimation. The experiments at both image and disparity estimation levels demonstrate the proposed approach achieves state-of-the-art results.

2. Related Work

Deep Learning Based Image Restoration: SRCNN [2] first constructs a 3-layer CNN for single image super-resolution (SISR) and leads to a dramatic leap. Zhang *et al*. proposed a 20-layer DnCNN [20] to tackle SISR, image denoising, and JPEG deblocking simultaneously. Following their steps, plenty of image restoration researches achieve continuous breakthrough by improving the network structure. Huge upsurge has also been witnessed in video and multi-frame restoration. VSRNet [7] and [9] warps the adjacent frames onto the central frame to utilize the consecutive degraded frames

Stereo Image Restoration: Jeon *et al*. [5] first proposed StereoSR to super-resolve the left image through a luminance SR and a chrominance SR networks. They compensated the parallax by shifting the right image 64 pixels horizontally. However, their network assumes the parallax in a stereo image is fixed to 64 and all pixels share same parallax without considering the variation. PASSRnet [18] super-resolves the left image via a parallax-attention mechanism, which learns a mask to fuse the most similar features of two views to incorporate global correspondence in a stereo pair. However, PASSRnet exploits the pixel correlation based on two original views, which are mismatched at pixel-level and limits reference meaning between these two views. The

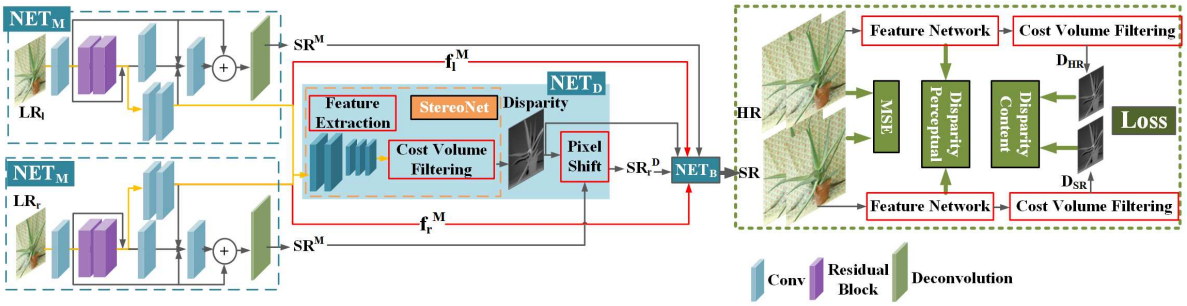

Figure 3: Overview of our StereoIRN, consisting of monocular network (NET_M), disparity flow network (NET_D) and binocular network (NET_B).

Left Right Disparity Aligned_Right

Figure 4: Visualization of warp process based on disparity. The left image, right image, disparity map, and warped right image respectively.

stereo image deblurring network, DAVNet [23], estimates bidirectional disparity based on blurry stereo images, aligns the features of two views, and fuses the features adaptively by learning a gate map. However the blurry stereo images limits the the disparity precision and the disparity information can be further exploited.

This work proposes a unified stereo image restoration framework to hunt for higher image quality and more accurate disparity estimation, which is new in literature.

3. Method

Before introducing the proposed method, we first analyze the stereo imaging process and explore the potential of disparity for recovering accurate image details.

3.1. Disparity prior analysis

The stereo imaging process, shown in Figure 2, illustrates the pixel relation between two views. The realistic scene is composed of multiple planes and the perspective projection on these planes between different viewpoints produces disparity. Hence, the disparity refers to apparent pixel difference or motion and represents the correspondences of pixel position between two views. We consider the disparity prior as a kind of knowledge, which can provide multiple sampling information with sub-pixel offsets to enhance the image quality.

Suppose $d(x, y)$ is the disparity of pixel (x, y) in left image I_l, we can calculate the corresponding position in right image I_r according to the relation of pixels between the left and right views represented as follow. Thus the disparity prior can assist to register the two views to provide more accurate information for stereo image reconstruction.

$$\begin{aligned} d(x, y) &= x_l - x_r, \\ I_l(x, y) &= I_r(x + d, y). \end{aligned} \tag{1}$$

3.2. Network

As shown in Figure 3, the proposed StereoIRN comprises monocular network (NET_M), disparity flow network (NET_D) and binocular network (NET_B). Specifically, the monocular network restores the spatial information for each view respectively and the disparity flow network learns the parallax and aligns different views to make sure the pixel accuracy. Combining the outputs of above sub-networks, the binocular network refines the final images by referring the cross-view information, exploiting outputs of previous networks in image and feature space, and incorporating the guidance of disparity. The importance of each component in StereoIRN will be investigated by performing ablation study later.

Monocular Network: As illustrated in Figure 3, we first restore each view independently and generate the corresponding feature by the monocular network (NET_M), the structure of which can be a common lightweight single image restoration network. For simplicity, we adopt a sequence of convolution layers and residual blocks [10] to extract and reconstruct the features of the degraded image, which is further delivered into two branches to reconstruct image and features respectively. The image reconstruction branch outputs image I_M by a convolution layer, which is replaced with a deconvolution layer to increase the spatial resolution of I_M for stereo image SR task. The feature reconstruction branch outputs feature f_M to extend the accuracy of the feature to the binocular network.

Disparity Flow Network: This section analyzes how to register the sub-pixel information by exploiting the disparity information. As discussed before, the pixel correlation between stereo views serves as a prior knowledge, which provides more accurate references for other view. Since

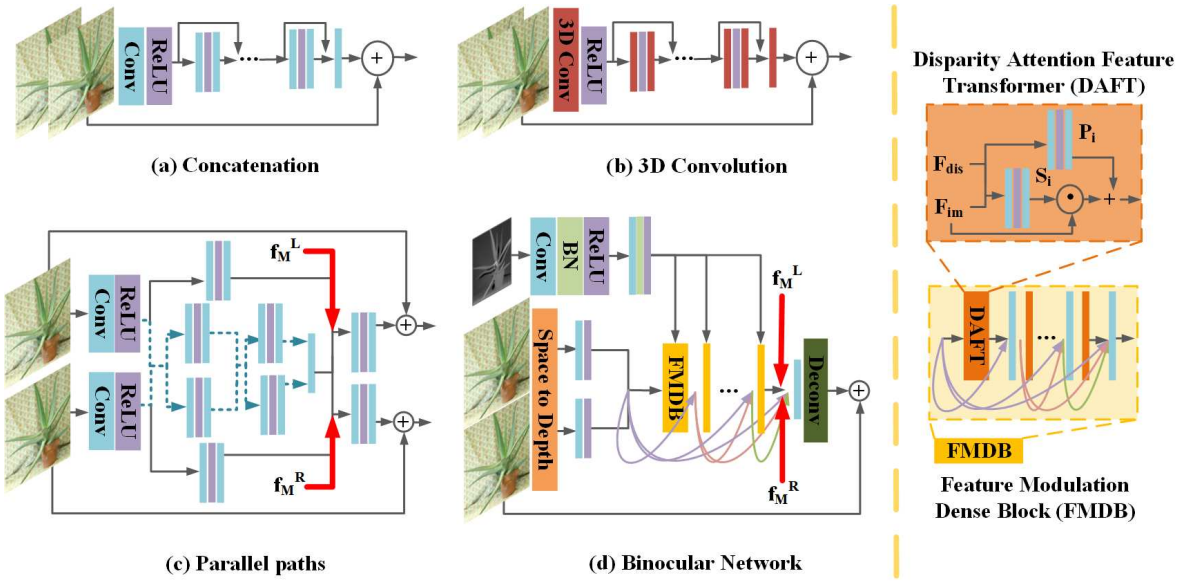

Figure 5: The different structures of the binocular network to explore the cross-view information. (a) Concatenates two views at the beginning directly. (b) 3D convolution. (c) Two parallel paths used to handle two views independently. (d) The network composed of our feature modulation dense block (FMDB).

disparity between different stereo images vary significantly, the disparity flow network NET_D is constructed to embed the disparity prior into the stereo image reconstruction.

For simplicity, the NET_D adopts the structure of StereoNet [8] to explicitly extract the disparity prior, called disparity flow. Inspired by multi-task learning strategy, we transfer the features of monocular network to the StereoNet instead of the stereo image. Thus, the StereoNet and the image reconstruction network share the convolution layers in feature extraction part. The shared features and representation of these networks can transfer the domain information between different tasks and improve generalization ability of networks. In our case, what is learned for disparity estimation task contribute to improve the quality of image restoration task by offering more pixel correspondence and vice versa.

Then, according to the disparity flow, we warp the right image I_M^R as the process illustrated in equation 1, and generate new stereo image pair (I_M^L, I_D^R), which are matched at pixel level and are shown in Figure 4.

Binocular Network: In addition the spatial information explored in the monocular network, the binocular network is proposed to incorporate the inter-viewpoints relationship for refining the image details. We first provide and compare different manners to exploit the stereo imaging nature, as shown in Figure 5, the performance of these methods will be illustrated in experiment.

To perceive the relationship between different views, the most simple and intuitional way is concatenating two views at the beginning of the network, as shown in Figure 5 (a). However, it is difficult for the network to learn the reference contents between two views. Since 3D convolution is proposed to extract features considering the inter-channel information, we replace the 2D convolution with 3D convolution and construct the network as shown in Figure 5 (b). The above two structures simultaneously treat two views without distinction by processing the two views through one path, which ignores the parallax between two views and seems little reasonable. Figure 5 (c) illustrates a two-path structure, which processes two views via parallel paths and transfers the features across these two views to better deal with the features of different views. Even if the features of two views are crossly delivered, these features offer limited positive effect to each other without utilizing the disparity prior. Besides, the two-path structure doubles the parameters and computation amounts.

To this end, we construct feature modulation dense block (FMDB) to incorporate the disparity prior into the whole binocular network and facilitate one single path to better utilize the cross-view nature. As shown in Figure 5 (d), the FMDB densely connects multiple disparity attention feature transformers (DFT), to enhance the guidance of disparity.The DFT learns a modulation parameter pair (s, p) based on image and disparity features (f_{im}^M, f_{dis}^M), which can adaptively influence the outputs by applying following affine transformation spatially to f_{im}^M.

$$DFT(f_{im}^M \mid s, p) = f_{im}^M \odot s + p. \tag{2}$$

The warp operation makes the shifted right view I_D^R contains inevitable pixel-wise error, which produces detrimen-

tal reference information to other view. Since feature maps contain more abundant information than a single image and keep consistent across the network. To mitigate these sub-pixel displacements and improve pixel accuracy, we explore the continuity of previous features by feeding f_M^L, f_M^R to the binocular network, which is represented as the red lines in Figure 3 and 5.

Our binocular network densely connects 6 FMDBs, which comprises 4 DFTs in total. The input image I_M^L is added to the output residual image. To reduce the computational cost with few accuracy loss, the spatial resolution of two images are decreased via a space-to-depth transformation and are increased via a deconvolution layer at the beginning and the end of our binocular network respectively. The kernel size of all convolution layers is 3×3. Since the optimal network structures of different tasks differ, the StereoIRN can deploy any architecture for each subnetwork to utilize the nature of different restoration tasks and amply its flexibility and the capacity.

Disparity Attention Loss: Most classic image restoration models can be formulated to solve the following problem :

$$x' = argmin_x \frac{1}{2\phi}|y - x|^2 + \lambda P(x), \tag{3}$$

where the first part $\frac{1}{2\phi}|y - x|^2$ is the data fidelity term, the second part P(x) is the regularization term. This equation only constrains the restored image *x* to be similar to the ground truth *y* at pixel level, without considering the whole structure and the global spatial consistency of the stereo image, which is critical for precise disparity estimation.

This observation motivates us to learn disparity perceptual constrain directly from the process of disparity estimation. In particular, we construct two disparity attention losses, including disparity content loss and disparity perceptual loss, on pixel level and feature level respectively to retrain our network with higher disparity precision.

The pixel-level disparity content loss $L_{dis_{acc}}$ is designed to push the restored stereo image similar to the natural stereo image manifold and to ensure the precision of disparity estimation. To achieve this, we constrain the similarity between the disparity $d_{SR}{}'$ generated on the restored stereo images and the ground truth disparity *d* by a two-parameter robust function [1] $\gamma(.)$, which approximates a smoothed L1 loss.

$$L_{dis_{acc}} = \gamma(d_{SR}{}' - d). \tag{4}$$

In addition, a more elegant constrain at feature level, named disparity perceptual loss L_{dis_p}, is introduced to make the restored images provide more accurate features for better disparity estimation and visual effect. The L_{dis_p} constrains the restored images to be similar to the ground truth images in feature space by minimizing the distance of the features at the middle layer of StereoNet.

$$L_{dis_p} = \|\Phi(HR) - \Phi(SR)||^2, \tag{5}$$

where Φ denotes the feature network in StereoNet. The proposed two disparity attention losses aid our StereoIRN to generate better results and disparity, which are hard to distinguish from real references.

Training Strategy: To achieve faster convergence and better performance, we employ a step-wise optimization to gradually train our models from easy to difficult. Specifically, we first train the monocular network and the disparity flow network with following constraints respectively:

$$\begin{aligned} L_{NET_M} &= ||SR_l^M - HR_l||^2 + ||SR_r^M - HR_r||^2, \\ L_{NET_D} &= \gamma(d' - d), \end{aligned} \tag{6}$$

where d' and d are the predicted and the ground truth disparity respectively.

Then all subnetworks are jointly updated, while fixing the parameters in NET_D.

$$\begin{aligned} L_{MSE} &= ||SR_l - HR_l||^2 + ||SR_r - HR_r||^2, \\ L_{all} &= \lambda_1 L_{MSE} + \lambda_2 L_{dis_{acc}} + \lambda_3 L_{dis_p}, \end{aligned} \tag{7}$$

where the parameter λ controls the contribution of different losses to our final loss and adopts 1 in our training process.

4. Experiments

4.1. Datasets and Training Settings

By following [5], all models are trained on 60 stereo pairs from the Middlebury dataset, the other 5 stereo images in which serve as testset, for different tasks. The disparity flow network is pretrained on SceneFlow dataset [12]. The training images are augmented by randomly down-scaling, flipping and rotating. We crop images into patches of size 80 and adopt 32 patches per batch. To train the SR model, we downscale the patches with scale factor 2, 3, 4. For denoising, we add additive white gaussian noise with noise level range [0, 40], to the clean patches. For deblurring, we convolve the clean image with blur kernel size 15 × 15 and σ sampled from [0.1, 4.0]. All models are trained on the machine with 2.20 GHz Intel (R) Xeon (R) CPU, and GTX1080Ti GPU (128G RAM) for 40 epochs with learning rate 1e-4. We adopt Adam optimizer with $\beta 1$ = 0.9, $\beta 2$ = 0.999, ε = 1e-8.

In addition to the Middlebury, the model, trained for stereo image SR task, is also evaluated on the first 15 images from Tsukuba [14], the first 20 images from KITTI2012 [3] and KITTI2015 [13] , and the A-000, B-000, C-000 sets of the FlyingThings3D subset in SceneFlow, including various disparities and occlusions. For simplicity, all results are calculated and demonstrated on the left view.

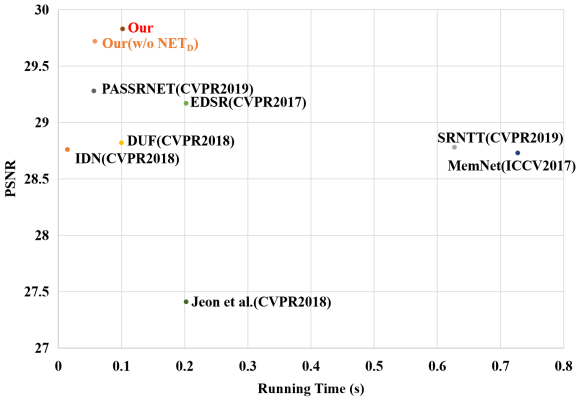

Figure 6: The trade-off between speed and accuracy on $\times 4$ SR task. The proposed model keeps a satisfactory balance between speed and accuracy.

model	NET_M	NET_B	feature	$PWCNet$	NET_D	PSNR/SSIM
NET_M	✓	×	×	×	×	28.987/0.8991
NET_B	×	✓	×	×	×	29.162/0.9016
NET_{MB}	✓	✓	×	×	×	29.423/0.9061
NET_{MB_f}	✓	✓	✓	×	×	29.722/**0.9104**
NET_{MBP_f}	✓	✓	✓	✓	×	29.625/0.9082
Our	✓	✓	✓	×	✓	**29.831**/0.9094

Table 1: Ablation study of different components of our network. Average PSNR/SSIM for $\times 4$ SR on Middlebury.

model	PSNR/SSIM	Parameters
Concatenation	29.16/0.902	936675
3D Convolution	27.79/0.879	942624
Parallel Paths	29.60/0.908	4598790
FMDB	**29.83/0.909**	1132932

Table 2: Comparison of structures of the binocular network.

MSE	$L_{dis_{acc}}$	L_{dis_p}	PSNR/SSIM	EPE
✓	×	×	**33.223/0.9612**	2.4982
✓	✓	×	32.757/0.9578	2.3665
✓	×	✓	32.673/0.9570	**2.2996**

Table 3: Ablation study of different losses. Average PSNR/SSIM and EPE for $\times 4$ SR on A-000 from SceneFlow.

4.2. Running Time

The computational efficiency analysis is conducted on stereo image SR task, and Figure 6 visualizes the comparison between the average running time for reconstructing a 640 × 480 HR stereo image pair from a 320 × 240 L-R stereo image pair and the reconstruction quality, which is represented by the PSNR for $\times 4$ upscaling on Middlebury. It is clear that the proposed algorithm has lower time complexity and maintains real-time when producing high-quality results.

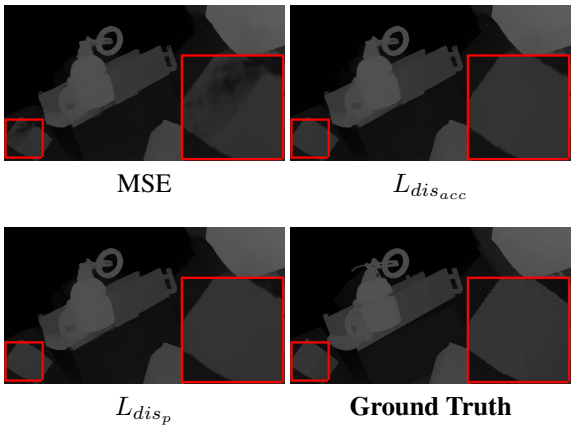

Figure 7: The disparity accuracy of different SR results.

4.3. Ablation Study

For clear illustration, all ablation studies are conducted on $\times 4$ stereo image SR task.

Network:

To illustrate the contribution of each component in our architecture and the feature transmission between NET_M and NET_B, we conduct experiment with different design options and report the results in Table 1. Similar to common SISR network, our monocular network NET_M super-resolves one image once without referring another view in stereo images. In the model without NET_M, the binocular network NET_B takes two LR images as input and removes the space-to-depth operation. The PSNR gains of NET_B over NET_M demonstrate that different views provide useful reference to each other even if only LR views are given. Compared to the LR reference, NET_M can provide better information for NET_{MB}.

Further improvement is obtained after adding the feature transmission between NET_M and NET_B, denoted as NET_{MB_f}, which indicates that the feature guidance works in correcting the deviation of pixel information. After enabling the disparity flow network, higher PSNR/SSIM are achieved, demonstrating that the disparity prior makes sense to recover more fine details. The final model, exploiting both dependently super-resolved views and features, obtains the best results.

To show the difference between the disparity and motion information, which is commonly considered in video restoration, we replace our disparity flow network with PWCNet [15]. And the results between NET_{MBP_f} and Our in Table 1 show that the disparity flow network captures more reliable stereo correspondence for image reconstruction.

To quantitatively compare the performance of several alternatives in exploiting the stereo imaging nature, which is discussed in Section 3.2, we construct different binocular

Dataset	Scale	SISR		Video SR		Ref_SR	Stereo SR			Our
		Bic	EDSR* [11]	SPMC [16]	DUF [6]	SRNTT [22]	StereoSR [5]	PASSRnet [18]	DAVNet* [23]	DASSR
Middlebury	s4	26.61/0.856	29.17/0.903	23.05/0.793	28.82/0.900	28.78/0.901	27.40/0.874	29.28/0.903	28.12/0.8829	**29.83/0.909**
	s3	28.35/0.900	32.03/0.944	-	-	-	30.37/0.926	-	-	**32.19/0.945**
	s2	31.49/0.949	36.14/0.976	28.88/0.929	-	-	34.28/0.967	-	-	**36.40/0.976**
KITTI 2012	s4	24.81/0.832	26.00/0.864	22.00/0.779	27.60/0.897	23.79/0.854	24.80/0.842	26.55/0.874	25.52/0.853	**26.96/0.882**
	s3	26.27/0.876	27.75/0.909	-	-	-	27.04/0.896	-	-	**27.97/0.909**
	s2	28.74/0.927	30.44/0.948	26.90/0.910	-	-	29.65/0.941	-	-	**30.73/0.950**
KITTI 2015	s4	23.37/0.814	24.41/0.852	20.34/0.741	25.14/0.871	24.20/0.858	23.15/0.823	24.97/0.865	24.11/0.842	**25.35/0.874**
	s3	24.92/0.867	25.87/0.902	-	-	-	25.49/0.889	-	-	**26.46/0.906**
	s2	27.46/0.928	28.90/0.949	25.41/0.905	-	-	28.09/0.941	-	-	**29.21/0.952**
Tsukuba	s4	30.83/0.923	34.03/0.959	26.73/0.869	33.68/0.957	33.51/0.956	30.92/0.934	34.52/0.962	31.66/0.943	**34.81/0.965**
	s3	33.23/0.955	37.11/0.980	-	-	-	36.05/0.975	-	-	**37.94/0.983**
	s2	37.36/0.982	43.87/0.995	33.49/0.968	-	-	41.88/0.993	-	-	**43.98/0.995**
SceneFlow	s4	29.29/0.916	31.59/0.945	25.36/0.867	24.63/0.853	31.55/0.944	29.46/0.922	32.22/0.951	30.85/0.935	**33.35/0.960**
	s3	31.18/0.946	34.622/0.971	-	-	-	33.53/0.965	-	-	**34.85/0.973**
	s2	34.31/0.974	38.78/0.989	30.97/0.950	-	-	37.81/0.987	-	-	**39.12/0.989**

Table 4: The average PSNR/SSIM comparisons between state-of-the-art SR methods on left images from benchmarks.

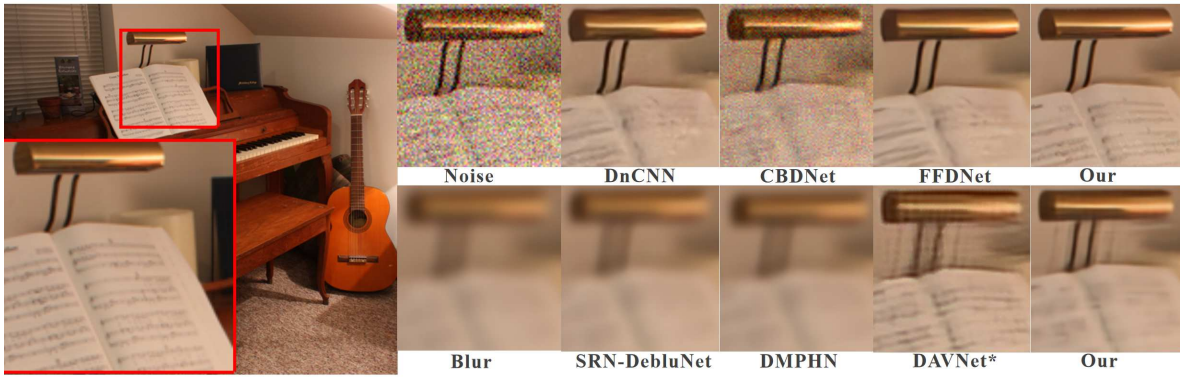

Figure 8: The qualitative comparison of two stereo image restoration tasks. The upper line denotes the denoising results with noise level 30 and the bottom line indicates the stereo image deblurring results with $\sigma = 3.6$.

networks and demonstrate results on SR task in Table 2.

As can be observed, the proposed FMDB yields best outputs. Naive input concatenation is not sufficient to exert the necessary cross-view information. Though good results are produced, parallel paths is not parameter efficient. 3D convolution cannot well handle the parallax existed in a stereo image. This supports the previous discussions.

Loss: One of our important contributions is introducing disparity attention losses at pixel and feature levels. To explore the performance of these losses, we train our models with different losses and show the PSNR/SSIM and end-point-error (EPE) in Table 3, which demonstrate the quantitative results and the disparity estimation accuracy respectively. The disparity results are visualized in Figure 7.

It is not surprising that the model optimized towards the MSE loss consistently achieves the best PSNR/SSIM. On the contrary, the accuracy of disparity estimation is gradually improved after adding disparity content loss and disparity perceptual loss. This is mainly because that PSNR/SSIM are calculated per-pixel and only show the similarity of the pixels without considering the global structure and the consistency correspondence between the stereo views. Both two disparity attention losses encourage images to generate reliable global and stereo experience, which is significant to disparity estimation.

4.4. Comparison with The State-of-The-Arts

To demonstrate the generalizability of StereoIRN, this section provides both quantitative and qualitative comparisons with various SR, denoising, and deblurring methods.

Stereo Image Super-Resolution For the stereo image SR task, the proposed model is compared with the state-of-the-art SISR (EDSR [11]), stereo image SR (StereoSR [5], PASSRnet [18]), RefSR (SRNTT [22]), and video SR methods (SPMC [16], DUF [6]). For fair comparison, we retrain the EDSR* with the same dataset as ours. We also retrain a DAVNet* for SR task by replacing the blurry inputs with the bicubic interpolated stereo images.

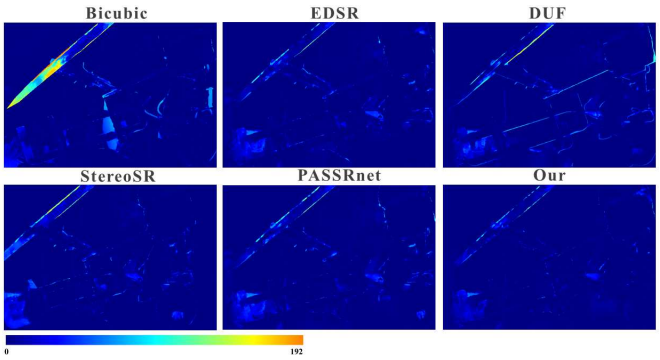

Figure 9: The absolute difference between the disparity of SR and that of HR stereo pairs.

σ	Blur	SRN-DeblurNet [17]	DAVNet*	Ours
1.3	29.06/0.916	29.38/0.92	32.36/0.955	**36.82/0.980**
2.0	26.43/0.857	26.92/0.868	30.10/0.927	**33.07/0.954**
3.6	23.82/0.776	24.48/0.795	27.54/0.872	**29.58/0.900**

Table 5: Deblurring comparison on Middlebury.

As depicted in Table 4, the proposed StereoIRN outperforms other methods by a large margin. PSNR results are improved by over 1 dB for $\times 4$ SR on SceneFlow compared with the second best one. The visual comparisons in Figure 1 show the super-resolved left images, indicating the proposed approach produces better structural details and more promising results.

Stereo Image Deblurring For gaussian deblurring task, our deblurring model is compared with state-of-the-art single image deblurring networks (SRN-DeblurNet [17], DMPHN [19]) and stereo image deblurring method (DAVNet*), retrained on our dataset. Table 5 provides quantitative evaluations on Middlebury with different blur kernel widths.

Stereo Image Denoising The stereo image denoising results of our and other state-of-the-art approaches (DnCNN [20], FFDNet [21], CBDNet [4]) on Middlebury are reported in Table 6, where the the obvious PSNR and SSIM gains of our model over the current best results indicate the advantages of our structure.

As the qualitative deblurring and denoising results depicted in Figure 8, the proposed StereoIRN can yield perceptually convincing outputs from both blur and noise condition. We can extend the proposed framework to other stereo image restoration tasks, such as deblocking, deraining, inpainting, and so on.

4.5. Disparity Perception

As discussed before, disparity estimation is a significant application of stereo images. However, the degraded stereo images, suffering from occlusions, noises and textureless regions, lead to significant artifacts in disparity estimation. Most impressively, besides the spatial precision, the richer information, which is provided by the restored stereo image, for accurate disparity estimation is also a main contribution of our method. To evaluate the disparity distortion, we measure the deviation between the disparity of restored results and that of clean stereo images (see Figure 9) and depict the end-point-error (EPE) in Table 7 for SR task. The EPE(HR) is calculated between the disparity of SR results and that of HR stereo pairs, and EPE(GT) is calculated between the disparity of SR results and the ground truth disparity respectively. Compared to state-of-the-art super-resolvers, the proposed model preserves the disparity of the new scene to be similar to the disparity of the original HR scene and leads to remarkably lower EPE.

noise	Noisy	DnCNN[20]	FFDNet [21]	CBDNet [4]	Ours
10	28.124/0.811	35.81/0.972	37.39/0.978	30.56/0.888	**39.12/0.985**
20	22.11/0.552	33.14/0.950	34.04/0.956	25.89/0.729	**36.38/0.973**
30	18.59/0.384	31.47/0.930	32.13/0.935	22.89/0.588	**34.73/0.964**

Table 6: Denoising comparison on Middlebury.

Models	EPE(HR)/EPE(GT)		
Dataset	A	B	C
HR	**0/2.238**	**0/1.098**	**0/2.624**
Bicubic	2.488/3.173	1.758/2.075	2.714/3.429
EDSR*	2.214/2.681	1.423/1.585	1.975/3.000
DUF	3.680/3.338	2.332/2.113	3.212/3.470
SRNTT	2.550/2.831	1.861/2.162	2.531/3.286
StereoSR	1.862/2.801	1.472/1.516	2.088/3.225
PASSRnet	2.851/2.724	1.532/1.588	2.153/2.983
Our	**2.248/2.498**	**1.296/1.432**	**1.917/2.889**

Table 7: The EPE of the disparity estimated on the super-resolved stereo pair from test sets in flyingthings3d dataset.

5. Conclusion

This work presents a unified stereo image restoration framework, composed of monocular, binocular, and disparity flow networks. The monocular and binocular network explore the spatial information and cross-view information to restore images respectively. To transfer the knowledge of disparity domain to image domain, the disparity flow network aligns two views to register the sub-pixel misplacement and the feature modulation dense block integrates the disparity prior into the whole pipeline. The experiment reveals the benefits of disparity to the stereo image restoration and evaluates the proposed approach in terms of reconstruction precision, efficiency and the accuracy of disparity estimation. The experimental results demonstrate the proposed approach achieves appealing performance over the state-of-the-arts on multiple stereo image restoration tasks.

References

[1] Jonathan T. Barron. A more general robust loss function. *ArXiv*, abs/1701.03077, 2017.

[2] Chao Dong, Change Loy Chen, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans Pattern Anal Mach Intell*, 38(2):295–307, 2014.

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, June 2012.

[4] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] D. S. Jeon, S. Baek, I. Choi, and M. H. Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1721–1730, June 2018.

[6] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos. Super-resolution of compressed videos using convolutional neural networks. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1150–1154, 2016.

[8] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 573–590, 2018.

[9] Tae Hyun Kim, Mehdi S. M. Sajjadi, Michael Hirsch, and Bernhard Schölkopf. Spatio-temporal transformer network for video restoration. In *Computer Vision – ECCV 2018*, pages 111–127, 2018.

[10] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, July 2017.

[11] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, volume 1, page 4, 2017.

[12] 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. arXiv:1512.02134.

[13] N. Mayer, E. Ilg, P. H?usser, 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, June 2016.

[14] M. Peris, S. Martull, A. Maki, Y. Ohkawa, and K. Fukui. Towards a simulation driven stereo vision system. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1038–1042, Nov 2012.

[15] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[16] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4472–4480, 2017.

[17] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018.

[18] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019.

[19] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2016.

[21] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for CNN based image denoising. *IEEE Transactions on Image Processing*, 2018.

[22] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7982–7991, 2019.

[23] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, Haozhe Xie, Jinshan Pan, and Jimmy Ren. Davanet: Stereo deblurring with view aggregation. In *CVPR*, 2019.