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 per-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 ×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
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
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 limit 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.

$$d(x, y) = x_L - x_R,$$
$$I_L(x, y) = I_R(x + d, y).$$

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.
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 NETD is constructed to embed the disparity prior into the stereo image reconstruction.

For simplicity, the NETD 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 $I^R_{LM}$ as the process illustrated in equation 1, and generate new stereo image pair $(I^L_{LM}, I^R_{LD})$, 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 $(f^M_{LM}, f^M_{dis})$, which can adaptively influence the outputs by applying following affine transformation spatially to $f^M_{LM}$:

$$DFT(f^M_{LM} | s, p) = f^M_{LM} ⊙ s + p. \quad (2)$$

The warp operation makes the shifted right view $I^R_{LD}$ 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 \( f_{M^l}^L \), \( f_{M^r}^R \) 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 \( I_{M^l}^L \) 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\( \times \)3. Since the optimal network structures of different tasks differ, the StereoIRN can deploy any architecture for each sub-network 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:

\[
x' = \arg\min_x \frac{1}{2\sigma^2} |y - x|^2 + \lambda P(x),
\]

where the first part \( \frac{1}{2\sigma^2} |y - x|^2 \) 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 \( \gamma(\cdot) \), which approximates a smoothed L1 loss.

\[
L_{dis_{acc}} = \gamma(d_{SR'} - d).
\]

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.

\[
L_{dis_p} = ||\Phi(HR) - \Phi(SR)||^2,
\]

where \( \Phi \) 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 monocural network and the disparity flow network with following constraints respectively:

\[
L_{NET_M} = ||SR_l^M - HR_l||^2 + ||SR_r^M - HR_r||^2;
L_{NET_D} = \gamma(d' - d),
\]

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

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

\[
L_{MSE} = ||SR_l - HR_l||^2 + ||SR_r - HR_r||^2;
L_{all} = \lambda_1 L_{MSE} + \lambda_2 L_{dis_{acc}} + \lambda_3 L_{dis_p},
\]

where the parameter \( \lambda \) 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 \(15 \times 15\) and \(\sigma\) sampled from \([0.1, 4.0]\). All models are trained on the machine with 2.20 GHz Intel (R) Xeon (R) CPU, and GTX1080 Ti GPU (128G RAM) for 40 epochs with learning rate 1e-4. We adopt Adam optimizer with \(\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1e^{-8}\).

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 ∗4 SR task. The proposed model keeps a satisfactory balance between speed and accuracy.

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

<table>
<thead>
<tr>
<th>model</th>
<th>NET_M</th>
<th>NET_B</th>
<th>feature</th>
<th>PWCNet</th>
<th>NET_B</th>
<th>PSNR/SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>NET_M</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>28.987/0.8991</td>
</tr>
<tr>
<td>NET_B</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>29.162/0.9016</td>
</tr>
<tr>
<td>NET_MB</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>29.423/0.9061</td>
</tr>
<tr>
<td>NET_MB_f</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>29.722/0.9104</td>
</tr>
<tr>
<td>Our</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>29.831/0.9094</td>
</tr>
</tbody>
</table>

Table 2: Comparison of structures of the binocular network.

<table>
<thead>
<tr>
<th>model</th>
<th>PSNR/SSIM</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concatenation</td>
<td>29.16/0.902</td>
<td>936675</td>
</tr>
<tr>
<td>3D Convolution</td>
<td>27.79/0.879</td>
<td>942624</td>
</tr>
<tr>
<td>Parallel Paths</td>
<td>29.60/0.908</td>
<td>4596790</td>
</tr>
<tr>
<td>FMDB</td>
<td>29.83/0.909</td>
<td>1132932</td>
</tr>
</tbody>
</table>

Table 3: Ablation study of different losses. Average P-SNR/SSIM and EPE for ∗4 SR on A-000 from SceneFlow.

<table>
<thead>
<tr>
<th>MSE</th>
<th>L_disacc</th>
<th>L_disp</th>
<th>PSNR/SSIM</th>
<th>EPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>32.22/0.9612</td>
<td>2.4982</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>32.75/0.9578</td>
<td>2.3665</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>32.67/0.9570</td>
<td>2.2996</td>
</tr>
</tbody>
</table>

4.3. Ablation Study

For clear illustration, all ablation studies are conducted on ∗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_MB_f and Our in Table 1 show that the disparity flow network captures more reliable stereo correspondence for image reconstruction.

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 LR stereo image pair and the reconstruction quality, which is represented by the PSNR for ∗4 upscaling on Middlebury. It is clear that the proposed algorithm has lower time complexity and maintains real-time when producing high-quality results.
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]), 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.

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scale</th>
<th>SISR</th>
<th>Video SR</th>
<th>Ref SR</th>
<th>Stereo SR</th>
<th>Our</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middlebury</td>
<td>s4</td>
<td>26.61/0.856</td>
<td>29.17/0.903</td>
<td>23.05/0.793</td>
<td>28.82/0.900</td>
<td>28.78/0.901</td>
</tr>
<tr>
<td></td>
<td>s3</td>
<td>26.35/0.900</td>
<td>32.03/0.944</td>
<td>-</td>
<td>-</td>
<td>30.37/0.926</td>
</tr>
<tr>
<td></td>
<td>s2</td>
<td>31.49/0.949</td>
<td>36.14/0.976</td>
<td>28.88/0.929</td>
<td>-</td>
<td>34.28/0.967</td>
</tr>
<tr>
<td>KITTI 2012</td>
<td>s4</td>
<td>24.81/0.832</td>
<td>26.00/0.864</td>
<td>22.00/0.779</td>
<td>27.60/0.897</td>
<td>23.79/0.854</td>
</tr>
<tr>
<td></td>
<td>s3</td>
<td>26.27/0.876</td>
<td>27.75/0.909</td>
<td>-</td>
<td>-</td>
<td>27.04/0.896</td>
</tr>
<tr>
<td></td>
<td>s2</td>
<td>28.74/0.927</td>
<td>30.44/0.948</td>
<td>26.90/0.910</td>
<td>-</td>
<td>29.65/0.941</td>
</tr>
<tr>
<td>KITTI 2015</td>
<td>s4</td>
<td>23.37/0.814</td>
<td>24.41/0.852</td>
<td>20.34/0.741</td>
<td>24.20/0.885</td>
<td>23.15/0.823</td>
</tr>
<tr>
<td></td>
<td>s3</td>
<td>24.92/0.867</td>
<td>25.87/0.902</td>
<td>-</td>
<td>-</td>
<td>25.49/0.889</td>
</tr>
<tr>
<td></td>
<td>s2</td>
<td>27.46/0.928</td>
<td>28.90/0.949</td>
<td>25.41/0.905</td>
<td>-</td>
<td>28.09/0.941</td>
</tr>
<tr>
<td>Tsukuba</td>
<td>s4</td>
<td>30.83/0.923</td>
<td>34.03/0.959</td>
<td>26.73/0.869</td>
<td>33.68/0.957</td>
<td>33.51/0.956</td>
</tr>
<tr>
<td></td>
<td>s3</td>
<td>33.23/0.955</td>
<td>37.11/0.980</td>
<td>-</td>
<td>-</td>
<td>36.05/0.975</td>
</tr>
<tr>
<td></td>
<td>s2</td>
<td>37.36/0.982</td>
<td>43.87/0.995</td>
<td>33.49/0.968</td>
<td>-</td>
<td>41.88/0.993</td>
</tr>
<tr>
<td>SceneFlow</td>
<td>s4</td>
<td>29.29/0.916</td>
<td>31.59/0.945</td>
<td>25.36/0.867</td>
<td>24.63/0.853</td>
<td>31.55/0.944</td>
</tr>
<tr>
<td></td>
<td>s3</td>
<td>31.18/0.946</td>
<td>34.62/0.971</td>
<td>-</td>
<td>-</td>
<td>33.53/0.965</td>
</tr>
<tr>
<td></td>
<td>s2</td>
<td>34.31/0.974</td>
<td>38.78/0.989</td>
<td>30.97/0.950</td>
<td>-</td>
<td>37.81/0.987</td>
</tr>
</tbody>
</table>

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 $\sigma = 3.6$. 

---

Table 4: The average PSNR/SSIM comparisons between state-of-the-art SR methods on left images from benchmarks.
Figure 9: The absolute difference between the disparity of SR and that of HR stereo pairs.

Table 5: Deblurring comparison on Middlebury.

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

Table 6: Denoising comparison on Middlebury.

<table>
<thead>
<tr>
<th>Models</th>
<th>EPE(HR)/EPE(GT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset A</td>
<td></td>
</tr>
<tr>
<td>Bicubic</td>
<td>2.488/3.173</td>
</tr>
<tr>
<td>EDSR*</td>
<td>2.214/2.681</td>
</tr>
<tr>
<td>DUF</td>
<td>3.680/3.338</td>
</tr>
<tr>
<td>SRNTT</td>
<td>2.550/2.831</td>
</tr>
<tr>
<td>StereoSR</td>
<td>1.862/2.801</td>
</tr>
<tr>
<td>PASSRnet</td>
<td>2.851/2.724</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>2.248/2.498</strong></td>
</tr>
</tbody>
</table>

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

Table 5: Deblurring comparison on Middlebury.

Table 6: Denoising comparison on Middlebury.

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