Learning Parallax Attention for Stereo Image Super-Resolution Supplemental Material

Longguang Wang[1], Yingqian Wang[1], Zhengfa Liang[2], Zaiping Lin[1], Jungang Yang[1], Wei An[1], Yulan Guo[1]*
[1]College of Electronic Science and Technology, National University of Defense Technology, China
[2]National Key Laboratory of Science and Technology on Blind Signal Processing, China
{wanglongguang15,yulan.guo}@nudt.edu.cn

Section 1 presents a detailed illustration of the proposed parallax-attention mechanism. Section 2 describes the details of the Flickr1024 dataset. Section 3 provides several additional analyses on the flexibility of our PASSRnet under different baselines and depths. Finally, Section 4 provides several additional comparative results between our PASSRnet and the state-of-the-art methods.

1. Parallax-attention Mechanism

1.1. Toy Example

The parallax-attention mechanism is illustrated with a toy example in Fig. 1. Given a stereo image pair $\mathbf{I}^{\mathrm{L}}_{left}$ and $\mathbf{I}^{\mathrm{L}}_{right}$ of size $\mathbb{R}^{30\times30}$, parallax-attention maps $\mathbf{M}_{left\rightarrow right}$ and $\mathbf{M}_{right\rightarrow left}$ of size $\mathbb{R}^{30\times30\times30}$ can be obtained by our parallax-attention module (PAM). Note that, each slice of the parallax-attention maps (*e.g.*, $\mathbf{M}_{right\rightarrow left}(i,:,:)$) delivers the dependency between corresponding rows (*i.e.*, $\mathbf{I}^{L}_{left}(i,:)$ and $\mathbf{I}^{L}_{right}(i,:)$). It can be observed from Fig. 1 (a) that parallax-attention maps are identity matrices if there is no disparity. That is because, the j^{th} pixel in $\mathbf{I}^{L}_{left}(i,:)$ corresponds to the j^{th} pixel in $\mathbf{I}^{L}_{right}(i,:)$. Therefore, position (j,j) in the parallax-attention map is focused on. For regions where disparities exist (*e.g.*, the red region in Fig. 1 (b) with disparity of 5), the j^{th} pixel in $\mathbf{I}^{L}_{left}(i,:)$ corresponds to the $(j-5)^{th}$ pixel in $\mathbf{I}^{L}_{right}(i,:)$. Therefore, position $(j,j-5)$ in the parallax-attention map is focused on. Consequently, stereo correspondence can be depicted by the positions of focused pixels in parallax-attention maps. Besides, occlusion can also be encoded. Specifically, it can be observed from Fig. 1 (c) that several horizontal regions are "discarded" without any position being focused on. That is because, these regions in $\mathbf{I}^{L}_{left}(i,:)$ are occluded in $\mathbf{I}^{L}_{right}(i,:)$, thus no correspondence should be focused on. Similar "discarded" vertical regions are also caused by occlusion. Moreover, occlusion can also be inferred from the cycle-attention maps.

It should be noted that, only integer disparities are considered in our toy example, which is not the real case. In practice, our PAM can focus on several adjacent pixels to address sub-pixel disparities. Due to the softmax layer used in PAM, several pixels in "discarded" horizontal regions may be incorrectly focused on. However, these occluded regions can be excluded using valid masks.

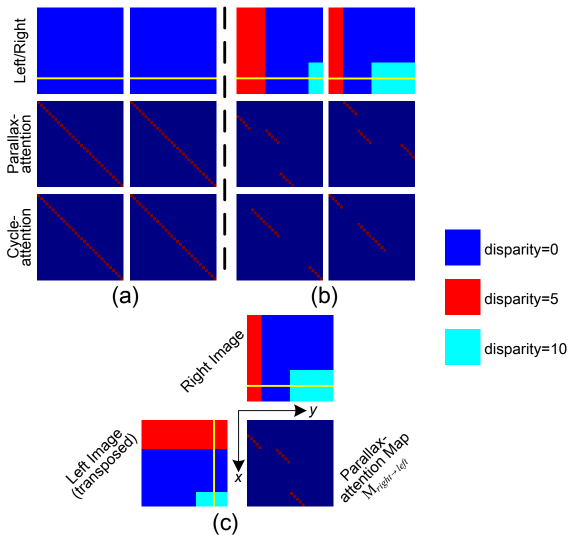
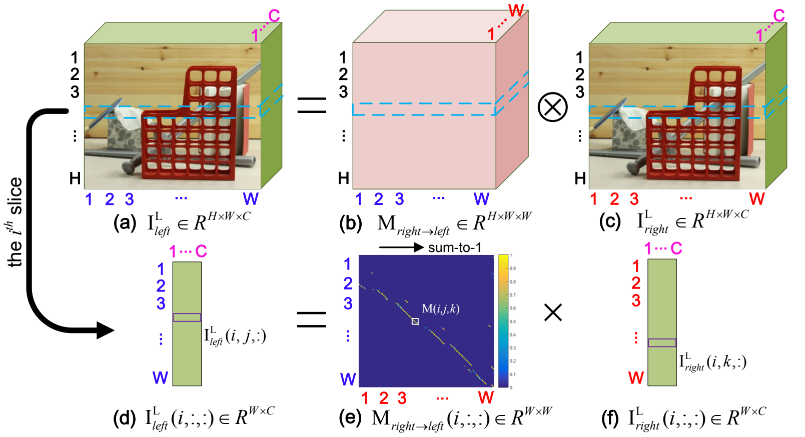


Figure 1: A toy example illustration of the parallax-attention and cycle-attention maps generated by our PAM. The attention maps (30×30) correspond to the regions (1×30) marked by a yellow stroke. In (a) and (b), the first row represents left/right stereo images, the second row stands for parallax-attention maps $\mathbf{M}_{right\rightarrow left}$ and $\mathbf{M}_{left\rightarrow right}$, and the last row represents cycle-attention maps $\mathbf{M}_{left\rightarrow right\rightarrow left}$ and $\mathbf{M}_{right\rightarrow left\rightarrow right}$.

1.2. Batch-wise Matrix Multiplication

$\otimes$ represents batch-wise matrix multiplication between two tensors[1]. Take Eq. (1) for an example (as shown

[1] $\otimes$ can be implemented using tf.matmul() or torch.matmul().


the i^{th} slice
1
2
3
⋮
H
1 2 3 ⋯ W
1 ⋯ C
1 ⋯ W
=
⊗
sum-to-1
M(i,j,k)
$\mathrm{I}^{L}_{left}(i,j,:)$
$\mathrm{I}^{L}_{right}(i,k,:)$
×

(a) $\mathrm{I}^{L}_{left} \in R^{H\times W\times C}$ (b) $\mathrm{M}_{right\rightarrow left} \in R^{H\times W\times W}$ (c) $\mathrm{I}^{L}_{right} \in R^{H\times W\times C}$

(d) $\mathrm{I}^{L}_{left}(i,:,:) \in R^{W\times C}$ (e) $\mathrm{M}_{right\rightarrow left}(i,:,:) \in R^{W\times W}$ (f) $\mathrm{I}^{L}_{right}(i,:,:) \in R^{W\times C}$

Figure 2: An illustration of batch-wise matrix multiplication $\otimes$.

in Fig. 2), the product of the i^{th} corresponding slices $\mathbf{M}_{right\rightarrow left}(i,:,:) \in \mathbb{R}^{W\times W}$ and $\mathbf{I}^{L}_{right}(i,:,:) \in \mathbb{R}^{W\times C}$ determines the i^{th} slice of $\mathbf{I}^{\mathrm{L}}_{left}$, *i.e.*, $\mathbf{I}^{L}_{left}(i,:,:) \in \mathbb{R}^{W\times C}$. All these slices are concatenated to obtain $\mathbf{I}^{\mathrm{L}}_{left} \in \mathbb{R}^{H\times W\times C}$.

1.3. Smoothness Loss

Take $\mathbf{M}_{right\rightarrow left}$ as an example, $\mathbf{M}_{right\rightarrow left}(i,j,k)$ measures the contribution of position (i,k) in $\mathbf{I}^{\mathrm{L}}_{right}$ to position (i,j) in $\mathbf{I}^{\mathrm{L}}_{left}$ using the similarity between them (*i.e.*, $S(\mathbf{I}^{L}_{left}(i,j),\mathbf{I}^{\mathrm{L}}_{right}(i,k))$). Our smoothness hypothesis argues that, $S(\mathbf{I}^{L}_{left}(i+1,j),\mathbf{I}^{L}_{right}(i+1,k))$ and $S(\mathbf{I}^{L}_{left}(i,j+1),\mathbf{I}^{L}_{right}(i,k+1))$ should be close to $S(\mathbf{I}^{L}_{left}(i,j),\mathbf{I}^{L}_{right}(i,k))$. That is, smoothness in correspondence (disparity) space can be encouraged.

2. The Flickr1024 Dataset

Although several stereo datasets such as Middlebury and KITTI are already available, these datasets are mainly proposed for stereo matching. Further, the Middlebury dataset only consists of close shots of man-made objects, while the KITTI 2012 and KITTI 2015 datasets only consist of road scenes. For stereo image super-resolution (SR) task, a large dataset which covers diverse scenes and consists of images with high quality and rich details is required. Therefore, we introduce a new Flickr1024 dataset for stereo image SR. The Flickr1024 dataset is available at: `https://yingqianwang.github.io/Flickr1024/`.

2.1. Data Collection

We manually collected 1024 RGB stereo image pairs from Flickr using tags such as stereophotography, stereoscopic and cross-eye 3D.

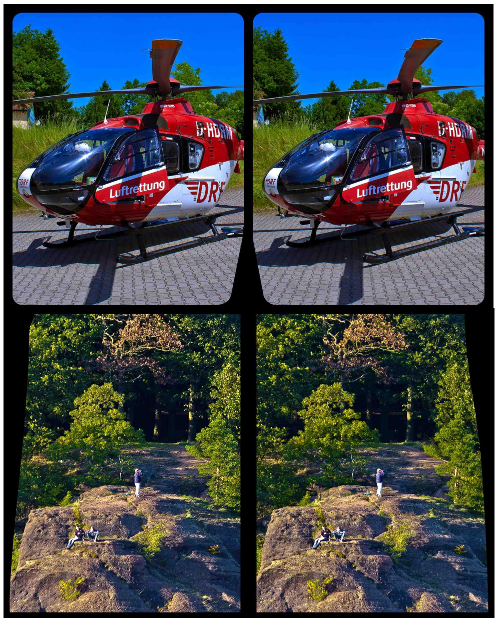


Figure 3: Two stereograms collected from Flickr.

2.2. Preprocessing

All of the 1024 images are taken by amateur photographers using dual lens or dual cameras. Since these images are stereograms (as shown in Fig. 3) provided by amateurs, prepocessing is required to generate our dataset. Specifically, we first cut each stereogram into an image pair and crop black margins. We exchange the left image and the right image of an image pair since the stereograms are provided in cross-eye mode. We then perform uncalibrated epipolar rectification and crop black margins again. Note that, the stereo image pairs are originally shifted to a common focus plane by the amateurs to produce a perception of 3D for viewers. In other words, both positive and negative disparities exist in the image pairs. Thus, we roughly shift these images back to ensure zero disparity corresponds to infinite depth. For close shots, since regions with infinite depth are unavailable, we just shift these images to make the minimum disparity over a threshold (empirically set to

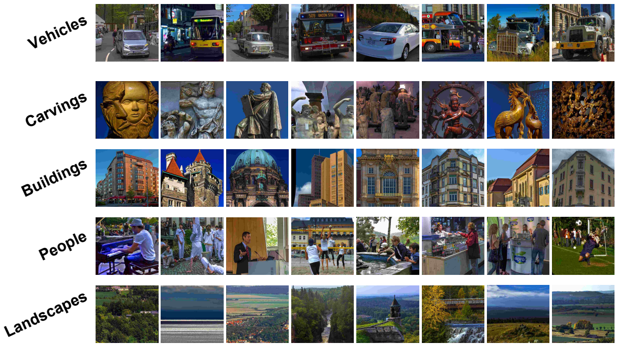


Figure 4: Samples of different scenes covered in the Flickr1024 dataset.

Table 1: Comparison between the Middlebury, KITTI 2012, KITTI 2015 and Flickr1024 datasets. Only the training sets of the KITTI 2012 and KITTI 2015 datasets are considered.

Dataset	image pairs	ppi	CNNIQA (↓)	entropy (↑)
Middlebury	65	3511605	20.18	**7.12**
KITTI 2012	194	462564	20.32	7.09
KITTI 2015	200	465573	22.86	7.02
Flickr1024	1024	800486	**19.75**	7.06

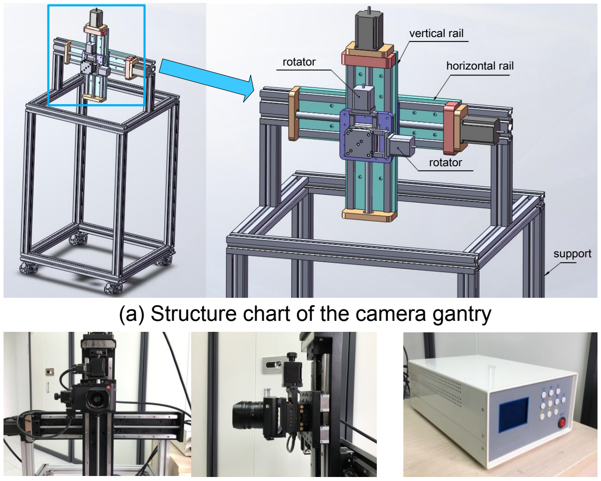
40 pixels in our dataset). Finally, we crop each resulting image to multiple of 12 pixels on both axes following [1].

2.3. Comparison to Existing Datasets

We compare our Flickr1024 dataset to three widely used stereo datasets including Middlebury, KITTI 2012 and KITTI 2015. Comparative results are shown in Table 1. It is clear that our Flickr1024 dataset is larger than other datasets by at least 5 times. Besides, the pixel per image (ppi) value of our Flickr1024 dataset is nearly 2 times that of the KITTI 2012 and KITTI 2015 datasets. Although the Middlebury dataset has the highest ppi value, the number of image pairs in this dataset is very limited. Further, the CNNIQA [2] and entropy of our Flickr1024 dataset is comparable to or even better than other datasets, which demonstrates the good image quality of our Flickr1024 dataset. Moreover, the Flickr1024 dataset covers a large diversity of scenes, including landscapes, urban scenes, people and man-made objects, as shown in Fig. 4.

3. Performance under Different Baselines and Depths

We furhter tested the flexibility of our PASSRnet and StereoSR [3] with respect to different baselines and depths.



(a) Structure chart of the camera gantry

(b) Real gantry with a camera

(c) Controller

Figure 5: An illustration of the device used for stereo image acquisition.

3.1. Baselines

We used a view-by-view scanning scheme as in [4] to obtain stereo image pairs with different baselines. The device used for image acquisition is shown in Fig. 5. Specifically, we first installed a camera on the gantry and acquired an image of a static scene. This image was referred to as the left image. Then, the camera was controlled to move rightward along the horizontal rail with different distances. The images acquired at different locations were referred to as right images with different baselines. Five different scenes were used for image acquisition. For each scene, we obtained three stereo image pairs with short baselines (with disparities around 20 pixels), medium baselines (with disparities around 60 pixels) and large baselines (with disparities around 180 pixels). Results achieved by StereoSR and our PASSRnet on these stereo image pairs with different baselines are shown in Table 2.

It can be observed that our PASSRnet outperforms StereoSR by 1.14 dB in terms of PSNR for stereo image pairs with short baselines. For stereo images pairs with large baselines, the improvement is increased to 1.30 dB. Compared to StereoSR, our PASSRnet effectively captures global correspondence for SR. Therefore, superior flexibility to disparity variations is achieved.

3.2. Depths

We collected 30 stereo image pairs from Flickr. These image pairs cover different scenes with different depths. We divided these image pairs into three groups with small depths (with disparities around 150 pixels), medium depths (with disparities around 80 pixels) and large depths (with disparities around 20 pixels). Results achieved by StereoSR

Table 2: Comparison between our PASSRnet and StereoSR [3] on stereo images with different baselines for 2× SR.

Baseline	StereoSR [3]		Ours	
	PSNR	SSIM	PSNR	SSIM
Large	37.13	0.9605	**38.43**(↑1.30)	**0.9690**(↑0.085)
Medium	37.35	0.9628	**38.50**(↑1.15)	**0.9692**(↑0.064)
Short	37.36	0.9628	**38.50**(↑1.14)	**0.9693**(↑0.065)

Table 3: Comparison between our PASSRnet and StereoSR [3] on stereo images with different depths for 2× SR.

Depth	StereoSR [3]		Ours	
	PSNR	SSIM	PSNR	SSIM
Small	37.60	0.9652	**39.03**(↑1.43)	**0.9749**(↑0.0097)
Medium	31.08	0.9145	**32.37**(↑1.29)	**0.9219**(↑0.0074)
Large	36.36	0.9596	**37.55**(↑1.19)	**0.9646**(↑0.0050)

and our PASSRnet on these stereo image pairs with different depths are shown in Table 3.

It can be observed that our PASSRnet achieves high improvement on stereo image pairs with small depths. That is because, the global receptive field of the parallax-attention mechanism facilitates our PASSRnet to capture global correspondence for performance improvement. In contrast, the fixed maximum disparity used in StereoSR hinders long-range correspondence to be employed. Therefore, the performance of StereoSR is limited.

4. Additional Visual Comparison

In this section, additional visual comparisons between our PASSRnet and the state-of-the-art methods are presented in Figs. 6 and 7. It can be observed that our PASSRnet recovers finer details with fewer artifacts, such as the stripe in Fig. 6 and the railings in Fig. 7. We further compare our PASSRnet with the state-of-the-arts on a stereo image pair acquired in our laboratory. The visual comparison is shown in Fig. 8. It can be observed from zoom-in regions that, the separate lines on the resolution test chart can be clearly distinguished in the SR images generated by our PASSRnet.

5. Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (No. 61602499 and No. 61605242), and Fundamental Research Funds for the Central Universities (No. 18lgzd06). The authors would like to thank *Sascha Becher* and *Tom Bentz* for the approval of using their cross-eye stereo photographs on *Flickr* to develop the Flickr1024 dataset.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR*, pages 1122–1131, 2017.

[2] Le Kang, Peng Ye, Yi Li, and David S. Doermann. Convolutional neural networks for no-reference image quality assessment. In *CVPR*, pages 1733–1740, 2014.

[3] Daniel S. Jeon, Seung-Hwan Baek, Inchang Choi, and Min H. Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *CVPR*, 2018.

[4] Yingqian Wang, Jungang Yang, Yulan Guo, Chao Xiao, and Wei An. Selective light field refocusing for camera arrays using bokeh rendering and superresolution. *IEEE Signal Processing Letters*, 26(1):204–208, 2019.

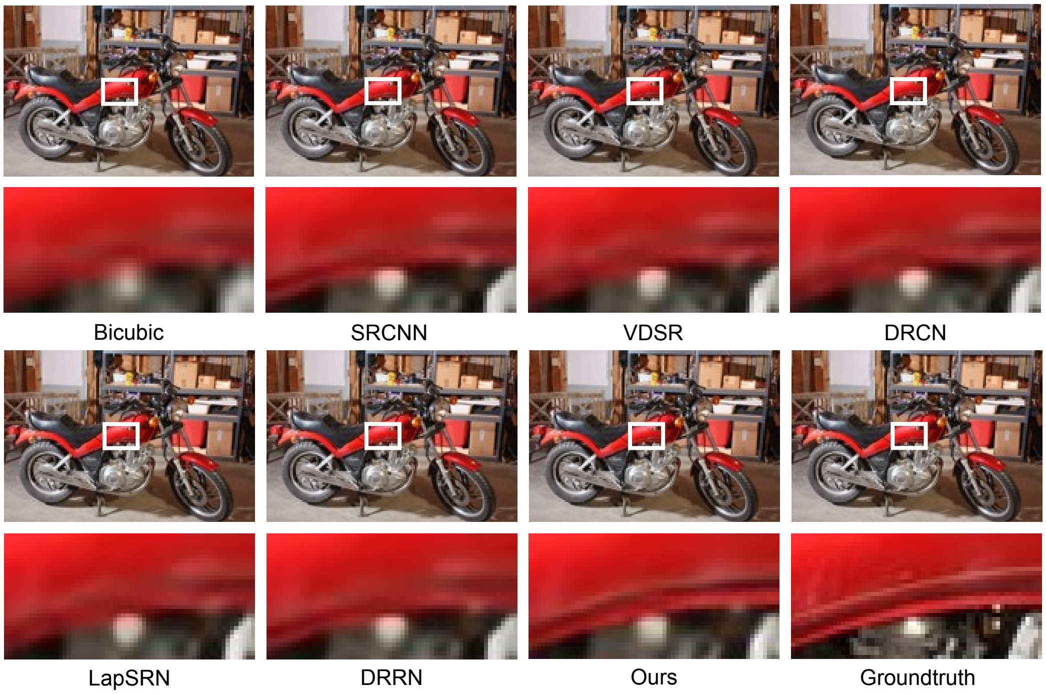


Figure 6: Visual comparison for 4× SR. These results are achieved on “Motorcycle” of the Middlebury dataset.

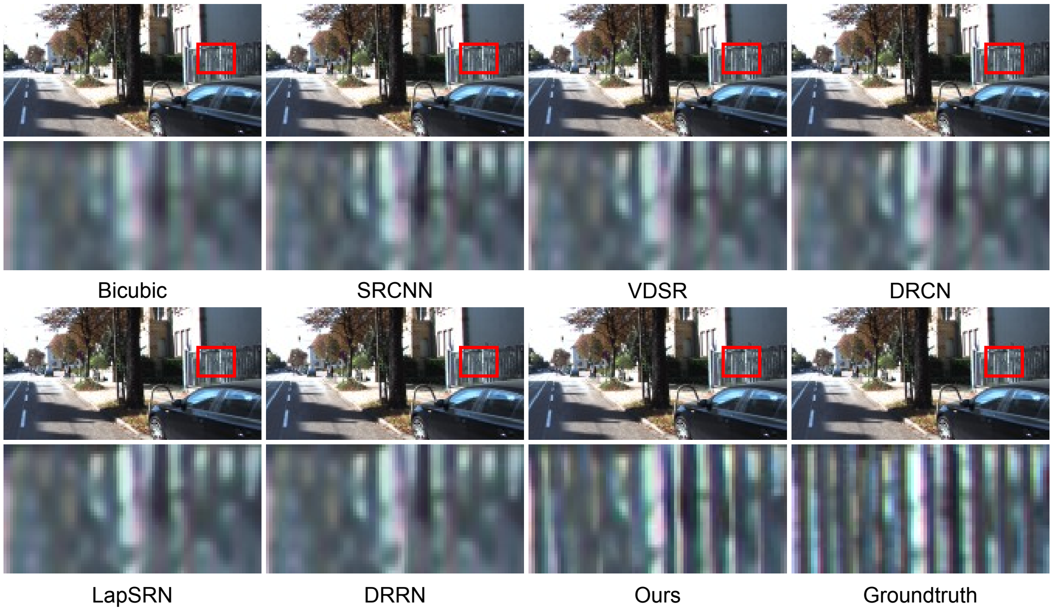
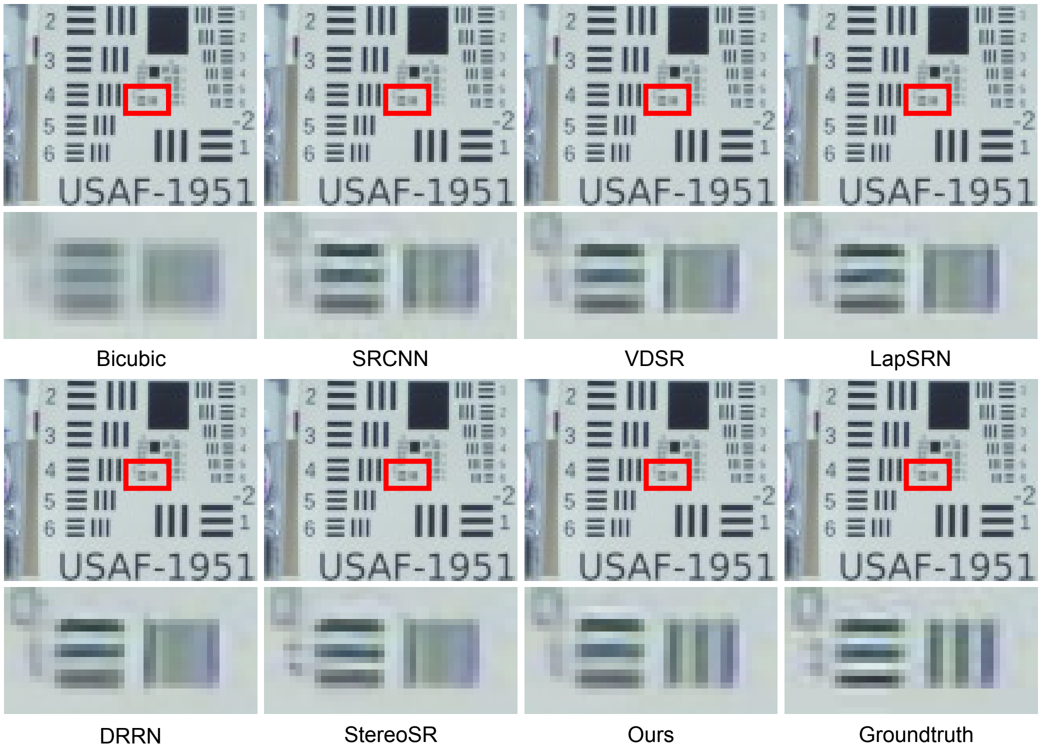


Figure 7: Visual comparison for 4× SR. These results are achieved on “test_image_004” of the KITTI 2015 dataset.



Bicubic SRCNN VDSR LapSRN

DRRN StereoSR Ours Groundtruth

Figure 8: Visual comparison for 2× SR. These results are achieved on a stereo image pair acquired in our laboratory.