Perception-Oriented Stereo Image Super-Resolution — Supplementary File

Anonymous Author(s)
Submission Id: 1393

ACM Reference Format:

Anonymous Author(s) and Submission Id: 1393. 2021. Perception-Oriented Stereo Image Super-Resolution — Supplementary File. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

This supplementary file first gives an example to illustrate the generalization of rankMOS in our StereoSRQA database, demonstrates the detailed structure of the StereoSR network, and further depicts more experimental results.

1 IMAGE QUALITY ASSESSMENT DATABASES

1.1 Existing Database

There are many image quality assessment (IQA) databases (such as LIVE [18], CSIQ [2], TID2008 [15] and TID2013 [6]), which mainly contain traditional distortion types, including Gaussian blur, motion blur, image compression, Gaussian noise, impulse noise, non eccentricity pattern noise, spatial warping, bilateral filter, comfort noise, and so on. These databases obtain the corresponding subjective scores, Mean Opinion Score (MOS) or differential Mean Opinion Score (DMOS), by laborious subjective experiment. Table 1 compares our database with several common IQA and StereoQA databases.

The subjective scores of distorted images, which requires many human judgement, limits the size of IQA database. Wu *et al.* [21] built a large-scale IQA database, which contains 10k reference image from MSCOCO and 21 distortion types, synthesizes a pseudo-MOS to represent the subjective quality of the distorted image. The pseudo-MOS is generated by following process: First, select the best FRIQA from five classical FRIQA metrics for each distortion type to compute a quality score for each image. Then, normalize the quality scores of all distorted images by building a nonlinear mapping function for each distortion type. There are also some perceptual similarity databases, PieAPP [16], and BAPPS [1].

To evaluate the image quality of SISR methods, Ma *et al.* [13] built a SISR quality assessment (SISRQA) database, which includes 30 reference images and 1620 super-resolved images from 9 SISR methods with different downsample factors and kernel width. Perceptual Image Processing Algorithms database (PIPAL) [7] contains

250 high-quality reference images and 29k distorted images corresponding 116 distortions, including traditional distortion types and several kinds of image restoration (IR) algorithms (traditional algorithms, GAN-based, PSNR-oriented methods), and employs the Elo rating system [17] to assign subjective scores.

As for the quality of stereo image, there are two publicly available 3D IQA databases, LIVE Phase I [14] and LIVE Phase II [3]. The LIVE I contains 20 reference stereo image pairs and 365 distorted stereo image pairs, which covers five symmetrical distortion types including JPEG, JPEG 2000 (JP2K), White Noise (WN), Gaussian Blur (Blur), and Fast-Fading (FF) with different distortion levels. The LIVE II, which also covers five distortion types, has 8 reference stereo images, 120 symmetrically distorted and 240 asymmetrically distorted stereo image pairs.

1.2 The proposed StereoSRQA database

The proposed StereoSRQA database generates the quality scores of super-resolved stereo images by combining three quality assessment methods to compute three quality scores for each stereo image pair. We sort the scores of all distorted versions for each reference stereo image pair, normalize the ranking scores, and merge them with a linear function to obtain the final quality score, rankMOS.

Here, we give a simple example to illustrate the generation of rankMOS. Supposing there are three voters and three SR methods for simplicity. The rankMOS of Image A can be obtained by following process. First, three voters predict three score for SR results (SR_1, SR_2, SR_3) of image A respectively, as shown in table 2.

Table 2: Original scores of different voters. This table is a simplified example, where three quality evaluation metrics, PSNR, IFC, and EPE denote three voters respectively. These three voters evaluate each super-resolved result of image A and predict a score.

image A	Voter#1 SISRQA $\uparrow$	Voter#2 StereoQA $\downarrow$	Voter#3 EPE $\downarrow$
SR_1	0.2	20	3
SR_2	0.3	40	6
SR_3	0.1	30	5

Then, as shown in table 3, we sort the scores of all SR results for each voter and assign the ranking ($Rank_n$ denotes the score ranking of current SR image is n -th among all SR images) as the new score S_j . Next, we average ranking of all voters for each SR image and get the averaged ranking score (RS), and the rank-MOS are obtained by normalizing the RS.

2 NETWORK STRUCTURE

The StereoSR network adopts the structure of DASSR [22], which consists of monocular network (NET_M), disparity flow network

Permission to make digital or hard copies of all or part of this work for personal or commercial use, or to republish, is granted by ACM, provided that the copyright notice, this permission notice, and the full citation on the first page are included. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference '17, July 2017, Washington, DC, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM, \$15.00
<https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

Table 1: Database comparison. We include the outputs of SISR and StereoSR algorithms as a novel distortion type. Traditional distortion types(Gaussian blur, motion blur, image compression, Gaussian noise, spatial warping, bilateral filter, comfort noise) and results of various image restoration (IR) methods, including SISR, denoising, and the mixture of denoising and SISR.

database	image type	reference image	distorted image	distortion type	Judgment type
LIVE	image	29	0.8k	5 [traditional]	MOS
CSIQ	image	30	0.8k	6 [traditional]	MOS
TID2008	image	25	1.7k	17 [traditional]	MOS(Swiss system)
TID2013	image	25	3.0k	24 [traditional]	MOS(Swiss system)
PieAPP	patch (256×256)	200	20.3k	75 [traditional]	Prob. of Preference
BAPPS	patch (256×256)	187.7k	375.4k	425[traditional]	Prob. of Preference
PIPAL [7]	patch (288×288)	250	29k	40 [IR methods, traditional]	MOS(Elo rating system)
Ma <i>et al.</i> [13]	image	30	1.62k	9 [SISR methods]	MOS(average)
Wu <i>et al.</i> [21]	image	10k	1050k	21 [traditional]	pseudo-MOS
LIVE I	stereo image pair	20	0.365k	5 [traditional]	DMOS
LIVE II	stereo image pair	8	0.36k	55 [traditional]	DMOS
Our	stereo image pair	66	3.366k	13 [SISR methods, StereoSR methods]	rankMOS

Table 3: The generalization of our rank-MOS.

	ordering			merging		
	Voter#1 SISRQA	Voter#2 StereoQA	Voter#3 EPE	Average (RS)	Normalize	rankMOS
SR_1	$Rank_2$	$Rank_1$	$Rank_1$	4/3	0.4444	4.444
SR_2	$Rank_1$	$Rank_2$	$Rank_3$	2	0.6667	6.667
SR_3	$Rank_3$	$Rank_3$	$Rank_2$	8/3	0.8889	8.889

Figure 1: Structure of the StereoSR network.

(NET_D) and binocular network (NET_B). As shown in Fig. 1, two

monocular networks, similar to common SISR networks, respectively super-resolve the left and right views by exploiting the dependent

spatial features inside each view. Then, the disparity flow network learns the parallax based on two roughly super-resolved stereo views (SR_L^M, SR_R^M) and aligns different views to the same viewpoint. The binocular network restores final stereo images by incorporating spatial information from NET_M and the information of adjacent view from NET_D . The detailed dynamic flow is demonstrated as follows:

Firstly, given a LR stereo image pair (LR_L, LR_R), two monocular networks respectively super-resolve two views. The main branch of the monocular network is composed of three convolution layers, two residual blocks [11], a deconvolution layer, and multiple skip connections. In addition, two convolution layers after the last residual block are added to extract the middle feature f^M , which is utilized to transfer spatial information from the monocular network to the following binocular network. At this point, we obtain two super-resolved views (SR_L^M, SR_R^M) and two middle features (f_L^M, f_R^M) from the two monocular networks.

Secondly, to estimate the disparity d between the left and right views, we adopt the disparity estimation network StereoNet [8] for simplicity. We warp the right view (SR_R^M) to left viewpoint based on d by regarding the disparity as a special kind of optical flow. Supposing $d(x, y)$ is the disparity of the left image $I_l(x, y)$ relative to the right image $I_r(x, y)$. The right view can be warped to the left viewpoint by replacing the pixel (x, y) in the right image with the pixel $(x + d, y)$, which is the inverse of the formation of disparity.

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

The aligned right view SR_R^w is concatenated with SR_L^M to get a new left view ($SR_L^D = [SR_L^M, SR_R^w]$). As demonstrated in [9], the human brain will fuse the stereo image pair into a single view, cyclopean image, before processing it. The concatenation of left view and the disparity-matched right view can be consider a cyclopean image, which can accurately simulate the processing of brain to a stereo image pair. The new right view (SR_R^D) can be obtained in a similar way. In this way, two views are both enhanced by adding an assistant image from another view, which provides more abundant information.

Thirdly, the SR_L^D and SR_R^D are transferred to two paths of the binocular network, where each path contains a space-to-depth transformation, two convolution layers and a residual block [11] to extract features (f^L, f^R). In order to deliver some shared information across two views, middle features (f_L^L, f_R^L) of the first convolution layers in two paths are concatenated with each other and passed to another two convolution layers followed with two residual blocks. The features from the two residual blocks are further concatenated and delivered to another two convolution layers and residual blocks, the output features of which are combined with a convolution layer to obtain a new feature f^{mid} . Then, the f^{mid} is concatenated with f^L (f^R) and f_L^M (f_R^M). Finally, two residual blocks, two convolution layers, and two deconvolution layers with residual learning are adopted at the end to reconstruct two views (SR_L, SR_R).

All convolution layers adopt kernel size 3 with 64 feature map, and the deconvolution layers adopt kernel size 8.

3 EXPERIMENTAL RESULTS

Figure 2 compares different constraints, used to optimize the StereoSR model. More comparisons with state-of-the-ars SR methods (including EDSR [11], SRFBN [10], DRN [4], ESRGAN [20], SP-SR [12], StereoSR [5], PASSRnet [19], DASSR [22]) are shown in Figure 3.

Figure 2: Visual comparison of StereoSR models trained with different losses. Better zoom in.

Table 4 demonstrates the results of PSNR/SSIM.

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

Database	EDSR*	SRFBN	SPSR	StereoSR	PASSRnet	DASSR	PSSR
Middlebury	29.17/0.903	29.32/0.903	27.53/0.869	27.40/0.874	29.28/0.903	29.83/0.909	29.27/0.902
	32.03/0.944	31.86/0.941	-	30.37/0.926	-	32.19/0.945	30.85/0.934
	36.14/0.976	36.05/0.975	-	34.28/0.967	-	36.40/0.976	34.64/0.969
KITTI 2012	26.00/0.864	26.52/0.874	24.71/0.833	24.80/0.842	26.55/0.874	26.96/0.882	26.19/0.869
	27.75/0.909	28.07/0.910	-	27.04/0.896	-	27.97/0.909	27.81/0.906
	30.44/0.948	30.83/0.951	-	29.65/0.941	-	30.73/0.950	30.42/0.947
KITTI 2015	24.41/0.852	24.93/0.865	22.77/0.807	23.15/0.823	24.97/0.865	25.35/0.874	24.68/0.858
	25.87/0.902	26.56/0.908	-	25.49/0.889	-	26.46/0.906	26.36/0.903
	28.90/0.949	29.28/0.953	-	28.09/0.941	-	29.21/0.952	28.84/0.948
Tsukuba	34.03/0.959	34.66/0.963	32.91/0.952	30.92/0.934	34.52/0.962	34.81/0.965	33.93/0.958
	37.11/0.980	37.91/0.982	-	36.05/0.975	-	37.94/0.983	36.32/0.976
	43.87/0.995	43.84/0.995	-	41.88/0.993	-	43.98/0.995	41.99/0.993
SceneFlow	31.59/0.945	32.63/0.954	30.62/0.931	29.46/0.922	32.22/0.951	33.35/0.960	32.04/0.950
	34.62/0.971	35.05/0.973	-	33.53/0.965	-	34.85/0.973	33.75/0.968
	38.78/0.989	39.20/0.989	-	37.81/0.987	-	39.12/0.989	37.40/0.985

Ablation study of the network design We construct different StereoSRQA models and demonstrate their performance in Table 5, where ‘NRStereoSRQA’ and ‘FRStereoSRQA’ represent a no-reference and a full-reference StereoSRQA models with the same network structure. The difference between the two models is the input, which is the degraded stereo image pair for the ‘NRStereoSRQA’ and the difference map between the degraded and the groundtruth stereo image pair for the ‘FRStereoSRQA’. To illustrate whether the stereo information is helpful for evaluating the StereoSR quality, we also build a no-reference SISR quality assessment model ‘NRSISRQA’, the structure of which is similar to one branch of our StereoSRQA model, by connecting several convolution layers, pooling layers, and a fully connected layer. Specifically, it outputs a score for each view independently and averages two scores to get the score

Figure 3: Comparison with state-of-the-art SR methods, including MSE-based SISR methods (EDSR [11], SRFBN [10], DRN [4]), GAN-based SISR (ESRGAN [20], SPSR [12]), and StereoSR (StereoSR [5], PASSRnet [19], DASSR [22]). Better zoom in.

Table 5: Validation of different designs for stereo image super-resolution quality assessment model.

	Stereo	reference	SROCC↑	PLCC↑	KROCC↑	RMSE↓
NRSISRQA	×	×	0.9598	0.9594	0.8275	0.7758
NRStereoSRQA	✓	×	0.9743	0.9738	0.8624	0.6249
FRStereoSRQA	✓	✓	0.9910	0.9918	0.9170	0.3505

of stereo image. From Table 5, ‘NRStereoSRQA’ and ‘FRStereoSRQA’ can effectively extract and utilize the stereo information and perform better than ‘NRSISRQA’.

Table 6 studies the weight of our IQP loss functions, where λ_0 keeps to be 1, and λ_1 equals to λ_2 . We can observe that less λ_1, λ_2 make the model tend to generate higher PSNR/SSIM, and larger λ_1, λ_2 will also degrades the subjective quality of the results for it limits the pixel constraint and brings some checkerboard artifacts. Therefore, a moderate weight is of vital importance to the final performance.

Running Efficiency Since only the StereoSR model is needed in inference stage, the running time of the proposed PSSR depends on the specific StereoSR model.

Table 6: Study of weight of IQP loss (λ_1, λ_2), $\lambda_0 = 1$.

$\lambda_1=\lambda_2$	PSNR	SSIM	SISRQA	StereoQA	StereoSRQA
0.01	29.83	0.909	0.61	21.02	6.75
0.1	29.10	0.901	0.60	18.24	8.73
1	28.60	0.890	0.58	18.39	6.77

REFERENCES

- [1] [n.d.]. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. ([n.d.]).
- [2] Larson Damon M. Chandler. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging* 19, 1 (2010), 011006.
- [3] Ming Jun Chen, Che Chun Su, Do Kyoung Kwon, Lawrence K. Cormack, and Alan C. Bovik. 2013. Full-reference quality assessment of stereopairs accounting for rivalry. *Image Communication* (2013).
- [4] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhong Cao, Zeshuai Deng, Yanwu Xu, and Minghui Tan. 2020. Closed-loop Matters: Dual Regression Networks for Single Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- [5] D. S. Jeon, S. Baek, I. Choi, and M. H. Kim. 2018. Enhancing the Spatial Resolution of Stereo Images Using a Parallax Prior. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1721–1730.
- [6] Jin, Lina, Lukin, Vladimir, Astola, Jaakko, Kuo, C., C., and Jay and. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing Image Communication A Publication of the the European Association for Signal*

- Processing (2015).
- [7] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S. Ren, and Dong Chao. 2020. PIPAL: A Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration. In *European Conference on Computer Vision*.
- [8] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. 2018. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision*. 573–590.
- [9] S. Li, X. Han, and Y. Chang. 2019. Adaptive Cyclopean Image-Based Stereoscopic Image-Quality Assessment Using Ensemble Learning. *IEEE Transactions on Multimedia* 21, 10 (2019), 2616–2624.
- [10] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu. 2019. Feedback Network for Image Super-Resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3862–3871.
- [11] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *The IEEE conference on computer vision and pattern recognition workshops*, Vol. 1. 4.
- [12] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. 2020. Structure-Preserving Super Resolution with Gradient Guidance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [13] Chao Ma, Chih Yuan Yang, Xiaokang Yang, and Ming Hsuan Yang. 2017. Learning a No-Reference Quality Metric for Single-Image Super-Resolution. *Computer Vision & Image Understanding* 158 (2017), 1–16.
- [14] Anush Krishna Moorthy, Che Chun Su, Anish Mittal, and Alan Conrad Bovik. 2013. Subjective evaluation of stereoscopic image quality. *Signal Processing Image Communication* 28, 8 (2013), 870–883.
- [15] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. 2009. TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics. *Advances of Modern Radioelectronics* 10 (01 2009), 30–45.
- [16] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. 2018. PieAPP: Perceptual Image-Error Assessment through Pairwise Preference. In *Computer Vision and Pattern Recognition*.
- [17] Arco Pub. 1978. Elo, A.E.: The rating of chessplayers, past and present. (1978).
- [18] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. 2006. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing* 15, 11 (2006), 3440–3451.
- [19] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. 2019. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 12250–12259.
- [20] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. *CoRR abs/1809.00219* (2018).
- [21] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin. 2020. End-to-End Blind Image Quality Prediction With Cascaded Deep Neural Network. *IEEE Transactions on Image Processing* 29 (2020), 7414–7426.
- [22] B. Yan, C. Ma, B. Bare, W. Tan, and S. C. H. Hoi. 2020. Disparity-Aware Domain Adaptation in Stereo Image Restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13176–13184.