Perception-Oriented Stereo Image Super-Resolution

Chenxi Ma  
Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University  
17210240039@fudan.edu.cn

Weimin Tan  
Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University  
wmtan14@fudan.edu.cn

ABSTRACT  
Recent studies of deep learning based stereo image super-resolution (StereoSR) have promoted the development of StereoSR. However, existing StereoSR models mainly concentrate on improving quantitative evaluation metrics and neglect the visual quality of super-resolved stereo images. To improve the perceptual performance, this paper proposes the first perception-oriented stereo image super-resolution approach by exploiting the feedback, provided by the evaluation on the perceptual quality of StereoSR results. To provide accurate guidance for the StereoSR model, we develop the first special stereo image super-resolution quality assessment (StereoSRQA) model, and further construct a StereoSRQA database. Extensive experiments demonstrate that our StereoSR approach significantly improves the perceptual quality and enhances the reliability of stereo images for disparity estimation.

CCS CONCEPTS
- Computing methodologies → Reconstruction.

KEYWORDS  
stereo image, super-resolution, quality assessment

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. 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.

ACM Reference Format:
Chenxi Ma, Bo Yan, Weimin Tan, and Xuhao Jiang. 2021. Perception-Oriented Stereo Image Super-Resolution. In Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475408

1 INTRODUCTION
Stereo image super-resolution (StereoSR), aiming to increase the spatial resolution of stereo image pairs, becomes a growing research direction and benefits a lot from the deep-learning. Though remarkable advances have been achieved, convolutional neural network

(CNN) based StereoSR methods [3, 7, 37, 42] are mainly designed to improve the common objective evaluation metrics (peak signal to noise ratio PSNR, structural similarity index SSIM [44]). These StereoSR models are trained with the pixel-level mean square error (MSE) or mean absolute error (MAE) loss. It is well known that MSE/MAE loss forces the restored pixels to be the average of all possible solutions, which leads to high PSNR/SSIM accompanied by poor perceptual performance.

In single image super-resolution (SISR), perceptual loss [10] and generative adversarial network (GAN) based loss [16, 38] are proposed to restore better visual effect. Due to the characteristic of GAN, some visual artifacts are generated. As shown in Figure 1, the StereoSR (PASSR [37]), MSE-based SISR (DRN [5]), and GAN-based SISR (SPSR [23]) methods output non-ideal visual effect with over-smooth or unrealistic texture of the wall.

To restore stereo images with better visual experience, we propose to achieve better perceptual StereoSR results by utilizing the

Figure 1: Comparison of MSE-based SISR (DRN [5]), GAN-based SISR (SPSR [23]), StereoSR (PASSR [37]), and StereoSR model (DASSR [42]) trained with GAN loss [16] and the proposed image quality perceptual (IQP) approach respectively, on image '000000_10' in KITTI2012.
Figure 2: The PSNR/SSIM and rankMOS scores (higher is better) of stereo images super-resolved by ESRGAN [38] (middle) and EDSR [20] (right). The left images denote the groundtruth. Though PSNR/SSIM of right images are higher than middle images, middle images are sharper and more consistent with human visual system (HVS), which can be figured out by our quality assessment score (rankMOS).

feedback from the evaluation of the perceptual quality. Therefore, how to effectively assess the subjective quality of the super-resolved stereo images becomes the key issue for StereoSR, and is still an unsolved problem.

As the most common evaluation metrics, PSNR and SSIM measure the distortion degree of images rather than human visual perception. Figure 2 shows the inconsistency between PSNR/SSIM and subjective quality of two stereo images, super-resolved by a generative adversarial network (GAN) based SR method (middle) and a MSE-based SR method (right). Traditional image quality assessment (IQA) [11, 17, 41] and stereo image quality assessment (StereoQA) [39] methods focus on distortion types, including JPEG compression, noise, Gaussian blur. These distortions are different from artifacts produced by different super-resolution methods. Therefore, we first propose a special stereo image super-resolution quality assessment (StereoSRQA) model, which focuses on degradation of both image structure, texture, and stereo effect.

This paper develops a universal perception-oriented StereoSR framework (PSSR), which learns the LR-to-HR mapping by utilizing the feedback information from the proposed quality assessment model. In specific, the proposed PSSR introduces a novel perception-oriented training constraint based on the prior knowledge from the StereoSRQA to enhance the visual quality and the stereo effect of the super-resolved stereo images. Note that, the proposed PSSR can be applied to different StereoSR networks. The main contributions of this paper are summarized as follows:

- We propose a perception-oriented StereoSR framework to restore stereo image with better subjective quality and stereo effect by employing the feedback from the quality assessment of the StereoSR results.
- We develop the first quality assessment model for StereoSR, which not only provides an unified method for assessing subjective StereoSR results, but also accurately constrains the optimization of StereoSR model.
- We construct the first StereoSRQA database (Mid3D_QA), covering a wide range of super-resolved stereo images. The quality score (rankMOS) of each stereo image pair can fairly reflect the comprehensive visual quality and stereo performance.

Extensive experiments demonstrate the proposed PSSR can significantly improve the subjective quality and practicability of the StereoSR results.

2 RELATED WORK

2.1 Super-Resolution

CNN-based Single Image Super-Resolution. Deep learning based methods have dominated the development of SISR. SRCNN [2] first introduces convolution neural network (CNN) in SISR to learn LR-to-HR mapping. Based on this work, following deep learning based SISR methods achieve continuous breakthroughs by proposing deeper, wider networks and other novel structures (VDSR [14], EDSR [20], SRFBN [19], DRN [5]).

To generate visually pleasing results, Johnson et al. [10] proposed a perceptual loss for SISR, which calculates the feature similarity between SR and HR images based on the VGG network [34]. SRGAN [16] exploits the generative adversarial network (GAN) to reconstruct more high-frequency details. SPSR [23] introduces gradient loss into GAN-based SR method to guide the restoration of image by employing structure prior of gradient map.

Stereo Image Super-Resolution. StereoSR [7], the first CNN-based StereoSR methods, achieves impressive results by taking the parallax prior of stereo image into consideration. In order to deal with various parallax, PASSRnet [37] and PSCASSRnet [3] present parallax-based spatial and spatial-channel attention mechanisms, which can extract the correspondence between two views of a stereo pair. To increase the correspondence between two views in a stereo image pair, DASSR [42] first proposes to align these two views based on the disparity map before restoring the degraded images.

2.2 CNN-based Image Quality Assessment

Mean Opinion Score (MOS), the common image quality assessment (IQA) score, is obtained by manually scoring to evaluate the image quality from the perspective of the human sense. Many IQA metrics (PSNR, SSIM [44], BRISQUE [36], NIQE [28]) use a mathematical model to evaluate the image quality. Deep learning has also promoted the development of IQA. Kang et al. [12] first constructed a no-reference (NR) IQA CNN, which does not require the clean image as reference. Li et al. [18] exploited ensemble learning and saliency map for stereo image quality assessment. StereoQA-Net [39] presents an end-to-end dual-stream interactive network for no-reference stereo image quality assessment.

To evaluate the image quality of SISR methods instead of traditional distortion types (blur, compression, noise, spatial warping),
Yan et al. [41] proposed a FR SISR quality assessment (SISRQA) CNN for assessing the visual perception of the super-resolved images. Since generating subjective scores (MOS) requires many human judgements, the size of IQA database is limited. Wu et al. [40] built a large-scale IQA database, and synthesized pseudo-MOS to represent the subjective quality of the distorted image by combining several classical IQA methods. RankIQA [22] also constructs a large IQA database without human annotation by generating distorted images with image processing transformation of different levels and types.

3 PERCEPTION-ORIENTED STEREOSR

This section mainly introduces the proposed perception-oriented StereoSR (PSSR) approach. Note that our goal is not proposing a novel StereoSR network structure but an effective framework, which can constrain the StereoSR network to generate promising visual quality. Therefore, our StereoSR network can adopt any structure, and we choose the DASSR [42] in this paper.

Traditional StereoSR methods treat pixel-wise mean square error (MSE) or mean absolute error (MAE) as their optimization objective, which forces the model output images with higher PSNR but over-smooth textures. To have a better understanding of the artifacts in StereoSR problem, we first propose a special quality assessment network for StereoSR, which will be introduced in detail latter. Given a well-trained StereoSQA model, which is able to predict the perceptual quality of StereoS results, we consider that an ideal StereoSR network can generate visually promising results with high StereoSQA scores.

Different from the MSE or MAS loss, which constrains the pixel similarity between the SR and HR images, we aim to constrain the perceptual similarity between the SR and HR images by utilizing the guidance of the StereSQA model. In this way, the StereoSR results enjoy promising visual effect, that is close to groundtruth and satisfies the human perception.

As shown in Figure 3, we propose the image quality perceptual (IQP) constraint to measure the feature similarity between groundtruth and super-resolved stereo images based on the StereoSQA network. In specific, we design two IQP losses, an image-level IQP loss ($L_{\text{IQP}_\text{im}}$), and a feature-level IQP loss ($L_{\text{IQP}_\text{f}}$). The two IQP losses calculate the similarity at image and feature levels respectively by feeding the output image and the middle feature from the StereoSR network to the StereoSQA network.

To obtain the image-level IQP loss ($L_{\text{IQP}_\text{im}}$), the super-resolved stereo image $I_{\text{SR}}$ and the ground-truth stereo image $I_{\text{GT}}$ are passed to the StereoSQA network. Therefore, two features from the first convolution layers in StereoSQA network ($f_{\text{SRQA}}$, $f_{\text{GT}}$) are obtained. Then, we adopt L2 norm function to calculate the feature similarity between $f_{\text{SRQA}}$ and $f_{\text{GT}}$. In this way, the $L_{\text{IQP}_\text{im}}$ is able to constrain the StereoSQA feature of super-resolved stereo image to be similar to that of the groundtruth.

$$L_{\text{IQP}_\text{im}} = \| f_{\text{SRQA}} - f_{\text{GT}} \|_2.$$  \hspace{1cm} (1)

The feature-level IQP loss ($L_{\text{IQP}_\text{f}}$) constrains features in StereoSR network to be able to reconstruct visually promising image. Specifically, as shown in Figure 3, we replace the first layer feature of StereoSQA network with the last layer feature of StereoSR network to output middle feature $f_{\text{SRQA}}$. Then, we also calculate the L2 norm between $f_{\text{SRQA}}$ and $f_{\text{GT}}$.

$$L_{\text{IQP}_\text{f}} = \| f_{\text{SRQA}} - f_{\text{GT}} \|_2.$$  \hspace{1cm} (2)

The IQP losses ($L_{\text{IQP}_\text{im}}$ and $L_{\text{IQP}_\text{f}}$) are used to calculate the IQP losses. The feature-level IQP loss is proposed based on following considerations: First, feature extraction of the first layer in StereoSQA shares same information with the image reconstruction feature of the last convolution layer in StereoSR network, and the feature-level IQP loss enhances the constraint strength of the feature similarity; Second, compared with the three-channel image, the feature map in convolutional neural network generally has more channels and consequently covers more information for image reconstruction. Therefore, optimizing the accuracy based on feature instead of image may enable stronger constraint; Third, since the IQP loss is calculated on the features of CNN, we believe that directly putting the output feature rather than the output image of the StereoSR network to the StereoSQA network can effectively bridges the gap between image and feature.

In this way, the proposed IQP constraints make sure the feature similarity related to the quality assessment, and further enforce the super-resolved stereo image share similar spatial details with the groundtruth, especially the details that are sensitive for StereoSQA to evaluate the visual quality of stereo images. In other words, the super-resolved stereo image has similar visual nature with the groundtruth image. The StereoSR network is trained on the combination of L2 and IQP losses.

$$\mathcal{L} = \lambda_0 \| f_{\text{SR}} - f_{\text{GT}} \|_2 + \lambda_2 L_{\text{IQP}_\text{f}},$$  \hspace{1cm} (3)

where $\lambda_0, \lambda_1, \lambda_2$ are the weights to balance different losses. We set $\lambda_0 = 1$ and $\lambda_1 = \lambda_2 = 0.1$.

By optimizing the StereoSR network with the IQP constraint in training stage, the knowledge from the StereoSQA network can be back-propagated to the StereoSR network.

4 STEREOSR QUALITY ASSESSMENT

Our perceptual-oriented StereoSR method requires a well-performing StereoSQA model. To our best knowledge, such a model has never been proposed before. This section constructs a model and a database for StereoSQA.
4.1 StereoSRQA Network

First, we train a StereoSRQA network, which simultaneously considers spatial distortion inside each view and disparity-related quality degradation cross different views, to predict the perceptual quality of different SR results. Different from IQA methods only considering spatial content of image, assessing a stereo image should additionally refer to the correlation between two views. Zhou et al. [39] indicated that integrating left and right views by summation and subtraction is consistent with the fusion and disparity information in human visual system (HVS). Inspired by their work, our StereoSRQA network takes the left view, the right view, and their difference map as input. Note that the main purpose of this paper is not proposing a novel QA network, but a suitable QA model that can provide helpful guidance for the StereoSR model. Therefore, our StereoSRQA model can flexibly adopt the structure of different StereoQA models.

As shown in Figure 4, the proposed StereoSRQA model is composed of three branches with similar structure. The upper and lower branches learn spatial features of two views by taking the left and the right views as input respectively.

The only difference between the middle and the upper/lower branches is the input at the beginning and middle convolution layers. The middle branch extracts the cross-view correlation by working on the difference map of image and the difference map of feature. In specific, we get the residual image of the left and right views as input respectively. Since the disparity estimation is a significant application of stereo image, we also measure the disparity estimation accuracy to represent the practicality of the super-resolved stereo image pair. Following most IQA methods, our StereoQA network is trained with MSE loss between the predicted score $y$ and the groundtruth score $z$.

$$\mathcal{L}_{\text{StereoSRQA}} = ||y - z||_2.$$  \hspace{1cm} (4)

Note that, to calculate the IQP loss, the feature similarity is calculated based on the concatenation of features from the first convolution layers of the three branches.

4.2 StereoSRQA Database with RankMOS

Existing stereo image quality assessment databases (e.g., LIVE Phase I [29], LIVE Phase II [1], NBU [31], MCL-3D [35]) focus on distortion types, including white noise, Gaussian blur, JPEG compressing, which differ from artifacts in SR. Toward this end, we build the first database, named Mid3D_QA, specifically for assessing the quality of StereoSr. The Mid3D_QA database contains 66 HR stereo image pairs from Middlebury stereo database [8] as reference, and each HR stereo pairs are degraded by 51 distortion types (including 13 SR methods under different scales, blur and noise levels) to generate 3366 (66×51) super-resolved versions served as the distorted stereo image pairs. Then, we split 3009 (59×51) stereo pairs for training and 357 (7×51) stereo pairs for testing.

As stated by Wu et al. [40], collecting MOS for images by subjective experiment is laborious and needs highly controlled conditions. Therefore, inspired by them, we generate quality scores of distorted images by measuring the quality of the SR results in three aspects to simulate the MOS. Consequently, the labeled data can be leveraged to train a stable StereoSRQA network.

**Distorted Images.** First, 66 high quality stereo image pairs from Middlebury stereo database [8] are used as reference. To generate diverse LR stereo images, we downsample the HR stereo images with scale factors (2, 3, 4, 5, 6, 8) by bicubic interpolation, convolve them with Gaussian blur ($\sigma \in [0.2, 1.2]$), add noise with noise level sampled from $[0, 30]$. Then, a wide range of super-resolved stereo images with diverse visual effects are generated by applying different SR algorithms, including interpolation-based SISR (bicubic, lanczos, nearest neighbor, bilinear), CNN-based SISR (SRCNN [2], EDSR [20], IDN [6], SRMD [43]), GAN-based SISR (SRGAN [16], ESRGAN [38]) and StereoSR (StereoSR [7], PASSRNet [37], DASSR [42]) methods. There are 51 public SR models for all scales in total. That is, each stereo image pair have 51 SR versions corresponding to 51 distortion levels.

**Generation of Subjective Score.** Inspired by the generation of the quality assessment data in [40] and RankIQA [22], we propose to synthesize the quality scores for distorted stereo image pairs by combining existing quality assessment metrics. In order to obtain a comprehensive evaluation score, we measure the quality of the super-resolved stereo image pairs from three aspects: the spatial quality inside single view, the 3D effect, and the practicality.

Supposing $ISR_{i,j}$ denotes the results of $i$-th SR method for $j$-th stereo image pair, where $i = 1 : 66, j = 1 : 51$, its quality score $rankMOS_{i,j}$ is generated by voting, ordering, and merging operations, as specified in Algorithm 1:

1. **Voting** First, we treat the SISRQA model [41] and the StereoQA model [39] as two voters to evaluate the spatial and stereo effect respectively. Since the disparity estimation is a significant application of stereo image, we also measure the disparity estimation accuracy to represent the practicality of the super-resolved stereo image pair $ISR_{i,j}$. The disparity, estimated by the StereoNet [13], is measured with endpoint-error (EPE), which serves as the third voter. Thus, each distorted stereo image pair has three values from three
Algorithm 1 Training Algorithm

Require:
1: 3 voters: StereoQA, SISRQA, disparity estimation accuracy (StereoNet [14]);
2: 66 reference stereo pairs \( I_{GT}^{i,j} \); 66 × 51 SR stereo pairs \( I_{SR}^{i,j} \);

Ensure: Subjective score of \( S_{i,j}^{SR} \); rankMOS\( _{i,j} \);

Voting
4: while \( k=1:3 \) // For k-th voter do
5: k-th voter outputs a value \( v_{i,j,k} \) for \( I_{SR}^{i,j} \);
6: while \( i=1:66 \) and \( k=1:3 \) // i-th reference, k-th voter do
7: Rank scores \( (v_{i,1,k},...,v_{i,51,k}) \);
8: Assign the ranking (1∼51) as the new score \( S_{i,j,k} \);

Merging
9: while \( i=1:66 \) // For i-th reference stereo image do
10: \( R_{S,i,j} = \frac{1}{3} \sum_{k=1}^{3} (S_{i,j,1},S_{i,j,2},S_{i,j,3}) \);
11: \( \text{rankMOS}_{i,j} = 10 \times \text{Normalize}(R_{S,i,j}) \);

Algorithm 1 Training Algorithm

Voters, which measure the image quality from different perspectives. The k-th voter gives a score value \( v_{i,j,k} \) for \( I_{SR}^{i,j} \).

2. Ordering Since quality values from different voters have disparate scales and variation. The quality values \( v_{i,j,k} \) need to be normalized into a unified range and re-scaled to the same distribution. Therefore, we adopt relative ranking value instead of the absolute output value as the new score, \( S_{i,j,k} \). In specific, for i-th stereo image pair and k-th voter, we order scores of all SR versions (51 in total) from worst to best and get rank 1 to rank 51. Rank 1 denotes the worst and rank 51 denotes the best. Then, the rank value of i-th distorted stereo image \( I_{SR}^{i,j} \) is assigned to the new score \( S_{i,j,k} \) from k-th voter.

3. Merging The single ranking score \( R_{S,i,j} \) for current SR stereo image \( I_{SR}^{i,j} \) is obtained by averaging the three scores \( S_{i,j,k} \). We further normalize the \( R_{S,i,j} \) and multiply it by 10 to generate the final quality score, \( \text{rankMOS}_{i,j} \in [1, 10] \).

\[
R_{S,i,j} = \frac{1}{3} \sum_{k=1}^{3} (S_{i,j,k}),
\]
\[
\text{rankMOS}_{i,j} = 10 \times \text{Norm}(R_{S,i,j}).
\] (5)

Note that we do not adopt more CNN-based IQA models. Since existing CNN-based IQA methods are trained on the same databases, their predicted scores are similar for same image, especially the rank of the images. Therefore, adopting more CNN-based IQA models makes our rankMOS be closer to results of existing IQA databases and lack concentration to SR features. We consider that a good score should merge more IQAs, that pay attention to different image distortion types.

Figure 5 demonstrates distributions of three voters and our generated score (rankMOS) cross different scales. Different lines denote results of different SR methods. As shown in Figure 5, the rankMOS can effectively incorporate three voters with different value ranges and clearly distinguish artifacts among different SR results under different scales. Furthermore, the rankMOS is able to reduce interference of metrics that do not match the subjective quality and better represent the subjective quality.

To more intuitively show the distribution of different voters (StereoQA [39], SISRQA [41], disparity accuracy (EPE) relative to the synthetic rankMOS, we normalize these scores and demonstrate relations in Figure 6, where different colors represent different voters. Points with same color correspond to results of different super-resolved stereo image pairs. We see that scores of different quality assessment methods are linearly related to our rankMOS, indicating the synthetic rankMOS is consistent with existing QA methods in terms of reflecting the super-resolution performance of stereo image. In particular, some outliers can be effectively weakened by ranking and averaging operations.

By adopting relative ranking score rather than absolute score, we can extract the preference of different QA methods from different aspects, and unify the scoring standards. The variance of same voter to different images and the standard difference among all voters can also be suppressed.

5 EXPERIMENTS AND ANALYSIS

This section first introduces the databases, evaluation metrics, and other experimental details. Next, ablation experiments are conducted to investigate the contributions of each design in our Perception-oriented StereoSR (PSSR) approaches. Then, our PSSR is compared with state-of-the-arts in terms of quantitative, qualitative, practicability, and running efficiency. Finally, we also report the performance of our proposed StereoSRQA model.

5.1 Database and Protocols

By following the StereoSR works [7, 42], our PSSR model is trained on 59 stereo image pairs from the Middlebury. Training images are augmented by randomly down-scaling, flipping and rotating. Each mini-batch contains 32 HR stereo image patches of size 120. The StereoSR models are evaluated on the same test sets in [42] (Middlebury [8], Tsukuba [30], KITTI2012 [4], KITTI2015 [27], and SceneFlow [26]). All SR models are evaluated with various quality evaluation methods and the StereoSRQA model.

The StereoSRQA models are trained and tested on the Mid3D_QA database, and work on stereo image patches of size 120. In test, suppose a stereo image pair is cropped into N patch pairs, the final score is obtained by averaging output scores of all patches. All StereoSRQA models are measured with common correlation criteria, including Pearson linear correlation (PLCC), Spearman rank order correlation coefficient (SROCC), Kendall rank-order correlation coefficient (KROCC), and Root mean square error (RMSE). The higher SROCC, PLCC, KROCC, and lower RMSE denote the better performance.

All models are based on the TensorFlow [25] implementation and optimized by Adam [15] with \( \beta_1 = 0.9, \beta_2 = 0.999 \). The learning rate is 1e-4. We conduct experiments on the machine with Nvidia GTX1080Ti GPU (128G RAM).

5.2 Ablation Study

The ablation study is conducted for \( \times 4 \) SR on the Middlebury database. Table 1 and Figure 7 compare different constraints, used to optimize the StereoSR model. We first train a baseline model (\( \mathcal{L}_{MSE} \)) with the MSE loss only, and then train other models with different combinations of losses. The \( \mathcal{L}_{\alpha} \) denotes that the model is trained with the MSE and \( \alpha \) loss.
to compute the feature similarity for the model ‘+ $L_{IQP_{SISR}}$’, which has better result on SFA. This is mainly because that $L_{IQP_{SISR}}$ enforces the SR results to have better subjective quality on single view, and the NR IQA method (SFA) pays more attention on quality of single view and prefers results of ‘+ $L_{IQP_{SISR}}$’. In comparison, our ‘+ $L_{IQP}$’ gets better visual results and StereoSRQA score.

5.3 Comparison with state-of-the-arts

Bicubic Degradation (BI). Existing StereoSR methods work for bicubic degradation (BI), and suppose LR image is generated by bicubic downscaling. Following their steps, we compare our PSSR with MSE-based SISR (EDSR [20], SRFBN [19], DRN [5]), GAN-based SISR (ESRGAN [38], SPSR [23]), and StereoSR (StereoSR [7], PASSRnet [37], DASSR [42]) methods on BI degradation in Table 2. Some methods only public ×4 SR models.

Table 2 demonstrates several NR-IQA scores (including NIQE [28], Perceptual Index (PI), the combination of Ma et al. [24] and NIQE, and our StereoSRQA model) to evaluate the perceptual performance of different SR results. Lower NIQE/PI and higher StereoSRQA scores denote better subjective quality.

We can observe that some GAN-based SISR methods have lower NIQE/PI than groundtruth, which goes against the common sense. As stated in [9], though NIQE and PI are relevant to human judgment, they cannot distinguish GAN generated noises and real details and prefer images with obvious unrealistic artifacts produced by GAN-based image restoration methods. Therefore, these perceptual evaluation metrics cannot fairly reflect the subjective SR performance. Even so, our PSSR achieves best NIQE/PI except the GAN-based SPSR, and makes significant progress on StereoSRQA.
Table 2: Quantitative comparison with state-of-the-art SR approaches in terms of perceptual quality evaluation metrics (NIQE/PI/StereoSRQA). ↑ denotes the higher is the better. The GAN-based SISR method performs better the groundtruth on NIQE and PI scores, indicating the two IQA metrics cannot well reflect the real quality of stereo image.

| Database | Scale | Bicubic | EDSR [20] | SRBNF [19] | DRN [5] |
|----------|-------|---------|-----------|------------|---------|
| Middlebury | ×4    | 5.67/0.49/8.49 | 5.32/5.38/8.26 | 5.96/5.76/8.41 | 5.31/5.86/8.16 |
| ×3       | 6.74/1.64/6.99 | 5.32/5.38/8.44 | 5.22/4.70/6.85 | 5.02/5.91/7.22 | 4.05/5.99/7.13 |
| ×2       | 5.15/2.55/8.55 | 4.17/3.57/9.17 | 4.59/5.35/9.18 | - | 4.33/5.15/7.77 |
|
| KITTI 2012 | ×4    | 7.49/0.87/5.10 | 5.84/5.72/5.17 | 5.97/5.73/6.42 | 5.55/4.22/8.11 |
| ×3       | 6.67/3.26/6.99 | 5.86/5.72/5.85 | 5.16/4.67/7.97 | 4.12/3.62/8.44 | 3.33/4.58/7.35 |
| ×2       | 5.04/4.88/9.98 | 4.12/3.79/5.92 | 4.18/3.21/7.81 | - | 4.19/3.54/7.22 |
|
| KITTI 2015 | ×4    | 7.42/0.88/6.54 | 5.70/5.88/8.35 | 6.00/5.96/8.35 | 5.55/4.22/8.11 |
| ×3       | 7.35/0.73/7.08 | 5.70/5.88/8.98 | 5.24/4.72/8.16 | - | 4.10/3.65/8.81 |
| ×2       | 5.09/5.10/9.19 | 4.21/3.57/6.86 | 4.13/3.71/9.79 | - | 4.11/3.59/9.32 |
|
| Tsukuba   | ×4    | 7.58/2.44/3.18 | 6.04/6.87/8.71 | 5.96/6.62/4.60 | 6.65/4.44/8.39 |
| ×3       | 6.94/5.44/5.84 | 6.04/6.83/8.00 | 5.54/1.07/7.21 | - | 4.14/3.67/6.83 |
| ×2       | 5.83/5.86/7.12 | 5.05/4.80/5.34 | 5.06/5.81/5.51 | - | 4.35/4.90/7.59 |
|
| SceneFlow | ×4    | 6.99/0.70/4.33 | 5.39/0.65/6.77 | 5.59/0.54/6.94 | 5.26/0.24/7.23 |
| ×3       | 7.05/0.63/6.05 | 5.39/0.55/7.86 | 4.99/0.95/7.23 | 2.96/0.24/7.24 | 3.96/2.71/6.87 |
| ×2       | 5.08/0.90/7.81 | 4.28/0.85/5.80 | 4.15/2.93/7.71 | - | 4.10/2.85/7.56 |

Table 3: The average quality scores of CNN-based IQA methods for ×4 StereoSR on Middlebury.

| IQA     | Bicubic | StereoSR [7] | PASSRNet [37] | DASSR [42] | PSRR | GT |
|---------|---------|--------------|----------------|------------|------|----|
| NIQE    | 55.98/40.81/43.88 | 41.83/41.71/32.13 | 32.21/29.79/24.05 |
| NIQE w/o SISRQA | 55.48/37.02/34.07 | 36.49/36.47/27.88 |
| NIQE w/ SISRQA | 55.36/37.02/33.96 | 36.46/36.47/27.88 |
| NIQE w/ SISRQA | 55.36/37.02/33.96 | 36.46/36.47/27.88 |
| NIQE w/ SISRQA | 55.36/37.02/33.96 | 36.46/36.47/27.88 |

In Figure 1 and 8, the PSRR remarkably improves the subjective quality and restores clearer textures.

Table 4: Perceptual scores (NIQE/PI/StereoSRQA) of applying the proposed training constraint on different SR models for ×4 StereoSR on Middlebury.

| Database     | Scale | EDSR [20] | StereoSR [7] |
|--------------|-------|-----------|--------------|
| Middlebury   | ×4    | 5.31/5.38/7.26 | 4.32/4.27/8.10 |
| KITTI 2012   | ×4    | 5.31/5.38/7.26 | 4.32/4.27/8.10 |
| KITTI 2015   | ×4    | 5.31/5.38/7.26 | 4.32/4.27/8.10 |
| Tsukuba      | ×4    | 5.31/5.38/7.26 | 4.32/4.27/8.10 |
| SceneFlow    | ×4    | 5.31/5.38/7.26 | 4.32/4.27/8.10 |

Table 5: ×4 StereoSR results (PSNR/SSIM/StereoSRQA) on BD degradation of models trained with L_MSE and L_IQP.

| σ   | Bicubic | L_MSE | Our L_IQP |
|-----|---------|-------|-----------|
| 1.0 | 28.22/0.84/6.40 | 28.22/0.84/6.40 | 28.22/0.84/6.40 |
| 2.0 | 24.42/0.79/5.24 | 24.42/0.79/5.24 | 24.42/0.79/5.24 |
| 3.0 | 23.52/0.76/4.82 | 23.52/0.76/4.82 | 23.52/0.76/4.82 |

Blurriness and blur degradation (BD) To further illustrate the robustness of our PSRR, we retrain StereoSR models with MSE loss and our training method respectively on a more difficult blur degradation (BD) dataset. The LR image is blurred with 15 × 15 Gaussian kernel, kernel width σ ∈ [0.8, 4.0]. Table 5 and Figure 9 demonstrate the StereoSR results under BD degradation. Both StereoSR models work well for small blur kernel, and the MSE loss can maximize PSNR/SSIM, as we stated. However, the StereoSR model, trained with MSE loss, suffers from significant performance drop facing severe blur. In comparison, our PSRR keeps better subjective.
performance and even higher PSNR/SSIM for larger blur kernels, demonstrating the $L_{IQP}$ is robust to more difficult degradation.

**Disparity Estimation** Since disparity estimation is a non-negligible application of stereo image, we evaluate the quality of the super-resolved stereo images on disparity estimation task by following [42]. Table 6 and Figure 10 compare the accuracy of disparity, estimated by StereoNet [13]. The PSSR can generate more textures and spatial details, which are valuable for predicting precise disparity.

![Figure 9: Visual comparison of StereoSR models trained with MSE and our IQP losses on BD degradation with different kernel widths.](image)

![Figure 10: Visual comparison of disparity accuracy.](image)

### 5.4 Performance of StereoSRQA model

By inputting the degraded stereo image pair or the difference map between the degraded and groundtruth stereo image pair to the StereoSRQA model, we can construct a no-reference and a full-reference StereoSRQA models (‘NRStereoSRQA’, ‘FRStereoSRQA’). Table 7 compares FRStereoSRQA and NRStereoSRQA with both full-reference IQA (PSNR, SSIM [44], IFC [33], VIF [32], SISRQA [41]) and no-reference IQA (BRISQUE [36], SSEQ [21], NIQE [28], CNNIQQA [11], SFGA [17], StereoQA [39]) methods, including CNN-based or non CNN-based, single image or stereo image based methods. Our FR-StereoSRQA achieves the best performance and is more consistent with human perception.

To further illustrate the contribution of our Mid3D_QA, Table 8 compares two SISRQA models and two StereoQA models, which are trained on their original database and our database respectively.

### 6 CONCLUSION

This paper proposes a perception-oriented stereo image super-resolution framework (PSSR). The proposed PSSR improves the visual quality for StereoSR by using the feedback guidance from the assessment of the StereoSR results. To provide accurate guidance, we construct the first quality assessment model (StereoSRQA) for StereoSR, and a special StereoSRQA database (Mid3D_QA). In specific, we propose a image quality perceptual constraint to calculate the feature similarity, used to evaluate the perceptual quality of StereoSR results, between the groundtruth and the super-resolved stereo images. Extensive experiments demonstrate our PSSR achieves state-of-the-art performance in terms of the quality, practicality, and flexibility.
