Image Super-Resolution Quality Assessment: Structural Fidelity Versus Statistical Naturalness

Wei Zhou[1,2], Zhou Wang[1], Zhibo Chen[2]
[1]Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON N2L3G1, Canada
[2]CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
University of Science and Technology of China, Hefei 230027, China
Email: {wei.zhou, zhou.wang}@uwaterloo.ca; chenzhibo@ustc.edu.cn

Abstract—Single image super-resolution (SISR) algorithms reconstruct high-resolution (HR) images with their low-resolution (LR) counterparts. It is desirable to develop image quality assessment (IQA) methods that can not only evaluate and compare SISR algorithms, but also guide their future development. In this paper, we assess the quality of SISR generated images in a two-dimensional (2D) space of structural fidelity versus statistical naturalness. This allows us to observe the behaviors of different SISR algorithms as a tradeoff in the 2D space. Specifically, SISR methods are traditionally designed to achieve high structural fidelity but often sacrifice statistical naturalness, while recent generative adversarial network (GAN) based algorithms tend to create more natural-looking results but lose significantly on structural fidelity, suggesting the desire to incorporate both types of methods. Furthermore, such a 2D evaluation can be easily fused to a scalar quality prediction. Interestingly, we find that a simple linear combination of a straightforward local structural fidelity and a global statistical naturalness measures produce surprisingly accurate predictions of SISR image quality when tested using public subject-rated SISR image datasets. Code of the proposed SFSN model is publicly available at https://github.com/weizhou-geek/SFSN.

Keywords—image super-resolution; quality assessment; image decomposition; structural fidelity; statistical naturalness

I. INTRODUCTION

Single image super-resolution (SISR) aims to recover a high-resolution (HR) image given a single low-resolution (LR) image as the input. SISR plays a significant role in a wide range of applications, from satellite imaging, web browsing, to video surveillance [1]. During the past decades, numerous SISR algorithms have been proposed, including interpolation-based [2], [3], dictionary-based [4]–[6], and deep learning-based methods [7]–[12]. The visual appearance and quality of SISR generated images vary dramatically when different SISR approaches are exploited. Nevertheless, there is still no consensus so far on how the quality of SISR created images should be assessed. This is critically important because image quality assessment (IQA) methods not only help evaluate and compare SISR algorithms, but also guide the development of future SISR methodologies.

In general, the most reliable quality assessment method is human subjective evaluation [13]–[15]. But subjective tests are usually expensive, time-consuming and hard to be integrated into SISR optimization frameworks. Therefore, it is highly desirable to design effective objective IQA models for SISR generated images. Depending on the availability of the original pristine image, full-reference (FR) IQA [16]–[22] and no-reference (NR) IQA approaches [23]–[27] may be applied.
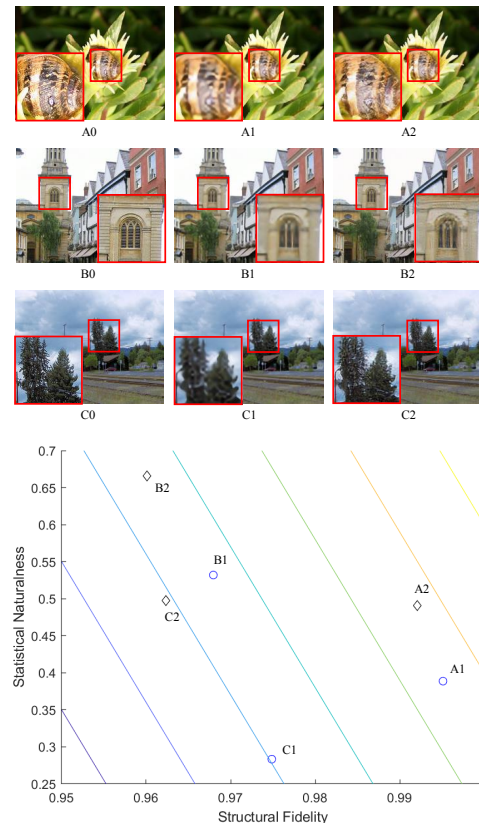

Fig. 1. SISR generated images and 2D quality assessment of statistical naturalness versus structural fidelity (*SF* vs *SN*). A0, B0, C0: original HR images; A1, B1, C1: reconstructed images by VDSR [7] at scaling factor 4; A2, B2, C2: reconstructed images by SRGAN [12] at scaling factor 4.

Additionally, since many SISR algorithms produce blurry reconstructed images, image sharpness assessment (ISA) or blur measures [28]–[30] may also be employed.

Despite the success in other IQA applications, existing FR-IQA, NR-IQA and ISA methods often fall short when evaluating the quality of SISR generated images. The gap is not only on the accuracy in predicting subjective scores, but also on effectively interpreting the nature of key quality degradation trends in SISR images. An example is given in Fig. 1, where traditional SISR methods such as VDSR [7] are highly effective at achieving high signal fidelity in terms of signal-to-noise ratio or structural similarity [16] when compared to the original images, but the resulting images often look artificial. On the other hand, recently proposed generative adversarial
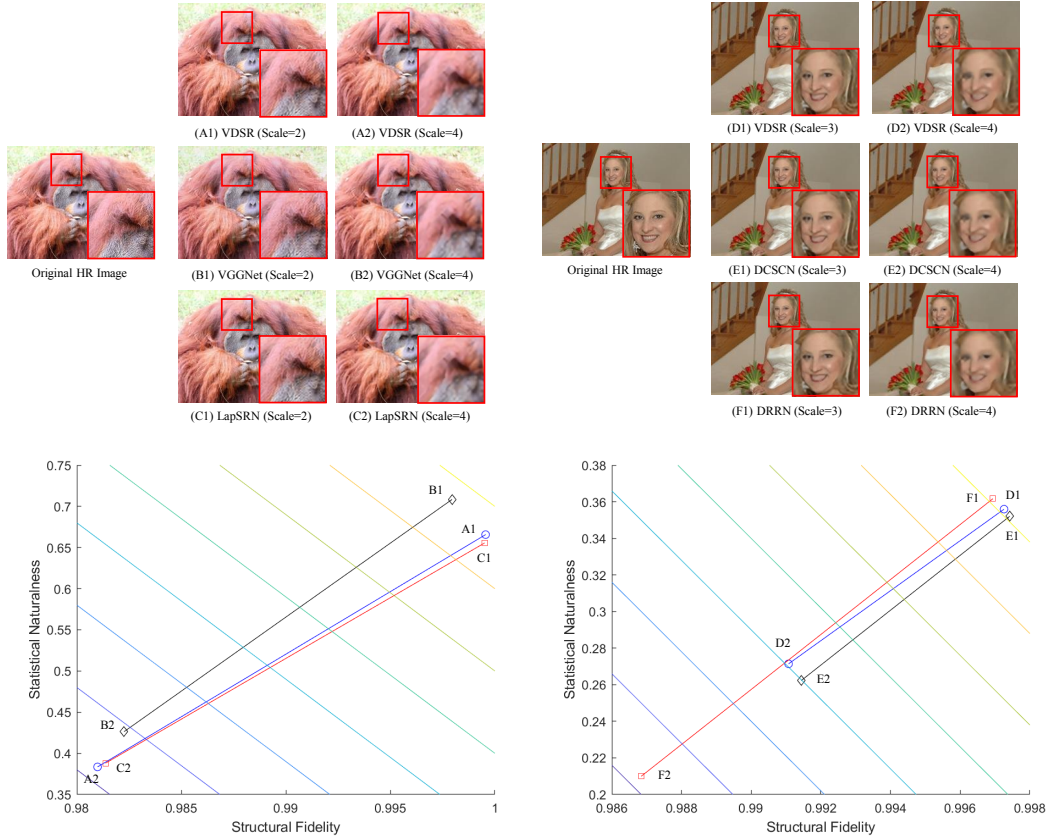
Fig. 2. Sample SISR generated images at different scaling factors (top) and their corresponding points in (*SF* vs *SN*) plots (bottom), where the level sets of the proposed quality prediction $Q(\mathbf{y})$ are also shown.

network (GAN) based approaches such as SRGAN [12] are impressive at producing natural-looking reconstructed images, but their signal fidelity measures are significantly lower. These observations motivate us to look at the problem in a two-dimensional (2D) space of structural fidelity versus statistical naturalness, as demonstrated at the bottom part of Fig. 1.

II. 2D QUALITY ASSESSMENT OF SISR IMAGES

Multi-scale image decomposition such as Laplace and wavelet transforms have been shown to be highly effective at both capturing the local perceptual degradation in terms of just noticeable differences or structural distortion, but also characterizing the statistical naturalness of images. Therefore, we apply a multi-scale image transform and construct local structural fidelity and global statistical naturalness measures both in the transform domain. Given the original HR image $\mathbf{x}$ as the reference, and the SISR generated test image $\mathbf{y}$, inspired by the success of the MS-SSIM algorithm [17], we define the subband patch level local structural fidelity measure as:

$$SF_{local}^k(x,y) = \frac{\sigma_{xy} + C}{\sigma_x \sigma_y + C},\qquad(1)$$

where x and y denote the patches extracted from the k-th subband from the reference and test images, respectively, σ_x and σ_y are their standard deviations, σ_{xy} represents the covariance between x and y, and C is a small positive stabilizing constant. The scale level structural fidelity measure

is then computed by spatial pooling:

$$SF^k(\mathbf{x},\mathbf{y}) = \frac{1}{M}\sum_{m=1}^{M} SF_{local}^k(x,y),\qquad(2)$$

where M denotes the number of local patches in the subband. Finally, we fuse across scales to obtain the overall structural fidelity between $\mathbf{x}$ and $\mathbf{y}$:

$$SF(\mathbf{x},\mathbf{y}) = \prod_{k=1}^{K}[SF^k(\mathbf{x},\mathbf{y})]^{\alpha_k},\qquad(3)$$

where K is the total number of scales/subbands, and α_k is the weight assigned to the k-th scale as in [17]. Furthermore, natural texture-rich content tends to have higher entropy in the transform domain [31]. Thus, we use the global entropy of transform coefficients as a statistical naturalness measure:

$$SN(\mathbf{y}) = -\sum P(c_y)\log(P(c_y)),\qquad(4)$$

where $P(c_y)$ denotes the probability of subband coefficients of the test image $\mathbf{y}$ and may be approximated with histograms.

Although the proposed pair of (*SF*, *SN*) measure is rather simple, it offers a meaningful 2D illustration of the behaviors of SISR algorithms. Fig. 1 shows three original HR images with their corresponding SISR images generated by VDSR [7] and SRGAN [12]. The (*SF* vs. *SN*) plot clearly indicates the relative advantage of VDSR over SRGAN on the *SF* measure, and conversely the advantage of SRGAN over VDSR

on the *SN* measure. The pattern is consistent over all three content, as indicated by the pairs of points (A1, A2), (B1, B2) and (C1, C2). Fig. 2 shows images generated by different SISR algorithms applied to LR images of different sizes and enhanced by different scaling factors. The (*SF* vs. *SN*) plots offer a platform to examine the behaviors of different SISR algorithms across scaling factors. It can be observed that the the general trend of any SISR algorithm is that both *SF* and *SN* measures drop with increasing scaling factors. However, the speed of change may vary depending on the algorithm and possibly the image content. For example, the DRRN [10] method appears to be much more sensitive to scaling factor change than VDSR [7] and DCSCN [9] for the right image.

III. FUSING 2D ASSESSMENT FOR 1D QUALITY PREDICTION

In practice, it is often desirable to obtain a single quality score indicating the overall quality of SISR generated images. This can be achieved by collapsing the proposed 2D measure into a scalar quality prediction, e.g., by a linear combination:

$$Q(\mathbf{y}) = w_F SF(\mathbf{x}, \mathbf{y}) + w_N SN(\mathbf{y}), \quad (5)$$

where the weighting factors w_F and w_N adjust the relative importance of the two measures, and are set empirically at 0.9 and 0.1, respectively, in the current implementation. We name Q the SFSN measure, which creates straight lines as level sets in the 2D space, as shown in the bottom plots of Figs. 1 and 2. This scalar quality prediction can then be validated by comparing against subject-ratings of SISR generated images.

TABLE I. SRCC PERFORMANCE COMPARISON OF OBJECTIVE MODELS ON WIND [13], CVIU [14] AND QADS [15] DATABASES.

Methods	WIND	CVIU	QADS	Average
PSNR	0.6320	0.5663	0.3544	0.5176
SSIM [16]	0.6125	0.6285	0.5290	0.5900
MS-SSIM [17]	0.8246	0.8048	0.7172	0.7822
FSIM [18]	0.8503	0.7481	0.6885	0.7623
CW-SSIM [19]	0.8626	0.7591	0.3259	0.6492
GSIM [20]	0.7649	0.6505	0.5538	0.6564
GMSD [21]	0.7966	0.8469	0.7650	0.8028
SPSIM [22]	0.8141	0.6698	0.5751	0.6863
BRISQUE [23]	0.7676	0.5863	0.5463	0.6334
NIQE [24]	0.6263	0.6525	0.3977	0.5588
BLIINDS-II [25]	0.5281	0.3705	0.3838	0.4275
DIIVINE [26]	0.5465	0.5479	0.4817	0.5254
LPSI [27]	0.6669	0.4883	0.4079	0.5210
S3 [28]	0.4455	0.5050	0.4636	0.4714
LPC-SI [29]	0.5375	0.5450	0.4902	0.5242
HVS-MaxPol-1 [30]	0.6166	0.6421	0.6170	0.6252
HVS-MaxPol-2 [30]	0.6309	0.6313	0.5736	0.6119
Proposed (*SF* only)	0.8642	0.8546	0.7867	0.8352
Proposed (*SN* only)	0.5873	0.6415	0.6115	0.6134
Proposed SFSN	**0.8867**	**0.8714**	**0.8407**	**0.8663**

We validate the proposed fused SFSN quality prediction method on three public SISR IQA databases, including WIND [13], CVIU [14], and QADS [15]. The WIND database considers 8 interpolation algorithms with scaling factors of 2, 4, and 8. It contains 312 SISR images corresponding to 13 reference images. The CVIU database consists of 30 reference HR images and 1,620 SISR generated images created by 9 algorithms with 6 pairs of (scaling factor, kernel width) combinations, where a larger scaling factor corresponds to a larger blur kernel width. The QADS database contains 20 original HR images and 980 images generated by 21 SISR algorithms, including 4 interpolation-based, 11 dictionary-based, and 6 deep learning (DL) based models applied for upsampling factors of 2, 3,

and 4. In all three databases, each SISR generated image is subject-rated and annotated by a value of mean opinion score (MOS). We compare the proposed method with 8 FR-IQA, 5 NR-IQA, and 4 ISA models. The Spearman Rank-order Correlation Coefficient (SRCC) comparison results are reported in Table I, where the best performances are highlighted in bold. Other commonly used evaluation criteria [32] produce very similar results but are not included due to space limit. Despite the straightforward implementations of both *SF* and *SN* assessment, and the simple linear fusion approach, SFSN achieves surprisingly competitive performance against state-of-the-art IQA and ISA models.

TABLE II. SRCC PERFORMANCE COMPARISON OF OBJECTIVE MODELS ON DIFFERENT SISR CATEGORIES ON QADS [15] DATABASE.

Methods	Interpolation	Dictionary	DL	Overall
PSNR	0.2972	0.3808	0.2656	0.3544
SSIM [16]	0.4015	0.5481	0.5121	0.5290
MS-SSIM [17]	0.6340	0.7425	0.7104	0.7172
FSIM [18]	0.5471	0.6846	0.6637	0.6885
CW-SSIM [19]	0.5254	0.4362	0.0986	0.3259
GSIM [20]	0.3946	0.5332	0.5661	0.5538
GMSD [21]	0.7054	0.7709	0.7363	0.7650
SPSIM [22]	0.4545	0.5518	0.5871	0.5751
BRISQUE [23]	0.5096	0.4951	0.4357	0.5463
NIQE [24]	0.4639	0.4547	0.4190	0.3977
BLIINDS-II [25]	0.1814	0.3628	0.6547	0.3838
DIIVINE [26]	0.4267	0.4175	0.5654	0.4817
LPSI [27]	0.2726	0.3309	0.6034	0.4079
S3 [28]	0.4016	0.3171	0.5458	0.4636
LPC-SI [29]	0.3301	0.3798	0.2558	0.4902
HVS-MaxPol-1 [30]	0.4584	0.5048	0.5032	0.6170
HVS-MaxPol-2 [30]	0.5318	0.4742	0.2991	0.5736
Proposed (*SF* only)	0.8273	0.7964	0.7766	0.7867
Proposed (*SN* only)	0.6210	0.5118	0.4975	0.6115
Proposed SFSN	**0.8979**	**0.8379**	**0.8004**	**0.8407**

Since different categories of SISR methods often generate drastically different appearance of the reconstructed images, it is intriguing to investigate how IQA methods perform for different SISR categories. The results on the QADS [15] database are reported in Table II, where the proposed method delivers superior performance in each of the interpolation-based, dictionary-based, and DL-based SISR categories, as well as when all three categories are evaluated together. Ablation test has also been conducted to assess the performance when only the *SF* or *SN* measure is employed. The results are shown in Tables I and II. It appears that both *SF* and *SN* measures make important contributions, but the best performance is achieved by the SFSN model that combines both of them.

IV. CONCLUSION

In this work, we opt to a 2D approach to assess the quality of SISR generated images as a tradeoff between structural fidelity and statistical naturalness. This allows us to better understand the nature of quality degradations and better observe the varying behaviors of different SISR algorithms. We also show that rather straightforward implementations of a local structural fidelity assessment, a global statistical naturalness measure, and a linear combination of the two, result in a SFSN model that achieves surprisingly high correlations with MOS when tested using public databases. In the future, better structural fidelity and statistical naturalness measures, and more sophisticated combination methods may be developed. The 2D assessment idea may also be integrated into novel SISR algorithms, aiming to achieve both goals at the same time and meanwhile balance between them.

REFERENCES

[1] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[2] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[3] Q. Wang and R. K. Ward, "A new orientation-adaptive interpolation method," *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 889–900, 2007.

[4] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[5] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2216–2223.

[6] W. Yang, Y. Tian, F. Zhou, Q. Liao, H. Chen, and C. Zheng, "Consistent coding scheme for single-image super-resolution via independent dictionaries," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 313–325, 2016.

[7] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[8] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[9] J. Yamanaka, S. Kuwashima, and T. Kurita, "Fast and accurate image super resolution by deep cnn with skip connection and network in network," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 217–225.

[10] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3147–3155.

[11] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 624–632.

[12] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

[13] H. Yeganeh, M. Rostami, and Z. Wang, "Objective quality assessment of interpolated natural images," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4651–4663, 2015.

[14] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[15] F. Zhou, R. Yao, B. Liu, and G. Qiu, "Visual quality assessment for super-resolved images: database and method," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3528–3541, 2019.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[17] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. IEEE, 2003, pp. 1398–1402.

[18] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[19] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2385–2401, 2009.

[20] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2011.

[21] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2013.

[22] W. Sun, Q. Liao, J.-H. Xue, and F. Zhou, "SPSIM: A superpixel-based similarity index for full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4232–4244, 2018.

[23] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[24] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[25] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[26] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[27] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *IEEE International Conference on Image Processing*, 2015, pp. 339–343.

[28] C. T. Vu, T. D. Phan, and D. M. Chandler, "S3: A spectral and spatial measure of local perceived sharpness in natural images," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 934–945, 2011.

[29] R. Hassen, Z. Wang, and M. M. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2798–2810, 2013.

[30] M. S. Hosseini, Y. Zhang, and K. N. Plataniotis, "Encoding visual sensitivity by maxpol convolution filters for image sharpness assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4510–4525, 2019.

[31] Z. Chen, W. Zhou, and W. Li, "Blind stereoscopic video quality assessment: From depth perception to overall experience," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 721–734, 2017.

[32] W. Zhou, Q. Jiang, Y. Wang, Z. Chen, and W. Li, "Blind quality assessment for image superresolution using deep two-stream convolutional networks," *Information Sciences*, vol. 528, pp. 205–218, 2020.