Deep Learning-based Distortion Sensitivity Prediction for Full-Reference Image Quality Assessment

Sewoong Ahn, Yeji Choi, and Kwangjin Yoon
SI analytics
441, Expo-ro, Yuseong-gu, Daejeon, Korea
{anse3832, yejichoi, yoon28}@si-analytics.ai

Abstract

Previous full-reference image quality assessment methods aim to evaluate the quality of images impaired by traditional distortions such as JPEG, white noise, Gaussian blur, and so on. However, there is a lack of research measuring the quality of images generated by various image processing algorithms, including super-resolution, denoising, restoration, etc. Motivated by the previous model that predicts the distortion sensitivity maps, we use the DeepQA as a baseline model on a challenge database that includes various distortions. We have further improved the baseline model by dividing it into three parts and modifying each: 1) distortion encoding network, 2) sensitivity generation network, and 3) score regression. Through rigorous experiments, the proposed model achieves better prediction accuracy on the challenge database than other methods. Also, the proposed method shows better visualization results compared to the baseline model. We submitted our model in NTIRE 2021 Perceptual Image Quality Assessment Challenge and won 12th in the main score.

1. Introduction

With the rapid increase of users in social media, it is essential to build systems that quickly transmit a large number of images to provide the best experiences to users. Therefore, various compression algorithms are applied to raw images, while various distortions are added due to wireless transmissions. Therefore, service engineers need to evaluate the distortions in images to provide high-quality images to users. When a person sees images, various operations are applied during the transmissions of visual information from the eyes to the brain. Accordingly, various image quality assessment (IQA) methods [30, 44, 50] have been proposed considering the human visual system (HVS) responses to image impairment.

Also, as the generative adversarial network (GAN) is successfully applied to various image generation applications [10, 16, 52], there are more needs to objectively assess the quality of images generated by GAN. In previous works [9, 36], inception score (IS) and Frechet inception distance (FID) are used to measure the quality and diversity of generated images. However, these scores do not match with human perceptions. In other words, high IS and FID does not guarantee that the generative models can make images with high perceptual quality. Therefore, some researchers [17, 24, 46] benchmarked their generative mod-

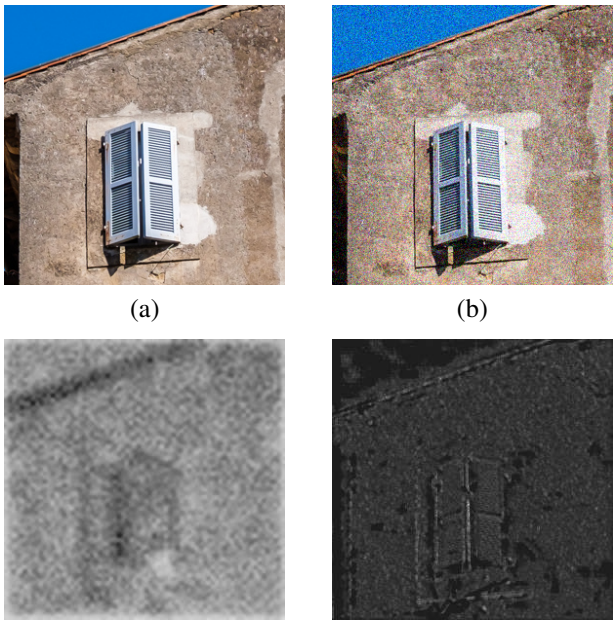

Figure 1. Examples of generated distortion sensitivity maps: (a) represents a reference image; (b) represents an image distorted by white noise; (c) represents a distortion sensitivity map generated by *Kim et al.* [19]; (d) represents a sensitivity map generated by the proposed models. In (c) and (d), darker regions mean higher distortion sensitivity.

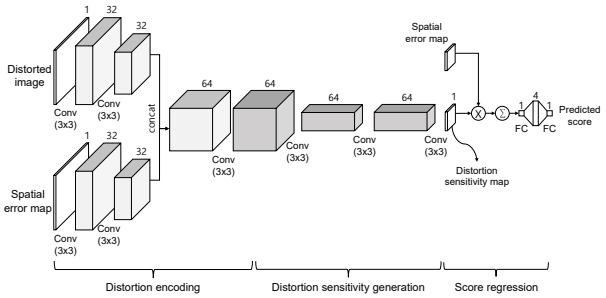

Figure 2. Overall framework of [19].

els using IQA metrics, such as SSIM [44] and NIQE [30]. However, these metrics are optimized to predict images distorted by synthetic distortions (e.g. JPEG, Gaussian blur, white noise, and so on), not the images distorted by generative models. Therefore, it is necessary to develop a model that evaluates the quality of images generated by generative models as well as synthetic distortions.

1.1. Limitations of *DeepQA*

Motivated by the work of *Kim et al.* [19], we use the model of [19] as a baseline to predict the quality of images, which is named as *DeepQA*. The baseline model generates distortion sensitivity maps as intermediate results, visualizing which areas are more sensitive to distortions. The model consists of three parts, as shown in Fig. 2: 1) distortion encoding network, 2) distortion sensitivity generation network, and 3) quality score regression network. In [19], it achieves state-of-the-art performance in various image quality assessment databases, but it has some limitations.

First, because of down-sampling operations, the resolution of distortion sensitivity maps are $1/4$ of distorted images, losing spatial information. Also, in the distortion encoding network, it uses two consecutive 3×3 convolution layers, which means that the effective receptive fields are 5×5. However, these receptive fields are not large enough for predicting an image quality because humans tend to assess images considering global and semantic information. After measuring the quality score by multiplying distortion sensitivity and spatial error maps, it passes fully connected (FC) layers for regression. However, we empirically found that these FC layers lower the performance.

To resolve these problems, we made some modifications to improve the performance of the baseline model [19]:

- Instead of using down-sampling operations for predicting distortion sensitivity maps, we use UNet structure [34] to conserve the spatial information of input images.
- In the distortion encoding network, we add more convolution layers to enlarge the receptive fields.
- We directly predict an image quality instead of using FC layers for regression.

The comparison between the baseline model and the proposed model is depicted in Fig. 1. As shown in Fig. 1, the white Gaussian noise is more sensitive in flat regions (sky in the left top side and window in the bottom side), which is well depicted in Fig. 1 (d). However, in Fig. 1 (c), the distortion sensitivity map losses the detail information, and it does not predict the high sensitivity in the sky region.

Also, the contributions are summarized as follows:

- By modifying the baseline model, we can get more spatial information from distortion sensitivity maps than the baseline model, resulting in performance improvements.
- We visualize the distortion sensitivity maps as intermediate results, enabling us to analyze the results.
- The proposed IQA model achieves high performance on a database with various distortion types, which is challenging. (NTIRE 2021 Perceptual Image Quality Assessment Challenge [12] : 12th in the main score)

The remainder of the paper is organized as follows. Section 2 introduces previous works of sensitivity in human perception and image quality assessment methods. Section 3 describes the overall flow of the proposed model. Section 4 shows rigorous experimental results, and conclusion are given in Section 5

2. Related works

2.1. Human perception regarding sensitivity

Several researchers proposed computational models of human visual sensitivity. Contrast sensitivity function (CSF) represents the varying sensitivity of human eyes according to the spatial frequency of images. According to the CSF, which acts as a band-pass filter, humans are not sensitive to signals which contains high or low frequencies [5]. Therefore, several studies are explaining that distortions are less noticeable if strong contrast or texture exists, which is called visual masking effects [5, 25].

Based on these observations, many IQA methods have been proposed. For example, *Wang et al.* [44] proposed a structural similarity index (SSIM) which assumed that humans are sensitive to contrast and structural distortions. Also, *Zhang et al.* [50] proposed feature similarity index (FSIM), assuming that phase congruency is crucial for human perception. In addition, many no-reference image quality assessment [29, 30, 31, 35] uses mean subtracted contrast normalization as pre-processing, assuming that a human primarily perceives structure information.

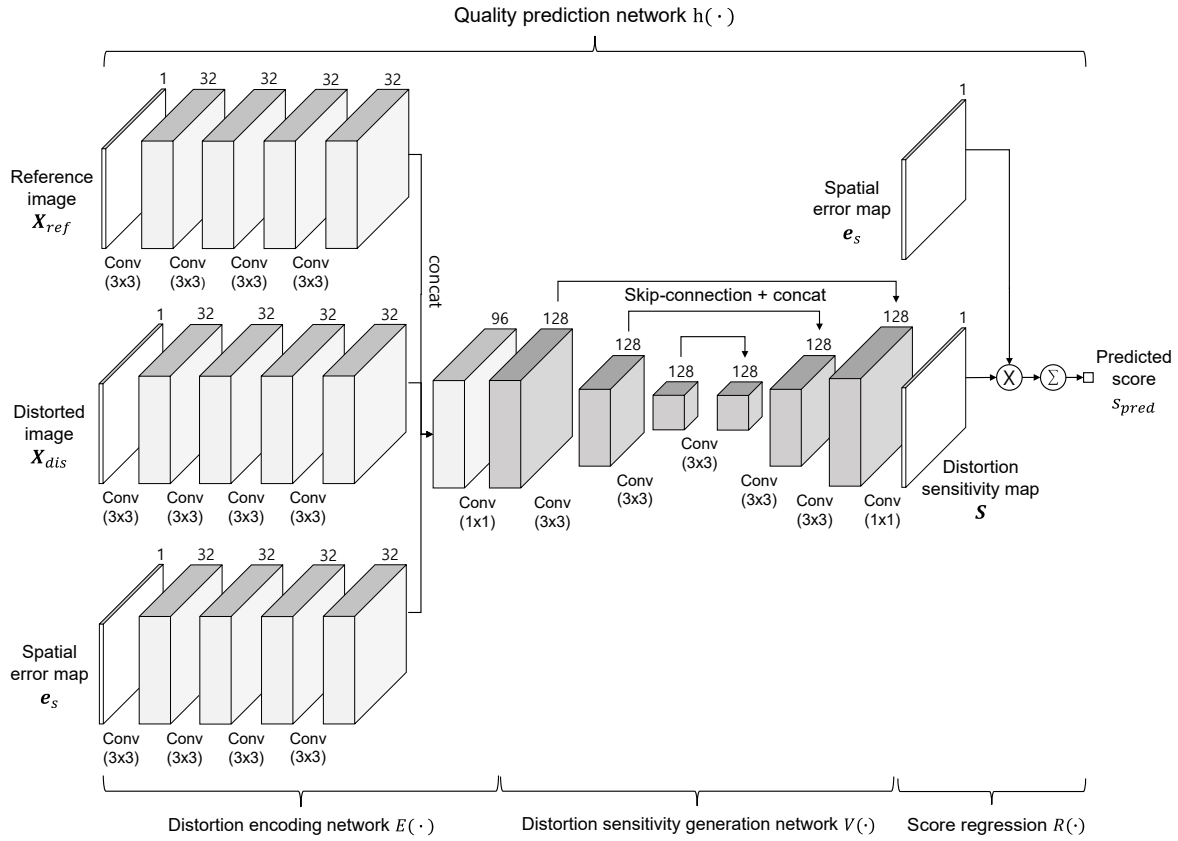

Figure 3. The architecture of the proposed model. The model takes a reference image, a distorted image, and a spatial error map and generates a distortion sensitivity map as an intermediate result. The sensitivity map is multiplied by the spatial error map, and the mean value is regarded as the predicted quality score.

2.2. Image quality assessment

Previous IQA methods can be categorized into: full-reference image quality assessment (FR-IQA) and no-reference image quality assessment (NR-IQA). FR-IQA methods use both references and distorted images to predict the quality scores of distorted images, while NR-IQA methods evaluate the image qualities without references.

Most previously proposed FR-IQA methods were developed based on the measurement of perceptual distances between distorted images and reference images [44, 50]. In the case of NR-IQA, researchers found different statistical characteristics between reference and distorted images, which is defined as natural scene statistics (NSS) [29, 30, 31, 35]. Although these methods show comparable performances, there are limitations that these hand-crafted features do not fully represent the HVS which explains the process of image perception.

Motivated by successful achievements in computer vision areas using convolutional neural networks (CNN) [13, 22, 41], many recent IQA studies apply CNN to extract features in a data-driven way. *Kang et al.* [15] firstly used CNN to predict the quality of images, applying a patch-based approach. *Kim et al.* [18] proposed a new approach for data augmentation and train the CNN model in two stages: 1) local quality prediction, 2) subjective score regression. *Lin et al.* [26] used GAN to recover pseudo-reference images from distorted images and predict the quality of distorted images with the aid of pseudo-reference images. *Su et al.* [40] proposed a hyper network which weights are decided by the semantic information extracted from a pre-trained ResNet [13] on the IMAGENET database [7].

3. Proposed framework

3.1. Model architecture

The overall flow of the proposed model $h(\cdot)$ is depicted in Fig. 3. The model uses a reference image $\mathbf{X}_{ref}$, a distorted image $\mathbf{X}_{dis}$, and a spatial error map $\mathbf{e}_s$, and it predicts a quality score s_{pred} of the distorted image. The model is trained end-to-end, and it generates a distortion sensitivity map $\mathbf{S}$ as an intermediate output, which explains the distortion sensitivity in terms of human perception. The model consists of three parts: 1) distortion encoding network $E(\cdot)$,

2) distortion sensitivity generation newtork $V(\cdot)$, and 3) score regression $R(\cdot)$. The details of each part are explained in Sections 3.1.1, 3.1.2, and 3.1.3.

3.1.1 Distortion encoding network

In the first part, feature maps containing distortion information are extracted from a reference image $\mathbf{X}_{ref}$, a distorted image $\mathbf{X}_{dis}$, and a spatial error map $\mathbf{e}_s$. In case of $\mathbf{X}_{ref}$ and $\mathbf{X}_{dis}$, gray-scale images are used, and pixel values are normalized to $[0, 1]$. Since the pixel differences between $\mathbf{X}_{ref}$ and $\mathbf{X}_{dis}$ yield many near-zero values, we use normalized log difference for measuring $\mathbf{e}_s$ as in [19, 20]:

$$\mathbf{e}_s = \frac{log(1/((\mathbf{X}_{ref} - \mathbf{X}_{dis})^2 + \epsilon/255^2))}{log(255^2/\epsilon)}, \quad (1)$$

where we use $\epsilon = 1$ in our model. In [19], two consecutive 3×3 convolution layers are used to extract feature maps. However, it means the model uses small receptive fields (similar to 5×5 filters), which is not appropriate for IQA because humans perceive an image quality considering global regions of an image. Therefore, we use more convolution layers to exploit larger receptive fields to extract distortion information. Furthermore, it leads to the representation with non-linearity. In our model, each input is individually fed into four consecutive 3×3 convolution layers (each layer has 32 filters), and the outputs are concatenated, producing 96-dimensional feature maps. And then, 1×1 convolution is applied to change the channel dimension from 96 to 128. Finally, the 128-dimensional feature maps are extracted in the first part of our model.

3.1.2 Distortion sensitivity generation network

In the second part, a distortion sensitivity map $\mathbf{S}$ is generated from the 128-dimensional feature maps of the distortion encoding layer $E(\cdot)$. In [19], the resolution of the distortion sensitivity map is $1/4$ of the input image because of down-sampling operations. Therefore, a model similar as UNet [34] ($V(\cdot)$ in Fig. 3) is used for generating the distortion sensitivity map. And then, 1×1 convolution is applied for channel conversion, yielding $\mathbf{S}$.

3.1.3 Score regression

In the last part, an image quality score is predicted. At first, the perceptual error map $\mathbf{e}_{percept}$ is defined as

$$\mathbf{e}_{percept} = \mathbf{e}_s \odot \mathbf{S}, \quad (2)$$

where $\odot$ means the element-wise product. The predicted quality score s_{pred} is the mean of perceptual error $\mathbf{e}_{percept}$:

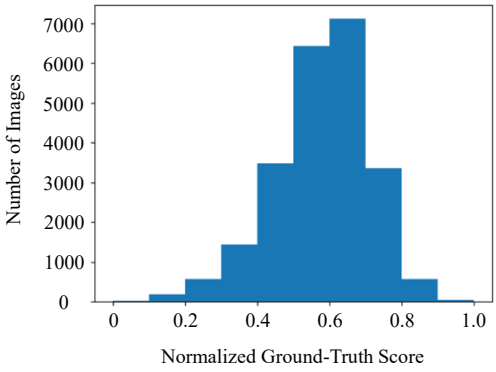

Figure 4. Histogram of PIPAL database (train phase) according to normalized ground-truth scores.

$$s_{pred} = \frac{1}{H_e \cdot W_e} \sum_{i=1}^{H_e} \sum_{j=1}^{W_e} e_{percept}(i, j) \quad (3)$$

where H_e and W_e means the height and width of perceptual error map $\mathbf{e}_{percept}$, respectively.

In [19], fully connected (FC) layers are used for regression. However, in our models, the FC layers degrade the performance, so we delete the FC layers, and s_{pred} is used directly for the loss function.

The loss function of the proposed model is mean-square error between the predicted quality score s_{pred} and subjective scores s_{subj}:

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{s}_{pred,i} - \mathbf{s}_{subj,i})^2 \quad (4)$$

where N means the total number of distorted images.

Also, L2 regularization is used to avoid over-fitting. Finally, total loss function L_{total} is given as:

$$L_{total} = L_{MSE} + \lambda \cdot L_2 \quad (5)$$

where L_2 is L2 regularization loss, and λ controls the relative importance of the two loss terms.

3.2. Training method

To optimize the total loss function L_{total}, the adaptive moment estimation optimizer (ADAM) [21] was used, and we use the hyperparameter suggested in [21]. The learning rate was initially set to 1×10^{-3}, and it was multiplied by 0.8 for every 20 epochs. Also, the relative weight was set to $\lambda = 1 \times 10^{-5}$.

As shown in Fig. 4, most of the images in the PIPAL database (train phase) are in the range of $[0.4, 0.8]$. Therefore, the images outside of this range were oversampled to

Table 1. SRCC and PLCC Comparisons of IQA Models on the PIPAL test dataset. *Italics* mean Deep Learning-Based Methods.

Type	Method	PIPAL (test)	
		SRCC	PLCC
NR	NIQE	0.0341	0.1317
	Ma	0.1405	0.1469
	PI	0.1036	0.1454
FR	PSNR	0.2493	0.2769
	NQM	0.3645	0.3954
	UQI	0.4195	0.4500
	SSIM	0.3614	0.3936
	MS-SSIM	0.4618	0.5007
	IFC	0.4851	0.5549
	VIF	0.3970	0.4795
	VSNR	0.3682	0.4107
	RFSIM	0.3037	0.3284
	GSM	0.4094	0.4646
	SRSIM	0.5728	0.6360
	FSIM	0.5038	0.5709
	FSIMc	0.5057	0.5727
	VSI	0.4584	0.5169
	MAD	0.5434	0.5804
	LPIPS-Alex	0.5658	0.5711
	LPIPS-VGG	0.5947	0.6331
	PieAPP	0.6074	0.5974
	WaDIQaM	0.5533	0.5408
	DISTS	**0.6548**	**0.6873**
	SWD	0.6243	0.6342
	Proposed	**0.6744**	**0.6535**

Table 2. SRCC and PLCC Comparisons of IQA Models on the PIPAL validation dataset.

Method	PIPAL (valid)	
	SRCC	PLCC
DeepQA	0.6966	0.7129
DeepQA + modify	0.7529	0.7623
DeepQA + modify + aug (proposed)	0.7951	0.7854

relieve score imbalance problems. In addition, the horizontal flip was also used for data augmentation. We did not use other augmentation methods such as resizing, rotating and cropping because these methods would change the quality score of images.

The proposed model was implemented in the Pytorch library on the Python 3.7 platform, using a single GPU (Tesla V100-DGXS-32GB). Using these setups, the training and inference time on the database in Section 4.1 was 12.25min/epoch and 0.0184s/image, respectively.

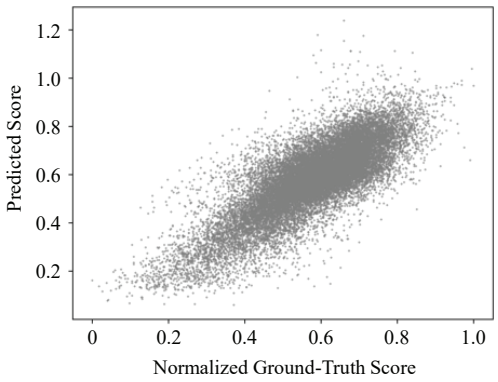

Figure 5. Scatter plot of normalized ground-truth score versus predicted scores. Each sample point represents one image.

4. Experimental results

4.1. Dataset

A dataset provided by NTIRE 2021 Perceptual Image Quality Assessment challenge [12] was used to evaluate the proposed algorithm: PIPAL [14]. The PIPAL database contains 250 reference images and about 29,000 distorted images impaired by 40 distortion types. These distortion types are categorized into four sub-types: traditional, super-resolution, denoising, and mixture restoration. Each image size is 288×288, and the ground truth scores are given as ELO rating scores. We normalized the score in the range of $[0, 1]$, where a value near 1 means high quality. The database is divided into three phases: train, validation, and test. Each phase contains 200, 25, 25 reference images, and these phases also have distorted images according to their references. In Sections 4.2 and 4.3, we trained our model on the train phase and tested our model on validation and test phases.

In the cross-dataset tests, we also used two well-known datasets: LIVE [39] and TID2013 [32] databases. The LIVE database consists of 29 reference images and 799 distorted images, containing five types of distortions: JPEG, JP2K, white noise, Gaussian blur, and fast-fading. The TID2013 database contains 25 reference images and 3,000 distorted images impaired by 24 distortion types at five levels of degradation. Since the ground-truth scores of the LIVE database are given as differential mean opinion scores (MOS), we normalize the differential MOS (DMOS) to $[0, 1]$ and converted to MOS (MOS = 1-DMOS).

4.2. Benchmark results

To compare the performances of the IQA algorithms, two correlation coefficients between predicted scores and ground-truths are employed: Spearman's rank-order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC). These coefficients are ranged to $[0, 1]$,

Table 3. Cross-Dataset Test Results on Various IQA Datasets

		Test					
		PIPAL (train)		LIVE		TID2013	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Train	PIPAL (train)	0.9002	0.9001	0.9263	0.9211	0.5371	0.6264
	LIVE	0.5051	0.5148	0.9807	0.9771	0.4496	0.5226
	TID2013	0.5950	0.6263	0.6074	0.7113	0.7783	0.7932

and the value closer to 1 means higher performance.

Table 1 shows the performances of FR and NR-IQA models on the PIPAL test dataset. We compared the proposed model against FR-IQA models (PSNR [1], NQM [6], UQI [43], SSIM [44], MS-SSIM [45], IFC [38], VIF [37], VSNR [4], RFSIM [49], GSM [27], SRSIM [47], FSIM [50], FSIMc [50], VSI [48], MAD [23], LPIPS-Alex [51], LPIPS-VGG [51], PieAPP [33], WaDIQaM [3], DISTS [8], SWD [11], and DeepQA [19]) and NR-IQA models (NIQE [30], Ma [28], and PI [2]). The bold fonts indicate the two top-performing models on the test dataset. Among the FR/NR-IQA models, the deep learning-based methods generally showed superior performance compared to previous hand-crafted methods. Also, the proposed model attained the highest correlation with subjective scores on the test dataset.

The scatter plot of normalized MOS versus output scores by the proposed model is shown in Fig. 5. As shown in Fig. 5, the proposed model well predicts the quality scores of distorted images, except for some outliers.

4.3. Ablation studies

In addition to performance comparisons, we also conducted ablation tests for model modification and data augmentation methods stated in Section 3. *DeepQA* [19] is the baseline model, *DeepQA+modify* means that three types of modifications (increasing convolution layers, using a UNet structure, and removal of FC layers) in Section 3.1 are applied to *DeepQA*, and *aug* means the data augmentation stated in Section 3.2. In other words, *DeepQA+modify+aug* in Table 2 is same as *proposed* in Table 1, but they use different phase of the PIPAL database when evaluating performances.

Compared to the performances of *DeepQA* and *DeepQA+modify* in Table 2, three types of modifications result performance improvements. Furthermore, compared the results of *DeepQA+modify* and *DeepQA+modify+aug*, data augmentation methods stated in Section 3.2 improves the performance of the proposed model. Since the distribution of normalized ground-truth scores on the PIPAL database is concentrated in a certain range, it causes score imbalance problems, resulting in the performance degradation on images in which ground-truth scores are rare.

4.4. Cross-dataset tests

To test the generalization ability of the proposed model, we conducted cross-dataset tests. In these experiments, we used the PIPAL, LIVE, and TID2013 databases. In Table 3, when train and test datasets are the same, we randomly divided reference images of the dataset into train and test sets (80% for training and 20% for testing) and divided distorted images according to their references. Also, in the PIPAL dataset, we only use the training phase because we can get the ground-truth score in the training phase only.

As shown in Tabel 3, except when the proposed model is trained on the PIPAL database, performances are the best when the train and test datasets are the same, and the performances degrade when the test dataset changes. It means that the proposed model tends to learn the distortion properties in the train dataset, so the distortions that do not exist in the train dataset degrade the performances. This tendency is most evident in the LIVE database, including only five types of distortions. When the proposed model is trained on the PIPAL database, test performances on other databases least degrade because the PIPAL database contains various types of distortions.

4.5. Distortion sensitivity visualization

Following the training step stated in Section 3, the model takes reference images $\mathbf{X}_{ref}$, distorted images $\mathbf{X}_{ref}$ and spatial error maps $\mathbf{e}_s$ as inputs, and the model generates distortion sensitivity maps $\mathbf{S}$ as intermediate outputs. To validate distortion sensitivity maps reflect the human perception, model inputs ($\mathbf{X}_{ref}$, $\mathbf{X}_{dis}$ and $\mathbf{e}_s$) and intermediate outputs ($\mathbf{S}$) are shown in Fig. 6. In Fig. 6, each column, from left to right, represents reference images $\mathbf{X}_{ref}$, distorted images $\mathbf{X}_{ref}$, spatial error maps $\mathbf{e}_s$ obtained by Eq. 1, and the corresponding generated distortion sensitivity maps $\mathbf{S}$, respectively. In third and fourth column of Fig. 6, darker regions indicate higher values, which means more spatial error and higher distortion sensitivity, respectively.

Fig. 6 (b) is an output of a generative model for super-resolution, named as ESRGAN [42]. In other words, Fig. 6 (a) is down-sampled ($\times 4$) using a bicubic method and up-sampled to the original resolution using ESRGAN. Comparing Figs. 6 (a) and (b), repeated patterns are collapsed in some parts of brick areas. While the spatial error map (Fig. 6 (c)) has similar values in all brick regions, the predicted

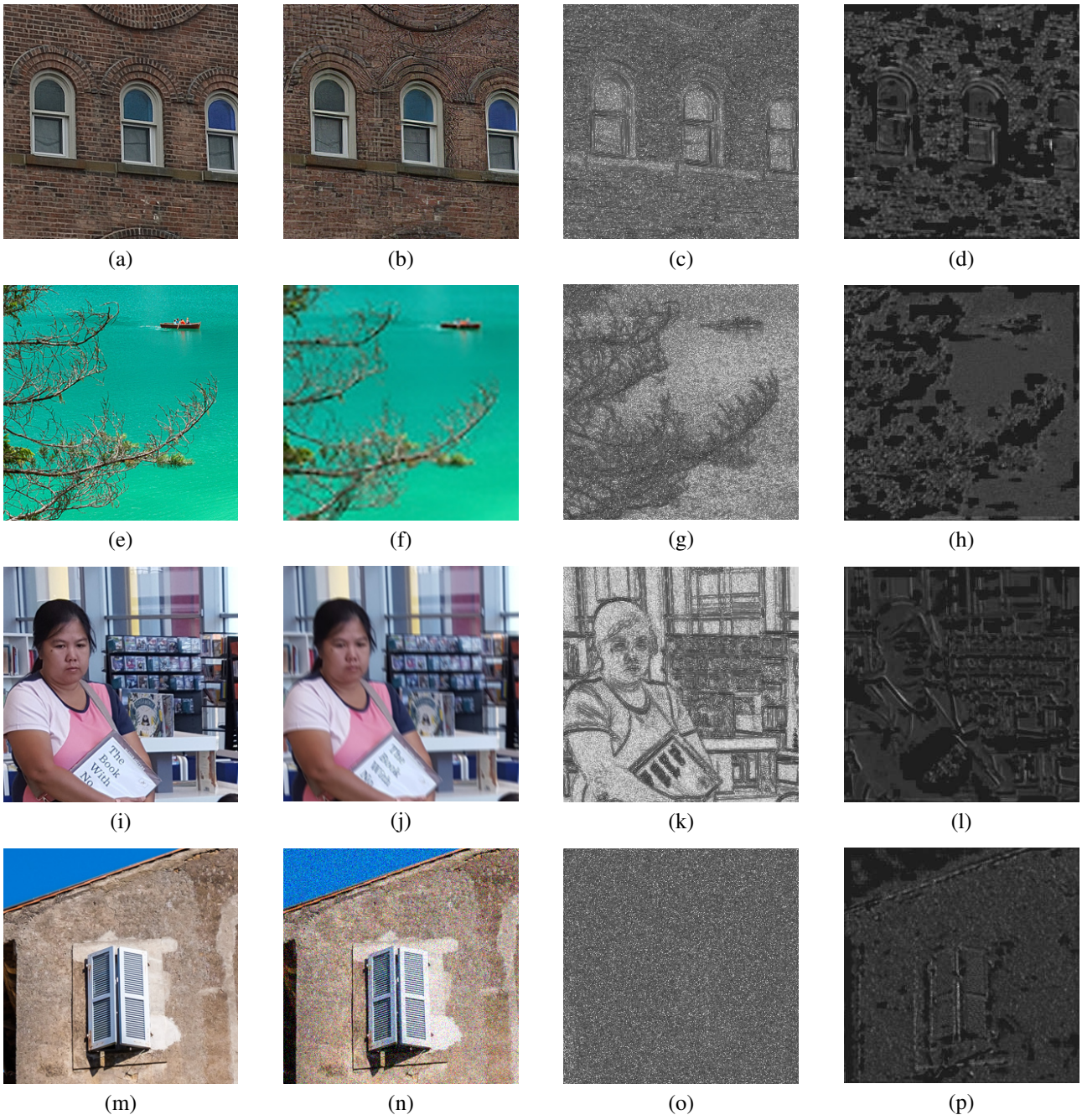

Figure 6. Visualization results of generated distortion sensitivity maps. (a), (e), (i), and (m) are reference images, while the distorted images are shown in (b), (f), (j), and (n). The spatial error maps of the distorted images are shown in (c), (g), (k), and (o), and the distortion sensitivity maps of distorted images are shown in (d), (h), (l), and (p).

distortion sensitivity map (Fig. 6 (d)) correctly predicts high distortion sensitivities where the patterns are destroyed.

In the second row of Fig. 6, a reference image is distorted by compression algorithms where high frequency components are suppressed. Interestingly, the corresponding distortion sensitivity map (Fig. 6 (h)) predicts that distortions occur at regions around branches and a boat where many details are deteriorated due to the compression.

In the third row, Fig. 6 (j) is impaired by blurring artifacts. As a result, people are more sensitive to distortions in edge regions than in flat regions. Therefore, the proposed model predicts that the distortion sensitivity of a book (in the front area) is higher than those of arms, depicted in Fig. 6 (l).

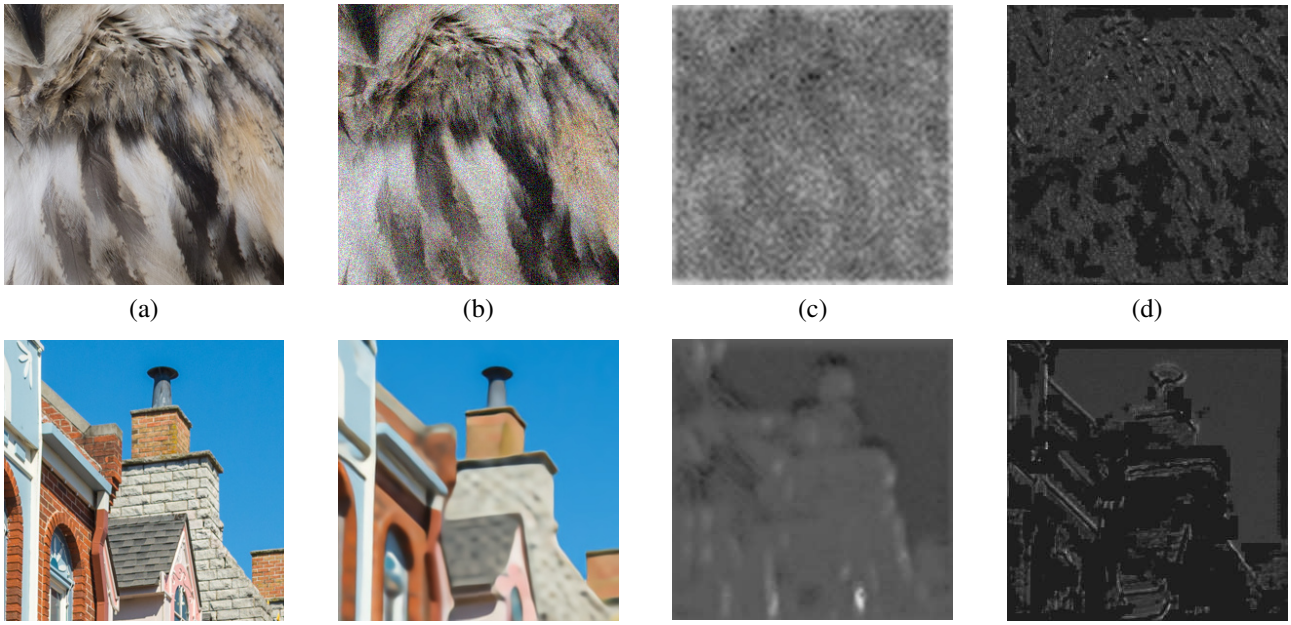

Figure 7. Comparisons of distortion sensitivity maps. (a) and (e) are reference images, while the distorted images are shown in (b) and (f). The distortion sensitivity maps generated by the baseline model [19] are are shown in (c) and (g), and the distortion sensitivity maps of the proposed model are shown in (d) and (h).

In the fourth row, Fig. 6 (n) is corrupted with white noise. In this case, humans are more sensitive to distortion in flat areas. As shown in Fig. 6 (p), the model appropriately predicts high distortion sensitivities in a flat sky (on the top left side) and window (on the bottom side) regions.

We also compared the distortion sensitivity maps of [19] and those of the proposed model in Fig. 7. Since the original resolution of Figs. 7 (c) and (g) is $1/4$ of Figs. 7 (b) and (f), we use bicubic up-sampling to make the size of Figs. 7 (c) and (g) same as Figs. 7 (d) and (h). In the third and fourth column in Fig. 7, darker regions indicate higher errors and distortion sensitivity, respectively.

According to the first row, an original image is distorted by white Gaussian noise, and people are sensitive to noise in flat regions. Therefore, people feel more distortions in the bottom areas, which is also observed in Fig. 7 (d) while Fig. 7 (c) is not. In the second row, a reference image is distorted by edge-preserving smoothing algorithms. Therefore, people feel more distortion in texture regions (bricks) than in a flat region (a sky), as depicted in Fig. 7 (h).

In conclusion, the distortion sensitivity maps of the proposed model well predict which areas are sensitive to noise. In previous works, most of the FR-IQA models [4, 6, 37, 43, 44] predict the quality of images using a difference between original and distorted images (in the third column of Fig. 6). It is based on the belief that people feel noise when errors are high. However, when comparing Figs. 6 (o) and (p), it is not always true. In other words, according to the visual masking effects [5, 25], some distortion is not visible to human eyes, as shown in Fig. 6 (p). Therefore, we predict the distortion sensitivity in a data-driven way with a triplet of inputs (reference/distorted images and spatial error maps). Finally, the proposed model successfully predicts which regions are sensitive to impairments.

5. Conclusion

We proposed a deep learning approach to the problem of FR-IQA. Motivated by the work of [19], the proposed model predicts the distortion sensitivity map, mimicking the process of HVS. Through various experiments, we demonstrated that the proposed model predicts the quality scores of distorted images well, and generates distortion sensitivity maps that agree with human perception. In addition, as compared to the benchmarks of [14], the proposed model gets better results by resolving various problems in [19]. However, in the real world problem, reference images do not usually exist, so we are going to study NR approaches as future works.

References

[1] Jochen Antkowiak, TDF Jamal Baina, France Vittorio Baroncini, Noel Chateau, France FranceTelecom, Antonio Claudio França Pessoa, FUB Stephanie Colonnese, Italy Laura Contin, Jorge Caviedes, and France Philips. Final

report from the video quality experts group on the validation of objective models of video quality assessment march 2000. 2000. 6

[2] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 6

[3] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017. 6

[4] Damon M Chandler and Sheila S Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE transactions on image processing*, 16(9):2284–2298, 2007. 6, 8

[5] Scott J Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, volume 1666, pages 2–15. International Society for Optics and Photonics, 1992. 2, 8

[6] Niranjan Damera-Venkata, Thomas D Kite, Wilson S Geisler, Brian L Evans, and Alan C Bovik. Image quality assessment based on a degradation model. *IEEE transactions on image processing*, 9(4):636–650, 2000. 6, 8

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[8] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020. 6

[9] William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: Gans do not need to decrease a divergence at every step. *arXiv preprint arXiv:1710.08446*, 2017. 1

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 1

[11] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric. *arXiv preprint arXiv:2011.15002*, 2020. 6

[12] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Yu Qiao, Shuhang Gu, Radu Timofte, et al. NTIRE 2021 challenge on perceptual image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2, 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[14] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651. Springer, 2020. 5, 8

[15] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014. 3

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 1

[18] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1):206–220, 2016. 3

[19] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1676–1684, 2017. 1, 2, 4, 6, 8

[20] Woojae Kim, Anh-Duc Nguyen, Sanghoon Lee, and Alan Conrad Bovik. Dynamic receptive field generation for full-reference image quality assessment. *IEEE Transactions on Image Processing*, 29:4219–4231, 2020. 4

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 3

[23] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010. 6

[24] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan 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*, pages 4681–4690, 2017. 1

[25] Gordon E Legge and John M Foley. Contrast masking in human vision. *Josa*, 70(12):1458–1471, 1980. 2, 8

[26] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 732–741, 2018. 3

[27] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, 2011. 6

[28] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 6

[29] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 2, 3

[30] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 1, 2, 3, 6

[31] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 2, 3

[32] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015. 5

[33] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 6

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 4

[35] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012. 2, 3

[36] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016. 1

[37] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006. 6, 8

[38] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12):2117–2128, 2005. 6

[39] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 5

[40] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 3

[41] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3

[42] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 6

[43] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002. 6, 8

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1, 2, 3, 6, 8

[45] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 6

[46] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. *arXiv preprint arXiv:2102.02808*, 2021. 1

[47] Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *2012 19th IEEE international conference on image processing*, pages 1473–1476. IEEE, 2012. 6

[48] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10):4270–4281, 2014. 6

[49] Lin Zhang, Lei Zhang, and Xuanqin Mou. Rfsim: A feature based image quality assessment metric using riesz transforms. In *2010 IEEE International Conference on Image Processing*, pages 321–324. IEEE, 2010. 6

[50] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 1, 2, 3, 6

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1