ARNIQA: Learning Distortion Manifold for Image Quality Assessment

Supplementary Material

Lorenzo Agnolucci Leonardo Galteri Marco Bertini Alberto Del Bimbo
University of Florence - Media Integration and Communication Center (MICC)
Florence, Italy
[name.surname]@unifi.it

S1. Analysis on Data Efficiency

In Sec. 4.4 we show that ARNIQA achieves state-of-the-art performance on several IQA datasets with both synthetic and authentic distortions. In addition, our approach proves to be more data-efficient than competing self-supervised methods, since it requires fewer training examples.

We recall that we rely on the 140K pristine images from the KADIS dataset [7] synthetically distorted with our degradation model to train our model. Given that we consider images both at full-scale and half-scale (see Sec. 3.2), we double the size of the training dataset. For all of the experiments, we train our model for 10 epochs. Therefore, for training, we use a total of 140K (training dataset) × 2 (scales) × 10 (epochs) = 2.8M images.

In contrast, CONTRIQUE [10] considers a combination of images with synthetic and authentic distortions for training, for a total of 1.3M. Specifically, the authors use the 700K synthetically distorted images from the KADIS dataset and the union of 4 datasets with realistic distortions: 255K images from AVA [13], 330K images from COCO [8], 2450 images from CERTH-Blur [11], 33K images from VOC [3]. The CONTRIQUE model employs both full-scale and half-scale images and is trained for 25 epochs. Therefore, the total number of training examples required by CONTRIQUE is given by 1.3M (training dataset) × 2 (scales) × 25 (epochs) = 65M.

Instead, Re-IQA uses two different datasets, both at full-scale and half-scale, as well as a diverse number of epochs, for the content-aware and the quality-aware encoder. In particular, the authors train the content-aware encoder on the 1.28 images of the ImageNet dataset [1] for 200 epochs. Thus, the total number of training examples for the content-aware encoder is given by 1.28M (training dataset) × 2 (scales) × 200 (epochs) = 512M. For the quality-aware encoder, Re-IQA uses the 140K pristine images from the KADIS dataset and the same combination of datasets with authentic distortions as CONTRIQUE, for a total of 760K images. Given that the authors train the quality-aware encoder for 25 epochs, the total number of training examples results in 760K (training dataset) × 2 (scales) × 25 (epochs) = 38M. Considering both the content-aware and the quality-aware encoders, Re-IQA requires a total of 550M images for training.

Ultimately, despite using only the 4.3% and 0.5% of the training examples compared, respectively, to CONTRIQUE and Re-IQA, ARNIQA manages to achieve state-of-the-art performance on several IQA datasets, thereby showing improved data efficiency.

S2. Additional Experimental Results

S2.1. Full-Reference Image Quality Assessment

We can easily extend our approach to the Full-Reference Image Quality Assessment (FR-IQA) task. FR-IQA aims to evaluate the quality of a distorted image in the setting in which a high-quality reference version is available. Similarly to [10, 15], we incorporate the information provided by the reference image with:

$$y = W\,|h_{ref} - h_{dist}| \tag{S1}$$

where y is the quality score, W indicates the trainable weights of the regressor, and h_{ref} and h_{dist} are the representations of the reference and distorted image, respectively. Therefore, the regressor predicts the quality score associated with the difference between the embeddings of the reference and the distorted image.

We follow the same evaluation protocol described in Sec. 4.3, thereby not fine-tuning the encoder weights for the FR-IQA task. Note that we can only evaluate the performance on FR-IQA with datasets consisting of synthetic distortions, given the unavailability of a reference image for datasets with realistic degradations. We report the results in Tab. S1. Despite being designed for NR-IQA, ARNIQA obtains competitive results also on FR-IQA, thus further prov-

Method	Type	LIVE SRCC	LIVE PLCC	CSIQ SRCC	CSIQ PLCC	TID2013 SRCC	TID2013 PLCC	KADID SRCC	KADID PLCC
PSNR	Traditional	0.881	0.868	0.820	0.824	0.643	0.675	0.677	0.680
SSIM [16]		0.921	0.911	0.854	0.835	0.642	0.698	0.641	0.633
FSIM [18]		0.964	0.954	0.934	0.919	0.852	0.875	0.854	0.850
VSI [17]		0.951	0.940	0.944	0.929	0.902	0.903	0.880	0.878
PieAPP [14]	Deep learning	0.915	0.905	0.900	0.881	0.877	0.850	0.869	0.869
LPIPS [19]		0.932	0.936	0.884	0.906	0.673	0.756	0.721	0.713
DISTS [2]		0.953	0.954	0.942	0.942	0.853	0.873	–	–
DRF-IQA [5]		**0.983**	**0.983**	0.964	0.960	**0.944**	**0.942**	–	–
CONTRIQUE-FR [10]	SSL + LR	0.966	0.966	0.956	0.964	0.909	0.915	**0.946**	**0.947**
Re-IQA-FR [15]		0.969	0.974	0.961	0.962	0.920	0.921	0.933	0.936
ARNIQA-FR	SSL + LR	0.969	0.972	**0.971**	**0.975**	0.898	0.901	0.920	0.919

Table S1. Comparison between the proposed approach and competing methods for the FR-IQA task. Best and second-best scores are highlighted in bold and underlined, respectively, – denotes results not reported in the original paper. SSL and LR stands for self-supervised learning and linear regression, respectively.

ing the effectiveness of our approach. Moreover, we observe that the additional information provided by the high-quality reference image leads to improved performance, compared to the NR-IQA setting reported in Tab. 1.

S2.2. Regressor Regularization Coefficient

We recall that during evaluation we freeze the encoder weights and map the image representations to quality scores using simple linear regression, as in Re-IQA [15]. Similarly to CONTRIQUE [10] and Re-IQA [15], we use the validation split of each dataset to identify the regularization coefficient of the Ridge regressor [4] via a grid search over values within the range $\left[10^{-3}, 10^{3}\right]$. To assess the robustness of both ARNIQA and Re-IQA with respect to the choice of the regularization coefficient of the Ridge regression, we conduct an evaluation considering various values in the range $\left[10^{-3}, 10^{3}\right]$. Table S2 shows the results for the SRCC metric on the validation set of the KADID dataset [7]. As explained in Sec. 4.3, we report the median of the results of 10 random training/validation/test splits. We observe that our approach is significantly more robust than Re-IQA. In fact, the difference Δ between the best and worst results obtained for the various values of the regularization coefficient is considerably lower compared to Re-IQA.

Coefficient	Method: Re-IQA†	Method: **ARNIQA**
$\alpha = 0.001$	0.499	0.900
$\alpha = 0.01$	0.565	0.907
$\alpha = 0.1$	0.690	0.912
$\alpha = 1$	0.763	0.914
$\alpha = 10$	0.842	0.907
$\alpha = 100$	0.862	0.894
$\alpha = 1000$	0.858	0.859
Best	0.862	**0.914**
Worst	0.499	**0.859**
Δ	0.368	**0.055**

Table S2. Results for varying regressor regularization coefficient α for the SRCC metric on the validation set of the KADID dataset [7]. Δ indicates the difference between the best and worst scores. † denotes results evaluated by us with the official pre-trained models. The best scores are highlighted in bold.

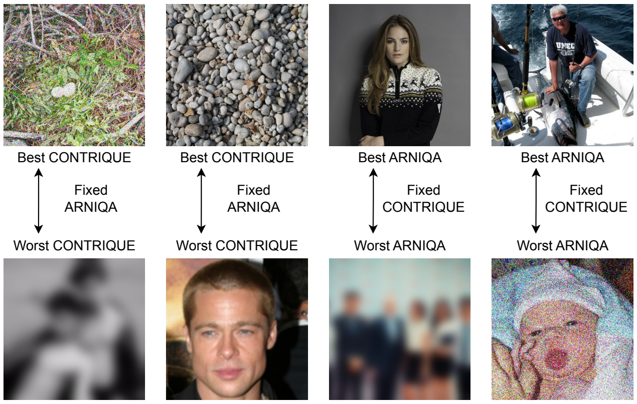
S2.3. gMAD Competition

We conduct the group maximum differentiation (gMAD) competition [9] between ARNIQA and CONTRIQUE [10] to evaluate the robustness of our model. See Sec. 4.4 for more details about gMAD. We report the results in Fig. S1. When we fix ARNIQA at a low-quality level (Fig. S1a), CONTRIQUE struggles to identify picture pairs with a clear quality disparity. On the contrary, when fixing ARNIQA at a high-quality level, the image pair found by CONTRIQUE shows a slight divergence in quality. However, when acting as the attacker (Figs. S1c and S1d), ARNIQA succeeds in highlighting the failures of CONTRIQUE by identifying image pairs exhibiting considerably different quality. Therefore, our method demonstrates superior robustness to that of CONTRIQUE.

S2.4. Manifold Visualization

We carry out an experiment to visualize the inherent structure of the distortion manifold learned by our model. Given two distortion types, our aim is to study the positions occupied in the manifold by images that exhibit both single and combined degradation patterns with varying levels of intensity. For a model that effectively learned the image distortion manifold, we expect images showing combined degradation patterns to occupy positions within the manifold that are intermediate to the locations associated with the single distortions themselves.

To conduct this study, we consider 1000 randomly se-



(a) (b) (c) (d)

Figure S1. gMAD competition results between ARNIQA and CONTRIQUE [10]. (a) and (b): Fixed ARNIQA at a low- and high-quality level, respectively. (c) and (d): Fixed CONTRIQUE at a low- and high-quality level, respectively.

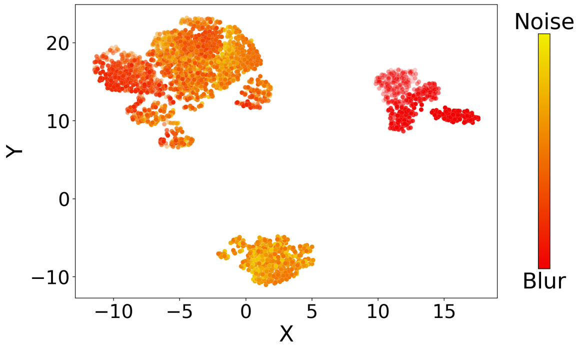
lected pristine images from the KADIS dataset [7] and the Gaussian blur and white noise distortions (see Sec. S3.2 for more details). First, we distort the images individually with each of the two degradations under consideration, using 5 different levels of intensity. Then, we consider all the possible combinations of the degrees of intensity of the Gaussian blur and white noise distortions, taken in this order. Finally, we distort each of the pristine images with each combination by applying the two distortions consecutively. Therefore, for each image, we obtain 5 + 5 embeddings corresponding to the single blur and noise distortions, and 5 × 5 representations for the combined ones.

Figures S2a and S2b shows the UMAP visualization [12] of the embeddings obtained with Re-IQA [15] and ARNIQA, respectively. As we can see, compared to Re-IQA, our approach leads to a smoother transition between the points corresponding to the single and combined degradations. Indeed, the stronger the intensity of the noise distortion, the closer the points are to the cluster of images degraded only with white noise. Note that most of the points corresponding to combined degradation patterns lie closer to the cluster of images distorted only with white noise as it was applied after the blur. Indeed, the final degradation in a distortion composition corresponds to more visible patterns, as they are not modified by subsequent degradations.

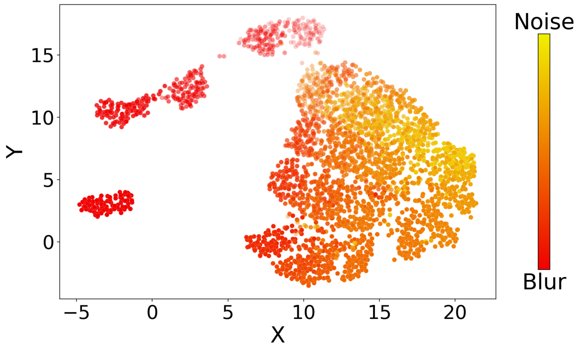
S3. Image Degradation Model

S3.1. Distortion Compositions

In Fig. S3 we report some examples of images belonging to the KADIS dataset [7] subjected to distortion compositions obtained through our image degradation model. We notice how the proposed degradation model leads to images showing a large variety of distortion patterns. In this way, our model is able to effectively learn the image distortion manifold.



(a) Re-IQA [15]



(b) ARNIQA

Figure S2. Manifold visualization with UMAP [12] of the embeddings of 1000 images degraded with Gaussian blur and white noise distortions, applied in this order. The color of each point is given by the weighted average between the colors of blur (red) and noise (yellow), based on the degradation intensity. A higher alpha value corresponds to a stronger degradation intensity.

S3.2. Distortion Types

Our image degradation model considers 24 different degradation types divided into the 7 distortion groups defined by the KADID dataset [7]. Each distortion has 5 levels of increasing intensity. Figures S4 to S10 shows the different levels of intensity for the degradations of each distortion group. The distortion types that we consider are mainly inspired by those of the KADID dataset and are described in the list below:

1. Brightness change:
    - *Brighten*: applies a sequence of color space transformations, curve adjustments, and blending operations to enhance the brightness of an input image, resulting in an output image with increased visual intensity;

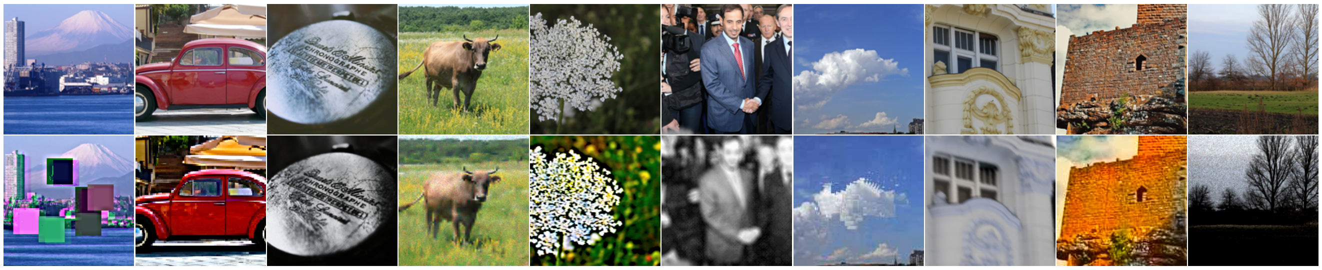


Figure S3. Comparison between pristine images from the KADIS dataset [7] and their distorted versions using the proposed degradation model. *Top*: Pristine images. *Bottom*: Distorted images.

- *Darken*: similar to brighten operation, but it leads to a decreased visual intensity;
- *Mean shift*: changes the average intensity of image pixels by adding a fixed amount to all the pixel values. Then, limits the resulting values to remain within the initial image range;

2. Blur:
    - *Gaussian blur*: filters every pixel of the image with a simple Gaussian kernel.
    - *Lens blur*: filters every pixel of the image with a circular kernel;
    - *Motion blur*: filters every pixel of the image with a linear motion blur kernel to simulate the effect of a moving camera or a moving object in the scene. Consequently, the image appears blurred in the direction of the motion;
3. Spatial distortions:
    - *Jitter*: randomly disperses image data by warping each pixel with small offsets;
    - *Non-eccentricity patch*: randomly extracts patches from the image and inserts them in random neighboring positions;
    - *Pixelate*: combines operations of downscaling and upscaling using nearest-neighbor interpolation;
    - *Quantization*: quantizes the image into N uniform levels. The thresholds are computed dynamically using Multi-Otsu's method [6];
    - *Color block*: randomly overlays homogeneous colored squared patches onto the image;
4. Noise:
    - *White noise*: adds Gaussian white noise to the image;
    - *White noise in color component*: converts the image to the YCbCr color space, then adds Gaussian white noise to each channel;
    - *Impulse noise*: adds salt and pepper noise to the image;
    - *Multiplicative noise*: adds speckle noise to the image;
5. Color distortions:
    - *Color diffusion*: converts the image to the LAB-color space, then applies Gaussian blur to each channel;
    - *Color shift*: randomly shifts the green channel and then blends it into the original image, masked by the normalized gradient magnitude of the original image;
    - *Color saturation 1*: converts the image to the HSV-color space and then multiplies the saturation channel by a factor;
    - *Color saturation 2*: converts the image to the LAB-color space, then multiply each color channel by a factor;
6. Compression:
    - *JPEG2000*: applies standard JPEG2000 compression to the image;
    - *JPEG*: applies standard JPEG compression to the image;
7. Sharpness & contrast:
    - *High sharpen*: sharpens the image in the LAB-color space using unsharp masking;
    - *Nonlinear contrast change*: calculates a nonlinear tone mapping operation to manipulate the contrast of the image;
    - *Linear contrast change*: calculates a linear tone mapping operation to manipulate the contrast of the image;

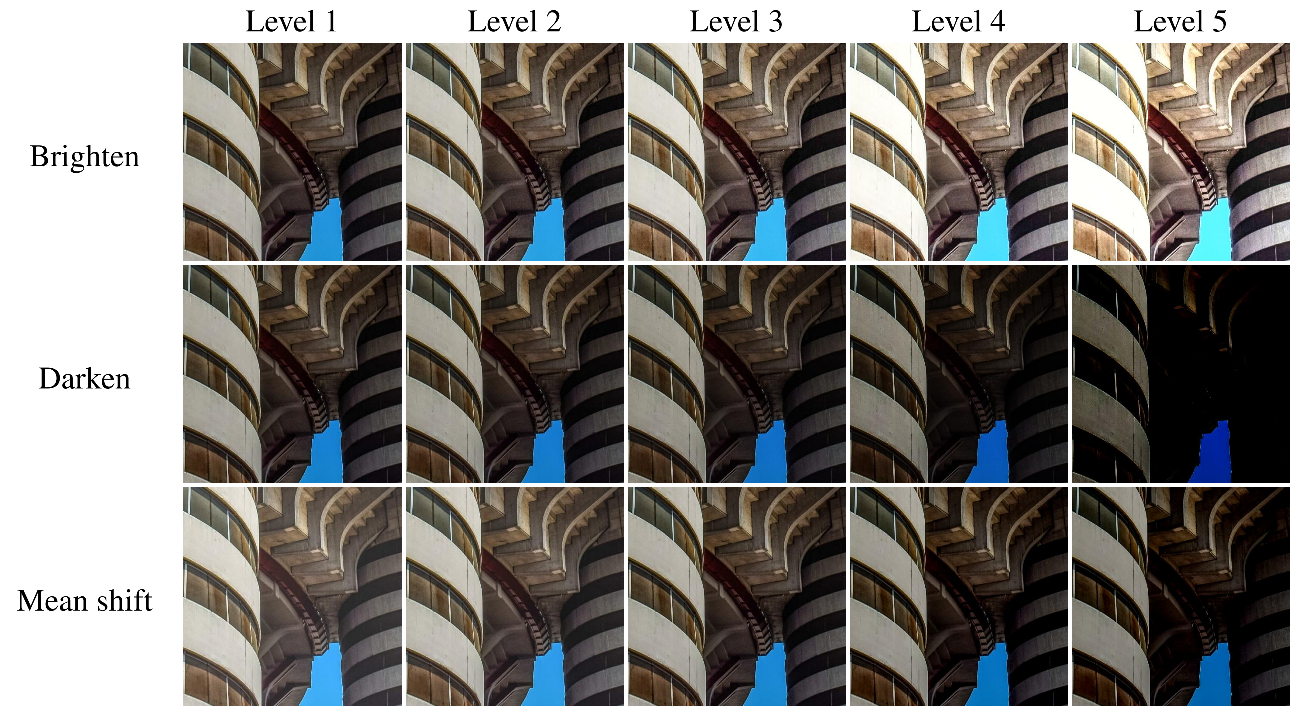


Figure S4. Visualization of the degradation types belonging to the *Brightness change* group for increasing levels of intensity.

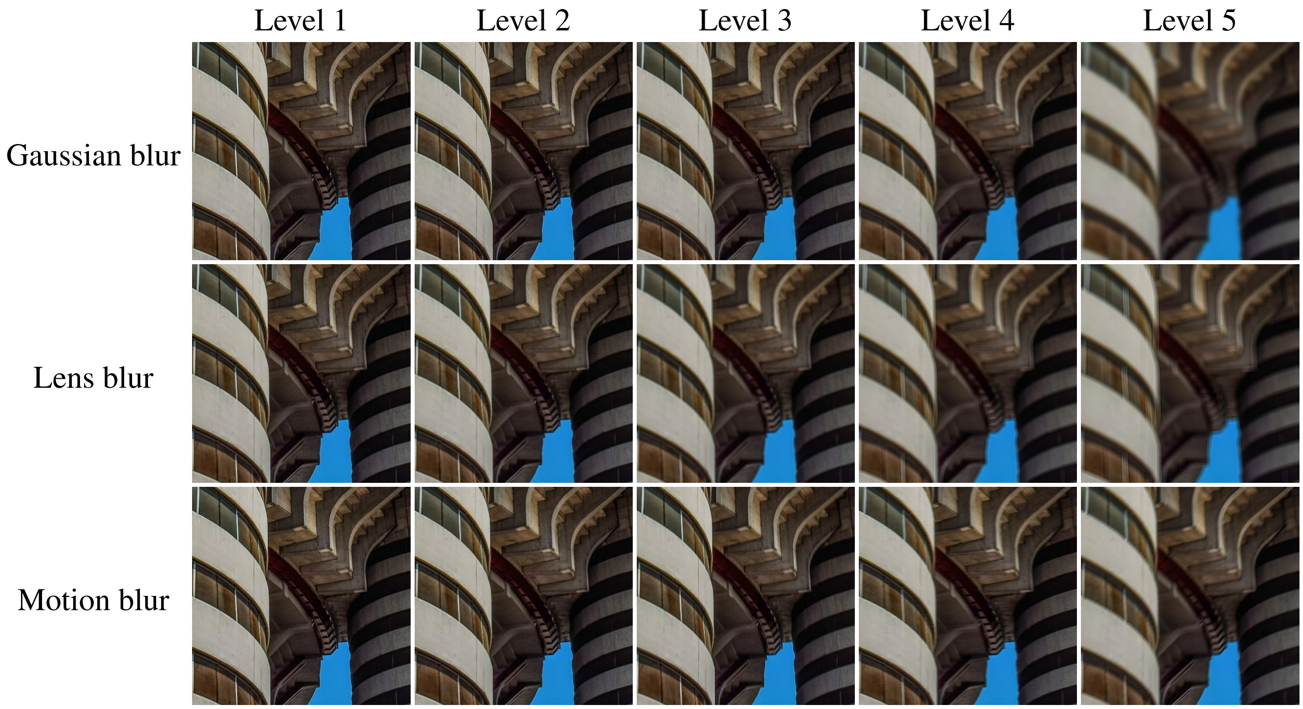


Figure S5. Visualization of the degradation types belonging to the *Blur* group for increasing levels of intensity.

References

[1] 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. 1

[2] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli.

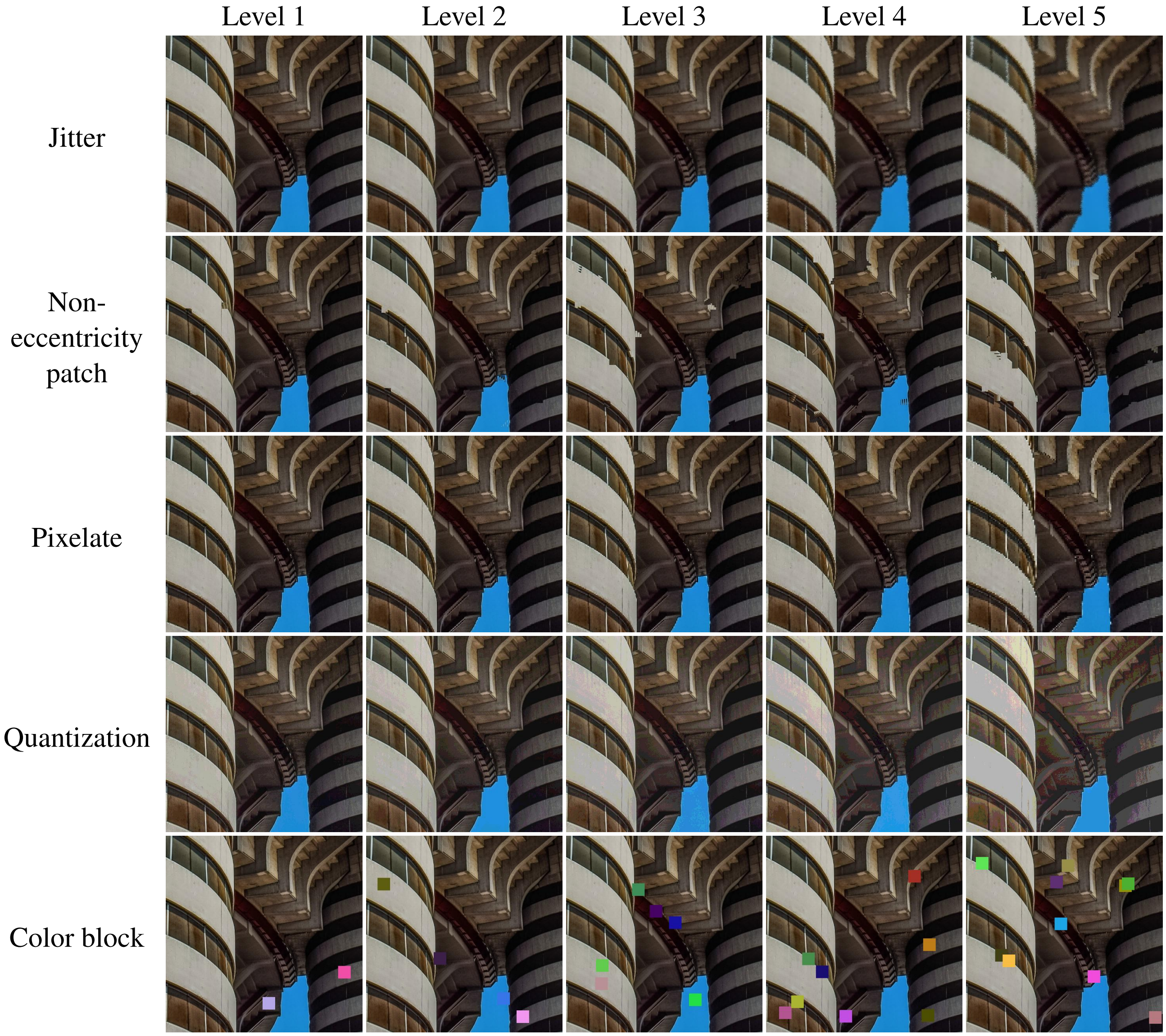


Figure S6. Visualization of the degradation types belonging to the *Spatial distortions* group for increasing levels of intensity.

Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 2

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1

[4] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 2

[5] 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. 2

[6] Ping-Sung Liao, Tse-Sheng Chen, Pau-Choo Chung, et al. A fast algorithm for multilevel thresholding. *J. Inf. Sci. Eng.*, 17(5):713–727, 2001. 4

[7] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 1, 2, 3, 4

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[9] Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group mad competition-a new methodology to compare objective im-

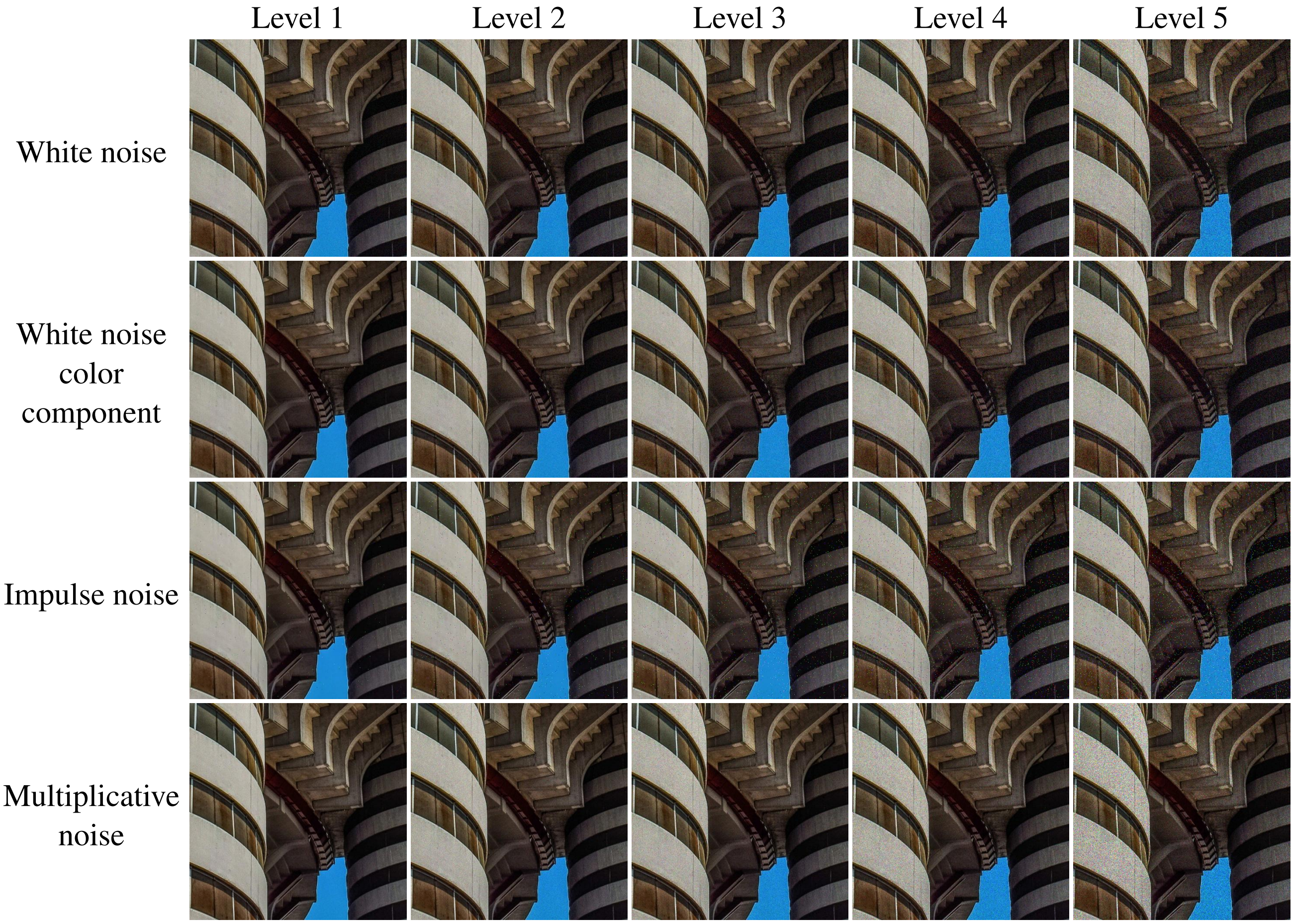


Figure S7. Visualization of the degradation types belonging to the *Noise* group for increasing levels of intensity.

age quality models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2016. 2

[10] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. 1, 2, 3

[11] Eftichia Mavridaki and Vasileios Mezaris. No-reference blur assessment in natural images using fourier transform and spatial pyramids. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 566–570. IEEE, 2014. 1

[12] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 3

[13] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012. 1

[14] 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. 2

[15] Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5846–5855, 2023. 1, 2, 3

[16] 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. 2

[17] 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. 2

[18] 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. 2

[19] 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. 2

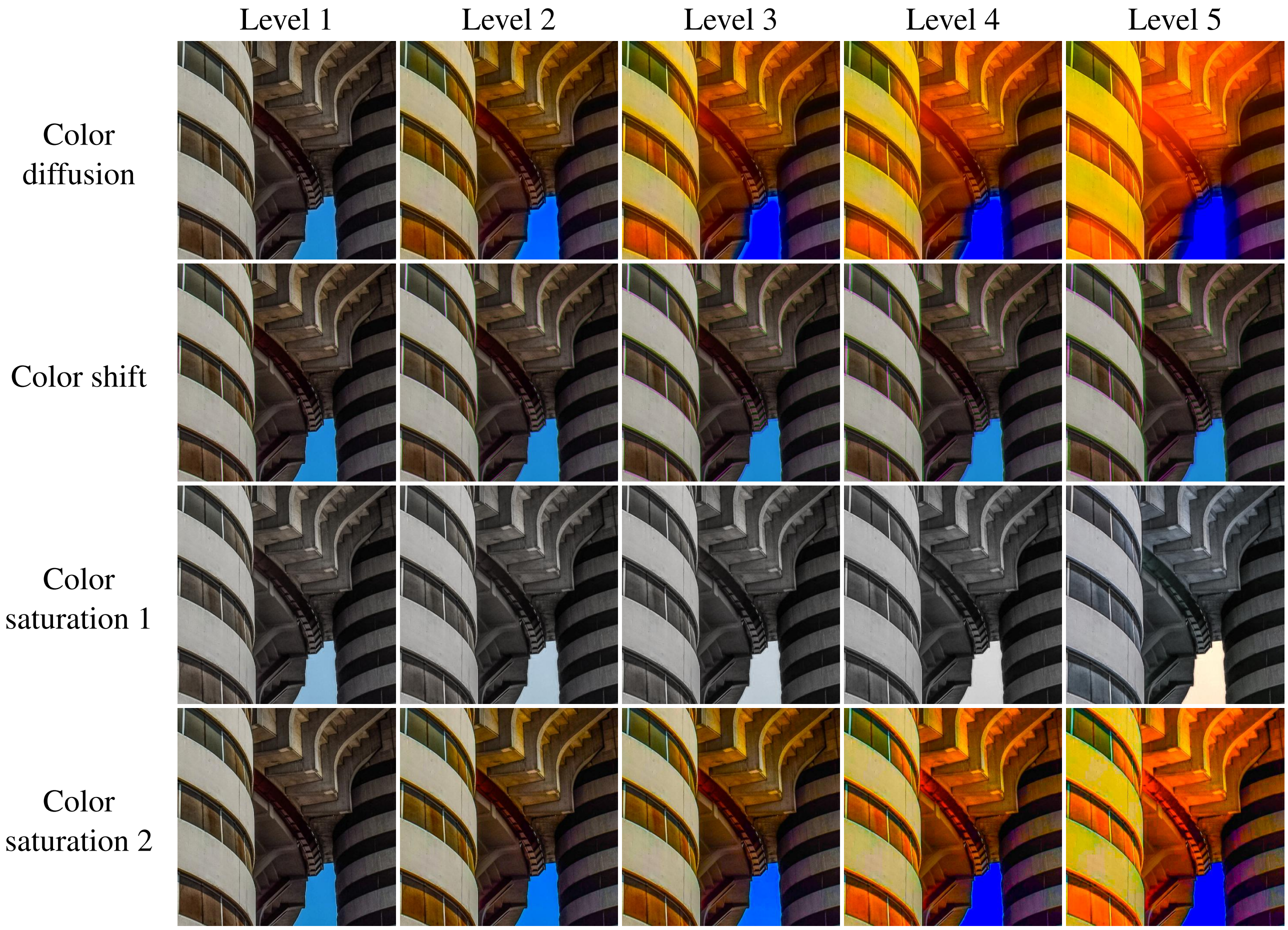


Figure S8. Visualization of the degradation types belonging to the *Color distortions* group for increasing levels of intensity.

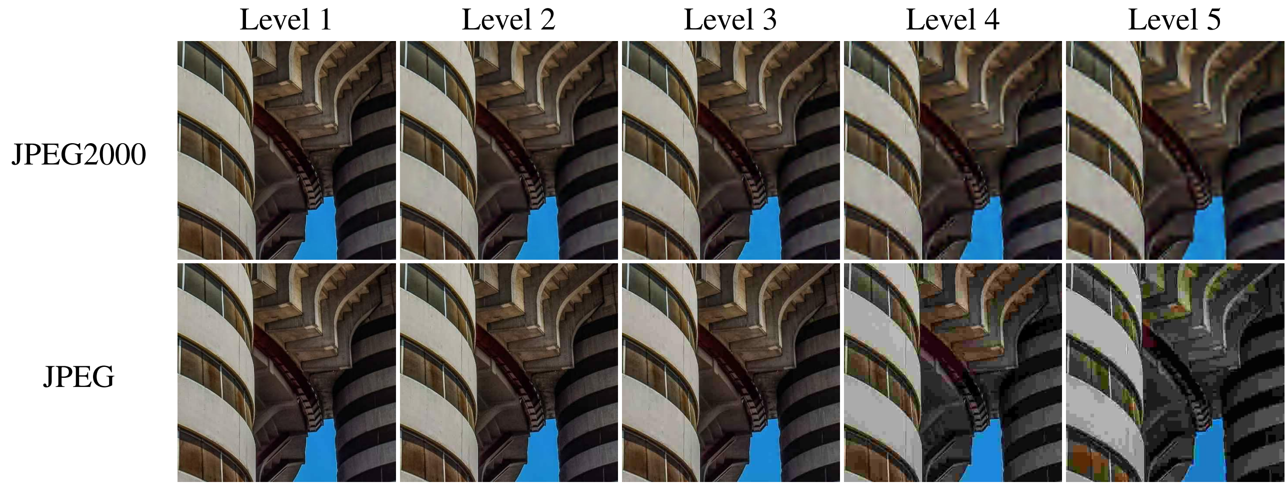


Figure S9. Visualization of the degradation types belonging to the *Compression* group for increasing levels of intensity.

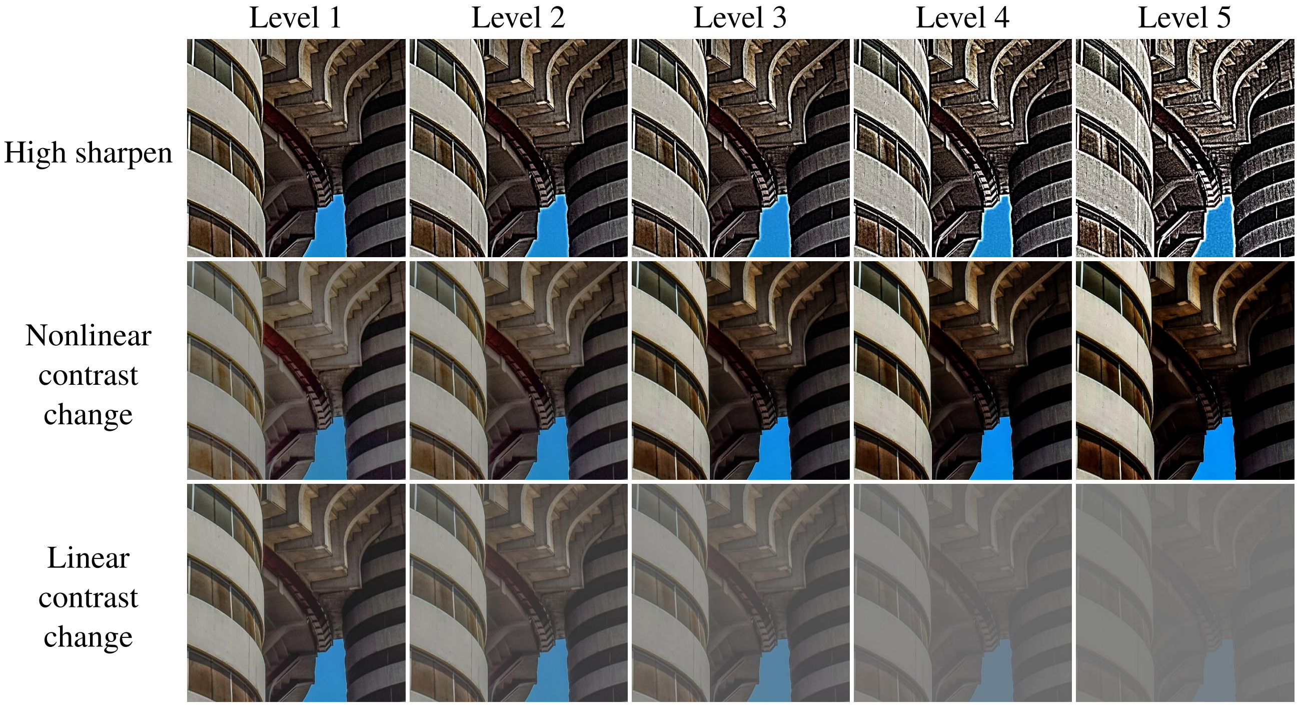


Figure S10. Visualization of the degradation types belonging to the *Sharpness & contrast* group for increasing levels of intensity.