Assessing a Single Image in Reference-Guided Image Synthesis

Jiayi Guo[1], Chaoqun Du[1], Jiangshan Wang[2], Huijuan Huang[3], Pengfei Wan[3], Gao Huang[1,4*]

[1]Department of Automation, BNRist, Tsinghua University, Beijing, China
[2]Beijing University of Posts and Telecommunications, Beijing, China
[3]Y-tech, Kuaishou Technology
[4]Beijing Academy of Artificial Intelligence, Beijing, China
{guo-jy20, dcq20}@mails.tsinghua.edu.cn, hill@bupt.edu.cn, {huanghuijuan, wanpengfei}@kuaishou.com,
gaohuang@tsinghua.edu.cn

Abstract

Assessing the performance of Generative Adversarial Networks (GANs) has been an important topic due to its practical significance. Although several evaluation metrics have been proposed, they generally assess the quality of the *whole* generated image distribution. For Reference-guided Image Synthesis (RIS) tasks, i.e., rendering a source image in the style of another reference image, where assessing the quality of a *single* generated image is crucial, these metrics are not applicable. In this paper, we propose a general learning-based framework, Reference-guided Image Synthesis Assessment (RISA) to quantitatively evaluate the quality of a single generated image. Notably, the training of RISA does not require human annotations. In specific, the training data for RISA are acquired by the intermediate models from the training procedure in RIS, and weakly annotated by the number of models' iterations, based on the positive correlation between image quality and iterations. As this annotation is too coarse as a supervision signal, we introduce two techniques: 1) a pixel-wise interpolation scheme to refine the coarse labels, and 2) multiple binary classifiers to replace a naïve regressor. In addition, an *unsupervised* contrastive loss is introduced to effectively capture the style similarity between a generated image and its reference image. Empirical results on various datasets demonstrate that RISA is highly consistent with human preference and transfers well across models.

Introduction

Reference-guided Image Synthesis (RIS) aims to utilize Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) to modify the style of a source image to that of a reference image. As described in recent image translation works (Lee et al. 2018; Huang et al. 2018; Choi et al. 2020), style refers to the unique appearance of a single image, while the underlying spatial structure is defined as content. Style also coincides with the definition of texture in some other works (Park et al. 2020). Nowadays, generative models are widely deployed to provide various RIS services, such as modifying a user's facial features to that of a super star, or changing a building's original appearance to another.

To enhance user experience in RIS applications, it is of great practical signficance to quantitatively evaluate the images generated by GANs. Although several sample-based GAN evaluation metrics have been proposed (Xu et al. 2018), e.g., Kernel MMD (Gretton et al. 2012), Inception Score (IS) (Salimans et al. 2016), Mode Score (MS) (Che et al. 2016), Wasserstein distance and Fréchet Inception Distance (FID) (Heusel et al. 2017), they mainly focus on assessing the *whole* generated image distribution. In specific, the discrepancy between feature distributions of real and generated images is computed as a quality measure.

However, these metrics are not applicable to evaluate a *single* generated image. In interactive RIS applications, a user may submit one source image (or a pair of source image and reference image) at a time and expects to obtain a satisfying generated image. Unfortunately, due to the notoriously unstable training procedure of GANs, it is challenging to guarantee that each generated image is synthesized with high quality, especially when the source image and the reference image have a large style discrepancy.

Hence, it is important to design an assessment metric for a single image in RIS to improve user experience. Once the quality of each generated image could be effectively assessed, we could simultaneously deploy several different models to generate images for a task and automatically render the image of the highest quality score to users. If all these images are synthesized with low quality, we could refuse to provide any image. However, recent works on single image assessment are either designed to report the average quality score with dozens of images (Shaham, Dekel, and Michaeli 2019) or not able to capture the style similarity between a generated image and its reference (Bosse et al. 2016; Talebi and Milanfar 2018; Zhang et al. 2018a; Gu et al. 2020).

In this paper, we propose a general learning-based framework, Reference-guided Image Synthesis Assessment (RISA), through which the quality of a single generated image can be effectively assessed. As illustrated in Figure 1, given a generated image and its reference image, RISA first extracts their style codes via the style encoder. Then the difference of style codes is calculated as the input of multiple binary classifiers. Finally, the quality score is obtained by averaging the predictions of all classifiers.

RISA works in a weakly supervised scheme, i.e., it does not require any human annotations. As illustrated in Figure

[*] Corresponding author.

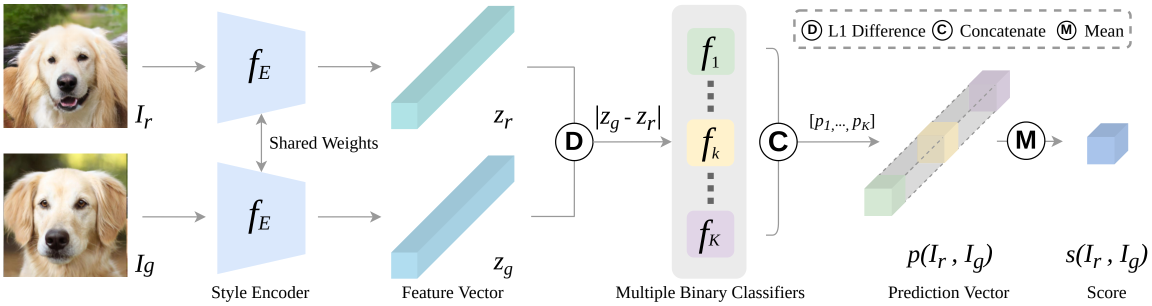


Figure 1: The pipeline of the proposed Reference-guided Image Synthesis Assessment (RISA). RISA consists of a style encoder and multiple binary classifiers. Given a generated image I_{g} and its reference image I_{r}, RISA first utilizes the style encode to extract their style codes z_{g} and z_{r}, respectively. Then the discrepancy (L1 difference) between z_{g} and z_{r} is calculated and fed into multiple binary classifiers. Finally, quality score is given by averaging the predictions of all classifiers.

2, the generated image quality generally increases with training iterations during the GAN training procedure. Therefore, we leverage images generated by intermediate models as the training data for RISA, and consider the number of model's training iterations as a pseudo quality label. Note that a naïve implementation of this idea leads to degenerated solutions, due to that the supervision signal is too coarse, as shown by the results in Table 3. To address this issue, we adopt a pixel-wise interpolation technique to generate images with different quality levels instead of directly utilizing images synthesized by intermediate models during the stable stage. To further suppress the label noise, we deploy multiple binary classifiers rather than a naïve regressor.

Moreover, RISA is optimized by a novel objective, containing 1) a weakly supervised loss to fit the quality label using binary cross entropy, 2) a contrastive loss to effectively capture the style similarity between a generated image and its reference image and 3) a supremum loss to learn the style consistency between two style-preserving augmentation views of a same real image.

We evaluate the effectiveness of RISA on various datasets. Compared with existing single image quality assessment metrics, empirical results demonstrate that our method achieves higher consistency with human preference, and transfers well across different models.

Related Work

Reference-guided image synthesis. In the context of neural style transfer, reference-guided image synthesis aims to render a source image in the style of a reference image. Gatys, Ecker, and Bethge (2015, 2016) utilize feature statistics of deep neural network to capture the style of an image for the first time. Huang and Belongie (2017) propose AdaIN to implement arbitrary style transfer. More recently, StarGAN (Choi et al. 2018) and StarGAN v2 (Choi et al. 2020) learn mapping between multiple domains with a single generator. MSGAN (Mao et al. 2019) proposes mode seeking regularization to resolve the mode collapse problem. DRIT (Lee et al. 2018), MUNIT (Huang et al. 2018) and Swapping Autoencoder (Park et al. 2020) focus on the style and content disentanglement in the feature space.

Image Quality Assessment (IQA). According to the availability of reference, IQA methods are generally divided into three categories: 1) *Full-Reference IQA* (FR-IQA) refers to estimating the quality of natural images with references. Widely-used FR-IQA metrics include MS-SSIM (Wang, Simoncelli, and Bovik 2003), SSIM (Wang et al. 2004), PSNR (Huynh-Thu and Ghanbari 2008), FSIM (Zhang et al. 2011) and LPIPS (Zhang et al. 2018a). 2) *Reduced-Reference IQA* (RR-IQA) tackles situations where the reference image is not fully accessible. Representative methods are local-harmonic based algorithm (Gunawan and Ghanbari 2003) and grouplet-based algorithm (Maalouf, Larabi, and Fernandez-Maloigne 2009). 3) *No-Reference IQA* (NR-IQA) assesses distorted image without any reference. Early works include support vector regression based methods (Moorthy and Bovik 2010, 2011) and probability based methods (Mittal, Soundararajan, and Bovik 2012). With the prevalence of deep learning, massive of network architectures (Bosse et al. 2016; Liu, Van De Weijer, and Bagdanov 2017; Talebi and Milanfar 2018; Lin and Wang 2018; Ren, Chen, and Wang 2018; Pan et al. 2018; Lim, Kim, and Ra 2018; Zhang et al. 2018b, 2021) are proposed. Unfortunately, most of the existing IQA methods aim to assess the quality of natural images, while they are limited when dealing with generated images.

Generative adversarial network assessment. Several sample based methods have been proposed to assess GAN performance (Xu et al. 2018). Among them, Fréchet Inception Distance (FID) (Heusel et al. 2017) is the most popular metric. There are also other proposed metrics like Kernel MMD (Gretton et al. 2012), Inception Score (IS) (Salimans et al. 2016), Mode Score (MS) (Che et al. 2016), and Wasserstein distance. Note that all these methods measure the deviation between the deep features distribution of generated images and that of real images. Single Image FID (Shaham, Dekel, and Michaeli 2019) aims to compare internal patch statistics difference between generated images and a single reference image, which also focuses on assessing the quality of a generated image distribution. To assess a single generated image, GIQA (Gu et al. 2020) is proposed as a NR-IQA metric from both learning-based and data-based perspectives. However, it can not evaluate whether the generated image inherits the style of its reference image.

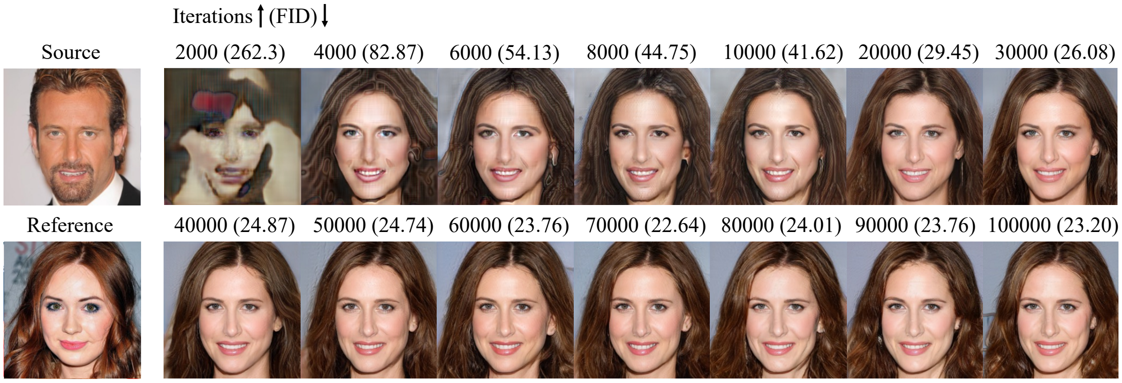


Figure 2: Visualizations and Fréchet Inception Distance (FID) variations of StarGAN v2 on CelebA-HQ dataset with the increase of training iterations. The first column shows the source image and the reference image of a specific image synthesis task, while remaining columns are images generated by intermediate models at different training iterations. The parentheses above each generated image gives FID score at corresponding training iterations.

Methodology

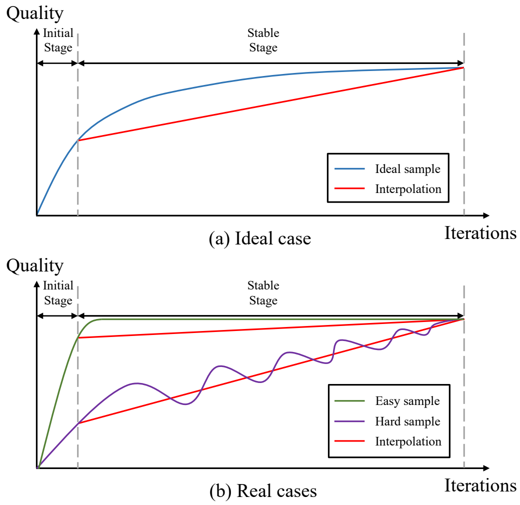
In this section, we first introduce the architecture of our RISA framework. Then we describe how to obtain labeled training data. Finally, our novel objective is presented.

Learning-based Framework

Given a triplet of $\{\{I_{\rm g}, I_{\rm r}\}, y\}$, where $I_{\rm g}$ and $I_{\rm r}$ refer to a generated image and its reference image, respectively, and $y \in [0, 1]$ is the target quality score, RISA aims to assess the quality of $I_{\rm g}$ according to $I_{\rm r}$ as a score $s(I_{\rm r}, I_{\rm g})$ under the supervision of y. As shown in Figure 1, RISA consists of a style encoder and multiple binary classifiers.

Style encoder. Following StarGAN v2 (Choi et al. 2020), we implement a convolutional neural network (CNN) with six pre-activation residual blocks (He et al. 2016) and one fully connected layer as our style encoder $f_{\rm E}$. Given an input image I, a style encoder learns to extract the style-specific attributes and represents them as the style code z. In our pipeline, we first encode $I_{\rm g}$ and $I_{\rm r}$ into $z_{\rm g}$ and $z_{\rm r}$ respectively. Then to guarantee the symmetry of RISA, we calculate the absolute value of element-wise subtraction $|z_{\rm g} - z_{\rm r}|$ as the difference of the style codes $z_{\rm g}$ and $z_{\rm r}$ and use it as the input of multiple binary classifiers.

Multiple binary classifiers. A straightforward solution to fit the target quality score y is to adopt a naïve regressor. Unfortunately, empirical results illustrate that a naïve regressor fails to converge in the setting where training images are coarsely annotated (Table 3). This is partially due to the image quality gap within the same iteration. To address this issue, inspired by previous works (Liu et al. 2016; Gu et al. 2020), we train a learning-based network with K binary classifiers instead of a regressor to learn the generated image quality score. To be specific, the k-th binary classifier is trained for classifying whether the image quality score (from 0 to 1) is greater than a certain threshold T_k, where $T_k = (k-1)/K, k = 1, 2, \cdots, K$. Denote p_k as the predicted probability of k-th binary classifier. The final predicted score of RISA is the mean of the prediction vector $p(I_{\rm r}, I_{\rm g}) = [p_1, p_2, \cdots, p_K]$ as shown in Figure 1. As a supervision signal of $p(I_{\rm r}, I_{\rm g})$, the target quality score y is converted to a *binary* label vector $t(I_{\rm r}, I_{\rm g}) = [t_1, t_2, \cdots, t_K]$ according to whether it is greater than T_k. For example, if we set K to 5, then the target quality score $y = 0.6$ will be converted to $[1, 1, 1, 0, 0]$. In our experiments, we set $K = 16$.



(a) Ideal case

(b) Real cases

Figure 3: An illustration of the image quality variation during the GAN training procedure. (a) shows an ideal case that the vanilla annotation method works well, while (b) indicates two cases that the vanilla annotation method is not suitable. Our proposed pixel-wise interpolation (red lines) can refine the annotations in all these settings.

Data Preparation

Based on the positive correlation between the quality of generated image and the number of training iterations, we use images generated by intermediate models from GAN training procedure to train RISA and consider the number of training iterations as a weak annotation. To refine this annotation, we propose a pixel-wise interpolation technique.

Coarsely labeled synthesized images. Figure 2 illustrates that the quality of generated images generally evolves with training iterations, in terms of both the visualization effect and FID. Based on this, a vanilla method to obtain training images for RISA is utilizing intermediate models in GAN training to synthesize images. Then each generated image $I_{\mathrm{g}}^{\mathrm{vanilla}}$ is annotated by the number of its corresponding model iterations. To meet the scale of RISA's output, we normalize the annotations into the range $[0,(K-1)/K]$ as y^{vanilla}. Here we suppose even synthesized by a converged model, the generated image is still not as perfect as a real image. In our experiments, we only annotate the different views of a same real image with the highest quality score 1, where the views are produced via style-preserving augmentation, such as scaling, cropping and clipping.

This vanilla method is reasonable for ideal cases as Figure 3(a), where the quality of the generated images monotonically increases with iterations. However, in real cases, it is unsuitable. As illustrated in Figure 3(b), the whole GAN training process could be separated into two successive stages, namely the initial stage and the stable stage. Empirically, we recommend the elbow point of FID curve as an appropriate stage boundary. During the initial stage, the quality of generated images all improves rapidly as ideal cases. However, during the stable stage, the quality of easy samples becomes stable and invariant after a few iterations, and the quality of hard samples presents an oscillatory convergence. As a result, the number of model iterations can not represent the image quality during the stable stage.

Pixel-wise interpolation. To tackle this problem, we introduce a pixel-wise interpolation technique (red lines in Figure 3) as an estimation approach to capture the quality changes during the stable stage. Given a pair of source image and reference image, we observe that the image generated by an intermediate model with iterations around the stage boundary have lower quality than the image synthesized by a finally converged model at the end of the whole GAN training procedure. For simplicity, $\{\{I_{\mathrm{g}}^{\mathrm{low}}, I_{\mathrm{r}}\}, y^{\mathrm{low}}\}$ refers to the former, and $\{\{I_{\mathrm{g}}^{\mathrm{high}}, I_{\mathrm{r}}\}, y^{\mathrm{high}}\}$ denotes the latter. To produce images with quality between $I_{\mathrm{g}}^{\mathrm{low}}$ and $I_{\mathrm{g}}^{\mathrm{high}}$, we implement linear interpolation in the pixel space:

$$I_{\mathrm{g}}^{\mathrm{inter}} = \epsilon I_{\mathrm{g}}^{\mathrm{high}} + (1-\epsilon) I_{\mathrm{g}}^{\mathrm{low}}, y^{\mathrm{inter}} = \epsilon y^{\mathrm{high}} + (1-\epsilon) y^{\mathrm{low}}, \tag{1}$$

where $I_{\mathrm{g}}^{\mathrm{inter}}$ and y^{inter} represent the interpolated image and its quality score, respectively. $\epsilon \in (0,1)$ is an interpolation factor. By varying ϵ, we could generate a series of images with different quality between the quality of $I_{\mathrm{g}}^{\mathrm{low}}$ and $I_{\mathrm{g}}^{\mathrm{high}}$.

A natural question to ask is why pixel-wise interpolation is effective to generate images of different quality. As illustrated in Figure 2, during the stable stage, generated images could all preserve the content of their source images perfectly and maintain the style of their reference images generally. The model mainly focuses on improving the generation of detailed textures. Pixel-wise interpolation with different ϵ could estimate these local texture variations while have no influence on the global structure and texture. Empirical results in Table 4 also indicate the performance improvements gained from the pixel-wise interpolation technique.

As a summary, generated images in RISA's training data are synthesized as:

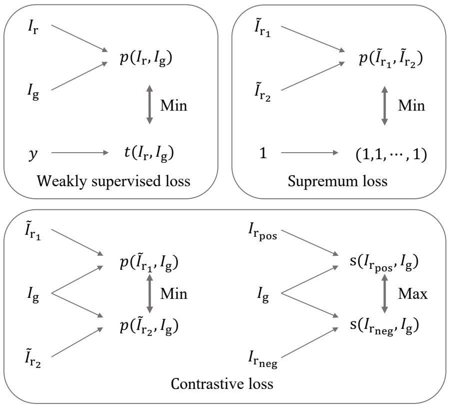

I_{r}
I_{g}
$p(I_{\mathrm{r}}, I_{\mathrm{g}})$
Min
y
$t(I_{\mathrm{r}}, I_{\mathrm{g}})$
Weakly supervised loss
$\tilde{I}_{\mathrm{r}_1}$
$\tilde{I}_{\mathrm{r}_2}$
$p(\tilde{I}_{\mathrm{r}_1}, \tilde{I}_{\mathrm{r}_2})$
Min
1
$(1,1,\cdots,1)$
Supremum loss
$\tilde{I}_{\mathrm{r}_1}$
I_{g}
$\tilde{I}_{\mathrm{r}_2}$
$p(\tilde{I}_{\mathrm{r}_1}, I_{\mathrm{g}})$
Min
$p(\tilde{I}_{\mathrm{r}_2}, I_{\mathrm{g}})$
$I_{\mathrm{r_{pos}}}$
I_{g}
$I_{\mathrm{r_{neg}}}$
$\mathrm{s}(I_{\mathrm{r_{pos}}}, I_{\mathrm{g}})$
Max
$\mathrm{s}(I_{\mathrm{r_{neg}}}, I_{\mathrm{g}})$
Contrastive loss

Figure 4: RISA's training objective. It contains a weakly supervised loss, a contrastive loss and a supremum loss.

$$I_{\mathrm{g}} = \begin{cases} I_{\mathrm{g}}^{\mathrm{vanilla}}, & \text{during the initial stage;} \\ I_{\mathrm{g}}^{\mathrm{inter}}, & \text{during the stable stage.} \end{cases} \tag{2}$$

Training Objective

As illustrated in Figure 4, our training objective includes three terms: 1) a weakly supervised loss to learn the *pseudo* quality score, 2) an unsupervised contrastive loss to effectively capture the style similarity between a generated image and its reference image, and 3) a supremum loss to learn the style consistency of two augmented views from a real image.

Weakly supervised loss. We implement the weakly supervised loss term utilizing the binary cross entropy:

$$\begin{aligned} &\mathcal{L}_{\sup}(p(I_{\mathrm{r}}, I_{\mathrm{g}}), t(I_{\mathrm{r}}, I_{\mathrm{g}})) \\ &= \sum\nolimits_{k=1}^{K} (t_k \log p_k + (1-t_k)\log(1-p_k)), \end{aligned} \tag{3}$$

where $p(I_{\mathrm{r}}, I_{\mathrm{g}})$ and $t(I_{\mathrm{r}}, I_{\mathrm{g}})$ refer to the prediction vector and the binary label vector, respectively, and their index is ignored for simplicity (the same below).

Contrastive loss. We employ the contrastive loss to capture the style similarity between a generated image and its reference image, which is essential to the generalization ability of our framework. In specific, we produce two views $\widetilde{I}_{\mathrm{r}_1}, \widetilde{I}_{\mathrm{r}_2}$ of reference image I_{r} via data augmentation. To preserve *style-relevant* information, the augmentation operations only consist of scaling, cropping and clipping. Then they are fed to our framework for producing the prediction vectors $p(\widetilde{I}_{\mathrm{r}_1}, I_{\mathrm{g}}), p(\widetilde{I}_{\mathrm{r}_2}, I_{\mathrm{g}})$ and the prediction scores $s(\widetilde{I}_{\mathrm{r}_1}, I_{\mathrm{g}}), s(\widetilde{I}_{\mathrm{r}_2}, I_{\mathrm{g}})$. We treat $p(\widetilde{I}_{\mathrm{r}_1}, I_{\mathrm{g}})$ and $p(\widetilde{I}_{\mathrm{r}_2}, I_{\mathrm{g}})$ as a positive pair and minimize the distance of them. Thus the positive part of constrastive loss is expressed as:

$$\mathcal{L}_{\mathrm{pos}}(p(\widetilde{I}_{\mathrm{r}_1}, I_{\mathrm{g}}), p(\widetilde{I}_{\mathrm{r}_2}, I_{\mathrm{g}})) = \|p(\widetilde{I}_{\mathrm{r}_1}, I_{\mathrm{g}}) - p(\widetilde{I}_{\mathrm{r}_2}, I_{\mathrm{g}})\|_2^2, \tag{4}$$

where $\|\cdot\|_2^2$ denotes the squared Euclidean norm.

In addition, we consider $s(I_{\mathrm{r_{neg}}}, I_{\mathrm{g}})$ and $s(I_{\mathrm{r_{pos}}}, I_{\mathrm{g}})$ as a negative sample and a positive sample, respectively, where $I_{\mathrm{r_{pos}}}$ could be $\widetilde{I}_{\mathrm{r}_1}$ or $\widetilde{I}_{\mathrm{r}_2}$ and $I_{\mathrm{r_{neg}}}$ refers to a randomly selected reference image. To enlarge the difference between

	CelebA-HQ	AFHQ	Yosemite	Church	Bedroom
	StarGAN v2		MSGAN	Swap Autoencoder	
NIQE	60.09±0.99%	60.12±0.97%	52.18±1.39%	54.48±2.39%	54.37±6.01%
Deep-IQA	50.22±4.66%	60.57±3.55%	53.27±0.58%	54.17±5.69%	50.32±4.34%
NIMA	52.39±0.65%	58.17±3.68%	47.82±3.33%	54.18±2.82%	48.00±2.27%
GMM-GIQA	52.25±1.79%	68.82±2.52%	57.09±1.14%	70.95±2.35%	64.42±4.27%
KNN-GIQA	50.94±7.25%	70.16±1.84%	49.82±1.48%	50.90±2.57%	49.67±1.18%
SSIM	55.01±1.60%	57.57±3.25%	45.45±3.61%	64.67±5.24%	77.51±4.11%
MS-SSIM	52.98±1.01%	61.77±5.52%	44.37±1.09%	60.78±1.62%	67.13±5.39%
PSNR	51.81±2.52%	55.47±2.36%	44.73±3.58%	65.56±2.37%	77.53±3.82%
FSIM	56.75±4.58%	61.17±5.52%	51.09±0.59%	58.68±2.01%	56.38±2.28%
LPIPS	54.86±1.18%	75.41±2.02%	57.45±0.99%	66.77±1.48%	71.15±1.57%
SIFID	57.76±1.55%	69.27±1.55%	58.55±3.42%	69.18±3.92%	72.81±1.75%
RISA(ours)	**70.54±1.84%**	**80.36±1.44%**	**64.36±1.79%**	**73.96±2.83%**	**83.55±3.37%**

Table 1: Consistency (in %) with human judgments (50% indicates a random guess). The training data for RISA and the samples for human judgments are generated by the models with the same architecture.

the negative sample and the positive sample, the negative part of the contrastive loss is defined as:

$$\begin{aligned}&\mathcal{L}_{\mathrm{neg}}(s(I_{\mathrm{r_{neg}}}, I_{\mathrm{g}}), s(I_{\mathrm{r_{pos}}}, I_{\mathrm{g}}))\\ =&\max(0, s(I_{\mathrm{r_{neg}}}, I_{\mathrm{g}}) - \gamma s(I_{\mathrm{r_{pos}}}, I_{\mathrm{g}})),\end{aligned} \tag{5}$$

where γ is a penalty factor to control the restraint strength. We set $\gamma = 0.5$ in our experiments.

Supremum loss is introduced as a supplement of the weakly supervised loss in which the pseudo quality score is strictly less than 1. Since different views of a reference image in contrastive loss preserve the style-relevant information, the quality score of $\widetilde{I}_{\mathrm{r_1}}$ and $\widetilde{I}_{\mathrm{r_2}}$ should be the highest score 1. Therefore, as an additional supervision signal, a supremum loss is represented as:

$$\mathcal{L}_{\mathrm{supre}}(p(\widetilde{I}_{\mathrm{r_1}}, \widetilde{I}_{\mathrm{r_2}}), \mathbf{1}) = \sum_{k=1}^{K} \log p_k, \tag{6}$$

where $\mathbf{1}$ denotes a vector of ones with length K.

Full objective. To sum up, our full objective is given as:

$$\begin{aligned}\mathcal{L} =&\mathcal{L}_{\mathrm{sup}}(p(I_{\mathrm{r}}, I_{\mathrm{g}}), t(I_{\mathrm{r}}, I_{\mathrm{g}}))\\ &+ \lambda_{\mathrm{p}}\mathcal{L}_{\mathrm{pos}}(p(\widetilde{I}_{\mathrm{r_1}}, I_{\mathrm{g}}), p(\widetilde{I}_{\mathrm{r_2}}, I_{\mathrm{g}}))\\ &+ \lambda_{\mathrm{n}}\mathcal{L}_{\mathrm{neg}}(s(I_{\mathrm{r_{neg}}}, I_{\mathrm{g}}), s(I_{\mathrm{r_{pos}}}, I_{\mathrm{g}}))\\ &+ \lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{supre}}(p(\widetilde{I}_{\mathrm{r_1}}, \widetilde{I}_{\mathrm{r_2}}), \mathbf{1}).\end{aligned} \tag{7}$$

In our experiments, we set λ_{p}, λ_{n} and λ_{s} to 1 for simplicity.

Experiments

In this section, we start by introducing the experimental setup. Then extensive results are reported to demonstrate the effectiveness and generalization ability of RISA. In addition, we carefully analyze each components of RISA and compare different training configurations as ablation studies.

Experimental Setup

Datasets and genarative models. We conduct our experiments on five datasets: Yosemite (Zhu et al. 2017), CelebA-HQ (Karras et al. 2017), AFHQ (Choi et al. 2020), LSUN Church and Bedroom (Yu et al. 2015), all at the resolution of 256×256. For multi-domain GAN training, we separate Yosemite into two domains of summer and winter, CelebA-HQ into two domains of male and female and AFHQ into three domains of cat, dog, and wildlife.

Following the original papers, we train DRIT (Lee et al. 2018) and MSGAN (Huang et al. 2018) on Yosemite and CelebA-HQ, StarGAN v2 (Choi et al. 2020) on CelebA-HQ and AFHQ, and Swap Autoencoder (Swap AE) (Park et al. 2020) on CelebA-HQ, LSUN Church and Bedroom.

Implementation details. For each generative model, we first choose 7 intermediate models (DRIT and MSGAN trained for 1, 10, 20, 40, 80, 200 and 1200 epochs, StarGAN v2 trained for 1k, 2k, 4k, 6k, 8k, 10k and 100k iterations and Swapping Autoencoder trained for 100k, 200k, 500k, 1M, 2M, 5M and 25M images), each of which is utilized to synthesize 1k images. Then we execute the pixel-wise interpolation using the last two models mentioned above, e.g., models at 10k and 100k iterations for StarGAN v2, with $\epsilon = 0.1, 0.2, \cdots, 0.9$. We synthesize 1k interpolated images under each ϵ. Finally, for each dataset, 16k images with 16 different quality scores ($k/16, k = 1, 2, \cdots, 15$) is obtained as the training images.

RISA is implemented in PyTorch (Paszke et al. 2019). The batch size is set to 4 and the model is trained for 100 epochs using a single NVIDIA RTX 2080Ti GPU. We use the Adam (Kingma and Ba 2014) with $\beta_1 = 0$ and $\beta_2 = 0.99$. The weight decay and the learning rate are set to 10^{-4}. The weights of all modules are initialized using He initialization (He et al. 2015) and all bias are set to zero.

Human evaluation. To compare the effectiveness of different metrics, we test the consistency of each metric with human judgments through various binary classification experiments. In specific, each testing sample is a triplet consisting of a *reference* image and two *generated* images synthesized by different generative models. Human observers are required to independently select the generated image of higher quality according to the reference image. To guarantee the experiments are nontrivial, the generative models are either two intermediate models during the stable stage (Table 1, 2) or directly two converged models with different architectures (Table 2). Samples that all observers reach

	CelebA-HQ				Yosemite	
	DRIT	MSGAN	MSGAN-Swap AE	StarGAN v2-Swap AE	DRIT	DRIT-MSGAN
NIQE	50.59±3.05%	48.77±4.44%	53.46±4.01%	46.97±4.69%	51.18±0.20%	55.13±3.54%
Deep-IQA	50.58±1.43%	46.60±4.37%	41.51±0.89%	51.14±2.90%	45.25±3.15%	53.53±5.51%
NIMA	49.42±3.02%	54.14±1.63%	58.28±1.19%	47.73±3.80%	52.35±5.03%	54.17±9.43%
GMM-GIQA	56.93±2.21%	59.49±5.23%	58.07±4.00%	49.81±1.87%	60.96±3.01%	60.90±5.94%
KNN-GIQA	51.46±4.48%	51.47±0.79%	53.67±2.14%	52.65±2.19%	51.17±2.82%	50.64±4.53%
SSIM	49.15±3.68%	52.93±1.30%	58.49±2.24%	55.49±5.11%	51.49±2.06%	48.40±2.97%
MS-SSIM	48.26±5.23%	56.35±1.71%	62.26±2.24%	59.47±3.60%	50.29±0.21%	45.19±2.08%
PSNR	50.58±2.36%	57.31±6.02%	66.04±4.39%	57.01±2.56%	49.39±4.63%	47.44±1.20%
FSIM	45.37±3.56%	58.29±0.73%	58.70±3.14%	53.22±6.05%	47.32±3.16%	54.49±2.76%
LPIPS	56.65±0.18%	64.15±1.79%	67.30±4.62%	60.04±1.49%	64.49±0.86%	57.37±0.45%
SIFID	55.78±1.94%	62.43±3.49%	62.89±1.36%	57.77±4.51%	64.51±4.39%	55.13±4.73%
RISA(ours)	**65.32±4.26%**	**70.73±1.20%**	**72.75±0.59%**	**68.94±1.63%**	**67.45±2.66%**	**63.46±2.36%**

Table 2: Cross-model consistency (in %) with human judgments. ”Cross-model” indicates that the training data for RISA and the samples for human judgments are generated by the models with differenet architectures. To be specific, for CelebA-HQ and Yosemite, the generative models to synthesis RISA’s training images are StarGAN v2 and MSGAN, respectively. The generative models used for human judgments are either intermediate models at different training iterations during the stable stage, e.g., DRIT, or two converged models with different architectures, e.g., MSGAN-Swap AE.

a consensus on (about 400 samples per setting) are used to evaluate metrics. We report the average consistency and the standard deviations by dividing samples into 3 equal parts.

Baselines. We compare our methods with 11 baselines:

NR-IQA methods: NIQE (Mittal, Soundararajan, and Bovik 2012) calculates the distance between the multivariate Gaussian model of the test image and a natural scene statistic model as a quality measure. Deep-IQA (Bosse et al. 2016) uses a deep network to assess the quality of various patches randomly sampled from the test image. NIMA (Talebi and Milanfar 2018) assesses the quality of images using a CNN trained with massive labeled images. GIQA (Gu et al. 2020) assesses an image from both learning-based and data-based perspectives. The recommended GMM-GIQA and KNN-GIQA are adopted in our experiments.

FR-IQA methods: SSIM (Wang et al. 2004) measures the discrepancy of two images’ luminance, contrast and structure. MS-SSIM (Wang, Simoncelli, and Bovik 2003) calculates the image’s SSIM in multiple scales. PSNR (Huynh-Thu and Ghanbari 2008) considers the ratio between the maximum possible power of a signal and the power of corrupting noise. FSIM (Zhang et al. 2011) is a variation of SSIM, using different weights to represent the importance of different regions in image. LPIPS (Zhang et al. 2018a) trains evaluation networks using 3 methods (named *lin*, *tune* and *scratch*). SIFID (Shaham, Dekel, and Michaeli 2019) applies FID by viewing features of a image as a distribution.

Results

A quick evaluation. We conduct an intuitive evaluation to verify the effectiveness of RISA. In particular, for each dataset, we manually select a series of images with visible quality gaps and utilize RISA to assess them. Empirical results in Figure 5 illustrate that images of higher quality can get a higher score via RISA, and vice versa.

Performance comparisons. Table 1 demonstrates the consistency of metrics with human judgments on various datasets and generative models, and we highlight the best performance in bold. From the results, our proposed RISA is more consistent with human judgments than other image assessment methods. Visualizations in Figure 6 also indicate that our proposed method could assess the quality of a generated image according to its reference image.

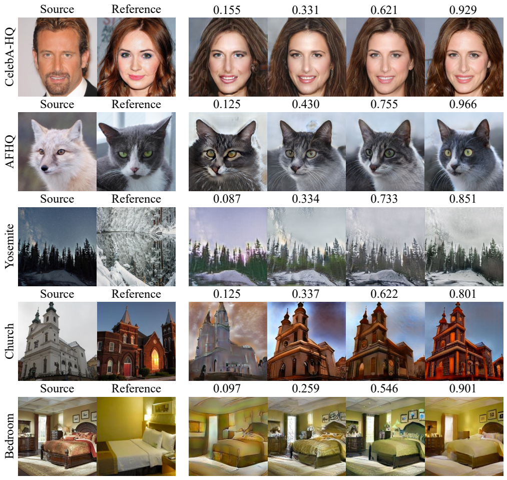


Figure 5: A quick evaluation on images with visible quality gaps. The left two columns are the source and reference images, and the right four columns show images generated by different models. The value above each generated image is the quality score assessed by RISA.

In addition, we report the favorable generalization performance of RISA in Table 2, where the RISA’s training images and the human judgement samples are synthesized by generative models with different architectures. Moreover, the settings in Table 2 can be further divided into two categories, according to the samples are generated by multiple intermediate models (single-model settings) or synthesized by two different converged models (double-model settings). Compared with baselines, results in single-model settings indicate that RISA transfers well across different models. Furthermore, results in double-model settings demonstrate that RISA effectively chooses the image with higher style similarity to its reference image.

	CelebA-HQ	
	StarGAN v2	StarGAN v2-Swap AE
Naïve Regressor	N/A	N/A
+ Multi-Classifiers	67.78±2.84%	53.22±2.98%
+ Contrastive Loss	69.52±1.00%	61.36±2.82%
+ Supremum Loss	**70.54±2.47%**	**68.94±1.93%**

Table 3: Consistency (in %) with human judgments corresponding to various components. The generative model utilized to build training set for RISA is StarGAN v2.

For baselines, the assessment of NR-IQA methods only leverage the generated image. They are not able to evaluate the image's style similarity to its reference image. Although FR-IQA methods calculate the discrepancy between the an image and its reference image, empirical results indicate that they are less suitable to measure the style similarity.

Ablation Study

Effectiveness of various components. We examine each individual component in our framework in Table 3, where each component is cumulatively added to a naïve regressor. We find that only with a naïve regressor, RISA fails to converge and gives the same prediction for all samples because the quality labels of training images are too coarse. In the "StarGAN v2" setting, it is intriguing that simply applying multiple binary classifiers instead of a naïve regressor makes RISA achieve a competitive performance compared with the standard setting. It demonstrates that the multiple binary classifiers effectively suppress the label noise. In the "StarGAN v2-Swap AE" setting, since the two generated images are synthesized by two powerful converged models, the style similarity is a more important aspect to assess. We can find that the contrastive loss is critical to capture the style similarity. In addition, using the supremum loss also lead RISA to achieve higher consistency with human preference.

Effectiveness of pixel-wise interpolation. Table 4 compares the performance of different ways to build the training set for RISA. As the samples used for human judgments are generated by intermediate models during the stable stage, it is natural to compare the setting that the training images are totally generated using models during the stable stage (Stable Stage Only) with the standard setting. To verify the effectiveness of pixel-wise interpolation, we also conduct an experiment which is trained on images generated by models from both the initial stage and the stable stage (+ Initial Stage). Empirical results show that combining images of low quality and of high quality promotes RISA to assess more accurately and refining the coarse labels with pixel-wise interpolation further improves RISA's performance.

Discussion

Assessing a single reference-guided synthesized image is an area of great practical significance but lacks of the research. Although our novel RISA framework performs well on various datasets and settings, there are still opening problems need to be further explored. RISA regards the generated image's style similarity to its reference image as a more significant component to assess, while ignores the content similarity to its source image. Although trained with few iterations, a generative model can synthesize the image which maintains the content (or spatial structure) of its source image perfectly as shown in Figure 2. In contrast, the style-relevant textures are continuously evolving through the whole training procedure. In addition, prior works (Gatys, Ecker, and Bethge 2015, 2016; Huang and Belongie 2017; Huang et al. 2018) also support our opinion since they mainly focus on designing effective objectives to render the style.

	CelebA-HQ	
	StarGAN v2	StarGAN v2-Swap AE
Stable Stage Only	61.10±4.45%	60.42±2.19%
+ Initial Stage	66.04±3.45%	63.07±1.23%
+ Interpolation	**70.54±2.47%**	**68.94±1.93%**

Table 4: Consistency (in %) with human judgments corresponding to different ways of building the training set. StarGAN v2 is utilized to build the training set for RISA.

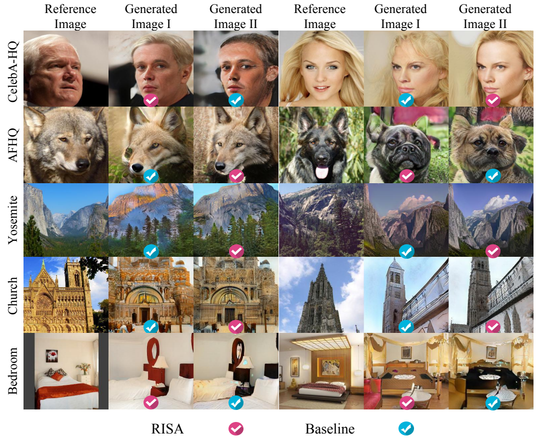


Figure 6: Compared with the most competitive baseline on each dataset (NIQE for CelebA-HQ, LPIPS for AFHQ, SIFID for Yosemite, GMM-GIQA for Church, and PSNR for Bedroom), RISA achieves better performance on selecting the generated image with higher quality.

Conclusion

In this paper, we propose RISA, a learning-based framework to assess a single reference-guided synthesized image. Notably, RISA works in a weakly supervised scheme without any human annotations. In specific, the training images are generated by intermediate models in RIS training and the corresponding labels are annotated by the number of models' iterations. To suppress the label noise, we propose a pixel-wise interpolation technique and adopt multiple binary classifiers. Moreover, an unsupervised contrastive loss is introduced to effectively capture the style similarity. Compared with existing single image assessment metrics, RISA achieves higher consistency with human preference on various datasets and transfers well across models. We believe that our work will contribute to improving user experience in real-world RIS applications and motivate future researches on developing more effective assessment metrics.

Acknowledgements

This work is supported in part by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grant 2018AAA0100701, the National Natural Science Foundation of China under Grants 61906106 and 62022048, and Beijing Academy of Artificial Intelligence (BAAI).

References

Bosse, S.; Maniry, D.; Wiegand, T.; and Samek, W. 2016. A deep neural network for image quality assessment. In *ICIP*.

Che, T.; Li, Y.; Jacob, A. P.; Bengio, Y.; and Li, W. 2016. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*.

Gatys, L.; Ecker, A. S.; and Bethge, M. 2015. Texture synthesis using convolutional neural networks. In *NeurIPS*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *CVPR*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*.

Gretton, A.; Sejdinovic, D.; Strathmann, H.; Balakrishnan, S.; Pontil, M.; Fukumizu, K.; and Sriperumbudur, B. K. 2012. Optimal kernel choice for large-scale two-sample tests. In *NeurIPS*.

Gu, S.; Bao, J.; Chen, D.; and Wen, F. 2020. Giqa: Generated image quality assessment. In *ECCV*.

Gunawan, I. P.; and Ghanbari, M. 2003. Reduced-reference picture quality estimation by using local harmonic amplitude information. In *London Communications Symposium*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity mappings in deep residual networks. In *ECCV*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *ECCV*.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *ECCV*.

Lim, H.-T.; Kim, H. G.; and Ra, Y. M. 2018. VR IQA NET: Deep virtual reality image quality assessment using adversarial learning. In *ICASSP*.

Lin, K.-Y.; and Wang, G. 2018. Hallucinated-IQA: No-reference image quality assessment via adversarial learning. In *CVPR*.

Liu, T.-J.; Liu, K.-H.; Liu, H.-H.; and Pei, S.-C. 2016. Age estimation via fusion of multiple binary age grouping systems. In *ICIP*.

Liu, X.; Van De Weijer, J.; and Bagdanov, A. D. 2017. Rankiqa: Learning from rankings for no-reference image quality assessment. In *ICCV*.

Maalouf, A.; Larabi, M.-C.; and Fernandez-Maloigne, C. 2009. A grouplet-based reduced reference image quality assessment. In *2009 International Workshop on Quality of Multimedia Experience*.

Mao, Q.; Lee, H.-Y.; Tseng, H.-Y.; Ma, S.; and Yang, M.-H. 2019. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*.

Moorthy, A. K.; and Bovik, A. C. 2010. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*.

Moorthy, A. K.; and Bovik, A. C. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*.

Pan, D.; Shi, P.; Hou, M.; Ying, Z.; Fu, S.; and Zhang, Y. 2018. Blind predicting similar quality map for image quality assessment. In *CVPR*.

Park, T.; Zhu, J.-Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A. A.; and Zhang, R. 2020. Swapping autoencoder for deep image manipulation. *arXiv preprint arXiv:2007.00653*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Ren, H.; Chen, D.; and Wang, Y. 2018. RAN4IQA: Restorative adversarial nets for no-reference image quality assessment. In *AAAI*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *NeurIPS*.

Shaham, T. R.; Dekel, T.; and Michaeli, T. 2019. Singan: Learning a generative model from a single natural image. In *ICCV*.

Talebi, H.; and Milanfar, P. 2018. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*.

Wang, Z.; Simoncelli, P. E.; and Bovik, C. A. 2003. Multiscale structural similarity for image quality assessment. *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference*.

Xu, Q.; Huang, G.; Yuan, Y.; Guo, C.; Sun, Y.; Wu, F.; and Weinberger, K. 2018. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*.

Yu, F.; Zhang, Y.; Song, S.; Seff, A.; and Xiao, J. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv preprint arXiv:1506.03365*.

Zhang, L.; Zhang, L.; Mou, X.; and Zhang, D. 2011. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2018b. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, W.; Ma, K.; Zhai, G.; and Yang, X. 2021. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.