Comparing the Robustness of Modern No-Reference Image- and Video-Quality Metrics to Adversarial Attacks

Anastasia Antsiferova[1,2], **Khaled Abud**[3], **Aleksandr Gushchin**[1,2,3], **Ekaterina Shumitskaya**[3], **Sergey Lavrushkin**[1,2], **Dmitriy Vatolin**[1,2,3]

[1]MSU Institute for Artificial Intelligence
[2]ISP RAS Research Center for Trusted Artificial Intelligence
[3]Lomonosov Moscow State University
{aantsiferova, khaled.abud, alexander.gushchin, ekaterina.shumitskaya, sergey.lavrushkin, dmitriy}@graphics.cs.msu.ru

Abstract

Nowadays, neural-network-based image- and video-quality metrics perform better than traditional methods. However, they also became more vulnerable to adversarial attacks that increase metrics' scores without improving visual quality. The existing benchmarks of quality metrics compare their performance in terms of correlation with subjective quality and calculation time. Nonetheless, the adversarial robustness of image-quality metrics is also an area worth researching. This paper analyses modern metrics' robustness to different adversarial attacks. We adapted adversarial attacks from computer vision tasks and compared attacks' efficiency against 15 no-reference image- and video-quality metrics. Some metrics showed high resistance to adversarial attacks, which makes their usage in benchmarks safer than vulnerable metrics. The benchmark accepts submissions of new metrics for researchers who want to make their metrics more robust to attacks or to find such metrics for their needs. The latest results can be found online: https://videoprocessing.ai/benchmarks/metrics-robustness.html.

Introduction

Nowadays, most new image- and video-quality metrics (IQA/VQA) employ deep learning. For example, in the latest NTIRE challenge on perceptual quality assessment (Gu et al. 2022), all winning methods were based on neural networks. With the increased sizes of datasets and availability of crowdsourced markup, deep-learning-based metrics started to outperform traditional approaches in correlation with subjective quality. However, learning-based methods, including IQA/VQA metrics, are more vulnerable to adversarial attacks. A simple metric like PSNR is more stable to image modifications that aim to manipulate quality scores (any changed pixel will decrease the score). In contrast, the behaviour of deep metrics is much more complex. The existing benchmarks evaluate metrics' correlation with subjective quality but do not consider their robustness. At the same time, *the possibility to manipulate IQA/VQA metrics scores is already being exploited in different real-life scenarios*. Below are some examples of such scenarios and potential negative impacts from using non-robust IQA/VQA.

Decrease of perceptual quality. Metrics-oriented optimization modes are already being implemented in video encoders. libaom (Deng, Han, and Xu 2020) and LCEVC (V-Nova 2023) have options that optimize bitstream for increasing a VMAF score. Such tuning was designed to improve the visual quality of the encoded video; however, as VMAF is a learning-based metric, it may decrease perceptual quality (Zvezdakova et al. 2019; Siniukov et al. 2021). Using unstable image quality metrics as a perceptual proxy in a loss function may lead to incorrect restoration results (Ding et al. 2021). For instance, LPIPS is widely used as a perceptual metric, but optimizing its scores leads to increased brightness (Kettunen, Härkönen, and Lehtinen 2019), which is unwanted or even harmful (for example, when analyzing medical images).

Cheating in benchmarks. The developers of image- and video-processing methods can use metrics' vulnerabilities to achieve better competition results. For example, despite LPIPS already being shown to be vulnerable to adversarial attacks, it is still used as the main metric in some benchmarks, e.g. to compare super-resolution methods (Zhang et al. 2021). In some competitions that publish the results of subjective comparisons and objective quality scores, we can see the vast difference in these leaderboards. For instance, the VMAF leaders in 2021 Subjective Video Codecs Comparisons differ from leaders by subjective quality (Comparison 2021).

Manipulating the results of image web search. Search engines use not only keywords and descriptions but also image quality measurement to rank image search results. For example, the developers of Microsoft Bing used image quality as one of the features to improve its output (Bing 2013). As shown in MediaEval 2020 Pixel Privacy: Quality Camouflage for Social Images competition (MediaEval 2020), there are a variety of ways to fool image quality estimators.

Our study highlights the necessity of measuring the adversarial robustness of contemporary metrics for the research community. There are different ways to cheat on IQA/VQA metrics, such as increasing or decreasing their scores. In our study, we focus on analyzing metrics' resistance to attacks that increase estimated quality scores, as this kind of attack has already appeared in many real-life cases. Also, by choosing to investigate metrics' stability to scores increasing, we do not limit the generability of the results. We believe that the existing image- and video-quality metrics benchmarks must be supplemented with metrics' robustness analysis. In this paper, we first attempt to do this and apply several types of adversarial attacks to a number of quality metrics. Our contributions are as follows: a new benchmark methodology, a leaderboard published online [1], and an analysis of currently obtained results. We published our code [2] for generating adversarial attacks and a list of open datasets used in this study, so the developers of IQA/VQA methods can measure the stability of their methods to attacks. For those who want their approach published on our website, the benchmark accepts new submissions of quality metrics. Try our benchmark using *pip install robustness-benchmark*.

Related Work

Depending on the availability of the undistorted image, IQA/VQA metrics can be divided into three types: no-reference (NR), full-reference (FR) or reduced-reference (RR). NR metrics have the broadest applications but generally show lower correlations with subjective quality than FR and RR metrics. However, recent results show that new NR metrics outperformed many existing FR methods, so we mainly focused on NR metric evaluation in this paper. The performance of IQA/VQA metrics is traditionally evaluated using subjective tests that measure the correlation of metric scores with perceptual ones. The most well-known comparisons were published within NTIRE Workshop (Gu et al. 2022), and two benchmarks currently accept new submissions: MSU Video Quality Metrics Benchmark (Antsiferova et al. 2022) and UGC-VQA (Tu et al. 2021). These studies show how well the compared metrics estimate subjective quality but do not reflect their robustness to adversarial attacks.

There are different ways to measure the robustness of neural network-based methods. It can be done via theoretical estimations, e.g. Lipschitz regularity. However, this approach has many limitations, including the number of parameters in the evaluated network. A more universal approach is based on applying adversarial attacks. This area is widely studied for computer vision models. However, not all methods can be adapted to attack quality metrics.

The first methods for measuring the robustness of IQA/VQA metrics were based on creating a specific situation in which the metric potentially fails. Ciaramello and Reibman (2011a) first conducted such analysis and proposed a method to reveal the potential vulnerabilities of an objective quality model based on the generation of image or video pairs with the intent to cause misclassification errors (Brill et al. 2004) by this model. Misclassification errors include false ordering (FO, the objective model rates a pair opposite to humans), false differentiation (FD, the objective model rates a pair as different but humans do not), and false tie (FT, humans order a pair as different, but the objective model does not). H. Liu and A. Reibman (2016) introduced a software called "STIQE" that automatically explores an image-quality metric's performance. It allows users to execute tests and then generate reports to determine how well the metric performs. Testing consists of applying several varying distortions to images and checking whether the metric score rises monotonically as the degree of the applied distortion.

Benchmark	# attacks / # metrics	Metrics type	Test datasets
Ciaramello and Reibman (2011a)	5 / 4	FR	10 images
Ciaramello and Reibman (2011b)	5 / 9	NR, FR	473 images
Liu and Reibman (2016)	5 / 11	NR, FR	60 images
Shumitskaya et al. (2022)	1 / 7	NR	20 videos
Zhang et al. (2022)	1 / 4	NR	12 images
Ghildyal and Liu (2023)	6 / 5	FR	12,227 images
Ours	9 / 15	NR, FR	3000 images, 1 video

Table 1: Comparisons of image- and video-quality metrics' stability to adversarial attacks.

Nowadays, metrics' adversarial robustness is primarily estimated by adapting attacks designed for computer vision tasks to image quality metrics. A more detailed description of existing attacks against metrics that we used in our study is given in the section "List of adversarial attacks". There are two recently published attacks that we aim to add to the benchmark shortly: a new CNN-based generative attack FACPA (Shumitskaya, Antsiferova, and Vatolin 2023), attack with human-in-the-loop by Zhang et al. (Zhang et al. 2022) and spatial attack that was adapted for metrics (Ghildyal and Liu 2023).

Recently, a new study on the adversarial robustness of full-reference metrics was published (Ghildyal and Liu 2023). The authors showed that six full-reference metrics are susceptible to imperceptible perturbations generated via common adversarial attacks such as FGSM (Goodfellow, Shlens, and Szegedy 2015), PGD (Madry et al. 2017), and the One-pixel attack (Su, Vargas, and Sakurai 2019). They also showed that adversarial perturbations crafted for LPIPS metric (Zhang et al. 2018) using stAdv attack can be transferred to other metrics. As a result, they concluded that more accurate learning-based metrics are less robust to adversarial attacks than traditional ones. We summarised the existing research on IQA/VQA metrics' robustness to adversarial attacks in Table 1.

Benchmark

List of Metrics

In this paper, we focused on the evaluation of only no-reference metrics for several reasons: firstly, there exists a similar evaluation of full-reference metrics (Ghildyal and

[1] https://videoprocessing.ai/benchmarks/metrics-robustness.html

[2] https://github.com/msu-video-group/MSU_Metrics_Robustness_Benchmark

Liu 2023); secondly, no-reference metrics have a more comprehensive range of applications and are more vulnerable to attacks; thirdly, these metrics are mostly learning-based. We considered state-of-the-art metrics according to other benchmarks and various other no-reference metrics. All tested metrics assess image quality, except for VSFA (Li, Jiang, and Jiang 2019) and MDTVSFA (Li, Jiang, and Jiang 2021), which are designed for videos.

RankIQA (Liu, Van De Weijer, and Bagdanov 2017) pretrains a model on a large dataset with synthetic distortions to compare pairs of images, then fine-tunes it on a small realistic dataset. **MetaIQA** (Zhu et al. 2020) introduces a quality prior model pre-trained on several dozens of specific distortions and fine-tuned on a smaller target dataset, similar to RankIQA. **WSP** (Su and Korhonen 2020) is concerned with Global Average Pooling feature aggregation used by most existing methods and replaces it with Weighted Spatial Pooling to distinguish important locations. **CLIP-IQA** (Wang, Chan, and Loy 2023) predicts the quality perception and image-provoked abstract emotions by feeding heterogeneous text prompts and the image to the CLIP network. **PAQ-2-PIQ** (Ying et al. 2020) introduces a large subjective picture quality database of about 40,000 images, trains a CNN with ResNet-18 backbone to predict patch quality and combines the predictions with RoI pooling. **HyperIQA** (Su et al. 2020) focuses on real-life IQA and proposes a hyperconvolutional network that predicts the weights of fully connected layers. **MANIQA** (Yang et al. 2022) assesses quality of GAN-based distortions. The model uses vision transformer features processed by proposed network modules to enhance global and local interactions. The final score prediction utilizes patch weighting. **TReS** (Golestaneh, Dadsetan, and Kitani 2022) proposes to compute local features with CNN and non-local features with self-attention, introduces a per-batch loss for correct ranking and a self-supervision loss between reference and flipped images. **FPR** (Chen et al. 2022) hallucinates pseudo-reference features from the distorted image using mutual learning on reference and distorted images with triplet loss. Attention maps are predicted to aggregate scores over patches. **VSFA** (Li, Jiang, and Jiang 2019) estimates video quality using ResNet-50 features for content awareness and differentiable temporal aggregation, which consists of gated recurrent units with min pooling. **MDTVSFA** (Li, Jiang, and Jiang 2021) enhances VSFA with explicit mapping between predicted and dataset-specific scores, supported by multi-dataset training. **NIMA** (Talebi and Milanfar 2018) predicts a distribution of scores instead of regressing a single value and considers both technical and aesthetic image scores. It is trained on the Aesthetic Visual Analysis database using squared earth mover's distance as a loss. **LINEARITY** (Li, Jiang, and Jiang 2020) invents the norm-in-norm loss, which shows ten times faster convergence than MSE or MAE with ResNet architecture. **SPAQ** (Fang et al. 2020) collects a database of 11,125 smartphone photos, proposes a ResNet-50 baseline model and three modified versions incorporating EXIF data (MT-E), subjective image attributes (MT-A) and scene labels (MT-S). **KonCept512** (Hosu et al. 2020) collects KonIQ-10k, a diverse crowdsourced database of 10,073 images and trains a model with InceptionResNetV2 backbone.

We also used MSE, PSNR and SSIM (Wang et al. 2004) as proxy metrics to estimate image quality degradation after attacks. The choice is motivated by their structure (full-reference and not learning-based), which makes them more stable to adversarial attacks.

List of Adversarial Attacks

In all attacks, we define the loss function as $J(\theta, I) = 1 - score(I)/range$ and minimize it by making small steps along the gradient direction in image space, which increases the attacked metric score. $range$ is computed as the difference between maximum and minimum metric values on the dataset and serves to normalize the gradient magnitude across different metrics.

FGSM-based attacks are performed for each image. The pixel difference is limited by ε. **FGSM** (Goodfellow, Shlens, and Szegedy 2015) is a basic approach that makes one gradient step: $I^{adv} = I - \varepsilon \cdot sign(\nabla_I J(\theta, I))$. **I-FGSM** (Kurakin, Goodfellow, and Bengio 2018) is a more computationally expensive method that uses T iterations and clips the image on each step: $I_{t+1}^{adv} = Clip_{I,\varepsilon}\{I_t^{adv} - \alpha \cdot sign(\nabla_I J(\theta, I_t^{adv}))\}$, where $t = 0, 1, \ldots T-1$, I_0 is the input image I, and α is the perturbation intensity. The clipped pixel value at position (x, y) and channel c satisfies $|I_t^{adv}(x, y, c) - I(x, y, c)| < \varepsilon$. PGD (Madry et al. 2017) is identical to I-FGSM except for the random initialization in the ε-vicinity of the original image; due to its similarity to I-FGSM, we didn't include it in the experiments. **MI-FGSM** (Dong et al. 2018) uses gradient momentum: $I_{t+1}^{adv} = Clip_{I,\varepsilon}\{I_t^{adv} - \alpha \cdot sign(g_t)\},\ t = 0, 1, \ldots T-1$, $g_t = \nabla_I J(\theta,\ I_t^{adv}) + \nu \cdot g_{t-1}, \quad g_{-1} = 0$, where ν controls the momentum preservation. **AMI-FGSM** (Sang et al. 2022) is identical to MI-FGSM, except the pixel difference limit ε is set to $1/NIQE(I)$ by computing the NIQE (Mittal, Soundararajan, and Bovik 2012) no-reference metric.

Universal Adversarial Perturbation (UAP)-based attacks generate adversarial perturbation for an attacked metric, which is the same for all images and videos. When UAP is generated, the attack process consists of the mere addition of an image with UAP. The outcome is the image with an increased target metric score. We used three methods to train UAPs. **Cumulative-UAP** is obtained by averaging non-universal perturbation on the training dataset. Non-universal perturbations are generated using one step of gradient descent. **Optimized-UAP** is obtained by training UAP weights using batch training with Adam optimizer and loss function defined as target metric with opposite sign. **Generative-UAP** is obtained by auxiliary U-Net generator training. The network is trained to generate a UAP from random noise with uniform distribution. The Adam optimizer is used for training, and the loss function is defined as the target metric with the opposite sign. Once the network is trained, a generated UAP is saved and further used to attack new images.

Perceptual-aware attacks use other image quality metrics to control attack imperceptibility to the human eye. **Korhonen et al.** (Korhonen and You 2022) proposes a method for generating adversarial images for NR quality metrics with perturbations located in textured regions. They use gradient descent with additional elementwise multiplication of gradients by a spatial activity map. The spatial activity map of an image is calculated using horizontal and vertical 3×3 Sobel filters. **MADC** (Wang and Simoncelli 2008) is a method for comparing two image- or video-quality metrics by constructing a pair of examples that maximize or minimize the score of one metric while keeping the other fixed. In our study, we fixed MSE while maximizing an attacked metric. The projected gradient descent step and binary search are performed on each iteration. Let $g1$ be the gradient with direction to increase the attacked metric and $g2$ the gradient of MSE on some iteration. The projected gradient is then calculated as $pg = g1 - \frac{g2^T \cdot g1}{g2^T \cdot g2} \cdot g2$. After projected gradient descent, the binary search to guarantee a fixed MSE is performed (with 0.04 precision). The binary search is the process that consists of small steps along the MSE gradient: if the precision is bigger than 0.04, then steps are taken along the direction of reducing MSE and vice versa.

Methodology

Datasets This study incorporated pre-trained quality metrics as a part of our evaluation benchmark. We did not perform metrics fine-tuning on any data. We used six datasets summarised in Table 2. These datasets are widely used in the computer vision field. We chose them to cover a diverse range of real-life scenarios, including images and video, with varying resolutions from 299 × 299 up to 1920 × 1080 (FullHD). All datasets have an open license that allows them to be used in this work. Our analysis categorized the adversarial attacks into trainable and non-trainable attacks. Three datasets were used to train adversarial attacks, and three were used for testing. We trained UAP attacks using each training dataset, resulting in three versions of each attack. These versions were subsequently evaluated on the designated testing datasets, and the results for different versions were averaged among each UAP-attack type and amplitude. Non-trainable attacks were directly evaluated on the testing datasets. We have analyzed the efficiency and generalization capabilities of both trainable and non-trainable adversarial attacks across various data domains while also considering the influence of training data on metric robustness. NIPS 2017: Adversarial Learning Development Set (2017) was also used to train metrics' domain transformations (described further in "Evaluation metrics").

Implementation Details We used public source code for all metrics without additional pretraining and selected the default parameters to avoid overfitting. The training and evaluation of attacks on the metrics were fully automated. We employed the CI/CD tools within a GitLab repository for our measurement procedures. We established a sophisticated end-to-end pipeline from the attacked metrics' original repositories to the resulting robustness scores to make the results entirely verifiable and reproducible. The pipeline scheme, the list of used attack's hyper-parameters and the hyperparameter choice justification are presented in the supplementary materials (Antsiferova et al. 2023). UAP-based attacks (UAP, cumulative UAP and generative UAP) were averaged with three different amplitudes (0.2, 0.4 and 0.8).

Quality metrics implementations were obtained from official repositories. We only modified interfaces to meet our requirements and used default parameters provided by the authors. Links to original repositories and a list of applied patches (where it was needed to enable gradients) are provided in supplementary materials (Antsiferova et al. 2023).

Calculations were performed on two computers with the following characteristics:

- 4 x GeForce RTX 3090 GPU, an Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz
- 4 x NVIDIA RTX A6000 GPU, AMD EPYC 7532 32-Core Processor @ 2.40GHz

All calculations took a total of about 2000 GPU hours. The values of parameters (ϵ, number of iterations, etc.) for the attacks are listed in the supplementary materials (Antsiferova et al. 2023).

Evaluation Metrics Before calculating metrics' robustness scores, metric values are transformed with min-max scaling so that the values before the attack lie in the range [0,1]. To compensate for the nonlinear dependence between metrics (Zhang et al. 2022), we converted all metrics to the same domain before comparison. MDTVSFA (Li, Jiang, and Jiang 2021) was used as the primary domain, as it shows the best correlations with MOS among tested metrics according to the MSU Video Quality Metrics benchmark results. We employed the 1-Dimensional Neural Optimal Transport (Korotin, Selikhanovych, and Burnaev 2023) method to build the nonlinear transformation between the distributions of all metrics to one general shape. We also present the results without the nonlinear transformation in the supplementary materials (Antsiferova et al. 2023).

Absolute and **Relative gain**. Absolute gain is calculated as the average difference between the metric values before and after the attack. Relative gain is the average ratio of the difference between the metric values before and after the attack to the metric value before the attack plus 1 (1 is added to avoid division problems, as values before the attack are scaled to [0,1]).

$$\begin{aligned} Abs.gain &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(f(x_i') - f(x_i)\right), \\ Rel.gain &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{f(x_i')-f(x_i)}{f(x_i)+1}, \end{aligned} \tag{1}$$

where n is the number of images, x_i is the clear image, x_i' — it's attacked counterpart, and $f(.)$ is the IQA metric function.

Robustness score (Zhang et al. 2022) R_{score} is defined as the average ratio of maximum allowable change in quality prediction to actual change over all attacked images in a logarithmic scale:

$$R_{score} = \frac{1}{n}\sum_{i=1}^{n} log_{10}\left(\frac{max\{\beta_1 - f(x_i'), f(x_i) - \beta_2\}}{|f(x_i') - f(x_i)|}\right). \tag{2}$$

As metric values are scaled, we use $\beta_1 = 1$ and $\beta_2 = 0$.

Wasserstein score (Kantorovich 1960) W_{score} and **Energy Distance score** (Szekely 2002) E_{score} are used to evaluate the statistical differences between distributions of metric values before and after the attack. Large positive values

Training datasets (for UAP attacks)	Type	Number of samples	Resolution	**Testing datasets**	Type	Number of samples	Resolution
COCO (2014)	Images	300,000	640 × 480	NIPS (2017)	Images	1,000	299 × 299
Pascal VOC (2012)	Image	11,530	500 × 333	Derf's collection (2001)	Video	24 ($\sim$ 10k frames)	1920 × 1080
Vimeo-90k Train set (2019)	Triplets of images	2,001	448 × 256	Vimeo 90k Test set (2019)	Triplets of images	11,346	448 × 256

Table 2: Summary of the datasets used in our study.

of these scores correspond to a significant upward shift of the metric's predictions, values near zero indicate the absence of the metric's response to the attack, and negative ones show a decrease in the metric predictions and the inefficiency of the attack. These scores are defined as corresponding distances between distributions multiplied by the sign of the difference between the mean values before and after the attack:

$$\begin{gathered} W_{score} = W_1(\hat{P}, \hat{Q}) \cdot sign(\bar{x}_{\hat{Q}} - \bar{x}_{\hat{P}}), \\ W_1(\hat{P}, \hat{Q}) = \inf_{\gamma \in \Gamma(\hat{P}, \hat{Q})} \int_{\mathbb{R}^2} |x - y| d\gamma(x, y) = \\ = \int_{-\infty}^{\infty} |\hat{F}_{\hat{P}}(x) - \hat{F}_{\hat{Q}}(x)| dx; \end{gathered} \tag{3}$$

$$\begin{gathered} E_{score} = E(\hat{P}, \hat{Q}) \cdot sign(\bar{x}_{\hat{Q}} - \bar{x}_{\hat{P}}), \\ E(\hat{P}, \hat{Q}) = (2 \cdot \int_{-\infty}^{\infty} (\hat{F}_{\hat{P}}(x) - \hat{F}_{\hat{Q}}(x))^2 dx)^{\frac{1}{2}}, \end{gathered} \tag{4}$$

where $\hat{P}$ and $\hat{Q}$ are empirical distributions of metric values before and after the attack, $\hat{F}_{\hat{P}}(x)$ and $\hat{F}_{\hat{Q}}(x)$ are their respective empirical Cumulative Distribution Functions, and $\bar{x}_{\hat{P}}$ and $\bar{x}_{\hat{Q}}$ are their respective sample means.

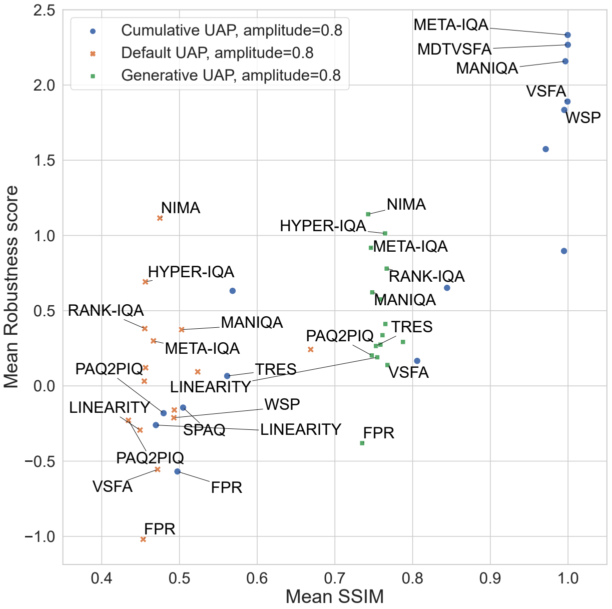

Figure 1: Metrics' robustness score for UAP-based adversarial attacks and SSIM measured between original and attacked images. The results are averaged for all test datasets.

Results

The main results of our study are aggregated across the different attack types, training and testing datasets. Tables and figures for other robustness measures, by specific datasets and attacks, are presented in the supplementary materials (Antsiferova et al. 2023) and on the benchmark webpage.

Metrics that are robust to UAP-based attacks. Despite the three types of implemented UAP-based attacks resulting in different attack efficiency, the most and least robust metrics for these attacks are similar. MANIQA showed the best robustness score for all amplitudes of Optimized UAP and is within top-3 metrics robust to Generative UAP. This metric uses ViT and applies attention mechanisms across the channel and spatial dimensions, increasing interaction among different regions of images globally and locally. HYPER-IQA showed high resistance to all UAP attacks. Besides FPR, the PAQ-2-PIQ showed the worst energy distance score. The robustness scores of analyzed attacks are provided in Table 3 and illustrated in Fig. 1. Annotations include only five best and five worst methods judged by robustness score for better visibility.

Metrics that are robust to iterative attacks. CLIP-IQA shows the best robustness to most iterative attacks, followed by RANK-IQA and MDTVSFA. RANK-IQA also offers the best resistance to perceptually oriented MADC and Korhonen attacks. These attacks use approaches to reduce the visibility of distortions caused by an attack, which makes it more difficult for them to succeed. The robustness score of analyzed attacks is shown in Table 3 and illustrated in Fig. 2. Annotations include only five best and five worst methods judged by robustness score for better visibility.

Metrics' robustness at different levels of perceptual quality loss. As described in the Benchmark section, we used SSIM, PSNR and MSE as simple proxies for estimating perceptual quality loss of attacks in this study. Fig. 3 shows an averaged robustness score depending on SSIM loss of attacked images for all attacks. It shows that all metrics become less robust to attacks when more quality degradation is allowed. HYPER-IQA's robustness is more independent from SSIM loss among all metrics. Otherwise, PAQ-2-PIQ, VSFA and FPR are becoming more vulnerable than other metrics with increasing SSIM degradation. Results for other proxy metrics (MSE and PSNR) are provided in the supplementary materials (Antsiferova et al. 2023) and on the

	O-UAP	G-UAP	C-UAP	FGSM	I-FGSM	MI-FGSM	AMI-FGSM	MADC	Korhonen et al.
CLIP-IQA	0.632	0.397	0.067	0.398	**0.836**	**0.821**	**0.819**	0.823	0.812
META-IQA	0.183	-0.029	0.003	0.529	1.307	1.285	1.287	0.934	0.997
RANK-IQA	0.295	0.064	0.180	0.285	1.063	0.891	0.893	**0.383**	**0.763**
HYPER-IQA	0.072	-0.094	0.086	**-0.406**	1.366	1.387	1.396	0.848	1.329
KONCEPT	0.419	0.187	0.435	0.574	1.248	1.066	1.066	0.753	1.042
FPR	1.705	0.846	0.966	0.682	3.344	3.210	3.215	1.703	3.018
NIMA	-0.024	0.046	0.018	0.258	1.203	1.147	1.148	0.959	1.041
WSP	0.784	0.155	0.012	0.405	1.260	1.251	1.257	0.760	0.894
MDTVSFA	0.756	0.359	0.005	0.185	1.011	0.983	0.983	0.914	0.805
LINEARITY	1.022	0.445	0.972	-0.220	1.284	1.218	1.224	0.816	1.204
VSFA	1.151	0.361	0.014	0.306	2.054	2.272	2.274	1.470	1.539
PAQ-2-PIQ	0.943	0.252	0.873	0.578	1.190	1.123	1.125	0.536	0.997
SPAQ	0.605	0.357	0.560	0.266	1.514	1.371	1.375	0.740	1.301
TRES	0.691	0.358	0.634	0.826	1.223	1.209	1.210	0.741	1.173
MANIQA	**-0.390**	**-0.174**	**-0.003**	0.499	1.403	1.225	1.226	0.698	0.843

Table 3: Metrics' robustness calculated using energy distance score measure to different types of attacks. The results are averaged across test datasets. O-UAP stands for "Optimised-UAP", G-UAP for "Generative-UAP", C-UAP for "Cumulative-UAP".

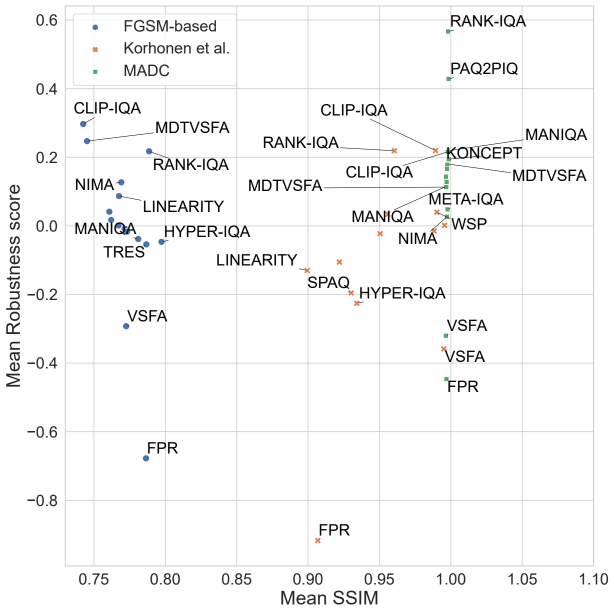

Figure 2: Metrics' robustness score for iterative adversarial attacks and SSIM measured between original and attacked images. The results are averaged for all test datasets.

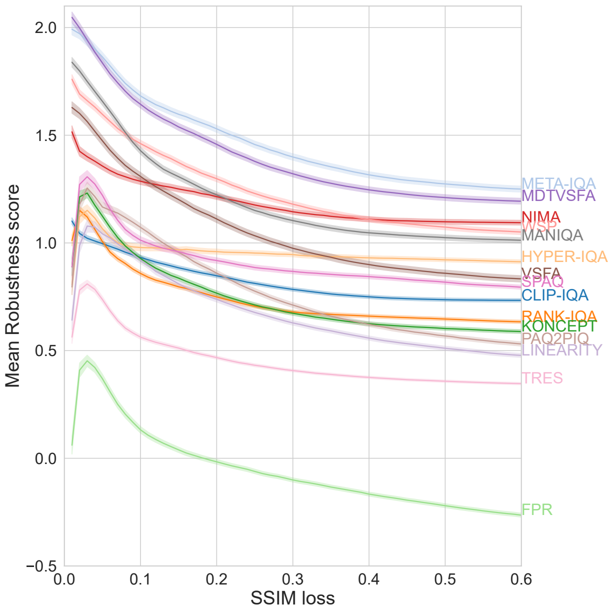

Figure 3: Dependency of metrics' robustness score of SSIM loss for attacked images (all types of attacks).

benchmark webpage.

Overall metrics' robustness comparison. Table 4 and Fig. 4 show the general results of our study. First, we see that iterative attacks are more efficient against all metrics. However, metrics' robustness is different for UAP and iterative attacks. We summarised the robustness of all attack types in the table and compared them using various measures. According to absolute and relative gain, the leaders are the same: MANIQA, NIMA and RANK-IQA, and they also perform well based on other measures. META-IQA and MDTVSFA have high robustness scores. Energy measures also show similar results. FPR is the least stable to adversarial attacks, considering all tests and measures.

One-sided Wilcoxon signed-rank tests. To study the statistical difference in the results, we conducted one-sided Wilcoxon tests on the values of absolute gains for all pairs of metrics. A table with detailed test results for different types of attacks can be found in the supplementary materials (Antsiferova et al. 2023). All metrics are statistically superior to the FPR metric, which means that FPR can be significantly increased under the influence of any of the con-

	$Abs.gain \downarrow$	$Rel.gain \downarrow$	$R_{score} \uparrow$	$E_{score} \downarrow$	$W_{score} \downarrow$
CLIP-IQA	0.256 (0.254, 0.258)	0.184 (0.182, 0.185)	0.702 (0.698, 0.707)	0.424	0.256
META-IQA	0.241 (0.238, 0.243)	0.182 (0.180, 0.184)	**1.168** (1.161, 1.176)	0.324	0.241
RANK-IQA	0.184 (0.183, 0.186)	0.12 (0.119, 0.122)	0.843 (0.839, 0.848)	0.285	0.184
HYPER-IQA	0.232 (0.228, 0.235)	0.151 (0.149, 0.153)	0.740 (0.735, 0.745)	0.277	0.237
KONCEPT	0.328 (0.326, 0.330)	0.227 (0.225, 0.228)	0.584 (0.579, 0.589)	0.489	0.328
FPR	2.591 (2.568, 2.615)	1.730 (1.714, 1.746)	-0.229(-0.234, -0.224)	1.409	2.591
NIMA	0.17 (0.168, 0.172)	0.115 (0.114, 0.117)	1.152 (1.146, 1.158)	0.239	**0.170**
WSP	0.380 (0.377, 0.384)	0.276 (0.273, 0.278)	0.893 (0.886, 0.901)	0.449	0.380
MDTVSFA	0.279 (0.277, 0.281)	0.186 (0.184, 0.187)	0.99 (0.983, 0.998)	0.447	0.279
LINEARITY	0.683 (0.679, 0.687)	0.447 (0.444, 0.450)	0.267 (0.263, 0.272)	0.780	0.683
VSFA	0.899 (0.891, 0.907)	0.611 (0.606, 0.617)	0.659 (0.650, 0.667)	0.739	0.899
PAQ-2-PIQ	0.521 (0.518, 0.524)	0.341 (0.338, 0.343)	0.449 (0.443, 0.454)	0.675	0.521
SPAQ	0.671 (0.665, 0.678)	0.536 (0.531, 0.542)	0.493 (0.488, 0.499)	0.637	0.671
TRES	0.433 (0.431, 0.435)	0.305 (0.304, 0.307)	0.320 (0.317, 0.323)	0.627	0.433
MANIQA	**0.104** (0.101, 0.107)	**0.078** (0.076, 0.08)	0.986 (0.979, 0.993)	**0.207**	0.175

Table 4: Metrics' robustness to tested adversarial attacks according to different stability measures. The results for abs. gain, rel. gain and R-score were averaged across different types of attacks and test datasets, so they are presented with confidence intervals. The E_{score} and W_{score} were calculated using the whole set of attacked results without averaging.

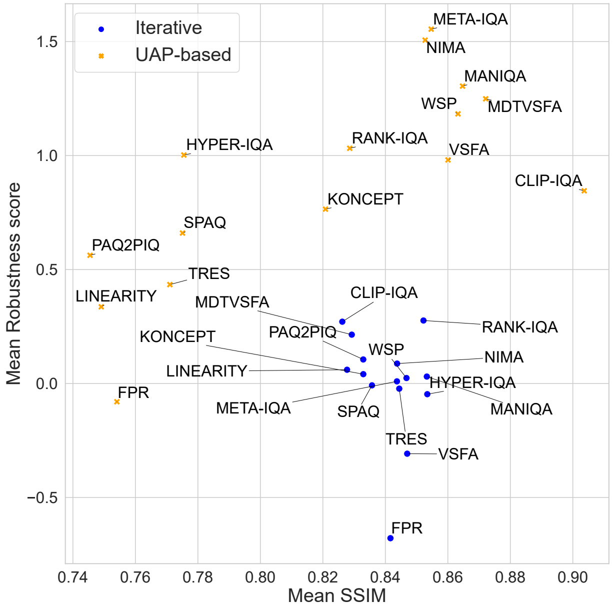

Figure 4: Mean robustness score of compared metrics versus SSIM averages for UAP-based and iterative attacks.

sidered attacks. MANIQA, on the contrary, turns out to be one of the most stable metrics for all attacks on average, but it is inferior to CLIP-IQA on FGSM-based attacks. Overall, the results of the Wilcoxon one-sided tests are consistent with our evaluations of the obtained results.

Stable metrics feature analysis. To analyze the relationship of metrics' architectures with robustness, we summarised the main features of tested metrics in Table 1 of the supplementary materials. A common feature of robust metrics is the usage of the input image cropping or resizing. High stability to attacks was also shown by META-IQA, which does not transform input images but uses a relatively small backbone network that leverages prior knowledge of various image distortions obtained during so-called meta-learning.

Conclusion

This paper analyzed the robustness of 15 no-reference image/video-quality metrics to different adversarial attacks. Our analysis showed that all metrics are susceptible to adversarial attacks, but some are more robust than others. MANIQA, META-IQA, NIMA, RANK-IQA and MDTVSFA showed high resistance to adversarial attacks, making their usage in practical applications safer than other metrics. We published this comparison online and are accepting new metrics submissions. This benchmark can be helpful for researchers and companies who want to make their metrics more robust to potential attacks.

In this paper, we revealed ways of cheating on image quality measures, which can be considered to have a potential negative social impact. However, as was discussed in the Introduction, the vulnerabilities of image- and video-quality metrics are already being exploited in some real-life applications. At the same time, only a few studies have been published. We open our findings to the research community to increase the trustworthiness of image/video processing and compression benchmarks. Limitations of our study are listed in the supplementary materials (Antsiferova et al. 2023).

Acknowledgments

The authors would like to thank the video group of MSU Graphics and Media Laboratory, especially Kirill Malyshev and Vyacheslav Napadovsky, for setting up the infrastructure and helping to receive computational results for this research. The work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming dated November 2, 2021 No. 70-2021-00142.

References

2001. Xiph.org Video Test Media [derf's collection]. https://media.xiph.org/video/derf/.

2017. NIPS 2017: Adversarial Learning Development Set. https://www.kaggle.com/datasets/google-brain/nips-2017-adversarial-learning-development-set.

Antsiferova, A.; Abud, K.; Gushchin, A.; Lavrushkin, S.; Shumitskaya, E.; Velikanov, M.; and Vatolin, D. 2023. Comparing the robustness of modern no-reference image- and video-quality metrics to adversarial attacks. arXiv:2310.06958.

Antsiferova, A.; Lavrushkin, S.; Smirnov, M.; Gushchin, A.; Vatolin, D.; and Kulikov, D. 2022. Video compression dataset and benchmark of learning-based video-quality metrics. In *Advances in Neural Information Processing Systems*, volume 35, 13814–13825.

Bing, M. 2013. A Behind the Scenes Look at How Bing is Improving Image Search Quality. https://blogs.bing.com/search-quality-insights/2013/08/23/a-behind-the-scenes-look-at-how-bing-is-improving-image-search-quality.

Brill, M. H.; Lubin, J.; Costa, P.; Wolf, S.; and Pearson, J. 2004. Accuracy and cross-calibration of video quality metrics: new methods from ATIS/T1A1. *Signal Processing: Image Communication*, 19(2): 101–107.

Chen, B.; Zhu, L.; Kong, C.; Zhu, H.; Wang, S.; and Li, Z. 2022. No-Reference Image Quality Assessment by Hallucinating Pristine Features. *IEEE Transactions on Image Processing*, 31: 6139–6151.

Ciaramello, F. M.; and Reibman, A. R. 2011a. Supplemental subjective testing to evaluate the performance of image and video quality estimators. In *Human Vision and Electronic Imaging XVI*, volume 7865, 249–257. SPIE.

Ciaramello, F. M.; and Reibman, A. R. 2011b. Systematic stress testing of image quality estimators. In *2011 18th IEEE International Conference on Image Processing*, 3101–3104. IEEE.

Comparison, M. V. C. 2021. MSU Video Codecs Comparison 2021 Part 2: Subjective. http://www.compression.ru/video/codec_comparison/2021/subjective_report.html.

Deng, S.; Han, J.; and Xu, Y. 2020. Vmaf based rate-distortion optimization for video coding. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 1–6. IEEE.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2021. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129: 1258–1281.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3677–3686.

Ghildyal, A.; and Liu, F. 2023. Attacking Perceptual Similarity Metrics. *arXiv preprint arXiv:2305.08840*.

Golestaneh, S. A.; Dadsetan, S.; and Kitani, K. M. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1220–1230.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Gu, J.; Cai, H.; Dong, C.; Ren, J. S.; Timofte, R.; Gong, Y.; Lao, S.; Shi, S.; Wang, J.; Yang, S.; et al. 2022. NTIRE 2022 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 951–967.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29: 4041–4056.

Kantorovich, L. V. 1960. Mathematical Methods of Organizing and Planning Production. *Management Science*, 6(4): 366–422.

Kettunen, M.; Härkönen, E.; and Lehtinen, J. 2019. E-lpips: robust perceptual image similarity via random transformation ensembles. *arXiv preprint arXiv:1906.03973*.

Korhonen, J.; and You, J. 2022. Adversarial Attacks Against Blind Image Quality Assessment Models. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, 3–11.

Korotin, A.; Selikhanovych, D.; and Burnaev, E. 2023. Neural Optimal Transport. In *International Conference on Learning Representations*.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Li, D.; Jiang, T.; and Jiang, M. 2019. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2351–2359.

Li, D.; Jiang, T.; and Jiang, M. 2020. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*, 789–797.

Li, D.; Jiang, T.; and Jiang, M. 2021. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129: 1238–1257.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; and Reibman, A. R. 2016. Software to stress test image quality estimators. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 1–6. IEEE.

Liu, X.; Van De Weijer, J.; and Bagdanov, A. D. 2017. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, 1040–1049.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

MediaEval. 2020. Pixel Privacy: Quality Camouflage for Social Images. https://multimediaeval.github.io/editions/2020/tasks/pixelprivacy/.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

Sang, Q.; Zhang, H.; Liu, L.; Wu, X.; and Bovik, A. 2022. On the Generation of Adversarial Samples for Image Quality Assessment. *Available at SSRN 4112969*.

Shumitskaya, E.; Antsiferova, A.; and Vatolin, D. S. 2022. Universal Perturbation Attack on Differentiable No-Reference Image- and Video-Quality Metrics. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press.

Shumitskaya, E.; Antsiferova, A.; and Vatolin, D. S. 2023. Fast Adversarial CNN-based Perturbation Attack of No-Reference Image Quality Metrics.

Siniukov, M.; Antsiferova, A.; Kulikov, D.; and Vatolin, D. 2021. Hacking VMAF and VMAF NEG: vulnerability to different preprocessing methods. In *2021 4th Artificial Intelligence and Cloud Computing Conference*, 89–96.

Su, J.; Vargas, D. V.; and Sakurai, K. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5): 828–841.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. 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*, 3667–3676.

Su, Y.; and Korhonen, J. 2020. Blind natural image quality prediction using convolutional neural networks and weighted spatial pooling. In *2020 IEEE International Conference on Image Processing (ICIP)*, 191–195. IEEE.

Szekely, G. J. 2002. E-statistics: The Energy of Statistical Samples. Technical Report 02-16, Bowling Green State University.

Talebi, H.; and Milanfar, P. 2018. NIMA: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011.

Tu, Z.; Chen, C.-J.; Wang, Y.; Birkbeck, N.; Adsumilli, B.; and Bovik, A. C. 2021. Video Quality Assessment of User Generated Content: A Benchmark Study and a New Model. In *2021 IEEE International Conference on Image Processing (ICIP)*, 1409–1413. IEEE.

V-Nova. 2023. FFmpeg with LCEVC. https://docs.v-nova.com/.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *AAAI*.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *Image Processing, IEEE Transactions on*, 13: 600 – 612.

Wang, Z.; and Simoncelli, E. P. 2008. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12): 8–8.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video Enhancement with Task-Oriented Flow. *International Journal of Computer Vision (IJCV)*, 127(8): 1106–1125.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1191–1200.

Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; and Bovik, A. 2020. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3575–3585.

Zhang, K.; Li, D.; Luo, W.; Ren, W.; Stenger, B.; Liu, W.; Li, H.; and Yang, M.-H. 2021. Benchmarking ultra-high-definition image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14769–14778.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, W.; Li, D.; Min, X.; Zhai, G.; Guo, G.; Yang, X.; and Ma, K. 2022. Perceptual Attacks of No-Reference Image Quality Models with Human-in-the-Loop. *arXiv preprint arXiv:2210.00933*.

Zhu, H.; Li, L.; Wu, J.; Dong, W.; and Shi, G. 2020. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14143–14152.

Zvezdakova, A.; Zvezdakov, S.; Kulikov, D.; and Vatolin, D. 2019. Hacking VMAF with video color and contrast distortion. In *CEUR Workshop Proceedings*, 53–57.