Black-box Universal Adversarial Perturbations for Image and Video Quality Assessment Methods

Georgii Bychkov[1,2], *Sergey Lavrushkin*[1], *Dmitriy Vatolin*[1,2,3]
[1]MSU Institute for Artificial Intelligence, Moscow, Russian Federation; [2]Lomonosov MSU, Moscow, Russian Federation
[3]Laboratory of Innovative Technologies for Processing Video Content, Innopolis University, Innopolis, Russian Federation
{georgy.bychkov,sergey.lavrushkin,dmitriy}@graphics.cs.msu.ru

***Abstract*—Most neural-network-based image- and video-quality-assessment metrics exhibit state-of-the-art performance; recent research, however, exposes their vulnerability to adversarial attacks. These attacks manipulate input data to inflate metric scores without enhancing visual quality. Although current research predominantly focuses on white-box adversarial attacks, black-box attacks remain underdeveloped despite their practical relevance, such as targeting metrics that are nondifferentiable or whose weights are unavailable. This paper adapts 10 black-box adversarial attacks, originally designed for image classifiers, to evaluate their effectiveness against image- and video-quality-assessment metrics. We tested them on 18 quality metrics and demonstrated their susceptibility to adversarial manipulations. Nonetheless, the practical utility of these attacks is limited by their computational complexity. To address this problem, we incorporated universal adversarial perturbations (UAPs) into four black-box attacks. Our proposed UAP attacks, despite providing less attack strength than the originals, still effectively increase the metric scores and are viable for real-time applications. Our code is available at https://github.com/georgebychkov/black-box-iqa.**

***Index Terms*—image and video quality assessment, black-box attacks, universal adversarial perturbations**

I. Introduction

Many computer vision algorithms are based on deep neural networks. The rapid spread of these networks has spurred research into trusted AI, as several studies [1]–[4] have demonstrated their susceptibility to adversarial attacks. Given that most neural networks are prone to such attacks, evaluating their robustness and developing defensive measures is crucial.

Broadly, adversarial attacks fall into two groups: white-box and black-box. In a white-box scenario, the attacker has complete access to the target model, including its architecture and weights. These attacks typically use gradients to generate adversarial examples. By contrast, the attacker in a black-box scenario has little knowledge of the target model and can only access its output. Black-box attacks rely solely on information about the inputs and their corresponding outputs to craft adversarial examples. This approach encompasses a wider range of applications than white-box attacks and is relevant to situations where white-box attacks are unfeasible.

This paper examines adversarial attacks on image-quality-assessment (IQA) and video-quality-assessment (VQA) methods. Machine-learning-based IQA and VQA have demonstrated superior correlation with subjective scores compared with traditional metrics, but several studies [5]–[7] have revealed their vulnerability to adversarial examples, limiting their applicability. Although white-box adversarial attacks on IQA metrics have undergone extensive study, black-box attacks remain underexplored. These latter ones, however, can be more perilous than the former:

- They can target metrics that are nondifferentiable (e.g., VMAF by Netflix [8]) and those whose model is inaccessible to the attacker.
- The objective function optimized to generate adversarial perturbations can be nondifferentiable. Nondifferentiable input-data transformations can serve when training adversarial perturbations with unique properties; for instance, compression can aid in creating perturbations that are resistant to such distortions.
- Defenses against white-box attacks, such as adversarial training, may be ineffective against certain black-box attacks.

Analyzing metric robustness is essential for several reasons. First, competitions and benchmarks [9] often rely on certain metrics to evaluate performance. If these metrics are vulnerable to adversarial attacks, competitors may exploit them to manipulate the results and compromise benchmark integrity. Second, the use of an unstable metric in the loss function when training image- or video-processing models can yield undesirable outcomes. Models trained under such conditions can generate adversarial examples that include visual artifacts [10], diminishing the subjective quality of the final output.

Our paper focuses on the vulnerabilities of IQA and VQA to black-box adversarial attacks. Its main contributions are as follows:

- Adaptation of 10 black-box attacks to IQA and VQA metrics, broadening their applicability to these domains.
- New methods, based on universal adversarial perturbations (UAPs), that incorporate those black-box attacks, thereby addressing the primary limitation of such approaches: high computational complexity during inference. We believe ours is the first black-box UAP for attacking IQA/VQA.
- Robustness evaluation of 18 common metrics against our proposed adversarial attacks, revealing their critical vulnerabilities.

II. Related Work

Several papers have addressed white-box adversarial attacks targeting IQA models. Shumitskaya et al. [11], [12] introduced

a UAP attack that generates universal perturbations for IQA models. In [13], they proposed an approach that generates image-specific UAPs employing a network architecture based on U-Net. Korhonen et al. [5] presented a gradient-descent-based attack where gradients receive weights from a spatial-activity map created through Sobel filtering and morphological operations. Zhang et al. [6] incorporated an FR IQA metric as a perceptual constraint in the loss function to guide their attack.

Several black-box adversarial attacks have targeted IQA and VQA. For FR metrics, earlier studies primarily focused on the vulnerabilities of nondifferentiable examples such as VMAF. Zvezdakova et al. [14] and Siniukov et al. [15] demonstrated attacks on VMAF and its variant, VMAF NEG [16], respectively. They uncovered the susceptibility of these metrics to straightforward transformations, including Drago's tonemap, CLAHE, gamma correction, histogram equalization, unsharp masking, and others. For NR metrics, Ran et al. [17] introduced a black-box attack based on a bidirectional loss function combined with random-search optimization to generate adversarial examples. Similarly, Yang et al. [18] proposed an adaptive, iterative black-box attack on NR metrics with initial attack-direction guidance.

Additional related work discussion can be seen in the supplementary material.

III. Black-box Attacks on IQA and VQA Models

To evaluate the adversarial robustness of quality-assessment methods, we propose adapting score-based black-box adversarial attacks originally designed for image classifiers. Furthermore, we introduce methods for generating UAP attacks in a black-box framework. Shumitskaya et al. [11] effectively applied UAPs to no-reference (NR) metrics in a white-box setting. Our work extends the UAP approach to black-box scenarios, enabling broader applicability and addressing the challenges of limited access to model details.

A. Formal Definition

Suppose $X = [0,1]^{C\times H\times W}$ is an image, where C is the number of channels, and H and W are the image height and width, respectively. An IQA or VQA metric is a function that takes images or video frames as input and produces a scalar value representing the predicted visual quality. For NR IQA metrics, the function takes a single image as input and is defined as $f : X \to \mathbb{R}$. For full-reference (FR) IQA metrics, it takes two images as input—the reference image and the distorted image—and is defined as $f : X \times X \to \mathbb{R}$. In the case of VQA metrics, the input consists of several consecutive frames from a video sequence. Therefore, an NR VQA metric evaluates the quality of k frames and is defined as $f : X^k \to \mathbb{R}$.

We formalize an attack on an arbitrary model as the following optimization problem:

$$x^{adv} = \underset{x:\|x-x^{clean}\|_p\leq\varepsilon}{\arg\min} J(x,y),$$

where $J(x,y)$ represents a loss function designed to modify the original input image $x^{clean} \in X$ such that the model is unable to predict the value y originally assigned to x^{clean}. The goal is to generate an adversarial example x^{adv} that minimally deviates from x^{clean} (within a constraint $\|x - x^{clean}\|_p \leq \varepsilon$) while disrupting the model's prediction. For example, in the case of attacks on image classifiers, $J(x,y)$ could be the cross-entropy loss, which shifts the model's predicted output away from the true class label y. The type of norm $\|\cdot\|_p$ can vary by attack; most attacks employ $p = 0, 1, 2, \infty$.

This paper focuses on increasing metric scores, as this goal aligns with real-world attacks that seek to do so without improving visual quality. This situation often arises when attempting to achieve higher benchmark rankings. However, our approach remains generalizable, as it can decrease metric scores by simply reversing the sign of the optimized objective function. To attack an NR metric, we define the optimization objective as

$$J(x) = 1 - k_f \frac{f(x)}{\text{range}(f)},$$

where $\text{range}(f) = \sup_{x,y} |f(x) - f(y)| = k_f(\max_x f(x) - \min_x f(x))$ is the metric's output-value range. Here, $b_u = \max_x f(x), b_l = \min_x f(x)$ are the metric's upper and lower bounds, respectively. The definition of the parameter k_f is based on the metric: $k_f = -1$ if higher metric scores indicate better subjective quality and $k_f = 1$ otherwise.

B. Adapting Black-box Adversarial Attacks to IQA and VQA

We selected several prominent black-box adversarial attacks, initially developed for classification, and adapted them for quality assessment. They include the **Square Attack** [19], **Parsimonious** [20], methods from **Sparse-RS** [21], and **NES** [22]. In addition, we incorporated transformations used to attack VMAF [15], including contrast-limited adaptive histogram equalization (**CLAHE**), **gamma correction**, **gamma + unsharp**, and **Drago's tonemap**. Each one, however, had a modified objective function in our study. Instead of using the original loss functions designed for image classifiers, we employed the metric-specific function from Section III-A.

Square Attack utilizes a random search. Although the original version is designed for both l_2 and l_∞ norms, we selected l_∞ for better compatibility and comparability with other attacks. We replaced the original vertical-stripe-based initialization with a zero initialization. This adjustment was necessary because the former approach decreased the SSIM and PSNR scores for the attacked image without yielding any metric gain. To make the attack workable to images of any dimension, we optimized a patch of size 256×256. This patch is tiled to match the target-image size, ensuring uniform applicability regardless of resolution.

We left **Parsimonious** unaltered. This attack also trains a patch of size 256×256 but resizes it proportionally to cover the entire image, unlike Square Attack.

NES, notably, depends on the number of optimized pixels in a patch, affecting its efficiency. Training on large patches

leads to inefficient gradient estimation, rendering the process impractical under reasonable time constraints. To address this problem, we trained on 32×32 patches and tiled them to match the full image size. The original algorithm incorporates a mechanism to easily vary the learning rate η and perturbation bound ε, but we removed that mechanism because it worsened the attacked image's metric score.

Our investigation used three Sparse-RS variations: l_0**-RS**, **Patch-RS**, and **Frame-RS**. These attacks remained unchanged from their original form except for the optimization function.

Finally, we adapted several image-processing methods—**CLAHE**, **gamma correction**, **gamma + unsharp**, and **Drago's tonemap**—to target metrics beyond VMAF and its variations. A genetic algorithm allowed us to optimize the parameters for these transformations and enhance the target-metric values.

Figure 1 shows examples of these adversarial attacks on PaQ-2-PiQ; note their effect on the image.

C. Black-box UAP Attack

Creating effective adversarial examples using black-box attacks requires numerous queries, rendering the process computationally expensive. To address this challenge, we propose training UAPs that apply to images in real time to artificially increase metric values. A universal perturbation can be formulated as the following optimization problem:

$$P = \underset{P:\|P\|_\infty \leq \varepsilon}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} J(\mathrm{Clip}(x_i + P, 0, 1)),$$

where n is the number of images in the training dataset.

The following outline describes the process for generating a UAP in a black-box setting:

1) Initialize the UAP following the same procedure the algorithm employs when attacking a single image.
2) Modify the target function in the base algorithm to evaluate its performance over the entire training dataset.
3) Run the algorithm for a predefined number of epochs.

We ran each algorithm for 10,000 epochs, corresponding to the number of queries in the original black-box attacks. Papers like [17], [19] have also used that number of queries when attacking a single image.

Our training dataset included 50 images from the training subset of Microsoft COCO [23]. Using the k-means algorithm, we clustered the images into 50 groups on the basis of their spatial information (SI). The images corresponding to the centers of these clusters were the representative samples.

Parsimonious, **NES**, and **Square Attack** were chosen as base attack methods due to their efficiency for IQA task described later in Section IV-B. Additionally, we adapted the UAP attack on classifiers, based on differential evolution (**DE**), that was introduced by Wang et al [24]. Training used black-box algorithms constrained by an l_∞ norm to ensure the trained UAPs could handle images of any size through tiling. Perturbations generated by algorithms with an l_0-norm constraint undergo resizing without norm preservation, making them inadequate. We trained our UAP attacks using $\varepsilon = 0.1$, consistent with prior work that trained white-box UAPs [11].

IV. Experiments

A. Experimental Setup

To evaluate non-UAP adversarial attacks, we selected 50 images from KonIQ-10k [25] with a resolution of 512×384 and AGIQA-3K [26] with a resolution of 512×512. The dataset size is deliberately small because black-box attacks are computationally complex. We divided the original datasets into 10 clusters using spatial information (SI), colorfulness (CF), and mean opinion score (MOS) as criteria. We selected 5 random images from each cluster, yielding a subsample of 50 images. These steps ensured the chosen datasets are representative despite their size.

UAP attacks can operate faster than non-UAP attacks, allowing their evaluation on additional datasets. To facilitate a comparison of these two attack types, we used the same subsample from KonIQ-10k. Further evaluation involved two additional datasets: NIPS 2017 [27](a widely used dataset for testing adversarial attacks) with a resolution of 299×299, and Vimeo-90k [28](a dataset containing small video sequences of three frames each) with a resolution of 448×256.

We evaluated the robustness of IQA and VQA metrics using 10 attacks defined in Section III-B. Details regarding all adapted attacks and their parameters appear in the supplementary materials. We chose the attacks to represent diverse black-box approaches, including random search (**Square Attack** and **Sparse-RS**), discrete optimization (**Parsimonious**), gradient estimation (**NES**), image processing (**CLAHE**, **gamma correction**, **gamma + unsharp**, and **Drago's tonemap**), sparse attacks (**l0-RS**, **Patch-RS**, and **Frame-RS**), and evolutionary approaches (**DE**). Our comparison also included the adversarial attack developed by **Ran et al.** [17], which is specifically designed to target IQA metrics and primarily utilizes random-search optimization.

We evaluated the robustness of 18 metrics against black-box adversarial attacks. Our choice of these metrics was based on their popularity, strong correlation with subjective scores, and robustness to white-box adversarial attacks as determined by Antsiferova et al. [7].

To assess metric robustness to our proposed adversarial attacks, we computed the normalized gain. This score represents the average metric change due to the attack, normalized by the metric's estimated range:

$$Norm.gain = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i^{adv}) - f(x_i^{clean})}{\mathrm{range}(f)},$$

where n denotes the number of images in the test dataset. This expression quantifies the average metric gain induced by the adversarial attack as a percentage of the metric's range. Lower normalized gains indicate greater robustness to the attack.

Detailed information on the chosen metrics, their metric ranges, attacks and their parameters is presented in the supplementary materials.

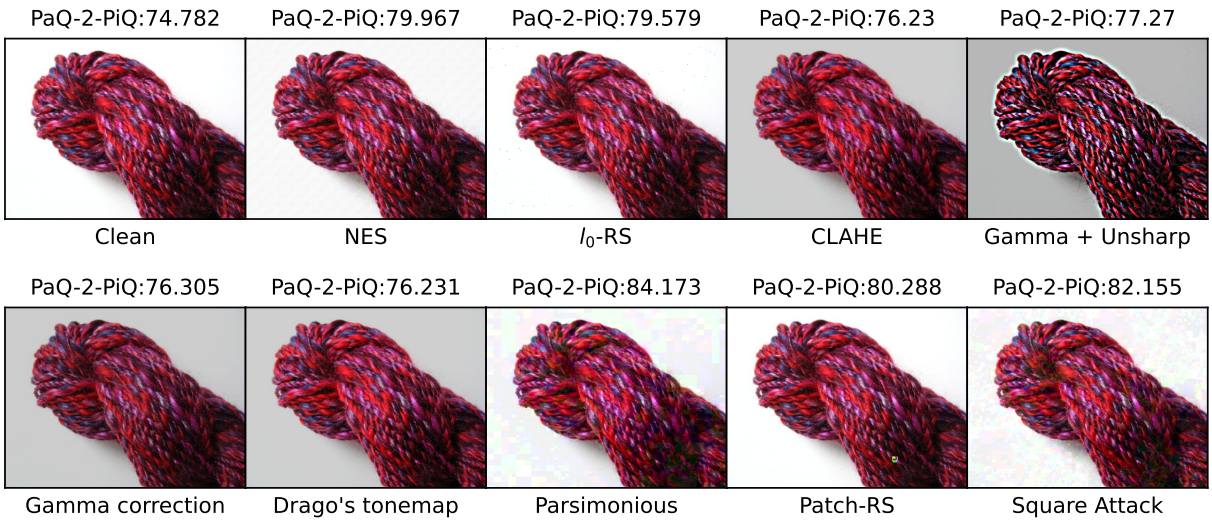

Fig. 1. Adversarial examples generated using adapted black-box adversarial attacks on PaQ-2-PiQ, applied to an image from KonIQ-10k.

B. Evaluation of Adapted Black-box Attacks

We fine-tuned the parameters of each attack for PaQ-2-PiQ and applied them to other metrics to reduce the computational complexity of the parameter selection.

Table I presents the normalized gain estimates for the NR metrics on the KonIQ-10k subsample. The results for the AGIQA-3K subsample are in the supplementary materials. The results demonstrate that all the NR metrics are susceptible to the adversarial attacks we considered. Among these attacks, **Ran et al.** achieves the highest performance, making it the most effective option for NR metrics, but it delivers lesser gains on AGIQA-3K than does **NES**, which was not designed to attack IQA/VQA metrics. Both **Square Attack** and **Parsimonious** consistently rank among the top three in normalized gain; they exhibit similar results and therefore can be considered equally effective. The experiments demonstrate that attacks based on gradient estimation generally achieve higher normalized gain, followed by the attacks based on random search and combinatorial optimization. Unrestricted attacks, originally designed to target VMAF, generally yield modest results; the exception is **Drago's tonemap**, which performs notably better. The experiment shows that most NR metrics are robust to **CLAHE**, **gamma correction**, and **gamma + unsharp**. However, most NR metrics are highly susceptible to sparse attacks, such as l_0**-RS**, **Patch-RS**, and **Frame-RS**. These attacks demonstrate the vulnerability of IQA/VQA metrics to adversarial patches, pixel replacement, and frame perturbations. Among the metrics, MDTVSFA stands out as the most robust to our proposed attacks, followed by MANIQA. They are similarly robust thanks to the use of several training datasets and a ViT backbone. Conversely, DB-CNN and PaQ-2-PiQ offer the highest normalized-gain estimates, making them particularly vulnerable.

The results for FR metrics, detailed in the supplementary materials, reveal that these metrics are more resistant to adversarial attacks than NR metrics are. Certain examples, such as traditional MS-SSIM and the No Enhancement Gain (NEG) version of VMAF, are robust against all our proposed attacks. Several FR models remain vulnerable, however, with ASNA-MACS and PieAPP standing out as the most unstable. Special methods from [15] also successfully attacked VMAF. Attacks with an l_∞-norm bound perform poorly because they rely heavily on adding perturbations with the maximum possible norm. Among the attacks we evaluated, **Square Attack** proves to be the most effective against FR metrics. By contrast, sparse attacks—including l_0**-RS**, **Patch-RS**, and **Frame-RS**—are largely ineffective owing to replacement of image pixels. **Drago's tonemap** and other unrestricted attacks perform particularly well when targeting VMAF, for which they were designed. Moreover, they also expose vulnerabilities in other FR metrics (PieAPP, for instance), similar to those in VMAF.

We also evaluated the computational complexity of the adapted black-box adversarial attacks on our KonIQ-10k subset. Table II presents the attack-speed performance measured in frames per second (FPS); we calculated it by dividing the total execution time of an attack by the number of images in the attacked dataset. The results highlight that adapted black-box adversarial attacks are more computationally intensive than white-box alternatives. Image-processing-based methods (**CLAHE**, **gamma correction**, **gamma + unsharp** and **Drago's tonemap**) exhibit relatively high FPS values. This efficiency is because their optimization problems have lower dimensionality, enabling faster convergence.

C. Evaluation of UAP Black-box Attacks

We compared UAPs trained on the dataset described in Section III-C with their corresponding non-UAP attacks, both constrained by the same bound l_∞ ($\varepsilon = 0.1$). Table III presents the normalized gains achieved by our proposed UAP attacks, evaluated using the NR metric on a subsample of KonIQ-10k. The results show that UAP attacks are generally less effective than their non-UAP counterparts. Notably, certain metrics such as MANIQA and TOPIQ-NR are robust against

TABLE I
THE ROBUSTNESS OF NR METRICS TO THE ADAPTED BLACK-BOX ADVERSARIAL ATTACKS IN TERMS OF THE NORMALIZED GAIN ESTIMATED ON THE SUBSAMPLE OF KONIQ-10K.

	Koncept512	Linearity	PaQ-2-PiQ	VSFA	MANIQA	MDTVSFA	DB-CNN	UNIQUE	TOPIQ-NR
Frame-RS	0.034	0.119	0.199	0.154	0.083	0.08	0.23	0.163	0.105
NES	**0.49**	0.372	0.304	0.203	0.132	0.105	0.46	**0.326**	0.353
l_0-RS	0.337	0.255	0.323	0.141	0.199	0.082	0.34	0.178	0.172
CLAHE	0.07	0.02	0.13	0.024	0.036	0.008	0.066	0.024	0.016
Gamma correction	0.028	0.035	0.011	0.008	0.024	0.007	0.054	0.052	0.049
Gamma + Unsharp	0.087	0.058	0.145	0.027	0.061	0.016	0.103	0.05	0.077
Drago's tonemap	0.2	0.135	0.185	0.067	0.089	0.037	0.178	0.085	0.13
Parsimonious	0.311	0.288	0.425	**0.329**	**0.233**	**0.131**	0.428	0.264	0.254
Patch-RS	0.376	0.184	0.295	0.08	0.077	0.052	0.304	0.142	0.128
Square Attack	0.365	0.27	0.441	0.319	0.229	0.127	0.436	0.263	0.231
Ran et al.	0.381	**0.419**	**0.509**	-0.037	-0.05	-0.007	**0.614**	0.321	**0.386**

TABLE II
THE FPS EVALUATION RESULTS FOR NON-UAP BLACK-BOX ADVERSARIAL ATTACKS AND I-FGSM CONDUCTED ON THE SUBSAMPLE OF KONIQ-10K. THE ATTACKS TARGETED PAQ-2-PIQ.

Method	Frame-RS	NES	l_0-RS	CLAHE	Gamma correction	Gamma + Unsharp	Drago's tonemap	Parsimonious	Patch-RS	Square Attack	Ran et al.	I-FGSM (white-box)	Black-box UAP (proposed)
Computational complexity (FPS)	0.026	0.031	0.014	0.305	0.294	0.216	0.227	0.024	0.029	0.027	0.027	5.020	$\approx$ **240**

our UAP methods. Metrics such as PaQ-2-PiQ, Linearity, DB-CNN, and UNIQUE, however, are susceptible, with gains being apparent for at least one UAP method. VQA metrics such as VSFA and MDTVSFA are vulnerable to all the algorithms we tested, though the estimated normalized gains remain modest. Notably, metrics that are robust to iterative attacks are also robust to UAP attacks. Among the various methods, perturbations generated using **Square Attack** are the most effective. This attack achieves the highest normalized gains for four metrics and delivers near-optimal results for the remainder. Additionally, it is the only UAP attack that can compromise MANIQA and TOPIQ-NR, although the gains in these cases are minor. By contrast, the **Parsimonious** UAP attack demonstrates the weakest performance, likely because it relies on differing assumptions about the optimized function (such as submodularity), potentially limiting its effectiveness on the metrics we considered.

Since the perturbations add directly to the original images with a fixed coefficient, UAP attacks are computationally efficient, achieving more than 200 FPS on KonIQ-10k images and thus enabling their application to larger datasets. The table in the supplementary materials summarizes the normalized gains of our proposed UAP methods, evaluated on the NR metrics applied to Vimeo-90k and NIPS 2017. Normalized-gain estimates for Vimeo-90k are slightly higher than those for KonIQ-10k, suggesting the former is more susceptible to adversarial attacks. This greater susceptibility may stem from Vimeo-90k's composition of video frames, which are more likely to include images with motion blur. As Zhang et al. [6] noted, blurred and low-quality images are inherently more prone to adversarial attacks. Among the attacks we tested, **Square Attack UAP** again proves to be the most effective for this dataset, although it fails to achieve positive gains on all metrics. By contrast, the attacks are less effective on NIPS 2017. This diminished efficiency may result from the properties of the dataset, which we curated to test adversarial attacks on image classifiers. Nevertheless, **Square Attack UAP** and **NES UAP** remain the most effective techniques. For all the datasets we evaluated, the NR metrics demonstrate consistent robustness. But we excluded FR metrics from additional dataset evaluations because UAP attacks did poorly on them when using KonIQ-10k.

The supplementary materials contain additional experimental results including visualizations of UAPs trained for various NR metrics, detailed analysis of UAP scaling, impact of adversarial attacks on subjective quality and defense performance.

V. LIMITATIONS AND SOCIETAL IMPACT

Our black-box adversarial attacks for IQA and VQA metrics contributed to our analysis of these metrics' adversarial robustness. But this paper has the following limitations and potential negative societal impacts:

- The current approach to selecting attack parameters lacks individual-tuning capability for each metric, owing to the computational complexity. As a result, optimum parameter configurations tailored to each metric could increase the attack effectiveness.
- Our proposed methods can attack IQA or VQA metrics in benchmarks and comparisons, potentially compromising their integrity. Such attacks could lead to misleading and thus unreliable leaderboards. Therefore, it is essential to check submitted results for signs of adversarial manipulation and to prioritize the development of robust means of adversarial detection and defense.

VI. CONCLUSION

This study assesses the robustness of neural-network-based IQA and VQA metrics against various black-box adversarial attacks. We adapted 10 black-box attacks originally designed

TABLE III
COMPARISON OF THE PROPOSED UAP ATTACKS WITH THE CORRESPONDING NON-UAP ATTACKS IN TERMS OF THE NORMALIZED GAIN FOR NR METRICS ON THE SUBSAMPLE OF KONIQ-10K.

	Koncept512	Linearity	PaQ-2-PiQ	VSFA	MANIQA	MDTVSFA	DB-CNN	UNIQUE	TOPIQ-NR
DE UAP	-0.05	**0.132**	0.263	0.002	-0.082	**0.024**	0.186	**0.112**	-0.014
NES UAP	**0.096**	-0.001	**0.55**	0.012	-0.012	0.007	0.057	0.074	-0.002
Parsimonious UAP	0.042	-0.017	0.269	0.01	-0.012	0.013	0.119	-0.015	-0.006
Square Attack UAP	0.053	0.112	0.364	**0.045**	**0.016**	0.02	**0.264**	0.022	**0.017**
Ran et al. UAP	-0.095	-0.013	0.206	-0.066	-0.056	-0.027	0.11	-0.076	-0.061
NES	0.596	0.395	0.362	0.212	0.157	0.11	0.483	0.334	0.38
Parsimonious	0.387	0.385	0.564	0.358	0.278	0.138	0.496	0.275	0.3
Square Attack	0.429	0.358	0.609	0.35	0.326	0.127	0.466	0.266	0.251
Ran et al.	0.467	0.508	0.681	-0.041	-0.097	-0.011	0.629	0.34	0.494

for image classifiers to target these metrics, and we introduced 5 UAP attacks to address the primary limitation of black-box approaches: computational complexity. Our study evaluated a total of 18 metrics against our proposed attacks. Although some metrics displayed substantial robustness, others were highly susceptible, exposing critical weaknesses in state-of-the-art models. FR metrics were more resistant to adversarial attacks than NR metrics, though certain attacks still achieved gains against certain FR metrics. Overall, this study highlights the pressing need for further research to develop more-robust and more-resilient IQA and VQA metrics. Ensuring their reliability and accuracy in practical applications is vital, particularly in the presence of adversarial manipulations.

ACKNOWLEDGMENT

The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

REFERENCES

[1] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli, "Evasion attacks against machine learning at test time," in *ECML PKDD 2013*.

[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *ICLR*, 2015.

[4] Shixiang Gu and Luca Rigazio, "Towards deep neural network architectures robust to adversarial examples," *ICLR workshop*, 2015.

[5] Jari Korhonen and Junyong You, "Adversarial attacks against blind image quality assessment models," in *2nd Workshop on Quality of Experience in Visual Multimedia Applications*, 2022, pp. 3–11.

[6] Weixia Zhang, Dingquan Li, Xiongkuo Min, Guangtao Zhai, Guodong Guo, Xiaokang Yang, and Kede Ma, "Perceptual attacks of no-reference image quality models with human-in-the-loop," *NeurIPS*, 2022.

[7] Anastasia Antsiferova, Khaled Abud, Aleksandr Gushchin, Ekaterina Shumitskaya, Sergey Lavrushkin, and Dmitriy Vatolin, "Comparing the robustness of modern no-reference image-and video-quality metrics to adversarial attacks," in *Proceedings of AAAI*, 2024, vol. 38, pp. 700–708.

[8] "Vmaf - video multi-method assessment fusion," https://github.com/Netflix/vmaf.

[9] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al., "Vbench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21807–21818.

[10] Egor Kashkarov, Egor Chistov, Ivan Molodetskikh, and Dmitriy Vatolin, "Can no-reference quality-assessment methods serve as perceptual losses for super-resolution?," *arXiv preprint arXiv:2405.20392*, 2024.

[11] Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy S Vatolin, "Universal perturbation attack on differentiable no-reference image- and video-quality metrics," in *BMVC*, 2022.

[12] Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy Vatolin, "Towards adversarial robustness verification of no-reference image-and video-quality metrics," *Computer Vision and Image Understanding*, vol. 240, pp. 103913, 2024.

[13] Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy S. Vatolin, "Fast adversarial cnn-based perturbation attack on no-reference image- and video-quality metrics," in *Tiny Papers @ ICLR*, 2023.

[14] A Zvezdakova, S Zvezdakov, D Kulikov, and D Vatolin, "Hacking vmaf with video color and contrast distortion," in *CEUR Workshop Proceedings*, 2019, pp. 53–57.

[15] Maksim Siniukov, Anastasia Antsiferova, Dmitriy Kulikov, and Dmitriy Vatolin, "Hacking vmaf and vmaf neg: vulnerability to different preprocessing methods," in *Artificial Intelligence and Cloud Computing Conference*, 2021, pp. 89–96.

[16] Zhi Li, "On vmaf's property in the presence of image enhancement operations," 2020.

[17] Yu Ran, Ao-Xiang Zhang, Mingjie Li, Weixuan Tang, and Yuan-Gen Wang, "Black-box adversarial attacks against image quality assessment models," *Expert Systems with Applications*, vol. 260, pp. 125415, 2025.

[18] Chenxi Yang, Yujia Liu, Dingquan Li, and Tingting Jiang, "Exploring vulnerabilities of no-reference image quality assessment models: A query-based black-box method," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[19] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein, "Square attack: a query-efficient black-box adversarial attack via random search," in *ECCV*, 2020, pp. 484–501.

[20] Seungyong Moon, Gaon An, and Hyun Oh Song, "Parsimonious black-box adversarial attacks via efficient combinatorial optimization," in *ICML*, 2019, pp. 4636–4645.

[21] Francesco Croce, Maksym Andriushchenko, Naman D Singh, Nicolas Flammarion, and Matthias Hein, "Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks," in *Proceedings of AAAI*, 2022, vol. 36, pp. 6437–6445.

[22] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin, "Black-box adversarial attacks with limited queries and information," in *ICML*, 2018, pp. 2137–2146.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[24] Sivy Wang, Yucheng Shi, and Yahong Han, "Universal perturbation generation for black-box attack using evolutionary algorithms," in *24th International Conference on Pattern Recognition*, 2018, pp. 1277–1282.

[25] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[26] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[27] "Nips 2017: Adversarial learning development set," www.kaggle.com/datasets/google-brain/nips-2017-adversarial-learning-development-set.

[28] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.