IOI: Invisible One-Iteration Adversarial Attack on No-Reference Image- and Video-Quality Metrics

Ekaterina Shumitskaya [1 2 3] **Anastasia Antsiferova** [2 1] **Dmitriy Vatolin** [1 2 3]

Abstract

No-reference image- and video-quality metrics are widely used in video processing benchmarks. The robustness of learning-based metrics under video attacks has not been widely studied. In addition to having success, attacks on metrics that can be employed in video processing benchmarks must be fast and imperceptible. This paper introduces an Invisible One-Iteration (IOI) adversarial attack on no-reference image and video quality metrics. The proposed method uses two modules to ensure high visual quality and temporal stability of adversarial videos and runs for one iteration, which makes it fast. We compared our method alongside eight prior approaches using image and video datasets via objective and subjective tests. Our method exhibited superior visual quality across various attacked metric architectures while maintaining comparable attack success and speed. We made the code available on GitHub: https://github.com/katiashh/ioi-attack.

1. Introduction

No-reference (NR) image- and video-quality assessment poses a significant challenge in computer vision. In contrast to full-reference (FR) quality metrics, NR metrics do not estimate the similarity of a distorted image or video to the original one but evaluate its visual appeal. The rapid integration of deep-learning-based NR image-and video-quality assessment metrics (Ying et al., 2020; Talebi & Milanfar, 2018; Su et al., 2020; Golestaneh et al., 2022) led to the importance of investigating their vulnerabilities to transformations of input. One of the most common types of input transformations is adversarial attacks. In the case of image- and video-quality metrics, adversarial attacks are modifications of input images or videos that change the predicted quality score without significant influence on perceptual quality. Several studies (Zhang et al., 2022a; Korhonen & You, 2022a; Sang et al., 2022; Shumitskaya et al., 2022; 2023b; Meftah et al., 2023; Yang et al., 2024) unveiled vulnerabilities in NR image quality metrics when exposed to adversarial attacks.

NR metrics are employed in various image- and video-processing benchmarks, such as super-resolution (PIRM-SR, 2019; Khrulkov & Babenko, 2021; Ma et al., 2017), video generation (VideoGeneration, 2024), video compression (Ghadiyaram et al., 2017). For some tasks, NR metrics show even better performance than FR metrics; for example, video super-resolution (SR-MSU, 2023) or video compression by new encoding standards such as H.266/VVC (Antsiferova et al., 2022). The developers of video-processing algorithms can integrate adversarial attacks on quality metrics into their methods to achieve higher positions in public benchmarks. Today, such cheating can be detected in benchmarks that publish subjective comparisons along with objective ones. For example, in MSU Codec Comparison 2021 (CC-MSU, 2021), the leaderboard by a learning-based metric VMAF, which was shown to be vulnerable to attacks (Siniukov et al., 2022; Zvezdakova et al., 2019), differs from a subjective one. Algorithms in these benchmarks compete in both visual quality and speed, and a high speed is essential for some real-life applications like universal encoding for video compression (from one to ten frames per second). Thus, to cheat in video processing benchmarks, it is profitable for an attacker to inject an imperceptible perturbation into the video without significantly decreasing the method speed.

In the literature, most of the existing approaches evaluate NR metrics robustness in the image domain. However, to evaluate the robustness of NR metrics for videos, an adversary has to satisfy several essential conditions, making such a task more challenging:

1. Quantitative success of an attack. For NR metrics, the success of an attack is measured by the amount of the metric's score change. Both decreasing and increasing

[1]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia [2]MSU Institute for Artificial Intelligence Moscow, Russia [3]Lomonosov Moscow State University, Moscow, Russia. Correspondence to: Ekaterina Shumitskaya <ekaterina.shumitskaya@graphics.cs.msu.ru>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

a metric's score can be considered an attack; however, making a score higher has more practical applications.

2. High speed of an attack. The attack must operate at high speed for practical viability. A slow attack holds limited practical significance, as its integration into video processing algorithms will greatly slow down the method and lower its position in benchmarks.

3. Temporal consistency of a transformed video. The per-frame implementation of adversarial attacks designed for images in videos leads to noticeable flickering effects that look suspicious in subjective comparisons.

Our research aims to investigate the potential of injecting fast, invisible and temporally consistent adversarial attacks on NR quality metrics. This paper introduces the Invisible One-Iteration (IOI) adversarial attack for images and videos. To achieve high attack speed, our method yields perturbation by calculating the gradient of an attacked model using one access to the model. We further show that a one-iteration attack for each frame is more efficient than a many-iteration attack applied to only some frames. To keep the temporal stability of a perturbed video, we make our attack invisible to the human eye by using weighting and frequency modules. The proposed attack operates in a white-box scenario, which does not limit its applicability for benchmarks: usually, a comparison methodology is known, and quality metrics are published for reproducibility. The primary contributions of this work can be summarized as follows:

- We propose an Invisible One-Iteration (IOI) adversarial attack that increases NR image- and video-quality metrics scores. It produces perturbations that are imperceptible and temporally stable in videos. The attack is fast: it does not require convergence and works efficiently with one iteration.
- We propose a methodology for comparing adversarial attacks at NR metrics. It is based on aligning attack speed and relative metrics increase after the attack, yielding to comparing only objective and subjective quality of adversarial videos.
- We conducted comprehensive experiments using two datasets and three NR models. Four quality metrics were used to demonstrate that the proposed attack generates adversarial images and videos of superior quality compared to prior methods.
- We conducted a subjective study on the proposed method's perceptual quality and temporal stability. A crowd-sourced subjective comparison with 685 subjects showed that the proposed attack produces adversarial videos of better visual quality than previous methods.

Our code is publicly available at https://github.com/katiashh/ioi-attack.

2. Related Work

Adversarial attacks usually have constraints on perturbations and minimize the L_∞ norm between adversarial examples and their original inputs. Other L_p norms (Su et al., 2019; Szegedy et al., 2013) are less common due to a higher computational complexity. We further consider methods that use only L_∞ constraint to work faster on high-dimensional data like videos. Among such kinds of attacks, Goodfellow et al. proposed the **FGSM** method (Goodfellow et al., 2015) that generates adversarial examples by leveraging gradients from the targeted model. Recently, **UAP** (Shumitskaya et al., 2022) and **FACPA** (Shumitskaya et al., 2023b) attacks on NR image quality metrics have been proposed. These methods also used L_∞ norm constraints at the training stage.

Numerous studies showed that L_p norms are not suitable as a distance metric to evaluate perceptual image quality (Sharif et al., 2018; Fezza et al., 2019; Wang et al., 2004; Johnson et al., 2016; Isola et al., 2017). Perturbations in images generated under L_p norm constraints often result in noisy pixels within smooth areas of the original image, which is easily perceptible to the human eye. Several adversarial attacks prioritize the visual quality of the generated adversarial images and use more sophisticated restrictions on perturbations than the bare implementation of L_p constraints.

Zhang et al. (2020) introduces a novel approach **AdvJND** by incorporating just noticeable difference (JND) (Yang et al., 2005) coefficients into the L_∞ norm constraint during adversarial example generation. These coefficients account for the human eye's ability to perceive the threshold of changes in an image. The authors employed I-FGSM and FGSM algorithms as baselines. To enhance the visual fidelity of adversarial images, they amplified original perturbations obtained from the FGSM or I-FGSM methods by scaled JND coefficients.

SSAH (Luo et al., 2022) adversarial attack targets two objectives: semantic similarity of images and low-frequency constraint. The first component is crafted with a focus on image classification tasks. The second component adds perturbations within high-frequency regions by minimizing the difference between low-frequency information of adversarial and original images. Usually, it requires many iterations to converge and produce good visual quality.

Korhonen et al. (2022b) introduced an iterative attack method targeting NR quality metrics, employing a Sobel filter to hide distortions within textured regions. In each iteration, the model's gradient under attack concerning the input is multiplied by a spatial activity map derived from

the original image via the Sobel filter. This map highlights areas with substantial texture, enhancing the visual fidelity of the resulting adversarial images.

Zhang et al. (2022a) introduced an iterative adversarial image crafting approach leveraging various FR metrics like Chebyshev distance, SSIM, LPIPS, and DISTS to manage visual distortions. Their method involved the iterative minimization of a loss function consisting of two components: the attacked loss and a loss based on some differential FR image quality metric.

Karli et al. (2021) introduced the **Normalized Variance Weighting (NVW)** method aimed at amplifying perturbations within high-variance regions of images. This technique can be utilized alongside gradient attacks like FGSM or I-FGSM. Additionally, the authors proposed the LPIPS-minimization method, aiming to enhance perceptual quality by minimizing the LPIPS distance between the original and adversarial images while ensuring the classifier remains deceived. However, this minimization method is exclusively effective for discrete tasks, such as classification or detection, while applying it to quality metrics will lead to eliminating relative gain.

Table 1 summarizes existing attacks on image- and video-quality metrics. The primary drawback of the prior methods is their requirement to run the attack via many iterations with small steps. This leads to low attack speed, particularly on high-resolution video data. They can be used efficiently to attack videos only when applied to each k-th frame, which, as we further show, reduces relative gain and temporal consistency.

Table 1. Comparison of existing adversarial attacks on image- and video-quality metrics regarding visual quality regulation features and the requirement to converge for an attack to succeed.

Method	Visual quality		Speed
	Weights	Freq. reg.	Needn't converg.
FGSM (2015)	×	×	√
UAP (2022)	×	×	√
FACPA (2023b)	×	×	√
AdvJND (2020)	JND map	×	√
SSAH (2022)	×	DWT	×
Korhonen et al. (2022b)	Sobel map	×	√
Zhang et al. (2022a)	×	×	×
NVW (2021)	Local STD	×	√
IOI (proposed)	Local STD-based	FFT	√

3. Proposed method

3.1. Problem formulation

The adversarial attack on the NR quality metric M is usually modelled as follows:

$$\underset{I^a}{\operatorname{argmax}}\{M(I^a) - M(I)\}, \|I^a - I\|_p \leq \epsilon, \quad (1)$$

where I is a clear video frame, I^a is an attacked frame, ϵ is a small constant.

To ensure a high visual quality of the perturbed image, instead of using l_p norms that are inefficient for this task, we formulate the problem as follows:

$$\begin{gathered}\underset{I^a}{\operatorname{argmax}}\{M(I^a) - M(I)\}, \\ \frac{L_f(I^a)\cdot L_f(I)}{||L_f(I^a)||_2||L_f(I)||_2} \geq 1 - \epsilon^*,\end{gathered} \quad (2)$$

where L_f is a low-frequency filter, ϵ^* is a small constant. When low-frequency components of the clear and adversarial images are closely aligned, the perturbations mainly affect high-frequency areas. As a result, distortions in adversarial images are nearly invisible to the human eye, according to the contrast masking theory of human vision (Legge & Foley, 1980). For measuring the difference between attacked and original images/frames during a perturbation construction, we use adversarial mean absolute error in the frequency domain (MAE^*), which is calculated as follows:

$$MAE^*(I^a, I) = \frac{1}{HW}\sum_{i=0}^{(H-1)}\sum_{j=0}^{(W-1)}|I^{a*}_{ij} - I^*_{ij}|, \quad (3)$$

where I^{a*}_{ij} and I^*_{ij} are Fast Fourier Transform (FFT) coefficients of attacked and original images correspondingly, H and W – image dimensions.

As discussed in the introduction, we focus on the scenario where metric scores increase after an attack. In some other studies, NR metrics are modified in two directions, and the attack success is measured by a decrease in the metric's correlation with subjective quality (Zhang et al., 2022b; 2023). While this approach holds theoretical significance in evaluating the stability of metric scores, it carries certain limitations. Only increasing the target metric score will keep the same correlation with subjective quality; thus, an attack remains undetected. Instead of using correlation as a measure of NR metrics' adversarial robustness, we evaluate attack success using relative gain (RG):

$$RG = \frac{M(I^a) - M(I)}{M_{range}} \quad (4)$$

where M_{range} is the range of scores produced by M.

3.2. One-iteration attack

Figure 1 provides an overview of our method. Initially, the proposed attack perturbs the image using a baseline gradient attack. Subsequently, it processes the perturbed image using two modules to enhance the visual quality of an adversarial image/video: the frequency module and the weighting module. For the baseline gradient attack, we employ FGSM (Goodfellow et al., 2015):

$$I^p = I + \epsilon * sign(\nabla_I M(I)) \tag{5}$$

3.3. Frequency module

The frequency module extracts features from the original I and perturbed I^p images for further processing. It decomposes them into low-frequency (LF) and high-frequency (HF) components using the Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT). The LF component keeps fundamental content information, while the HF component contains noise and texture details. A threshold tr^* for dividing frequencies in f% highest and (1-f)% the lowest coefficients is used to calculate indexes $d_f(I)$ of the highest FFT coefficient of the original image I (Equation 6).

$$\begin{gathered} tr^* = \underset{tr}{\operatorname{argmin}} |f - \frac{\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\mathbb{1}[I^*_{u,v} > tr]}{HW}| \\ d_f(I) = \{(u,v) : I^*_{u,v} > tr^*\} \end{gathered} \tag{6}$$

We use $d_f(I)$ to extract LF and HF components from the original image, as shown in Equation 7. These components are further utilized in the weighting module to enhance the visual quality of adversarial video frames. HF components of the perturbed image I^p are extracted using indexes $d_f(I)$ from the original image. Thus, the perturbations will be transferred into HF areas of the original image (Equation 8). The parameter f allows fine-tuning the visibility of perturbations and relative attack gain at different levels. Lower values of f result in more visible perturbations within the frequency module and higher relative gain.

$$\begin{aligned} L^{d_f(I)}_{u,v}(I) &= \begin{cases} I^*_{u,v} & \text{if } (u,v) \in d_f(I) \\ 0 & \text{otherwise} \end{cases} \\ H^{d_f(I)}_{u,v}(I) &= \begin{cases} I^*_{u,v} & \text{if } (u,v) \notin d_f(I) \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{7}$$

$$H^{d_f(I)}_{u,v}(I^p) = \begin{cases} (I^p)^*_{u,v} & \text{if } (u,v) \notin d_f(I) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

3.4. Weighting module

Various image processing applications (Lin et al., 2005; Liu et al., 2010) operate under the assumption that distortions in low-variance regions are more visible to the human eye than in high-variance areas. The weighting module generates a weighting map for the input image based on the variance map. A similar approach was used in previous studies (Croce & Hein, 2019; Karli et al., 2021) for generating adversarial examples. We introduce additional features to enhance the weights obtained from the variance map method. The proposed weights map generation method is described in Equation 9. Initially, we compute local variance and local mean maps for an image x, determining standard deviations and means for both axes using a window size of 3 for each colour channel. Next, we derive the relative local variance of the image by dividing the local variance map by the local mean map and normalizing the weights to the range [0, 1]. In the next step, we zero out 1% and compute the square root of the weights.

$$\begin{gathered} \sigma_{i,j} = \sqrt{\frac{\sum x^2_{i,j}}{n} - (\frac{\sum x_{i,j}}{n})^2}, m_{i,j} = \frac{\sum x_{i,j}}{n} \\ \gamma = \frac{\sigma}{m},\ \gamma_{max} = \max_{i,j}(\gamma_{i,j}),\ \gamma_{norm} = \frac{\gamma}{\gamma_{max}} \\ w_{i,j} = \begin{cases} \sqrt{(\gamma_{norm})_{i,j}}\text{ , if } (\gamma_{norm})_{i,j} \geq 0.01 \\ 0\text{ , if } (\gamma_{norm})_{i,j} < 0.01 \end{cases} \end{gathered} \tag{9}$$

Figure 2 illustrates the weights map derived by the proposed method and three prior methods (Korhonen et al. (2022a), AdvJND (Zhang et al., 2020), and NVW (Karli et al., 2021)). NVW and AdvJND methods assign non-zero weights for a noisy background, often resulting in visible distortions within smooth regions. The image area covered by non-zero weights in Korhonen et al.'s map is small, potentially limiting the strength of the attack it can generate.

We use the proposed weights map to guide the weighting of adversarial image HF components. As a result, the ultimate perturbation remains absent in smooth areas of an image. The construction of the final adversarial image involves the composition of three elements: the original LF component, the perturbed HF component multiplied by the weights map, and the original HF component multiplied by inverse weights:

$$I^a = L^{d_f(I)}_c(I) + wH^{d_f(I)}_c(I^p) + (1-w)H^{d_f(I)}_c(I) \tag{10}$$

3.5. Mathematical properties

This section provides theoretical restrictions of the generated adversarial images or video frames by the proposed method.

Theorem 1. *Let I and I^p be original and perturbed image correspondingly, I^a – adversarial image after IOI attack that is based on I^p with truncating parameter f. Then inequality 11 is correct, where $MAE^*(\cdot,\cdot)$ is given by Equation 3.*

$$||I^a - I||_\infty \leq (1-f)MAE^*(I^p, I) \tag{11}$$

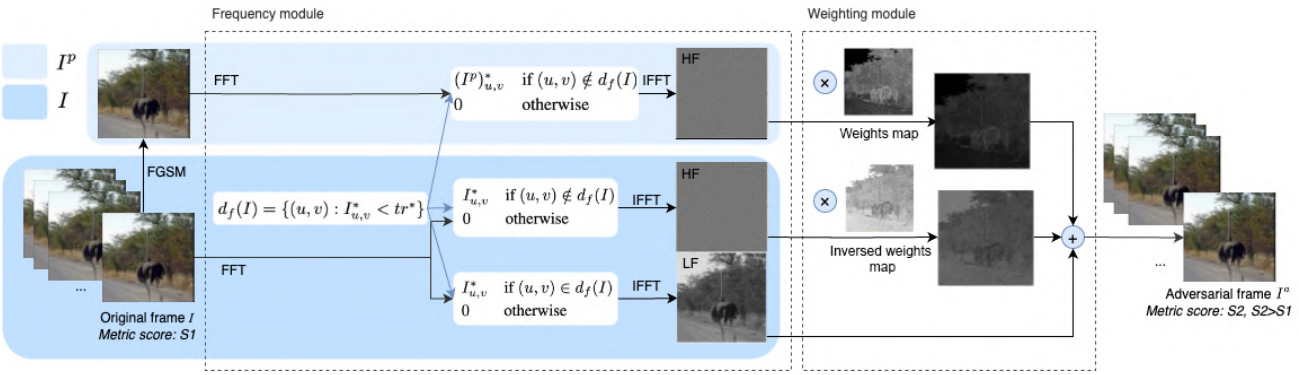

Figure 1. An overview of the proposed IOI adversarial attack. I is stands for input image, I^p – FGSM attacked image and I^a – the final IOI attacked image. Weights map is calculated using formula 9.

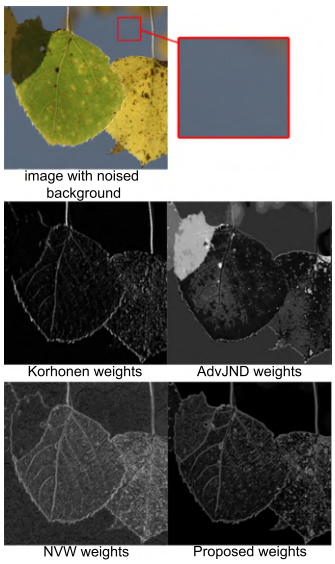

Figure 2. Comparison of weights used in prior and proposed methods. NVW and AdvJND assign non-zero weights for a background. Korhonen et al.'s weight total area is relatively small.

Proof. Each element of the $(I^a - I)$ is estimated using Equation 10 and representation of high-frequency FFT components as re-transformed 2D discrete Fourier transform without f% of highest frequencies. The complete proof is presented in the Appendix A.

The statement above demonstrates that the proposed method guarantees theoretical restriction of l_∞ norm of the adversarial image, which depends on initial attack strength expressed in MAE^* and parameter f for truncating frequencies. Higher f leads to saving more frequencies and providing better visual quality of the generated IOI adversarial image/video.

4. Experiments

We compared our method to the previous approaches targeting three learning-based NR image- and video-quality metrics on two datasets of images and videos.

Datasets. NIPS2017 image dataset (2017) was used to evaluate attacks on three NR metrics. It includes 1,000 images of a 299×299 resolution. For evaluating methods on videos, we used 12 videos with 1280×720 resolution from the DERF dataset (2001). Descriptions of these videos are available in the Appendix J. We extracted 75 frames from each original video and saved an attacked video with a frame rate of 25 frames per second, resulting in 3-second videos. The datasets licenses allow usage for research purposes.

Attacked models. For experiments on images, we selected PaQ-2-PiQ (Ying et al., 2020), Hyper-IQA (Su et al., 2020), and TReS (Golestaneh et al., 2022) NR models. For videos, we attacked the PaQ-2-PiQ (Ying et al., 2020) metric. These metrics were chosen to cover different architectures. PaQ-2-PiQ (Ying et al., 2020) employs RoIPool layers, which allows the flexibility to aggregate at different scales. Hyper-IQA (Su et al., 2020) utilizes ResNet50 for semantic feature extraction with further processing in the proposed Content Understanding Hyper Network. TReS (Golestaneh et al., 2022) is based on transformer architecture.

Methodology. To evaluate the efficiency of adversarial attacks, we considered three factors: relative gain (Equation 4), speed and the objective or subjective quality of perturbed images/videos. To compare these three factors, we fixed two of them by aligning the speed and relative gain of all methods and compared the objective and subjective visual quality of adversarial images/videos. Since most of the attack time is spent on backpropagation, we executed each attack for one iteration to standardize the speed of all attacks. We also conducted additional experiments with multiple

iterations, results presented in the Appendix C. In a one-iteration comparison, we combined the results of Zhang et al. and SSAH attacks with FGSM attacks in all tables because one iteration of these attacks is equivalent to one iteration of FGSM. Zhang et al. use an FR quality metric to preserve image fidelity, and SSAH uses the distance between low-frequency information of two images. Since the distorted image used in both methods appears only during the first iteration, these fidelity-preserving components yield zero gradients.

We employed an automatic process described in Algorithm 1 to ensure equal relative gain. Each attack has a parameter to regulate its strength. We denote this parameter as lr. Initially, we ran the proposed method with fixed parameters ($lr = 0.1$ and $f = 0.07$ for image data and $lr = 0.1$ and $f = 0.05$ for video data). Then, for each other attack and each image/video, we searched the minimal lr parameter to achieve the same relative gain. The search process also halted if the reached relative gain did not improve for $n = 5$ search iterations. For videos, the quality score was calculated as the mean of quality scores on each frame (PaQ-2-PiQ (Ying et al., 2020), Hyper-IQA (Su et al., 2020), and TReS (Golestaneh et al., 2022) are metrics that run per-frame on videos).

Algorithm 1 Relative gain aligning

Inputs: data element $\mathbf{X}$, target relative gain RG_t, adversarial attack $\mathbf{Adv}$, attacked model $\mathbf{M}$, search step d, stop parameter n, range of $\mathbf{M}$ M_{range}
Output: attacked data element $\mathbf{X_{adv}}$
$lr = 0$, $counter = 0$, $RG_{prev} = 0$, $flag = False$
while not $flag$ **do**
  $\mathbf{X_{adv}} = \mathbf{Adv}(\mathbf{X}, \mathbf{M}, lr)$
  $RG = \frac{\mathbf{M}(\mathbf{X_{adv}}) - \mathbf{M}(\mathbf{X})}{M_{range}}$
  if $RG \geq RG_t$ **do** $flag = True$
  if $RG \leq RG_{prev}$ **do** $counter = counter + 1$
  if $counter == n$ **do** $flag = True$
  $lr = lr + d$, $RG_{prev} = RG$
end while

Quality metrics and subjective study. We compared the objective quality of adversarial images and videos using four FR image- and video-quality metrics: PSNR, SSIM (Wang et al., 2004), VIF (Sheikh & Bovik, 2006), and LPIPS (Zhang et al., 2018).

We conducted a crowd-sourced subjective comparison using Subjectify.us (sub, Accessed: Jan 2024) to get subjective scores for adversarial videos. Original and adversarial videos were compressed using the x264 video codec with a CRF value 16 (preset "Medium"). Each participant was asked to choose the video of the superior visual quality from a random pair of videos shown sequentially. An option "Can't choose" was also available for them. Videos were pre-downloaded in the browser to prevent delays in playback, and participants had the flexibility to replay the videos multiple times. Each participant compared 12 video pairs; two of the 12 pairs were special verification pairs with apparent differences in visual quality. Answers from 200 participants who failed the verification were excluded. We collected 8220 responses from 685 participants who passed verification and calculated subjective scores using the Bradley-Terry model (Bradley & Terry, 1952). More details about the subjective experiment setup are presented in the Appendix K.

5. Results

Table 11 compares the proposed IOI adversarial attack and eight prior attacks at one iteration on the NIPS2017 image dataset (2017). The comparison involves three attacked image quality models, and the relative gain is aligned using the proposed Algorithm 1. On average, the relative gain achieved by all attacks was 7.7% for all models.

The proposed IOI method showed higher SSIM, VIF, and LPIPS scores for all attacked NR metrics. The PSNR score of our attack method is lower than that of other methods, which means that IOI changes more information in images; however, the perturbations are hidden in the images' texture/contrast regions. AdvJND method showed promising results for attacking Hyper-IQA and TReS models but failed on PaQ-2-PiQ. NVW performed well on PaQ-2-PiQ and TReS.

Table 3 contains the results of objective comparison on 12 videos from the DERF dataset (2001) and attacking the PaQ-2-PiQ (Ying et al., 2020) metric. We applied attacks with greater intensity on videos to enhance distinguishability for further subjective comparison, so the average relative gain was 15.3%. The objective results are similar to the comparison on the images: IOI outperformed prior methods on SSIM, VIF, and LPIPS metrics and showed comparable PSNR scores. The per-video results are in the Appendix L.

Subjective comparison. The subjective scores obtained from pairwise comparisons showed that the IOI attack generates adversarial videos of better visual quality: it holds a quality of 2.97, while other methods' scores are below 2.16. Confidence intervals for IOI and other methods do not intersect. The intervals intersect for some prior methods, since their adversarial videos are noisy and flickering, making their subjective quality difficult to rank. The video with the highest distinguishability was "Blue Sky" with tree branches swaying against the smooth sky. The video with the lowest distinguishability was "Rush Hour" with a highly noisy background. As shown in Figure 3, FGSM, NVW, and Korhonen et al. methods generate imperceptible distortion on the stone region but fail to suppress the

Table 2. The objective quality of adversarial images generated by existing and proposed methods on the NIPS2017 image dataset (2017) for three attacked models: PaQ-2-PiQ (Ying et al., 2020), Hyper-IQA (Su et al., 2020), and TReS (Golestaneh et al., 2022). The table presents FR metrics scores for adversarial images averaged across the dataset, with aligned relative gain and 95% confidence intervals. Each attack run for one iteration.

Attacked model	Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓
PaQ-2-PiQ (2020)	FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.884±0.007	33.6±0.3	0.635±0.010	0.134±0.009
	NVW (2021)	0.897±0.007	**34.7±0.5**	0.648±0.011	0.120±0.008
	Korhonen et al. (2022b)	0.872±0.008	33.1±0.3	0.617±0.011	0.151±0.011
	AdvJND (2020)	0.740±0.008	29.5±0.2	0.384±0.008	0.208±0.007
	UAP (2022)	0.737±0.004	26.3±0.2	0.371±0.004	0.314±0.005
	FACPA (2023b)	0.863±0.003	30.5±0.2	0.539±0.005	0.182±0.004
	IOI (ours)	**0.950±0.002**	33.4±0.2	**0.695±0.005**	**0.059±0.003**
Hyper-IQA (2020)	FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.746±0.017	30.6±0.6	0.542±0.019	0.326±0.023
	NVW (2021)	0.801±0.015	33.4±0.7	0.610±0.019	0.255±0.021
	Korhonen et al. (2022b)	0.765±0.016	31.1±0.6	0.562±0.019	0.303±0.022
	AdvJND (2020)	0.909±0.004	**37.1±0.3**	0.660±0.011	0.073±0.005
	UAP (2022)	0.545±0.010	21.4±0.3	0.192±0.007	0.447±0.008
	FACPA (2023b)	0.627±0.008	24.8±0.2	0.270±0.007	0.299±0.007
	IOI (ours)	**0.952±0.002**	33.5±0.2	**0.722±0.005**	**0.058±0.003**
TReS (2022)	FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.876±0.011	35.9±0.4	0.719±0.015	0.134±0.013
	NVW (2021)	0.902±0.010	37.7±0.5	0.754±0.014	0.107±0.011
	Korhonen et al. (2022b)	0.888±0.011	36.3±0.4	0.734±0.015	0.123±0.013
	AdvJND (2020)	0.915±0.006	**39.1±0.4**	0.736±0.013	0.064±0.006
	UAP (2022)	0.445±0.008	17.5±0.1	0.120±0.003	0.715±0.008
	FACPA (2023b)	0.611±0.007	23.4±0.2	0.221±0.007	0.530±0.011
	IOI (ours)	**0.945±0.002**	33.4±0.2	**0.756±0.005**	**0.059±0.003**

perturbation on the sky background. AdvJND, UAP and FACPA cause visible distortions on the whole image. To produce the same relative gain as IOI using one iteration, lr for other methods, was high, yielding visible perturbations. As shown in the Appendix C, all methods (except FGSM, UAP and FACPA) produce almost equivalent results at 20 iterations. But at one iteration, there is a crucial difference. The videos used for the comparison are available at https://github.com/katiashh/ioi-attack.

6. Discussion

Different frame frequency and attack success. We conducted additional experiments to show the importance of a one-iteration setup when attacking NR quality metrics for videos. This section demonstrates that employing a single iteration for each frame produces superior results compared to the sporadic application of multiple iterations, such as ten iterations for every tenth frame. We selected the PaQ-2-PiQ (2020) NR metric, "Controlled Burn" video from the DERF dataset (2001), and applied the I-FGSM attack (Kurakin et al., 2018). I-FGSM is an extension of FGSM, involving multiple iterations.

Let n represent the number of iterations, ϵ a small constant, I the original video frame, and M the target NR quality metric. The k-th iteration of I-FGSM is formulated as shown in Equation 12 ($k \in [0, n]$, $I_0^p = I$).

$$I_{k+1}^p = I_k^p + \frac{\epsilon}{n} * sign(\nabla_{I_k^p} M(I_k^p)) \tag{12}$$

We executed I-FGSM for different iteration counts: $n = 1$, $n = 2$, $n = 4$, $n = 6$, $n = 8$, and $n = 10$. Only some frames underwent attack in each experiment, specifically $\frac{1}{n}$. The selection of frames for the attack was done uniformly. The results of these experiments are illustrated in Figure 4. As the number of iterations in the attack increases, there is a corresponding rise in the attack's relative gain on each particular frame. However, evaluating the overall relative gain involves averaging relative gains across all frames. The optimal averaged relative gain occurs when each frame is attacked with just one iteration, and this gain decreases monotonically with the increase in the parameter n. We also measured the computation time for attacks with different values of n, which remained nearly constant at 3 seconds for all attacks. This consistency arises from the fact that each experiment's total number of adversarial iterations was the same.

Compared with other values of n, we can see that a one-

Table 3. Subjective comparison results on 12 videos from the DERF dataset (2001). Adversarial videos generated for PaQ-2-PiQ model (Ying et al., 2020) at equal speed and relative gain of all attacks. The table presents averaged quality metrics and subjective scores with 95% confidence intervals. Each attack runs for one iteration on each video frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.859±0.005	33.1±0.2	0.555±0.007	0.195±0.006	1.95±0.16
NVW (2021)	0.871±0.005	33.4±0.2	0.570±0.007	0.178±0.006	2.16±0.16
Korhonen et al. (2022b)	0.855±0.005	33.0±0.2	0.550±0.007	0.204±0.007	2.06±0.16
AdvJND (2020)	0.848±0.005	**34.5±0.2**	0.516±0.008	0.153±0.006	1.76±0.16
UAP (2022)	0.809±0.003	29.8±0.2	0.450±0.003	0.301±0.004	0.19±0.19
FACPA (2023b)	0.887±0.002	32.9±0.2	0.578±0.004	0.207±0.003	0.87±0.17
IOI (ours)	**0.941±0.016**	34.3±1.7	**0.669±0.046**	**0.098±0.030**	**2.97±0.16**

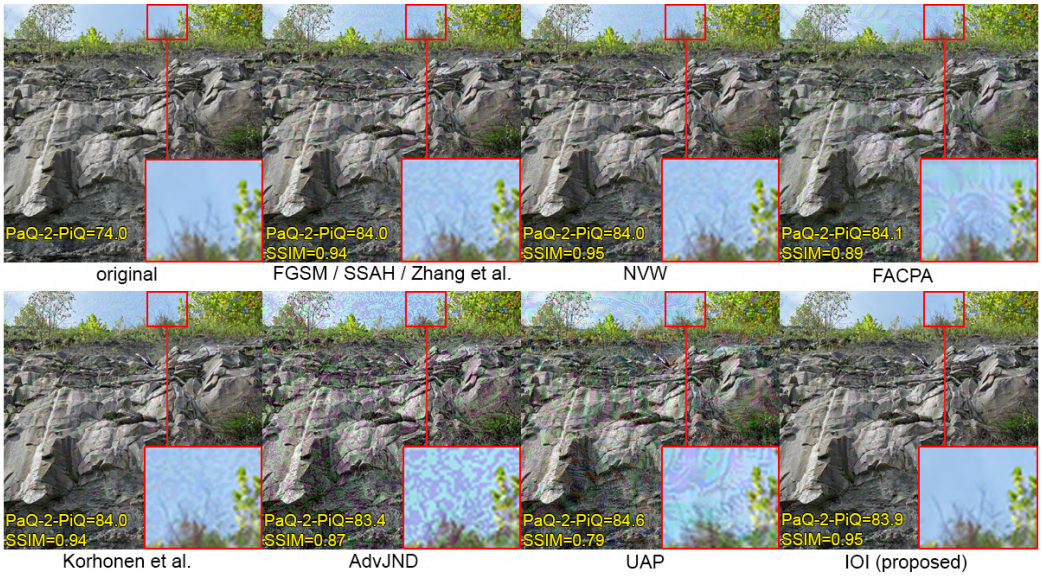

Figure 3. Comparison of adversarial images generated using FGSM (2015), SSAH (2022), Zhang et al. (2022b), NVW (2021), Korhonen et al. (2022b), AdvJND (2020), UAP (2022), FACPA (2023b) and IOI (ours) attack methods when attacking PaQ-2-PiQ (2020) NR quality metric at one iteration with relative gain aligned by Algorithm 1.

iteration attack yields superior averaged relative gain within the same attack time. This highlights that attack on video quality metrics differ from the classification task, where an attacker can affect only several frames to fool the classifier. From the results of this experiment, we can conclude that the effectiveness of adversarial attacks for video quality metrics is defined by their effectiveness at a one-iteration setup.

Speed of the proposed method. The PyTorch realization of the IOI attack allows reaching 8 fps on the NVIDIA Tesla T4 GPU. Details presented in Appendix B.

IOI performance under defences. We did additional experiments (Table 4) to check the robustness of the proposed method to three adversarial defences: video compression (Shaham et al., 2018), random crop and resize used in (Shumitskaya et al., 2023a) for NR metrics. Defences were evaluated for videos from the DERF dataset. Although the proposed method affects only high-frequency information, video compression reduced relative gain only by 2.4%. Random cropping confuses the attack and reduces relative gain approximately two times. Frames resizing almost completely mitigates the relative gain from 14.6% to 1%; however, an NR metric increase by 1% is still significant for benchmarks; sometimes, teams compete to achieve a 0.1% metric increase to win the competition. More details are presented in the Appendix E.

Limitations. Our method works in a white-box scenario that implies that an attacker knows and has access to the target model. The white-box scenario is less universal than

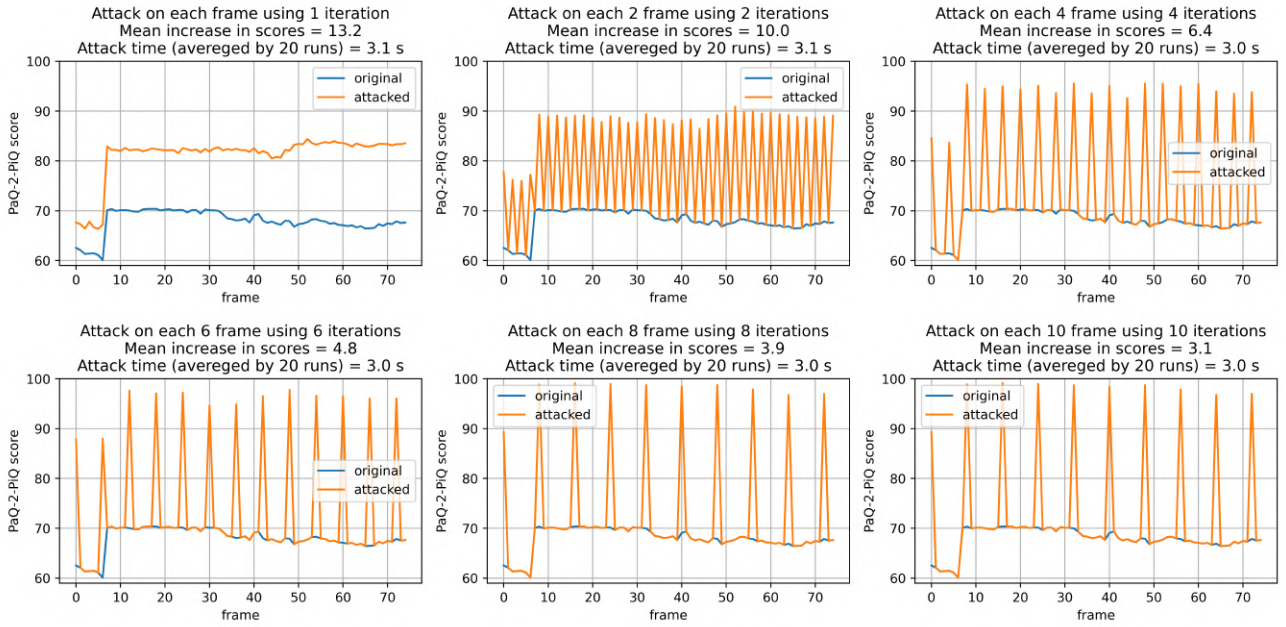

Figure 4. Results of experiments when attacking PaQ-2-PiQ (2020) NR metric on the "Controlled Burn" video through I-FGSM attack (2018) with different numbers of iterations and altered frames.

	IOI	IOI + compress.	IOI + random crop	IOI + resize
RG	14.6%	12.2%	6.30%	0.98%

Table 4. IOI performance under adversarial purification defences (video compression, random cropping, and resizing). Relative gain averaged for 12 videos.

a black-box; however, as described in the introduction, NR quality metrics are usually published as part of the benchmark methodology. We made additional experiments of analysing black-box transferability (Appendix F) and IOI performance in black-box settings (Appendix G). We found out low transferability across different models and significant difference in operation speed of white-box and black-box attacks. Black-box attacks are unlikely to be injected into video processing algorithms that compete in quality and speed, which is the scenario we target in our work. We considered only NR metrics, as FR metrics are much more difficult to attack in real-life scenarios. The robustness of FR metrics has been studied in (Ghildyal & Liu, 2023).

Additional experiments. We made the following additional experiments: experiment to compare the proposed method with prior methods when applying different parameters (Appendix D), metric score decreasing experiment (Appendix H), IOI attack on segment-level video quality model (Appendix I). We also measured time spending for one-iteration for all methods tested in this paper (Appendix B).

7. Conclusion

This paper introduces the IOI adversarial attack on NR image- and video-quality metrics. Its primary objective is to generate imperceptible perturbations for images or videos using only one iteration. Through extensive experiments, we showed that existing methods fail to produce high-fidelity adversarial videos in near real-time scenarios (1 – 10 fps). In contrast, our proposed method demonstrates better effectiveness at high speed. Subjective and objective comparisons showed that the proposed method produces adversarial images and videos of superior visual quality, achieving the same attack success and speed as prior methods. The proposed attack is a potent tool for experimentally assessing the vulnerability of NR quality metrics. By publishing our method, we provide a tool for verification of NR metrics robustness for benchmark organizers and contribute to the future development of robust image- and video-quality metrics. The proposed method can be used as a part of an adversarial training technique to improve the robustness of image- and video-quality metrics. Our code is openly accessible at `https://github.com/katiashh/ioi-attack`.

Acknowledgements

The authors would like to thank the CMC Faculty of MSU, especially A.V. Gulyaev for providing the necessary computing resources, which enabled us to undertake some of the calculations for this paper. We also would like to express our gratitude to Mikhail Pautov for discussing the results of this research and assistance in theoretical statements.

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 of dated November 2, 2021 No. 70-2021-00142.

Impact Statement

In this paper, we propose a new method that can be used to exploit vulnerabilities in video quality assessment metrics. NR quality assessment is widely used for public competitions. However, little research has been published in this area. Using NR metrics vulnerabilities is profitable for benchmark participants, so they are unlikely to publish their findings. Also, no robust NR metrics have been proposed so far due to the complexity of the task: adversarial training and purification methods reduce the performance of defended methods, significantly reducing the usability of NR metrics as a substitution for subjective tests.

References

Xiph.org video test media [derf's collection]. https://media.xiph.org/video/derf/, 2001.

Nips 2017: Adversarial learning development set. https://www.kaggle.com/datasets/google-brain/nips-2017-adversarial-learning\-development-set, 2017.

Pixel Privacy: Quality Camouflage for Social Images . https://multimediaeval.github.io/editions/2020/tasks/pixelprivacy/, 2020. Accessed: 2024-05-26.

Subjectify.us: Crowd-sourced subjective quality evaluation platform. https://www.subjectify.us/, Accessed: Jan 2024.

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pp. 484–501. Springer, 2020.

Antsiferova, A., Lavrushkin, S., Smirnov, M., Gushchin, A., Vatolin, D., and Kulikov, D. Video compression dataset and benchmark of learning-based video-quality metrics. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13814–13825. Curran Associates, Inc., 2022.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

CC-MSU. Msu video codecs comparison 2021 part 2: Subjective. https://www.compression.ru/video/codec_comparison/2021/subjective_report.html, 2021.

Croce, F. and Hein, M. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4724–4732, 2019.

Fezza, S. A., Bakhti, Y., Hamidouche, W., and Déforges, O. Perceptual evaluation of adversarial attacks for cnn-based image classification. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6. IEEE, 2019.

Ghadiyaram, D., Chen, C., Inguva, S., and Kokaram, A. A no-reference video quality predictor for h.264 compression and scaling artifacts. In *IEEE International Conference on Image Processing*, 2017.

Ghildyal, A. and Liu, F. Attacking perceptual similarity metrics. *Transactions on Machine Learning Research*, 2023.

Golestaneh, S. A., Dadsetan, S., and Kitani, K. M. 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*, pp. 3209–3218, 2022.

Goodfellow, I. J., Shlens, J., and Szegedy, C. 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*, 2015. URL http://arxiv.org/abs/1412.6572.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.632. URL http://ieeexplore.ieee.org/document/8100115/.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.),

Computer Vision – ECCV 2016, Lecture Notes in Computer Science, pp. 694–711, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46475-6. doi: 10.1007/978-3-319-46475-6_43.

Karli, B. T., Sen, D., and Temizel, A. Improving Perceptual Quality of Adversarial Images Using Perceptual Distance Minimization and Normalized Variance Weighting. In *The AAAI-22 Workshop on Adversarial Machine Learning and Beyond*, 2021.

Khrulkov, V. and Babenko, A. Neural side-by-side: Predicting human preferences for no-reference super-resolution evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4988–4997, June 2021.

Korhonen, J. and You, J. Adversarial attacks against blind image quality assessment models. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, pp. 3–11, 2022a.

Korhonen, J. and You, J. Adversarial Attacks Against Blind Image Quality Assessment Models. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, pp. 3–11, Lisboa Portugal, October 2022b. ACM. ISBN 978-1-4503-9499-4. doi: 10.1145/3552469.3555715. URL https://dl.acm.org/doi/10.1145/3552469.3555715.

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

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Legge, G. E. and Foley, J. M. Contrast masking in human vision. *Josa*, 70(12):1458–1471, 1980.

Li, D., Jiang, T., and Jiang, M. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129(4):1238–1257, 2021.

Li, T., Li, M., Yang, Y., and Deng, C. Frequency domain regularization for iterative adversarial attacks. *Pattern Recognition*, 134:109075, 2023.

Lin, W., Dong, L., and Xue, P. Visual distortion gauge based on discrimination of noticeable contrast changes. *IEEE transactions on circuits and systems for video technology*, 15(7):900–909, 2005.

Liu, A., Lin, W., Paul, M., Deng, C., and Zhang, F. Just noticeable difference for images with decomposition model for separating edge and textured regions. *IEEE Transactions on Circuits and Systems for Video Technology*, 20 (11):1648–1652, 2010.

Long, Y., Zhang, Q., Zeng, B., Gao, L., Liu, X., Zhang, J., and Song, J. Frequency domain model augmentation for adversarial attack. In *European conference on computer vision*, pp. 549–566. Springer, 2022.

Luo, C., Lin, Q., Xie, W., Wu, B., Xie, J., and Shen, L. Frequency-driven Imperceptible Adversarial Attack on Semantic Similarity. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15294–15303, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.01488. URL https://ieeexplore.ieee.org/document/9879877/.

Ma, C., Yang, C.-Y., Yang, X., and Yang, M.-H. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

Meftah, H. F. B., Fezza, S. A., Hamidouche, W., and Déforges, O. Evaluating the vulnerability of deep learning-based image quality assessment methods to adversarial attacks. In *2023 11th European Workshop on Visual Information Processing (EUVIP)*, pp. 1–6. IEEE, 2023.

Meng, F., Huang, K., Li, H., and Wu, Q. Class activation map generation by representative class selection and multi-layer feature fusion. *arXiv preprint arXiv:1901.07683*, 2019.

PIRM-SR. Image super-resolution on pirm-test. https://paperswithcode.com/sota/image-super-resolution-on-pirm-test, 2019.

Sang, Q., Zhang, H., Liu, L., Wu, X., and Bovik, A. On the generation of adversarial samples for image quality assessment. *Available at SSRN 4112969*, 2022.

Shaham, U., Garritano, J., Yamada, Y., Weinberger, E., Cloninger, A., Cheng, X., Stanton, K., and Kluger, Y. Defending against adversarial images using basis functions transformations. *arXiv preprint arXiv:1803.10840*, 2018.

Sharif, M., Bauer, L., and Reiter, M. K. On the suitability of lp-norms for creating and preventing adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1605–1613, 2018.

Sheikh, H. R. and Bovik, A. C. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

Shumitskaya, E., Antsiferova, A., and Vatolin, D. S. 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, 2022. URL https://bmvc2022.mpi-inf.mpg.de/0790.pdf.

Shumitskaya, E., Antsiferova, A., and Vatolin, D. Towards adversarial robustness verification of no-reference image- and video-quality metrics. *Computer Vision and Image Understanding*, pp. 103913, 2023a.

Shumitskaya, E., Antsiferova, A., and Vatolin, D. S. Fast adversarial cnn-based perturbation attack on no-reference image- and video-quality metrics. In Maughan, K., Liu, R., and Burns, T. F. (eds.), *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023*. OpenReview.net, 2023b. URL https://openreview.net/pdf?id=xKf-LSD2-Jg.

Siniukov, M., Antsiferova, A., Kulikov, D., and Vatolin, D. Hacking vmaf and vmaf neg: Vulnerability to different preprocessing methods. In *Proceedings of the 2021 4th Artificial Intelligence and Cloud Computing Conference*, AICCC '21, pp. 89–96, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450384162. doi: 10.1145/3508259.3508272. URL https://doi.org/10.1145/3508259.3508272.

SR-MSU. Msu video super-resolution quality metrics benchmark 2023. https://videoprocessing.ai/benchmarks/super-resolution-metrics.html, 2023.

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

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., and Zhang, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

VideoGeneration. Video generation on paperswithcode.com. https://paperswithcode.com/task/video-generation, 2024.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yang, C., Liu, Y., Li, D., et al. Exploring vulnerabilities of no-reference image quality assessment models: A query-based black-box method. *arXiv preprint arXiv:2401.05217*, 2024.

Yang, X., Ling, W., Lu, Z., Ong, E. P., and Yao, S. Just noticeable distortion model and its applications in video coding. *Signal processing: Image communication*, 20(7): 662–680, 2005.

Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., and Bovik, A. 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*, pp. 3575–3585, 2020.

Zhang, A., Ran, Y., Tang, W., and Wang, Y.-G. Vulnerabilities in video quality assessment models: The challenge of adversarial attacks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

Zhang, W., Li, D., Min, X., Zhai, G., Guo, G., Yang, X., and Ma, K. Perceptual attacks of no-reference image quality models with human-in-the-loop. *arXiv preprint arXiv:2210.00933*, 2022a.

Zhang, W., Li, D., Min, X., Zhai, G., Guo, G., Yang, X., and Ma, K. Perceptual Attacks of No-Reference Image Quality Models with Human-in-the-Loop. In *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=3AV_53iRfTi.

Zhang, Z., Qiao, K., Jiang, L., Wang, L., Chen, J., and Yan, B. AdvJND: Generating Adversarial Examples with Just Noticeable Difference. In Chen, X., Yan, H., Yan, Q., and Zhang, X. (eds.), *Machine Learning for Cyber Security*, volume 12487, pp. 463–478. Springer International Publishing, Cham, 2020. ISBN 978-3-030-62459-0 978-3-030-62460-6. doi: 10.1007/978-3-030-62460-6_42. URL http://link.springer.com/10.1007/978-3-030-62460-6_42. Series Title: Lecture Notes in Computer Science.

Zvezdakova, A., Zvezdakov, S., Kulikov, D., and Vatolin, D. Hacking vmaf with video color and contrast distortion, 2019.

A. Proof of Theorem 1

Theorem 1. *Let I and I^p be original and perturbed image correspondingly, I^a – IOI adversarial image based on I^p with truncating parameter f. Then inequality 13 is correct, where $MAE^*(\cdot,\cdot)$ is given by Equation 3.*

$$||I^a - I||_\infty \leq (1-f)MAE^*(I^p, I) \tag{13}$$

Proof. We can estimate the difference between I^a and I. Since $||w||_\infty \leq 1$, we can write the following:

$$\begin{aligned} &I^a = L_c^{d_f(I)}(I) + wH_c^{d_f(I)}(I^p) + (1-w)H_c^{d_f(I)}(I) \\ &I = L_c^{d_f(I)}(I) + H_c^{d_f(I)}(I)) \\ &||I^a - I||_\infty = ||w(H_c^{d_f(I)}(I^p) - H_c^{d_f(I)}(I))||_\infty \leq ||w||_\infty ||(H_c^{d_f(I)}(I^p) - H_c^{d_f(I)}(I))||_\infty \leq \\ &\leq ||(H_c^{d_f(I)}(I^p) - H_c^{d_f(I)}(I))||_\infty \end{aligned} \tag{14}$$

Since FFT and IFFT are linear transformations and indexes $d_f(I)$ for truncating I^p and I are the same:

$$H_c^{d_f(I)}(I^p) - H_c^{d_f(I)}(I) = H_c^{d_f(I)}(I^p - I) \tag{15}$$

We can write high-frequency component as re-transformed two-dimensional discrete Fourier transform without f% of highest frequencies and estimate the module of each element of $H_c^{d_f(I)}(I^p - I)$, where k_r, l_s – indexes of the sorter FFT coefficients, such that $|(I^p - I)^*_{k_n,l_n}| >= |(I^p - I)^*_{k_{n+1},l_{n+1}}| \; \forall n$:

$$\begin{aligned} &|H_c^{d_f(I)}(I^p - I)_{u,v}| = \tfrac{1}{HW}|\sum_{s=f(H-1)(W-1)}^{(H-1)(W-1)} (I^p - I)^*_{k_s,l_s} e^{i2\pi(\frac{k_r u}{H} + \frac{l_s v}{W})}| \leq \\ &\leq \tfrac{1}{HW} \sum_{s=f(H-1)(W-1)}^{(H-1)(W-1)} |(I^p - I)^*_{k_s,l_s}| = \frac{1}{HW} \sum_{s=f(H-1)(W-1)}^{(H-1)(W-1)} |(I^p)^*_{k_s,l_s} - (I)^*_{k_s,l_s}| = \beta - \alpha \end{aligned} \tag{16}$$

where α and β are given by Equation 17.

$$\begin{aligned} \beta &= \tfrac{1}{HW} \sum_{s=0}^{(H-1)(W-1)} |(I^p)^*_{k_s,l_s} - (I)^*_{k_s,l_s}| \\ \alpha &= \tfrac{1}{HW} \sum_{s=0}^{f(H-1)(W-1)} |(I^p)^*_{k_s,l_s} - (I)^*_{k_s,l_s}| \end{aligned} \tag{17}$$

Considering the facts that $\beta = MAE^*(I^p, I)$ (by definition) and $\alpha >= fMAE^*(I^p, I)$ (since α is the sum of modules of $f\%$ highest coefficients and $MAE^*(I^p, I)$ is the sum of all coefficients), we get the resulting estimate:

$$||I^a - I||_\infty \leq MAE^*(I^p, I) - fMAE^*(I^p, I) = (1-f)MAE^*(I^p, I) \tag{18}$$

The equation above demonstrates that the proposed method guarantees theoretical restriction of l_∞ norm of the adversarial image, which depends on initial attack strength and f parameter for truncating frequencies. It is worth noting that there was a rough estimate of $||w||_\infty \leq 1$ in Equation 14. In practice, multiplication on weights highly improves the l_2 norm of an adversarial image.

B. Speed for one iteration

Table 5 presents the calculation times of attacks at one iteration that were used for comparison in this paper when targeting the PaQ-2-PiQ (Ying et al., 2020) NR metric on one image from the NIPS2017 dataset (2017) and one video from the DERF

dataset (2001). We measured the calculation time on a server with an NVIDIA Tesla T4 GPU and averaged the results over 20 runs. The AdvJND is notably slower than others due to the computational complexity of calculating JND coefficients.

Table 5. GPU calculation times of attacks at one iteration when attacking PaQ-2-PiQ (2020) NR metric on images and videos.

METHOD	ONE ITERATION TIME ON IMAGE	ONE ITERATION FPS ON VIDEO
FGSM (2015), SSAH (2022), ZHANG ET AL. (2022B)	0.025 SEC	8.92 FPS
NVW (2021)	0.059 SEC	2.78 FPS
KORHONEN ET AL. (2022B)	0.037 SEC	7.05 FPS
ADVJND (2020)	10.38 SEC	0.01 FPS
IOI (OURS) $_{PyTorch}$	0.028 SEC	7.81 FPS

C. Experiment with multiple iterations

We conducted an additional experiment to evaluate the performance of visual-oriented methods employed in this paper in the context of multiple iterations. For $n = 10$ and $n = 20$ iterations, we run the proposed method with $\epsilon = 0.1$ and step size of $\frac{2\epsilon}{n}$. These experiments used the PaQ-2-PiQ NR model (Ying et al., 2020); the relative gain was 13% for both 10 and 20 iterations. Utilizing Algorithm 1, we searched for the minimal lr parameter in other attacks to achieve the same relative gain. Subsequently, we evaluated the visual quality of the resulting adversarial images using four FR metrics: SSIM, PSNR, VIF, and LPIPS. The results for $n = 10$ and $n = 20$ iterations are presented in Tables 6 and 7 respectively. At 20 iterations, all methods produce nearly identical results, highlighting the primary strength of the proposed IOI method in its effectiveness in one-iteration settings.

From the results of this experiment, we can conclude that in the setup of multiple iteration attack, there are no significant differences in which method from Table 7 to use (except SSAH and AdvJND – they need more than 20 iterations for convergence). The primary strength of the proposed IOI method is its effectiveness and superiority in one-iteration settings, but in multi-iteration setup, it also shows competitive results.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓
SSAH (2022)	0.890±0.005	33.6±0.4	0.680±0.009	0.106±0.005
Zhang et al. (2022b)	0.938±0.003	35.7±0.2	0.700±0.007	0.073±0.004
NVW (2021)	0.957±0.001	37.1±0.4	0.725±0.006	0.055±0.002
Korhonen et al. (2022b)	**0.974±0.001**	**37.5±0.2**	0.757±0.005	**0.035±0.002**
AdvJND (2020)	0.889±0.003	34.2±0.2	0.546±0.007	0.097±0.004
IOI (ours)	0.965±0.001	34.6±0.2	**0.758±0.004**	0.043±0.002

Table 6. Comparison results for 10 iterations with relative gain aligning.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓
SSAH (2022)	0.949±0.002	36.6±0.2	**0.790±0.005**	0.048±0.002
Zhang et al. (2022b)	0.971±0.001	38.1±0.2	0.786±0.005	0.034±0.002
NVW (2021)	0.970±0.001	**38.7±0.4**	0.778±0.005	0.038±0.002
Korhonen et al. (2022b)	**0.978±0.001**	38.2±0.2	0.781±0.004	**0.029±0.001**
AdvJND (2020)	0.916±0.003	35.9±0.2	0.613±0.007	0.076±0.004
IOI (ours)	0.972±0.001	35.5±0.2	0.779±0.004	0.035±0.002

Table 7. Comparison results for 20 iterations with relative gain aligning.

D. Different parameter's comparison

We made an additional experiment to compare the proposed method with prior methods when applying different parameters. This allows us to compare methods in slightly different attack strengths. Results in the Table 8 showed that the proposed

method generates images with better visual quality when achieving the same increase in metric score for three different increase levels.

Table 8. Experiment of methods comparison under different relative gains, corresponding to different ϵ levels in the proposed IOI method.

ϵ	Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓
0.08	FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.934±0.006	36.4±0.2	0.733±0.009	0.082±0.008
	NVW (2021)	0.940±0.006	**37.4±0.4**	0.745±0.009	0.072±0.007
	Korhonen et al. (2022b)	0.932±0.005	36.2±0.2	0.727±0.009	0.083±0.007
	AdvJND (2020)	0.812±0.005	31.9±0.1	0.466±0.007	0.151±0.006
	UAP (2022)	0.792±0.004	27.9±0.2	0.432±0.005	0.251±0.004
	FACPA (2023b)	0.888±0.003	31.7±0.2	0.586±0.005	0.151±0.003
	IOI (ours)	**0.966±0.002**	35.5±0.1	**0.786±0.004**	**0.037±0.002**
0.1	FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.884±0.007	33.6±0.3	0.635±0.010	0.134±0.009
	NVW (2021)	0.897±0.007	**34.7±0.5**	0.648±0.011	0.120±0.008
	Korhonen et al. (2022b)	0.872±0.008	33.1±0.3	0.617±0.011	0.151±0.011
	AdvJND (2020)	0.740±0.008	29.5±0.2	0.384±0.008	0.208±0.007
	UAP (2022)	0.737±0.004	26.3±0.2	0.371±0.004	0.314±0.005
	FACPA (2023b)	0.863±0.003	30.5±0.2	0.539±0.005	0.182±0.004
	IOI (ours)	**0.950±0.002**	33.4±0.2	**0.695±0.005**	**0.059±0.003**
0.12	FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.789±0.016	30.7±0.5	0.548±0.015	0.274±0.022
	NVW (2021)	0.795±0.016	31.5±0.6	0.555±0.016	0.264±0.022
	Korhonen et al. (2022b)	0.774±0.016	30.0±0.5	0.524±0.016	0.295±0.022
	AdvJND (2020)	0.696±0.007	28.4±0.1	0.340±0.006	0.239±0.008
	UAP (2022)	0.705±0.004	25.4±0.1	0.342±0.004	0.348±0.007
	FACPA (2023b)	0.758±0.003	26.9±0.2	0.392±0.004	0.291±0.005
	IOI (ours)	**0.936±0.003**	**32.2±0.2**	**0.681±0.005**	**0.077±0.004**

E. IOI performance under defences

We conducted additional experiments to evaluate the robustness of the proposed IOI attack to three defences: compression (Shaham et al., 2018), random crop and resize (Shumitskaya et al., 2023a). Random crop and resize defences were previously studied in (Shumitskaya et al., 2023a) for evaluation of UAP (Shumitskaya et al., 2022) against NR quality metrics. Authors showed that random crop and resize to 80% of the original image/video size help to improve NR quality metric robustness to adversarial attacks without significant loss in correlations with subjective scores. Because of that, we chose parameter 80% for our experiments. Defended transformation in the case of random crop defence was selecting a random crop of the frame with 80% of the original frame size. An image was resized to 80% of the original frame size for resize-based defence. Shaham et al. showed that compression can be used as defence (Shaham et al., 2018). In our experiment, defended transformation for compression defence was compression with a CRF value 16 using x264 video codec (preset "Medium").

For the experiment, we used 12 adversarial videos generated against PaQ-2-PiQ (source videos from the DERF dataset (der, 2001)) and original ones. We measured relative gain for the IOI attack with and without defences. For relative gain measurement under defence, we calculated PaQ-2-PiQ scores for original and adversarial videos after defended transformation and then, based on these scores, calculated relative gain. Results are presented in the Table 9. Compression defence reduced relative gain only by 2.4%. Random crop confuse attack and reduce relative gain approximately two times. Resize defence almost completely mitigate the relative gain, however gain in 1% still can be significant in benchmarks. From the results of these experiments, we can conclude that the proposed IOI attack is robust to compression and random crop defences but vulnerable to the resize defence.

F. Black-box transferability

We conducted an additional experiment to analyse the applicability of IOI in transferable black-box setting. The experiment was organised as follows: for all generated adversarial images from NIPS2017 dataset, we measured PaQ-2-PiQ, Hyper-IQA

Video	Relative gain			
	IOI	IOI + compression	IOI + random crop	IOI + resize
Blue Sky	13.1%	11.2%	6.17%	0.88%
Rush Hour	23.9%	18.0%	11.3%	0.77%
Old Town Cross	15.6%	13.9%	5.33%	1.29%
Crowd Run	15.2%	12.4%	4.10%	0.64%
Aspen	9.50%	6.70%	4.11%	0.51%
Sunflower	18.9%	15.8%	13.3%	1.33%
Life	10.0%	7.64%	1.76%	-0.09%
Controlled Burn	16.2%	15.4%	7.12%	1.85%
Red Kayak	16.3%	14.3%	7.08%	2.40%
Ducks Take Off	7.94%	6.24%	2.80%	0.30%
Tractor	11.6%	9.57%	6.00%	0.74%
Park Joy	16.8%	14.8%	6.47%	1.10%
Mean	14.6%	12.2%	6.30%	0.98%

Table 9. Results of performance proposed IOI attack under compression, random crop and resize defences.

and TReS, i.e. for adversarial images created to attack PaQ-2-PiQ we also measured quality scores by Hyper-IQA and TReS. Based on these metrics scores, we calculated relative gains. Results are presented in the Table 10. All eight methods tested in the paper showed low transferability to unseen models. Low transferability can be explained by these metrics having completely different architectures: PaQ-2-PiQ employs RoIPool layers, HyperIQA utilizes ResNet50 and TReS is based on transformer architecture. Also, it's important to note that improving transferability for image quality models is more challenging than for classifiers or detectors. Transferability can occur in classification and detection tasks because different classifiers "look" at the same regions of images where classified objects are located (Meng et al., 2019). In contrast, different image quality metrics can look at different regions to estimate the score, which inherently complicates the achievement of transferability. Thus, we can infer that developing imperceptible and, at the same time, transferable one-iteration attacks on video-quality models is a challenging problem that we will consider for further research.

Table 10. Results of experiment on transferability of methods used in the paper in one-iteration setting.

Attack	**PaQ-2-PiQ**			**Hyper-IQA**			**TReS**		
Test	**PaQ-2-PiQ**	Hyper-IQA	TReS	PaQ-2-PiQ	**Hyper-IQA**	TReS	PaQ-2-PiQ	Hyper-IQA	**TReS**
FGSM, SSAH, Zhang	**7.40%**	-8.76%	-14.93%	0.16%	**1.03%**	-18.99%	0.01%	-0.39%	**4.07%**
NVW	**7.17%**	-8.27%	-14.19%	0.13%	**0.70%**	-15.32%	0.08%	0.09%	**4.50%**
Korhonen	**7.36%**	-8.85%	-14.81%	0.23%	**0.42%**	-15.97%	0.07%	0.29%	**5.04%**
AdvJND	**6.61%**	-0.57%	-19.84%	-0.18%	**4.09%**	-3.96%	0.04%	-0.43%	**7.02%**
UAP	**11.62%**	-6.03%	-15.55%	3.13%	**4.83%**	6.65%	2.44%	-0.52%	**0.85%**
FACPA	**9.68%**	-5.01%	-9.14%	-0.30%	**1.48%**	9.52%	-6.02%	-0.51%	**1.14%**
IOI (ours)	**7.33%**	-3.56%	-6.25%	0.26%	**7.86%**	-1.73%	0.39%	2.34%	**7.40%**

G. IOI performance in black-box setting

We conducted an additional experiment aimed to show that the proposed method can be adapted for use in a black-box setting. To do this, we replaced an FGSM-generated perturbation with a black-box-generated perturbation, e.g. Square Attack (Andriushchenko et al., 2020). We will call this modification BB-IOI. It's important to note that such modification transforms the method into a multi-iteration. To verify the efficiency of BB-IOI, we conducted an additional experiment involving the implementation of the BB-IOI attack that consists of two stages: 1) generating the adversarial perturbation in a black-box manner using the Square Attack (Andriushchenko et al., 2020) method and 2) processing this perturbation using frequency and weighting modules. As a result, BB-IOI provides high-quality adversarial images by objective metrics. Although the property of imperceptibility remains, it's important to note that the computation complexity of the method has increased, which is common for black-box methods. For attacking PaQ-2-PiQ on the NIPS2017 dataset, BB-IOI achieved

an average 1.21% gain operating at 40 seconds per image. Results are presented in the Table 11. This performance is 6.12% lower and 1400 times slower than IOI. Furthermore, the proposed IOI method can be adapted for use in transfer-based settings by combining it with transfer-based perturbation generation techniques (Long et al., 2022), (Li et al., 2023).

Table 11. Comparison of the proposed method used in white-box setting (IOI) and black-box setting (BB-IOI).

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Relative gain ↑	Time on one image ↓
IOI	0.945±0.002	33.4±0.2	0.756±0.005	0.059±0.003	7.33%	0.028 sec
BB-IOI	0.988±0.001	36.9±0.1	0.815±0.003	0.024±0.001	1.21%	40 sec

H. Metric score decreasing experiment

In this section, we show the possibility of metrics score decreasing. Given that the proposed method consists of two parts (generating a perturbed image using FGSM and subsequent processing in frequency and weighting modules), changing the optimization direction in FGSM leads to guiding an attack in the opposite direction. We conducted additional experiments to show that the proposed method can decrease PaQ-2-PiQ metric scores. We applied the proposed method to attack it on the same NIPS2017 dataset we used to increase this metric. The results revealed that increasing metric scores yielded a 7.32% score increase, while decreasing metric scores resulted in a 7.61% score decrease (almost the same). However, it's worth noting that we focused on increasing metrics' scores because decreasing quality metrics' scores holds less practical significance. An attacker can decrease the metrics for quality camouflage (2020), and it is the only real-life scenario known to the authors.

I. IOI attack on complex VQA metric

The proposed method applies to segment-level video quality models. We conducted an additional experiment targeting the segment-level VQA metric MDTVSFA (Li et al., 2021) to show this. To apply the proposed attack, we first slightly modified MDTVSFA to get access to its gradient (we removed $torch.no\_grad()$ context-manager from the feature extraction module and modified the forward process in the inference model to process batches rather than dictionaries). Then, we applied IOI to attack three videos on a per-frame basis, processing one frame at a time. Subsequently, we constructed adversarial videos from these frames and calculated quality scores using the original segment-level MDTVSFA on these videos. Results are presented in the Table 12. Remarkably, this approach yielded a significant relative gain, with a 15% increase in scores. This experiment showed that attacking only spatial features of a VQA metric without accounting for temporal and other features is enough to achieve high attack success.

Table 12. Experiments of IOI attack targeting segment-level video-quality metric MDTVSFA. The attack was performed per-frame. Resulting gain was calculated using original segment-level MDTVSFA.

Video	MDTVSFA clean	MDTVSFA IOI attacked
Blue Sky	0.544	0.659 (↑11.5%)
Crowd Run	0.555	0.759 (↑20.4%)
Pedestrian Area	0.584	0.737 (↑15.3%)

J. Video sequences

We used the following 12 videos with a resolution of 1280×720 from the DERF dataset (2001):

1. "Blue Sky": https://media.xiph.org/video/derf/y4m/blue_sky_1080p25.y4m
2. "Aspen": https://media.xiph.org/video/derf/y4m/aspen_1080p.y4m
3. "Sunflower": https://media.xiph.org/video/derf/y4m/sunflower_1080p25.y4m
4. "Crowd Run": https://media.xiph.org/video/derf/y4m/crowd_run_1080p50.y4m

5. “Old Town Cross”: https://media.xiph.org/video/derf/y4m/old_town_cross_1080p50.y4m
6. “Life”: https://media.xiph.org/video/derf/y4m/life_1080p30.y4m
7. “Controlled Burn”: https://media.xiph.org/video/derf/y4m/controlled_burn_1080p.y4m
8. “Rush Hour”: https://media.xiph.org/video/derf/y4m/rush_hour_1080p25.y4m
9. “Red Kayak”: https://media.xiph.org/video/derf/y4m/red_kayak_1080p.y4m
10. “Ducks Take Off”: https://media.xiph.org/video/derf/y4m/ducks_take_off_1080p50.y4m
11. “Tractor”: https://media.xiph.org/video/derf/y4m/tractor_1080p25.y4m
12. “Park Joy”: https://media.xiph.org/video/derf/y4m/park_joy_1080p50.y4m

We extracted 75 frames from each original video and saved an attacked video with a frame rate of 25 frames per second, resulting in videos with a duration of 3 seconds. Figure 5 contains spatial and temporal information for these videos. Figure 6 contains the first frames of videos.

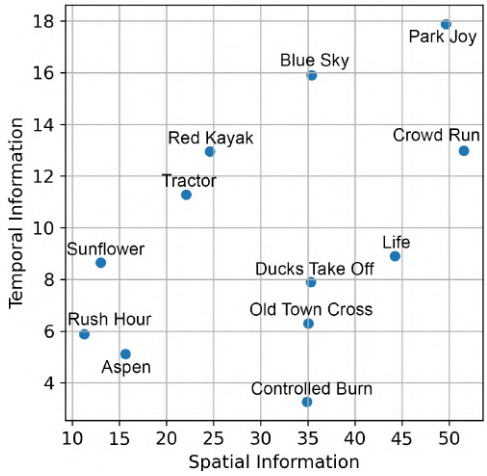

Figure 5. Spatial and temporal information for videos.

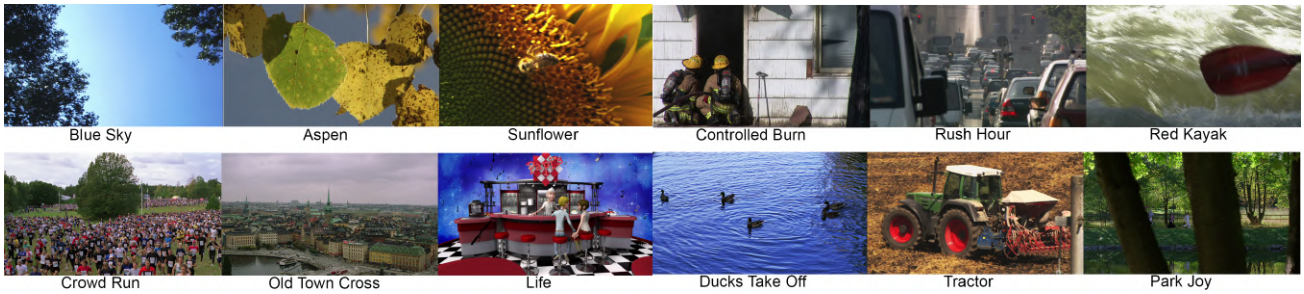

Figure 6. First frames of videos.

K. Subjective experiment setup

To derive subjective scores for adversarial videos, we conducted a crowd-sourced subjective comparison on the Subjectify.us service (sub, Accessed: Jan 2024).

For comparison, we compressed all videos, including the original ones, using the x264 video codec with a CRF value of 16 (preset "Medium"). Each pair shown to participants consisted of two samples of the same source video attacked by various attack methods. Each participant was presented with a random pair of videos sequentially and was asked to choose the video with the superior visual quality. An option "Can't choose" was also provided. Videos were pre-loaded in the browser to prevent delays in playback, and participants had the flexibility to replay the videos multiple times. Each participant compared 12 video pairs, of which two were for verification. Answers from 200 participants who failed the verification were excluded.

We collected 8220 responses from 685 successful participants and calculated subjective scores using the Bradley-Terry model (Bradley & Terry, 1952). The average payment to crowdworkers per a pair of sequences was \$0.05. We estimate the overall cost of the subjective tests was \$410. Figure 7 presents the subjective experiment's general process.

Details about the crowdsourced study:

1. Screen resolutions were from 360×800 to 3440×1440. Table 13 shows the most popular.
2. Participants were from 31 countries.
3. Participant ages ranged from 18 to 93, with an average of 39. Figure 8 shows the distribution.

Command line for encoding. Given the directory of video frames in PNG format (set of images 000.png, 001.png, ..., 074.png) we run the following FFmpeg command line:

```
ffmpeg -pattern_type glob -i *.png -c:v libx264 -crf 16 -pix_fmt yuv420p res.mp4
```

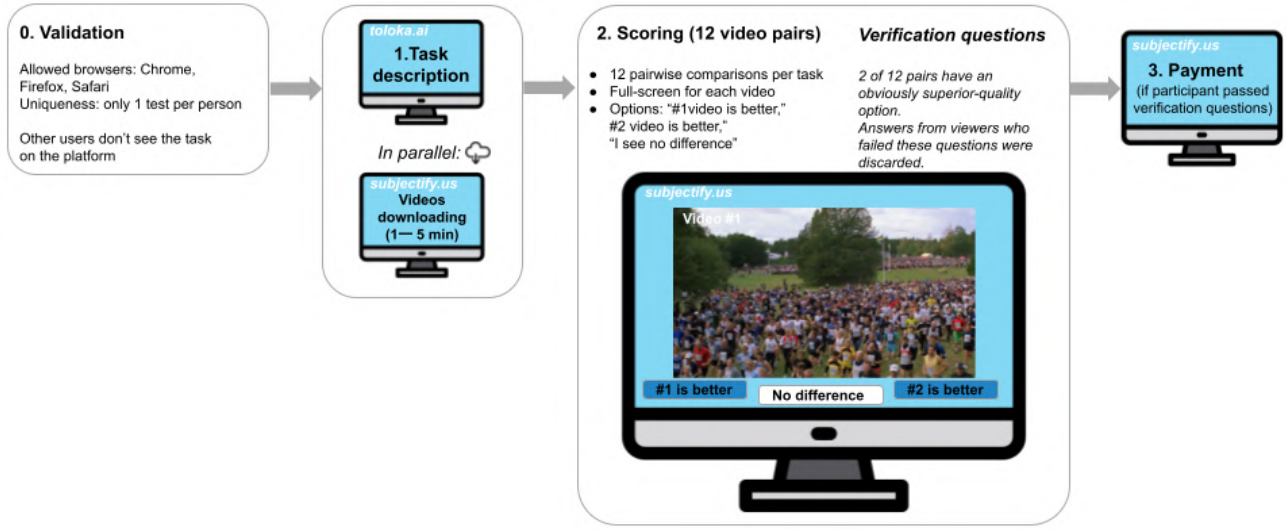

Figure 7. Subjective-assessment scheme.

L. Per-video results

Tables 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 contain results of subjective comparison of proposed IOI adversarial attack with eight prior attacks for each video.

M. Attack examples on images and videos

IOI adversarial images and videos are available in the zip archive: https://drive.google.com/file/d/1nrvV70Q4W0vh-2FdWrXHUhMDzYcI6zY1/view?usp=sharing.

Resolution	Number of users
1920×1080	194
1366×768	167
1536×864	100
1280×1024	39
1600×900	35
1280×720	22
1440×900	19
1024×768	11
2560×1440	10
1680×1050	10
1360×768	10

Table 13. Most popular screen resolutions among crowdworkers.

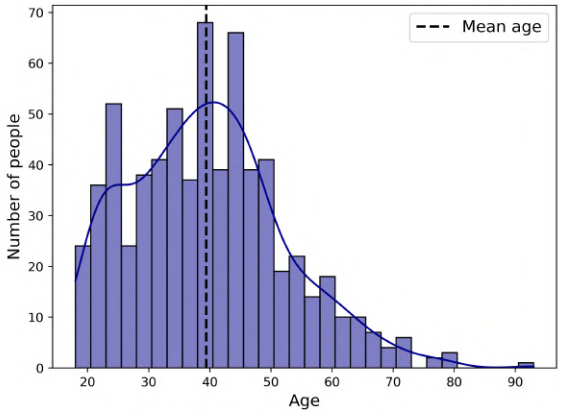

Figure 8. Age distribution of crowdworkers.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.739	30.8	0.390	0.352	1.44±0.66
NVW (2021)	0.719	30.1	0.370	0.364	1.91±0.64
Korhonen et al. (2022b)	0.699	29.8	0.361	0.380	1.56±0.65
AdvJND (2020)	0.795	33.9	0.422	0.247	2.20±0.63
UAP (2022)	0.809	31.6	0.453	0.338	0.70±0.70
FACPA (2023b)	0.890	**34.4**	0.557	0.253	1.11±0.67
IOI (ours)	**0.956**	33.9	**0.649**	**0.048**	**4.05±0.62**

Table 14. Comparison results on the "Blue Sky" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.956	40.0	0.720	0.064	2.88±0.47
NVW (2021)	0.959	40.3	0.733	0.057	3.13±0.48
Korhonen et al. (2022b)	0.957	40.2	0.725	0.062	2.91±0.48
AdvJND (2020)	0.950	**41.1**	0.687	0.052	2.57±0.48
UAP (2022)	0.842	32.8	0.465	0.264	0.65±0.65
FACPA (2023b)	0.915	36.0	0.605	0.174	1.18±0.58
IOI (ours)	**0.981**	38.3	**0.823**	**0.044**	**3.33±0.48**

Table 15. Comparison results on the "Aspen" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.919	37.2	0.604	0.206	2.71±0.52
NVW (2021)	0.924	**37.4**	0.619	0.195	2.96±0.52
Korhonen et al. (2022b)	0.844	34.0	0.465	0.326	2.28±0.53
AdvJND (2020)	0.741	32.7	0.328	0.365	1.57±0.56
UAP (2022)	0.846	32.8	0.463	0.345	0.65±0.65
FACPA (2023b)	0.922	35.9	0.618	0.209	1.20±0.60
IOI (ours)	**0.946**	36.9	**0.690**	**0.181**	**3.39±0.52**

Table 16. Comparison results on the "Sunflower" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.921	33.3	0.689	0.071	3.61±0.51
NVW (2021)	0.924	**33.4**	0.694	0.068	3.16±0.51
Korhonen et al. (2022b)	0.922	**33.4**	0.693	0.070	3.51±0.51
AdvJND (2020)	0.844	31.7	0.525	0.089	2.12±0.56
UAP (2022)	0.798	27.2	0.465	0.204	0.73±0.73
FACPA (2023b)	0.885	30.5	0.604	0.140	1.44±0.64
IOI (ours)	**0.951**	32.4	**0.695**	**0.039**	**4.00±0.51**

Table 17. Comparison results on the "Crowd Run" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.854	31.8	0.512	0.151	2.77±1.16
NVW (2021)	0.856	31.9	0.517	0.147	2.92±1.16
Korhonen et al. (2022b)	0.854	31.9	0.513	0.150	3.05±1.16
AdvJND (2020)	0.844	33.5	0.471	0.120	1.90±1.18
UAP (2022)	0.827	30.4	0.473	0.237	1.21±1.21
FACPA (2023b)	**0.913**	**34.1**	**0.623**	0.168	2.59±1.16
IOI (ours)	0.908	31.1	0.539	**0.118**	**3.61±1.15**

Table 18. Comparison results on the "Old Town Cross" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.737	29.7	0.419	0.216	2.44±0.66
NVW (2021)	0.818	31.4	0.497	0.153	2.83±0.64
Korhonen et al. (2022b)	0.873	33.8	0.588	0.109	4.14±0.59
AdvJND (2020)	0.832	34.1	0.512	0.092	3.57±0.61
UAP (2022)	0.717	28.1	0.396	0.264	0.79±0.79
FACPA (2023b)	0.814	30.7	0.502	0.188	1.50±0.72
IOI (ours)	**0.936**	**34.8**	**0.668**	**0.063**	**4.89±0.59**

Table 19. Comparison results on the "Life" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.833	30.7	0.532	0.256	1.64±0.52
NVW (2021)	0.853	31.1	0.553	0.225	1.93±0.52
Korhonen et al. (2022b)	0.777	28.8	0.458	0.328	1.18±0.54
AdvJND (2020)	0.833	32.7	0.508	0.176	1.54±0.52
UAP (2022)	0.798	28.8	0.476	0.314	0.57±0.57
FACPA (2023b)	0.903	**32.9**	0.648	0.195	1.01±0.55
IOI (ours)	**0.932**	32.5	**0.652**	**0.117**	**2.90±0.53**

Table 20. Comparison results on the "Controlled Burn" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.956	40.0	0.685	0.093	3.04±0.49
NVW (2021)	0.960	40.4	0.704	0.082	3.08±0.49
Korhonen et al. (2022b)	0.959	40.4	0.701	0.084	3.39±0.49
AdvJND (2020)	**0.974**	**44.1**	**0.767**	**0.035**	**3.51±0.49**
UAP (2022)	0.838	32.8	0.421	0.357	0.71±0.71
FACPA (2023b)	0.913	36.0	0.557	0.247	1.58±0.59
IOI (ours)	0.952	38.5	0.695	0.118	3.14±0.49

Table 21. Comparison results on the "Rush Hour" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.790	30.7	0.466	0.333	1.36±0.56
NVW (2021)	0.843	32.3	0.535	0.249	2.20±0.54
Korhonen et al. (2022b)	0.792	30.7	0.469	0.329	1.84±0.54
AdvJND (2020)	0.759	31.7	0.393	0.272	1.42±0.56
UAP (2022)	0.840	31.5	0.517	0.280	0.61±0.61
FACPA (2023b)	0.900	34.0	0.638	0.219	1.20±0.57
IOI (ours)	**0.942**	**34.3**	**0.675**	**0.116**	**3.38±0.55**

Table 22. Comparison results on the "Red Kayak" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.904	30.6	0.549	0.172	3.42±0.56
NVW (2021)	0.906	30.7	0.551	0.170	3.62±0.56
Korhonen et al. (2022b)	0.905	30.7	0.553	0.168	3.42±0.56
AdvJND (2020)	0.895	32.0	0.515	0.103	3.04±0.57
UAP (2022)	0.826	26.0	0.400	0.346	0.77±0.77
FACPA (2023b)	0.880	28.6	0.491	0.194	1.48±0.68
IOI (ours)	**0.963**	**34.1**	**0.693**	**0.055**	**4.09±0.55**

Table 23. Comparison results on the "Ducks Take Off" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.893	33.3	0.598	0.182	3.48±0.56
NVW (2021)	0.902	33.5	0.612	0.168	4.29±0.54
Korhonen et al. (2022b)	0.899	33.7	0.615	0.168	3.56±0.56
AdvJND (2020)	0.892	**35.0**	0.584	**0.111**	3.66±0.55
UAP (2022)	0.800	28.8	0.444	0.333	0.86±0.86
FACPA (2023b)	0.877	31.5	0.561	0.221	1.75±0.73
IOI (ours)	**0.944**	34.7	**0.696**	0.119	**4.72±0.54**

Table 24. Comparison results on the "Tractor" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.

Method	SSIM ↑	PSNR ↑	VIF ↑	LPIPS ↓	Subjective score ↑
FGSM (2015), SSAH (2022), Zhang et al. (2022b)	0.807	29.7	0.491	0.240	2.55±0.65
NVW (2021)	0.784	28.8	0.458	0.263	2.07±0.66
Korhonen et al. (2022b)	0.780	28.8	0.456	0.269	2.23±0.65
AdvJND (2020)	0.812	**31.8**	0.485	0.175	1.94±0.66
UAP (2022)	0.764	27.3	0.424	0.325	0.74±0.74
FACPA (2023b)	0.834	29.7	0.525	0.273	1.41±0.69
IOI (ours)	**0.881**	30.2	**0.549**	**0.161**	**3.29±0.63**

Table 25. Comparison results on the "Park Joy" video and PaQ-2-PiQ attacked model (Ying et al., 2020) with relative gain aligning. FR quality metric score for video is calculated as a mean of quality scores on each frame.