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

Ekaterina Shumitskaya, Anastasia Antsiferova, Dmitriy Vatolin
Video Group
CS MSU Graphics&Media Lab

Motivation

Problem

No-reference (NR) image- and video-quality metrics are widely used in video benchmarks. However, recent studies unveiled vulnerabilities in NR image quality metrics when exposed to adversarial attacks

attack

increases metric

does not increase visual quality

Motivation
Real-life scenarios

There are several real-life scenarios for attacks on the image- or video-quality metric:

- Cheating in public benchmarks
- Video quality control fooling in streaming services
- Manipulating results of web search

Motivation

Attacks on videos

Criteria for attacks integrated into video processing methods:

1. Quantitative success of an attack
2. High speed of an attack
3. Temporal consistency of an adversarial video

Our research investigates the potential of injecting ***fast***, ***invisible*** and ***temporally consistent*** adversarial attacks on NR metrics in videos

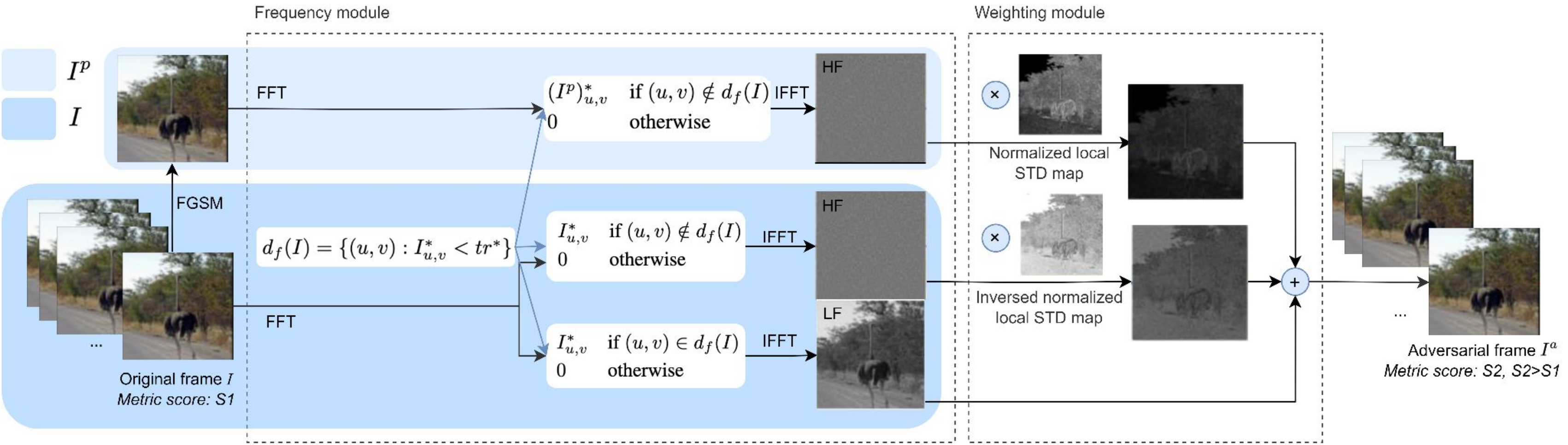
Proposed method

Overview

Initially, we perturb the image using a baseline gradient attack:

$$I^p = I + \epsilon * sign(\nabla_I M(I))$$

Then we process the perturbed image using frequency and weighting modules to enhance the visual quality of an adversarial image/video



Proposed method

Mathematical properties

We provide upper bound of adversarial perturbation added by our method

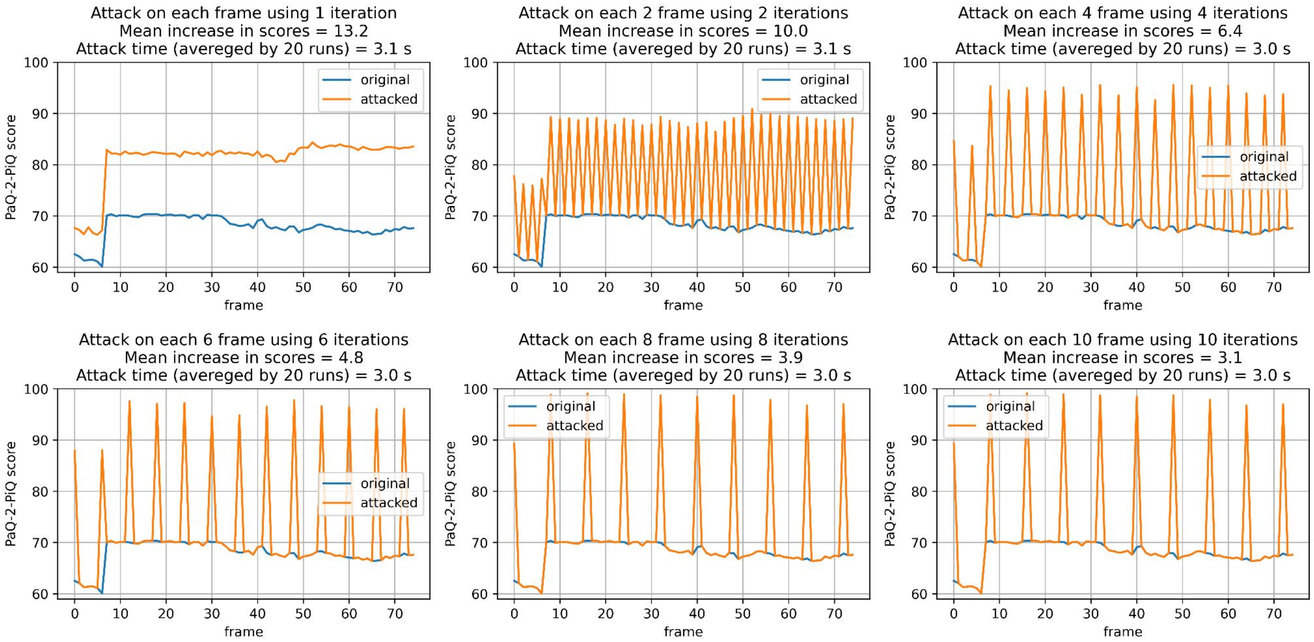
Theorem. Let I and I^p be original and perturbed image correspondingly, I^a – adversarial image after IOI attack that is based on I^p with truncating parameter f. Then inequality (1) is correct, where is given by Equation (2)

$$||I^a - I||_\infty \leq (1 - f)MAE^*(I^p, I) \qquad (1)$$

$$MAE^*(I^p, I) = \frac{1}{HW} \sum_{i=0}^{(H-1)} \sum_{j=0}^{(W-1)} |I_{ij}^{p*} - I_{ij}^*| \qquad (2)$$

Proposed method

Why one-iteration attack?



We conducted additional experiments to show the importance of a one-iteration setup when attacking NR quality metrics for videos. Compared with other values of n, we can see that a one-iteration attack yields superior averaged relative gain within the same attack time

Comparison methodology

Overview

Factors to evaluate the attack efficiency

- Relative gain (attack success)
- Speed
- Visual quality of adversarial images

Comparison methodology

Relative gain

Factors to evaluate the attack efficiency

Relative gain (attack success)	Speed	Visual quality of adversarial images

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

I – original image/frame

I^a – adversarial image/frame

M – target quality metric

M_{range} – range of M scores

Comparison methodology

Visual quality

Factors to evaluate the attack efficiency

Relative gain (attack success)	Speed	Visual quality of adversarial images

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

I – original image/frame

I^a – adversarial image/frame

M – target quality metric

M_{range} – range of M scores

Measuring the time for constructing I^a from I

Comparison methodology

Visual quality

Factors to evaluate the attack efficiency

Relative gain (attack success)

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

I – original image/frame

I^a – adversarial image/frame

M – target quality metric

M_{range} – range of M scores

Speed

Measuring the time for constructing I^a from I

Visual quality of adversarial images

$Q(I, I^a)$

- **Objective**: SSIM, PSNR, VIF, LPIPS
- **Subjective**: Subjective comparison

Comparison methodology

Fixed and compared factors

Factors to evaluate the attack efficiency

Relative gain (attack success)	Speed	Visual quality of adversarial images
fixed factor	fixed factor	compared factor
align using search in attack parameters space	align using a run with an equal number of iterations	

Experiments

Setup (1)

Two datasets:

- 100 images from NIPS2017 (299×299 resolution)
- 12 videos from DERF2001 (1280×720 resolution)

Three target quality models:

- PaQ-2-PiQ
- Hyper-IQA
- TReS

Nine compared methods:

FGSM, SSAH, Zhang et al., NVW, Korhonen et al., AdvJND, UAP, FACPA, IOI (ours)

Experiments
Setup (2)

Objective comparison

Comparison using four FR quality metrics: SSIM, PSNR, LPIPS, VIF

Subjective comparison

- Conducted using Subjectify.us service
- Pair comparison with verification questions
- Each participant compared 12 video pairs
- Collected 8220 responses from 685 participants

Experiments

Objective results

The proposed IOI method showed higher SSIM, VIF, and LPIPS scores for all attacked NR metrics. The PSNR score of IOI is lower than that of other methods, which means that IOI changes more information in images

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**

The objective quality of adversarial images generated by existing and proposed methods averaged across the NIPS2017 dataset

The subjective scores 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

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**

Subjective comparison results on 12 videos from the DERF2001 dataset. Adversarial videos generated for PaQ-2-PiQ model at equal speed and relative gain of all attacks. Each attack runs for one iteration on each video frame

Results

Video examples. Original

Original video. PaQ-2-PiQ = 68.4

Attacked video. PaQ-2-PiQ = 79.5

“Blue Sky” video from the DERF dataset Xiph.org video test media [derf’s collection]. https://media.xiph.org/video/derf/, 2001

Conclusion

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