Состязательные атаки на метрики качества и прекрасные нейросетевые артефакты ближайшего будущего

Dmitriy Vatolin

MSU Institute for Artificial Intelligence
ISP RAS Research Center for Trusted AI
CS MSU Graphics&Media Lab

Об авторе

- Зав Graphics&Media Lab ВМК МГУ, Video Lab ИИИ МГУ
- Создатель сайтов по алгоритмам
  - https://compression.ru/video
  - https://videoprocessing.ai
  - https://videoprocessing.github.io/
- Области интересов: современное сжатие видео, измерение качества видео, четырехмерное видео
- Руководил 40+ проектами с компаниями **Intel, Cisco, Samsung, Huawei, Broadcom** и др.
- Автор №1* на Habr.com в хабах **«AR и VR»**, **«Искусственный интеллект»**, **«Работа с видео»** и **«Видеотехника»**
- Сомневается в разумности Homo Sapiens

* Без учета редакторов

Наши партнеры

- **90% наших проектов** финансируются компаниями
- **Долгосрочное сотрудничество** с Intel, Samsung, Huawei и другими
- Наши исследования **максимально практичны**

и многие другие...

My former student — author of VGG

Karen Simonyan is the first author of VGG — a revolutional object-recognition model

VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

Karen Simonyan* **& Andrew Zisserman**+
Visual Geometry Group, Department of Engineering Science, University of Oxford
{karen,az}@robots.ox.ac.uk

Karen Simonyan

robots.ox.ac.uk/~karen

Affiliation: Google Inc.
Citations: 81,390
h-index: 43
Research interests: Deep Learning

VirtualDub MSU Motion Estimation Filter

MSU Graphics & Media Lab (Video Group)

Project, idea: Dr. Dmitriy Vatolin
Algorithm: Karen Simonyan, Sergey Grishin
Implementation: Karen Simonyan

INSIDER

Subscribe

TECH

The AI 100 2023: The top people in artificial intelligence

Dario Amodei

Anthropic

READ MORE

Clem Delangue

Hugging Face

READ MORE

Sarah Na

Seek AI

READ MOR

Karén Simonyan

Inflection AI

READ MORE

Robin Li

Baidu

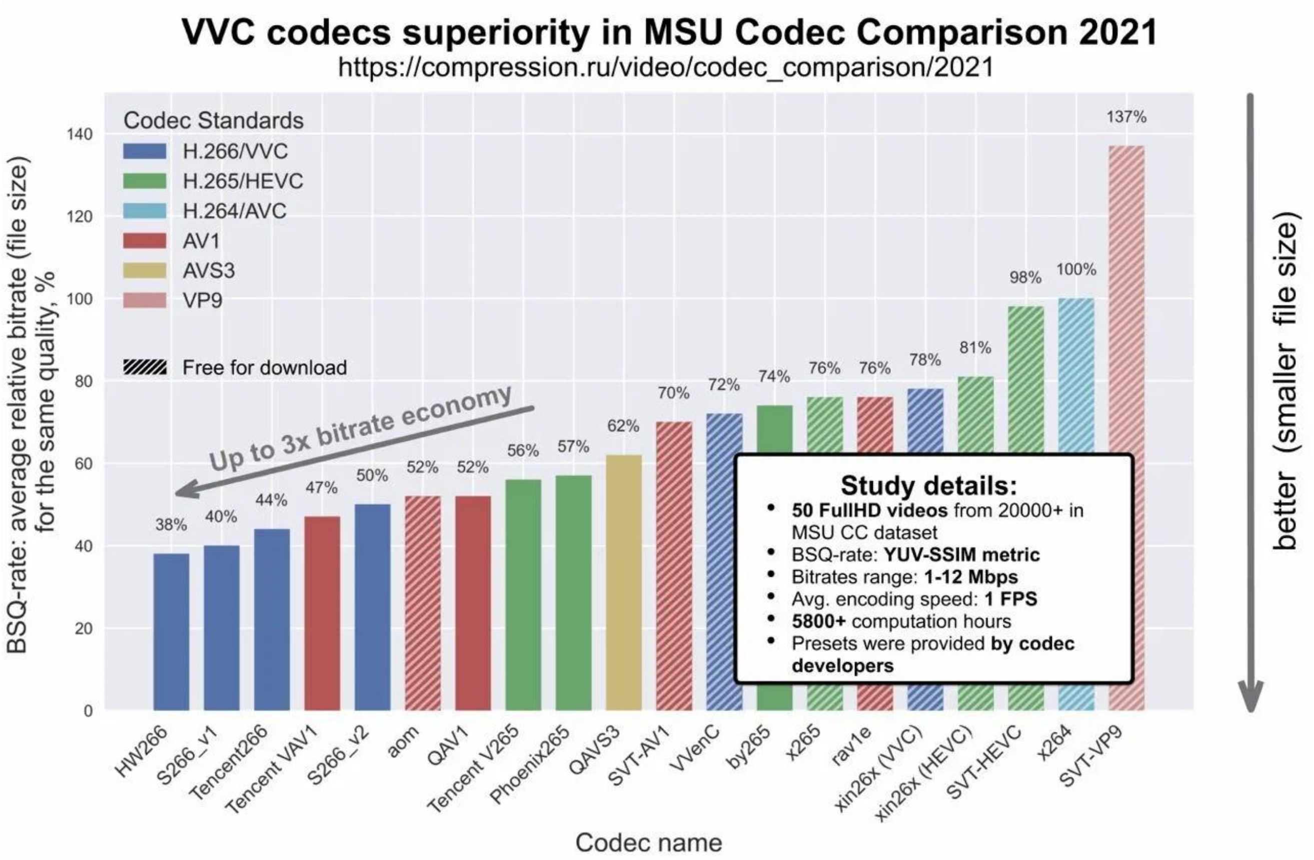
READ MORE

Aidan Go

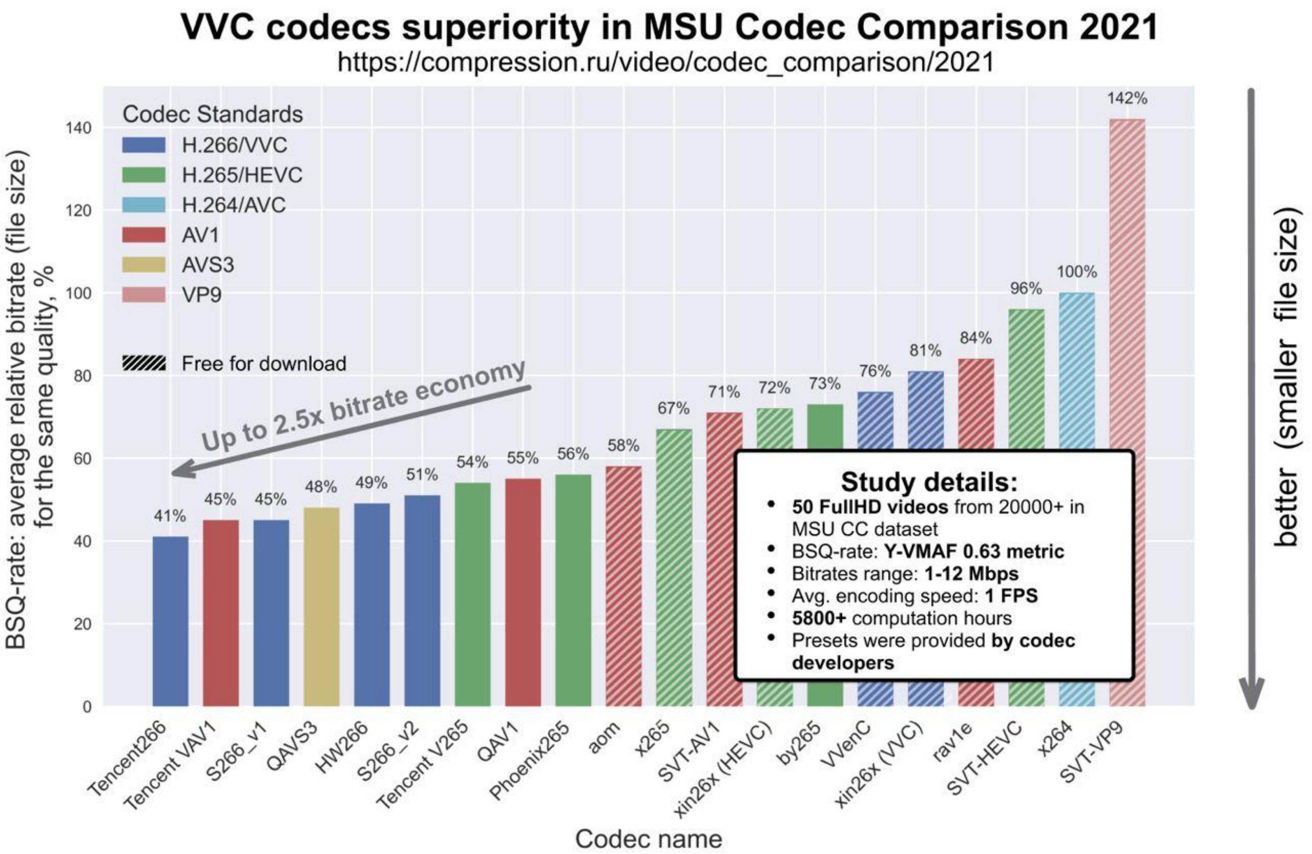
Cohere

READ MOR

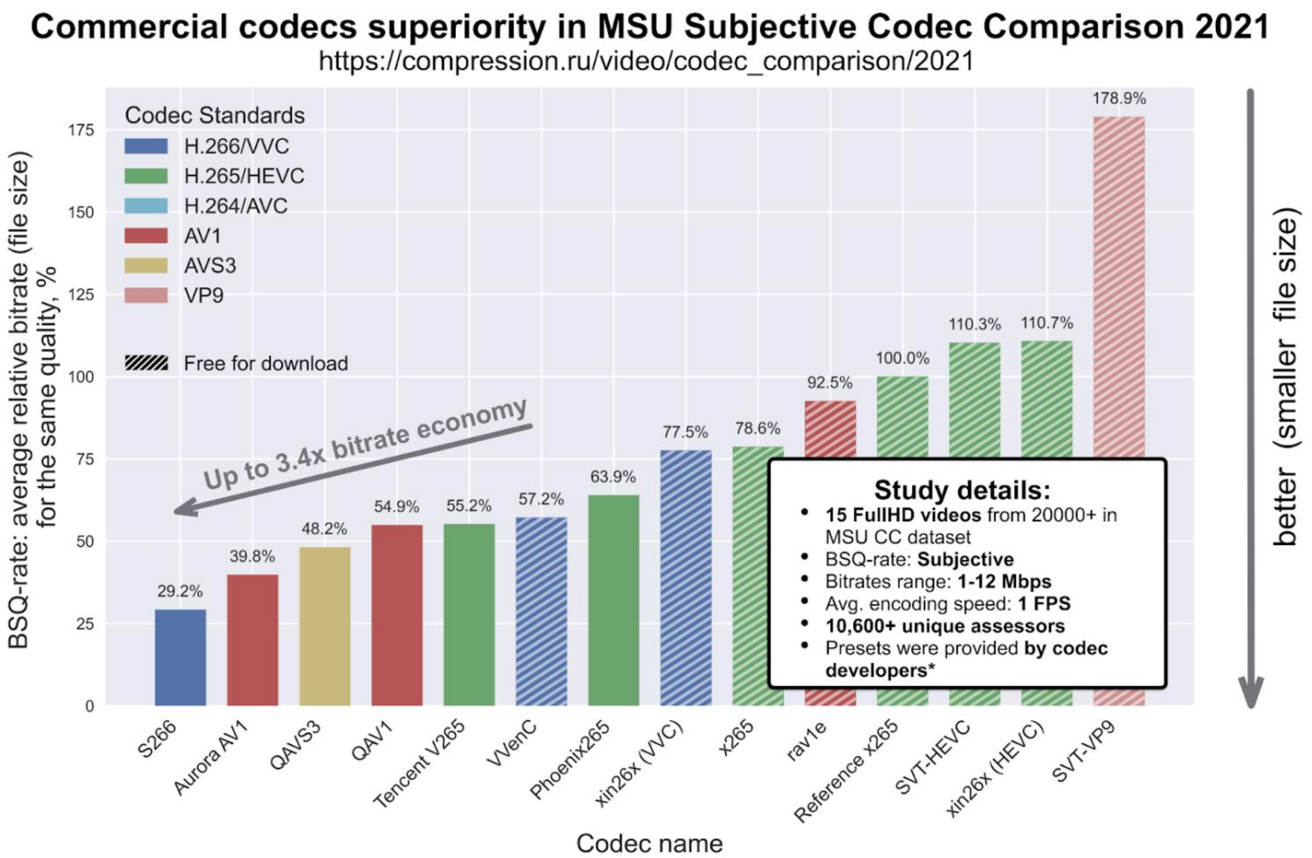
SSIM comparison



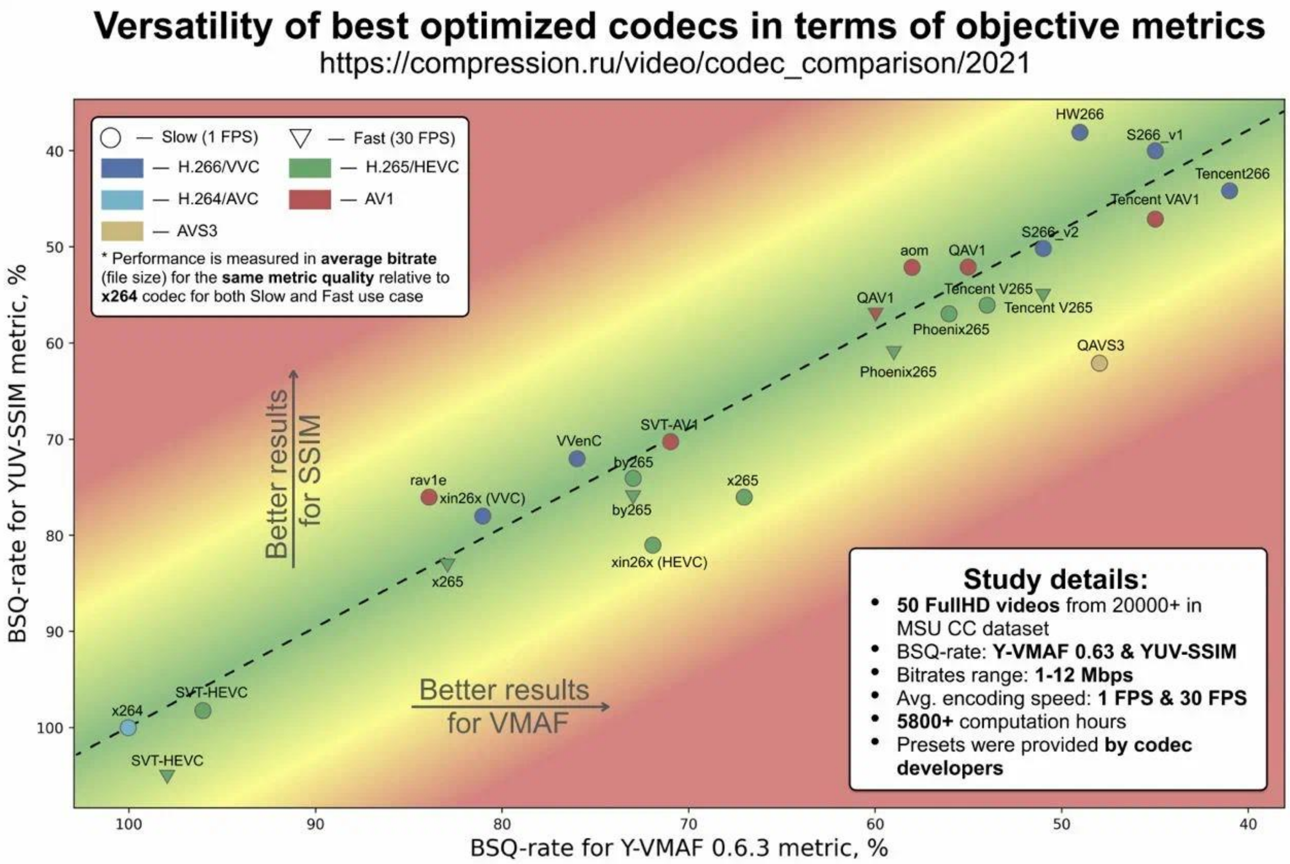
VMAF comparison



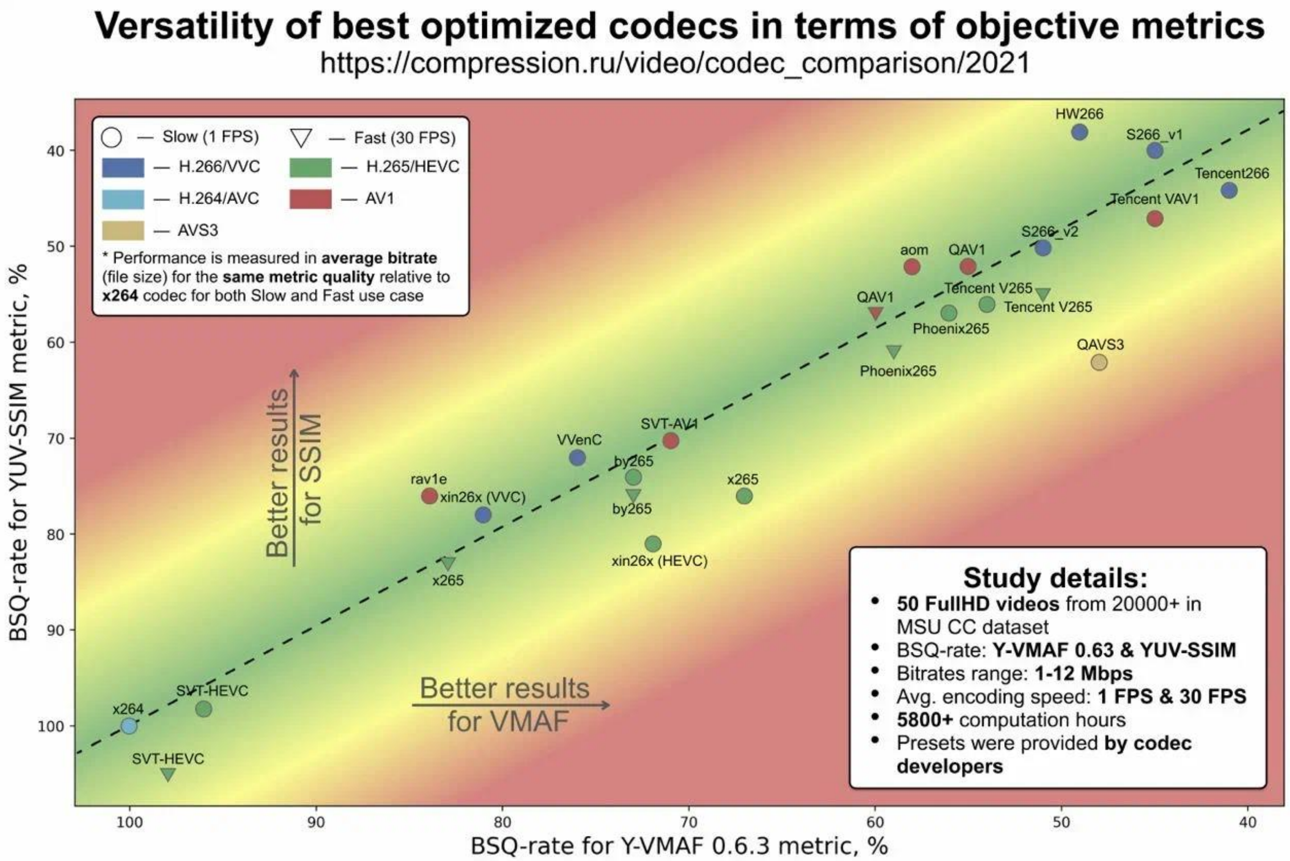
Subjective comparison



Different encoders optimise different metrics (1)

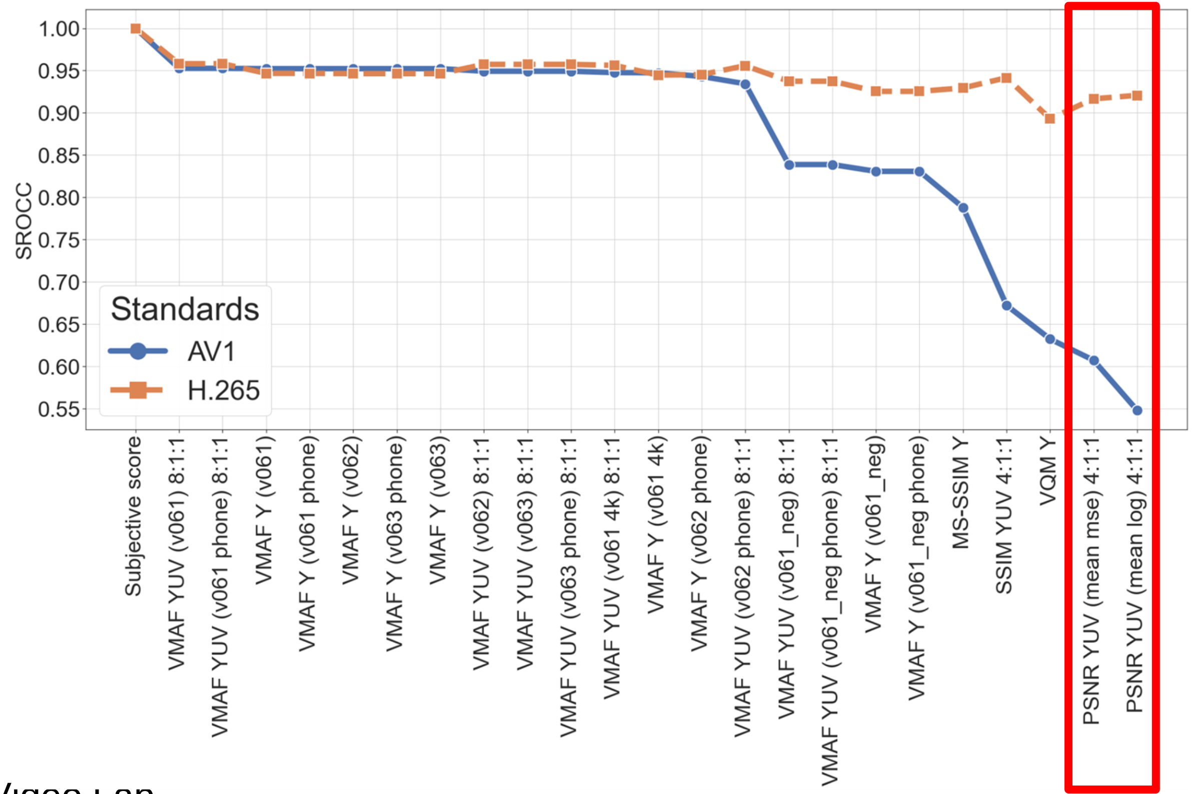


Different encoders optimise different metrics (2)



Video Quality Metrics Benchmark

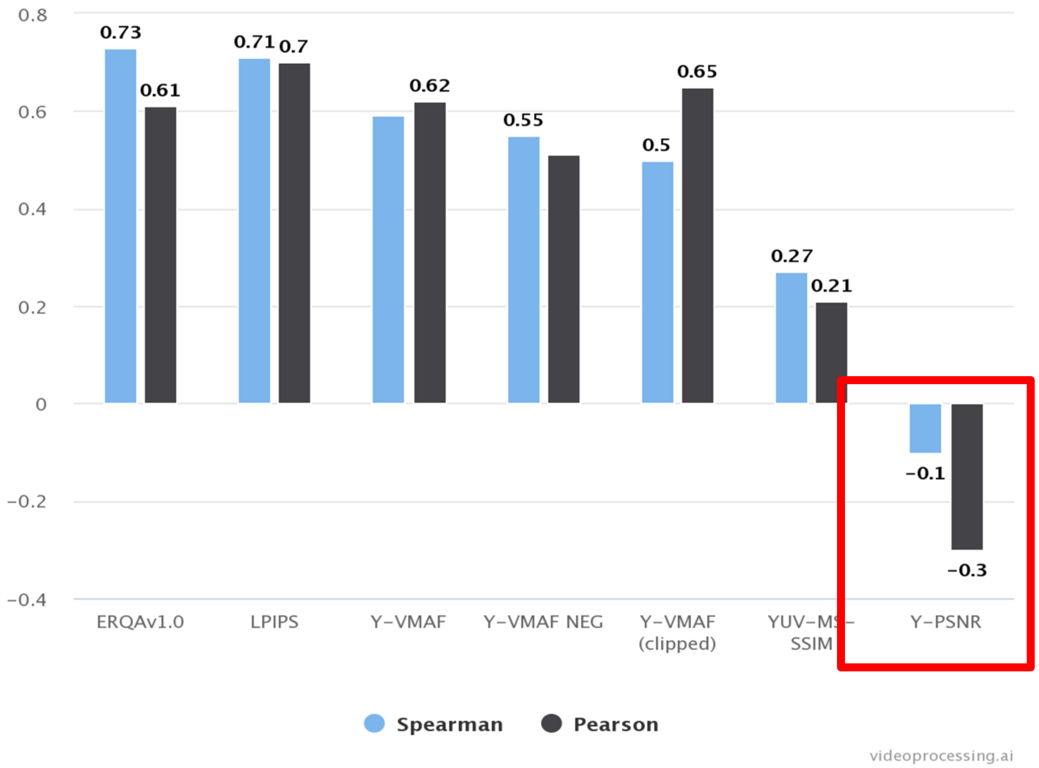
Dramatic PSNR degradation on AV1 vs H.265



SR quality metrics

Correlations of metrics with subjective assessments

BSQ-rate was calculated via subjective results extrapolation using the most similar objective metric



Metric benchmarks

Biggest dataset

Dataset	Original videos	Average duration (s)	Distorted videos	Distortion	Subjective framework	Subjects	Answers
MCL-JCV (2016) [42]	30	5	1,560	Compression	In-lab	150	78K
VideoSet (2017) [43]	220	5	45,760	Compression	In-lab	800	-
UGC-VIDEO (2020) [25]	50	> 10	550	Compression	In-lab	30	16.5K
CVD-2014 (2014) [36]	5	10-25	234	In-capture	In-lab	210	-
LIVE-Qualcomm (2016) [14]	54	15	208	In-capture	In-lab	39	8.1K
GamingVideoSET (2018) [9]	24	30	576	Compression	In-lab	25	-
KUGVD (2019) [8]	6	30	144	Compression	In-lab	17	-
KoNViD-1k (2017) [16]	1,200	8	1,200	In-the-wild	Crowdsource	642	205K
LIVE-VQC (2018) [39]	585	10	585	In-the-wild	Crowdsource	4,776	205K
YouTube-UGC (2019) [44]	1,500	20	1,500	In-the-wild	Crowdsource	>8,000	600K
LSVQ (2020) [50]	39,075	5-12	39,075	In-the-wild	Crowdsource	6,284	5M
MSU Compression Dataset (2022)	36	10, 15	2,486	Compression (83 codecs)	Crowdsource	10,800	766K

Comparison of quality metrics

Video compression dataset and benchmark of learning-based video-quality metrics

Anastasia Antsiferova[1,2], **Sergey Lavrushkin**[1,2], **Maksim Smirnov**[3],
Alexander Gushchin[3], **Dmitriy Vatolin**[1,2,3]
ISP RAS Research Center for Trusted Artificial Intelligence[1]
MSU Institute for Artificial Intelligence[2]
Lomonosov Moscow State University[3]
{aantsiferova, sergey.lavrushkin, maxim.smirnov.2025, alexander.gushchin, dmitriy}@graphics.cs.msu.ru

Abstract

Video-quality measurement is a critical task in video processing. Nowadays, many implementations of new encoding standards — such as AV1, VVC, and LCEVC — use deep-learning-based decoding algorithms with perceptual metrics that serve as optimization objectives. But investigations of the performance of modern video- and image-quality metrics commonly employ videos compressed using older standards, such as AVC. In this paper, we present a new benchmark for video-quality metrics that evaluates video compression. It is based on a new dataset consisting of about 2,500 streams encoded using different standards, including AVC, HEVC, AV1, VP9, and VVC. Subjective scores were collected using crowdsourced pairwise comparisons. The list of evaluated metrics includes recent ones based on machine learning and neural networks. The results demonstrate that new no-reference metrics exhibit high correlation with subjective quality and approach the capability of top full-reference metrics.

439 **Video compression dataset and benchmark of learning-based video-quality metrics**

Download PDF

Anastasia Antsiferova, Sergey Lavrushkin, Maksim Smirnov, Aleksandr Gushchin, Dmitriy Vatolin

Show details

5 Reviews Submitted

Reviewer Wdzx: Rating: 7: Good paper, accept / Confidence: 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature
Read Review

Reviewer YC5y: Rating: 8: Top 50% of accepted papers, clear accept / Confidence: 3: The reviewer is fairly confident that the evaluation is correct
Read Review

Reviewer q4yp: Rating: 7: Good paper, accept / Confidence: 3: The reviewer is fairly confident that the evaluation is correct
Read Review

Reviewer oqgs: Rating: 7: Good paper, accept / Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper
Read Review

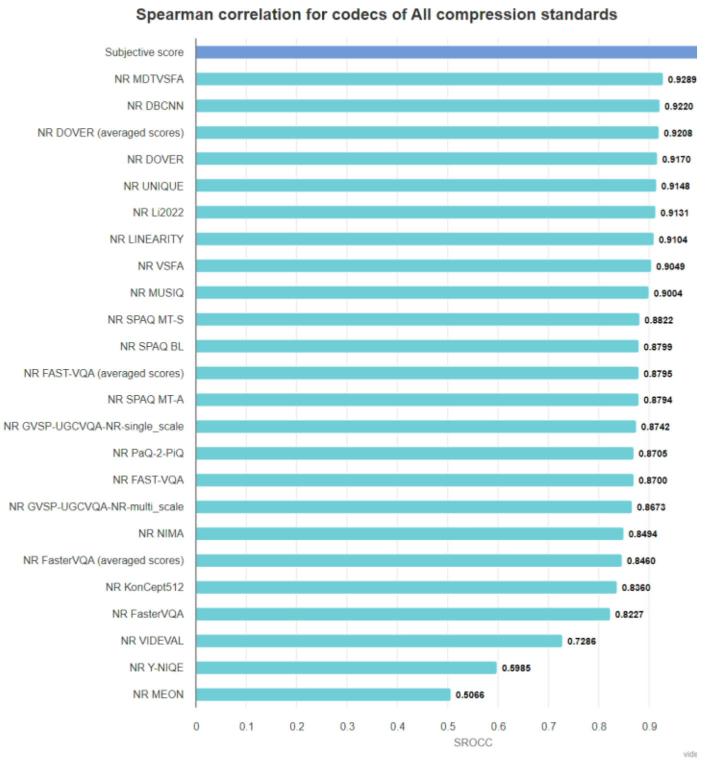
Reviewer Wvxs: Rating: 6: Marginally above acceptance threshold / Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct
Read Review

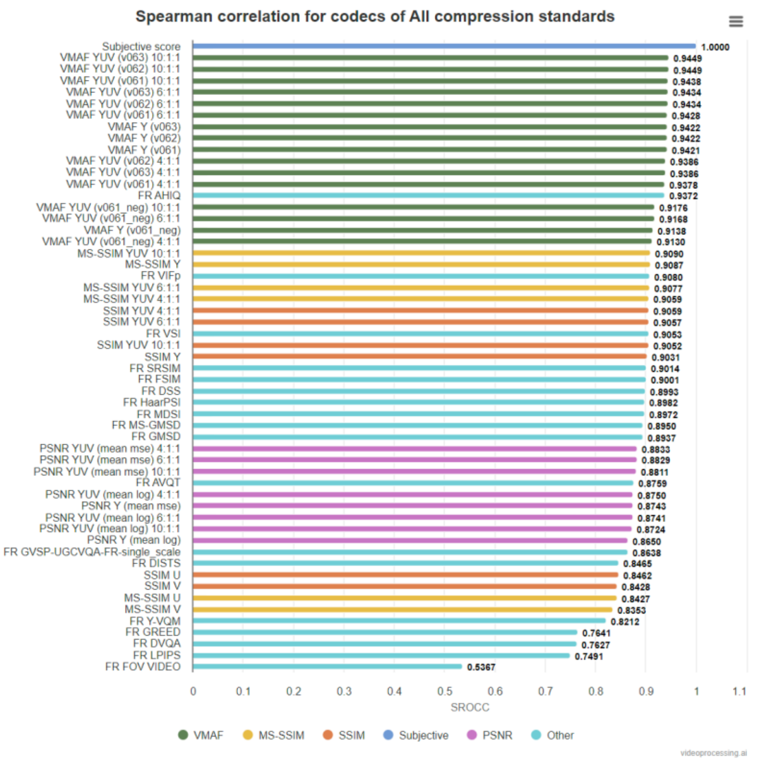
Average Rating: 7 (Min: 6, Max: 8)
Average Confidence: 3.4 (Min: 2, Max: 5)

Recommendation:
Accept
Read

Metrics benchmark for video compression

- **40** different video codecs of 10 compression standards
- **2500+** compressed streams
- **780.000+** subjective scores
- **10.000+** viewers
- open and hidden parts





Video Quality Metrics Benchmark

Community reaction (1)

Alan Bovik

Bovik received a Primetime Emmy Award in 2015 for his development of perception-based video quality measurement tools that are now standards in television production. He also received a Technology and Engineering Emmy Award in 2021 for the "development of perceptual metrics for video encoding optimization."

"I saw this with great interest. I notice that (of course) the database is all compression distortions, and the trainable models (which have been trained on other distortions, like UGC), have not been retrained on the MSU data."

	All	Since 2017
Citations	131584	65258
h-index	126	79
i10-index	537	324

Video Quality Metrics Benchmark

Community reaction (2)

“We really would like to contribute our quality measures to MSU. Actually, I am closely following what MSU is doing. You have done many things that have changed the society.”

“Thanks for bring your wonderful benchmark to my attention, on which I will definitely test our VQA models.”

“I am from the QoE team working on video quality assessment on UGC at ByteDance inc. Could you kindly share the public samples from the MSU VQA benchmark dataset with us? We would like to submit our no-reference quality assessment method.

Video Quality Metrics Benchmark

Community reaction (3)

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

“Thanks for pointing us to this benchmark. We look forward to participating in this study.”

深圳大学
SHENZHEN UNIVERSITY

“Thank you for your email, your work sounds interesting.”

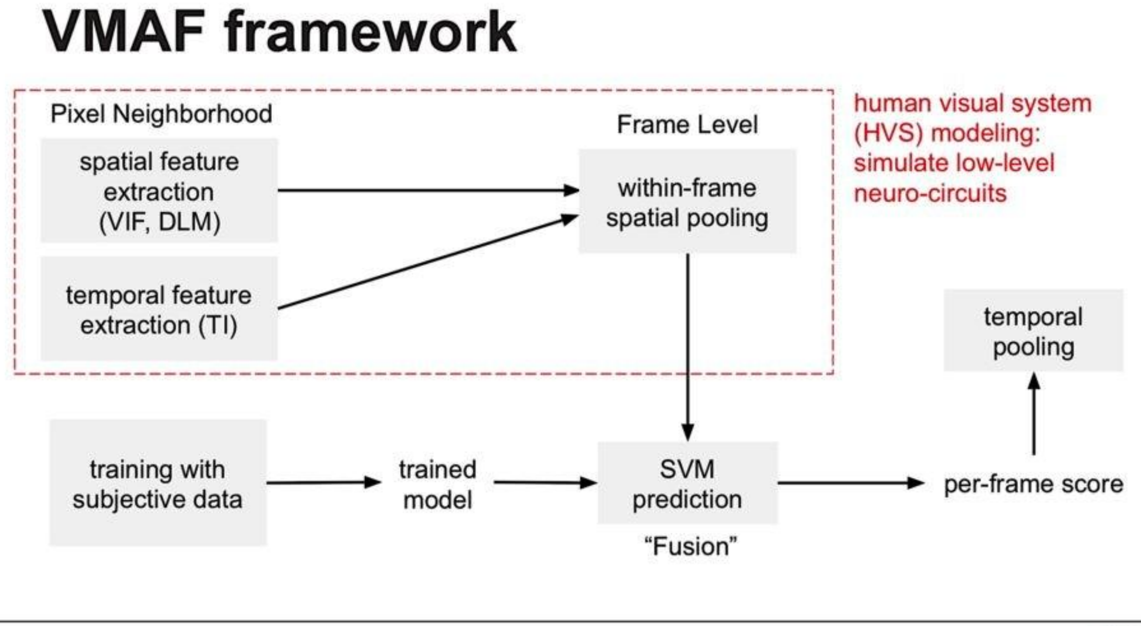
YouTube

“Very nice benchmark. Appreciate your team's contribution in video quality assessment.
Our team (YouTube Media Algorithms) has built multiple VQA metrics, and we also plan to open source our models. Could you please share the dataset for us to do some preliminary analysis? Thank you very much.”

Hacking VMAF: adversarial attack at VMAF NEG and improving VMAF

Introduction

VMAF – the most popular modern video quality metric



https://thebroadcastknowledge.com/2020/11/19/videomeasuring-video-quality-with-vmaf-why-you-should-care/

Hacking VMAF

Impact of VMAF stability research

1. Our team revealed VMAF vulnerability

Hacking VMAF with Video Color and Contrast Distortion

A. Zvezdakova[1], S. Zvezdakov[1], D. Kulikov[1,2], D. Vatolin[1]
azvezdakova@graphics.cs.msu.ru|szvezdakov@graphics.cs.msu.ru|dkulikov@graphics.cs.msu.ru|dmitriy@graphics.cs.msu.ru
[1]Lomonosov Moscow State University, Moscow, Russia;
[2]Dubna State University, Dubna, Russia

2. Jan Ozer, Streaming Media leading expert, reproduced it

STREAMING LEARNING CENTER

About Jan Ozer | Courses | Books | Help | Blog | Contact

VMAF is Hackable: What Now?

3. Our VMAF tuning integrated to the AV1 official code

aomedia / aom / master / . / av1 / encoder / **tune_vmaf.c**

```
    commit  615dc24579d531cb3a2c9627ab25a3026f9e2b47        [log] [tgz]
    author  sdeng <sdeng@google.com>                        Tue Feb 04
 committer  Sai Deng <sdeng@google.com>                     Thu Feb 06
      tree  1cb2a23b5f66527222bb3aca7b5d07594b1392ea
    parent  0de21d3cf211283deb87ac20174148d14abbc9de [diff]

Add new mode tune=vmaf

This mode enables block based video pre-processing, RDO Lagrange
multiplier scaling using VMAF and VMAF motion based Q-index adjustment
to maximize encoder's VMAF performance.
```

4. Netflix released new more stable VMAF version

On VMAF's property in the presence of image enhancement operations

Zhi Li, Video Algorithms, Netflix
July 13, 2020

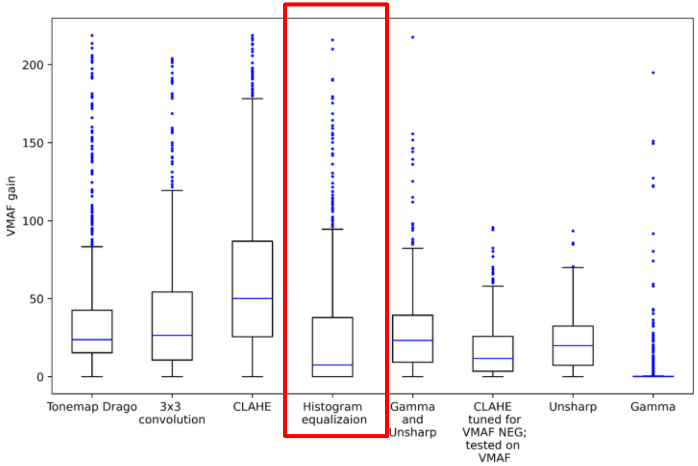
Jan Ozer "VMAF is Hackable: What Now?"
https://streaminglearningcenter.com/blogs/vmaf-is-hackable-what-now.html

https://docs.google.com/document/d/1dJczEhXO0MZjBSNyKmd3ARiCTdFVMNPBykH4_HMPoyY/edit#heading=h.oaikhnw46pw5

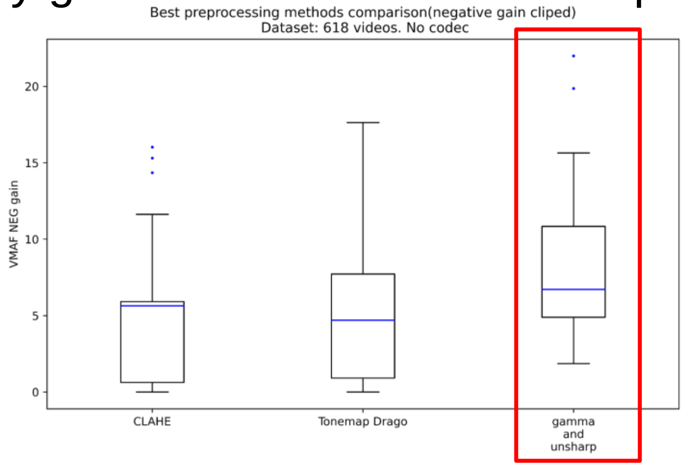
Results

Black-box hacking VMAF and VMAF NEG

- Average VMAF increase is about 50 by CLAHE preprocessing
- Average VMAF NEG increase is 7 by gamma correction + unsharp masking



VMAF gain by different preprocessing



VMAF NEG gain by different preprocessing

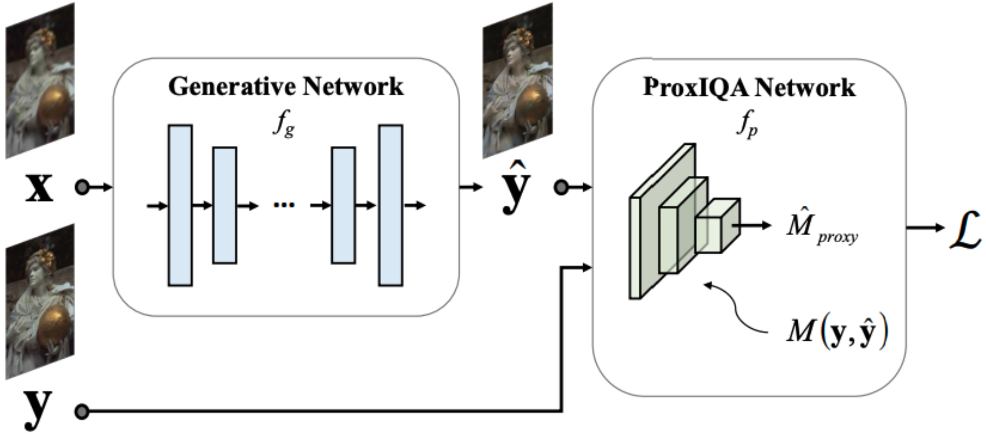
Siniukov M. et al., “Hacking VMAF and VMAF NEG: vulnerability to different preprocessing methods”, in *AICCC'21: 2021 4th Artificial Intelligence and Cloud Computing Conference*, 2021.

Hacking VMAF: adversarial attack using distillation

Adversarial attack using distillation

Base method: VMAF-aware neural compression

Compressor network
is learned using differentiable approximation of VMAF (ProxIQA Network)

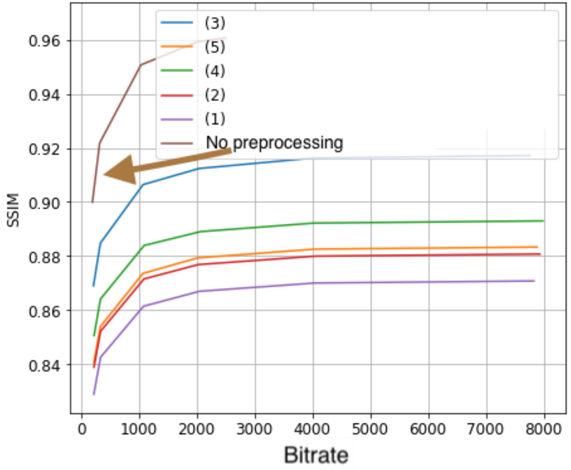


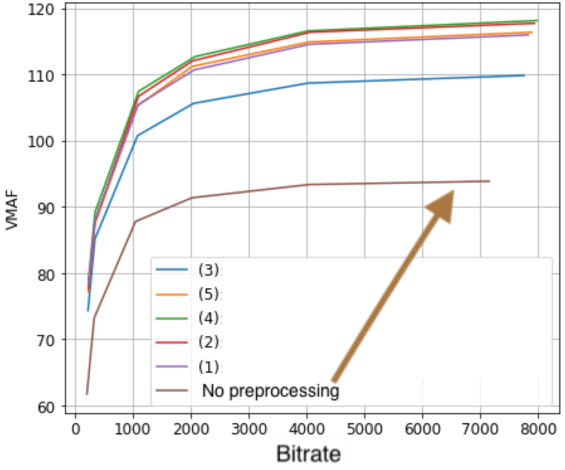
Adversarial attack using distillation

Our method: video preprocessing for VMAF increase.

VMAF gain persists after compression with common codecs (RD-curves for x264 for different model versions)

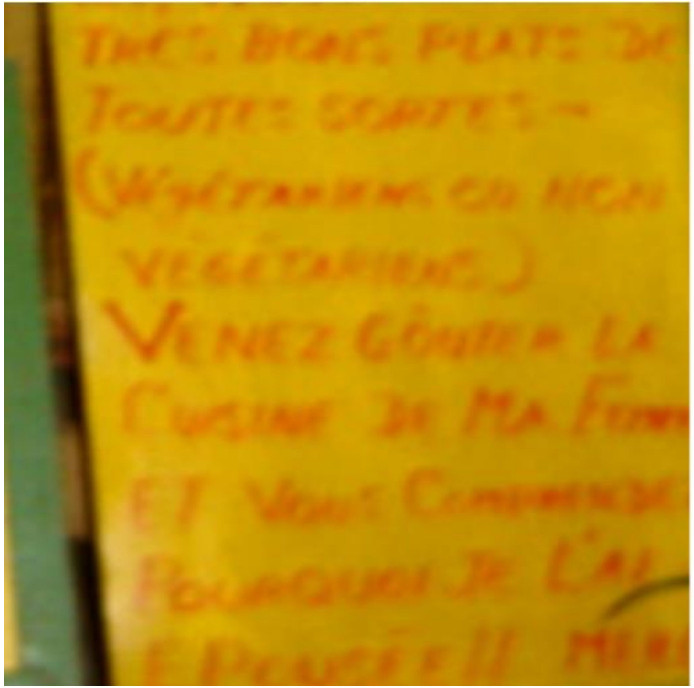
On average, VMAF increased on 21%





Adversarial attack using distillation

Example



Source image

(VMAF: 97.4)

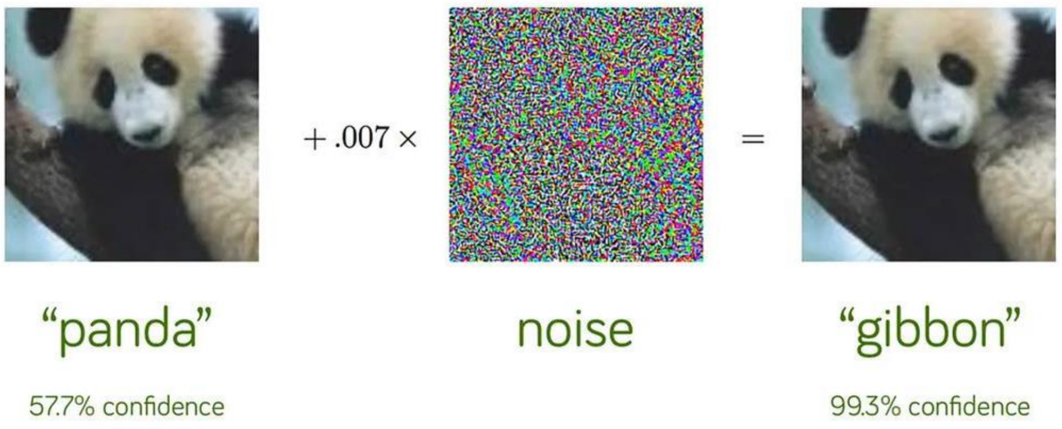
Preprocessed image

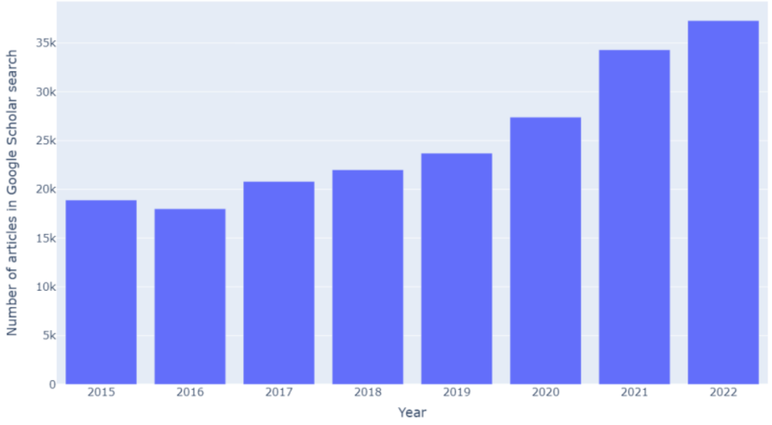
(VMAF: 160.4)

Adversarial attacks on image and video quality assessment methods

Adversarial attacks

Adversarial attacks – preprocessing of model input data forcing it to make incorrect predictions





Number of papers on Google Scholar by "adversarial attacks" request

I. Goodfellow et al., "Explaining and Harnessing Adversarial Examples", in ICLR, 2014

Adversarial attacks on metrics

Adversarial attacks on metrics – preprocessing of metrics input data to increase or decrease its values without corresponding change in visual quality

Changing RANK-IQA values with Korhonen et al. attack, image from NIPS2017 dataset

Types of adversarial attacks

Non-targeted

Adversarial attack

“Cat”
Confidence 88%

“Guacamole”
Confidence 90%

Targeted

Adversarial attack

Target class: “Airplane”

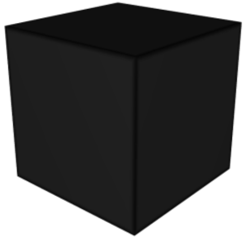
“Cat”
Confidence 88%

“Airplane”
Confidence 99%

Types of adversarial attacks

“Black Box”

Input

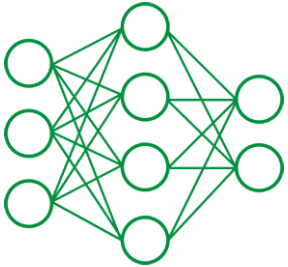


Output

Model architecture unknown

“White Box”

Input



Output

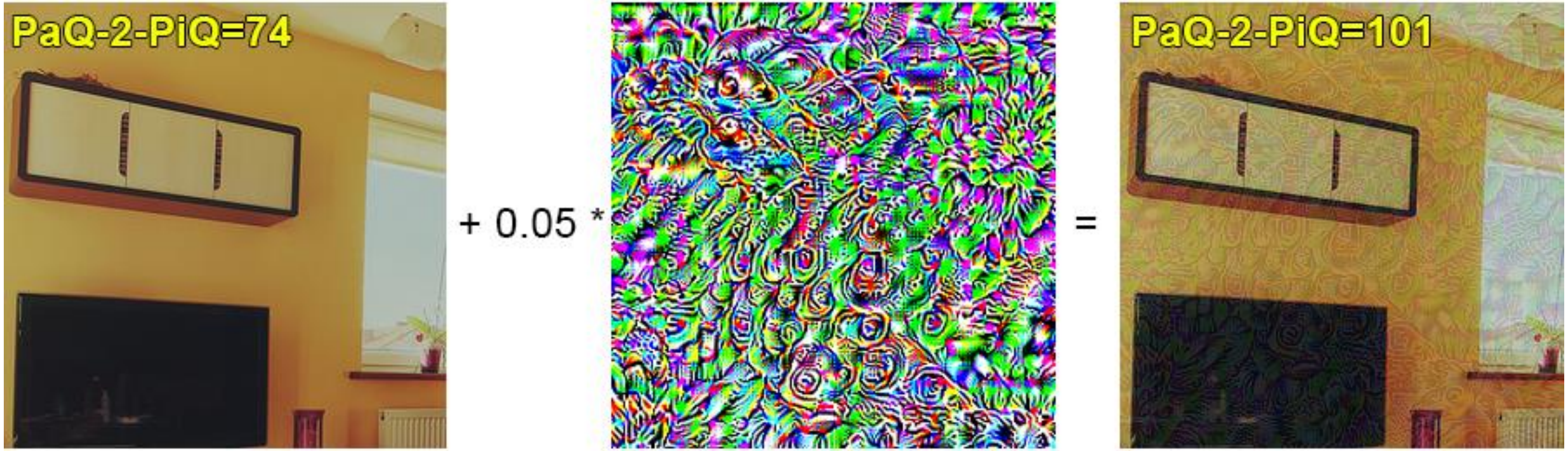
Model architecture and
weights are available

Gradients w.r.t. input

Proposed attacks on metrics
Universal adversarial perturbation

Trainable perturbation that increases attacked metric values when added to any image

- Low computational complexity
- Independent of input
- Quite noticeable distortions



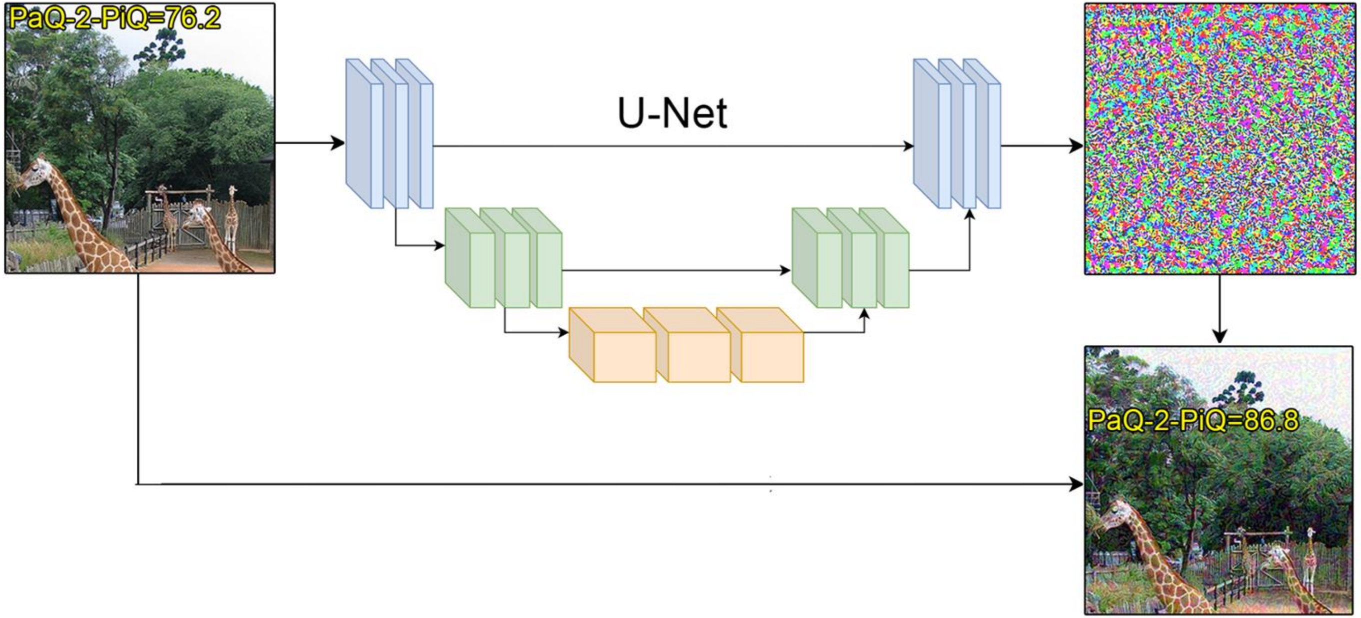
Attacking PaQ-2-PiQ with UAP

Proposed attacks on metrics

CNN attack

Trainable CNN that generates perturbation for input image

- Medium computational complexity
- Distortions are often introduced in low-frequency regions

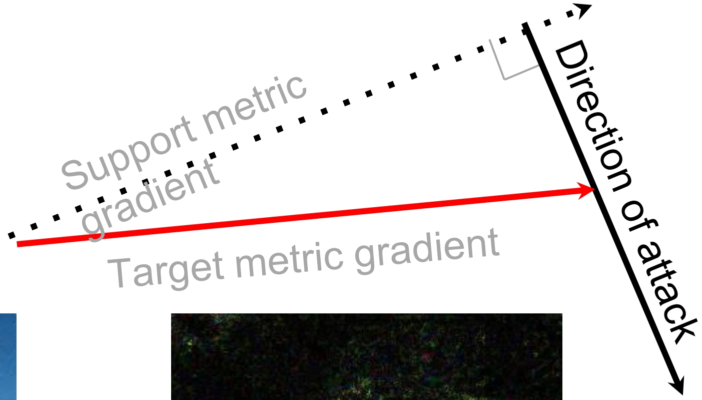


Proposed attacks on metrics

Attack with metrics preservation

Uses target metric gradients, projected on subspace orthogonal to the gradients of preserved metrics, that is built with Gram-Schmidt orthogonalization process

- Able to preserve arbitrary number of metrics
- Need to calculate all metrics gradients



Target image,
PSNR=31, SSIM=0.89,
SPAQ=83.4, **PAQ2PIQ=76.9**

Attacked image,
PSNR=31, SSIM=0.88,
SPAQ=82.1, **PAQ2PIQ=109.09**

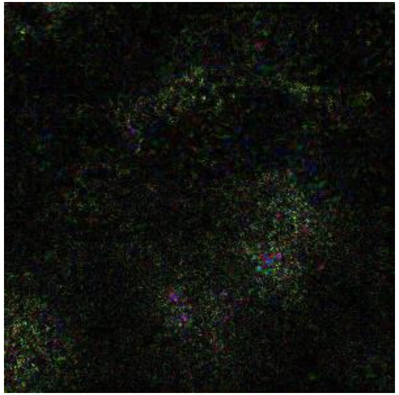
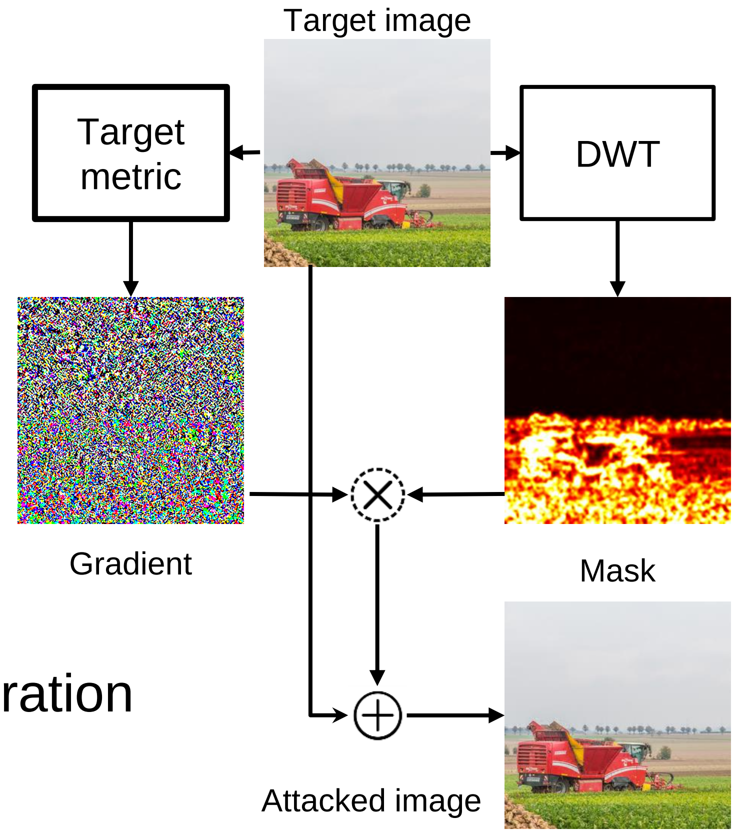


Image difference increased by
5 times

Proposed attacks on metrics
Attack with DWT

Hide adversarial perturbation in high frequency regions using DWT as a mask

- Almost invisible
- Can be used with other attacks
- Decreases attack gains
- Additional computations



One iteration

Current state

- **All tested** learning-based metrics **are vulnerable** to adversarial attacks (**big problem!**), so:
  - **many benchmarks** of image and video processing algorithms **will be compromised**
  - **vulnerable metrics** in loss can cause **fake results**
- **Correlation of metrics more important than robustness** during their development
- There are no full-fledged benchmarks of metrics robustness (we are trying to fix this)

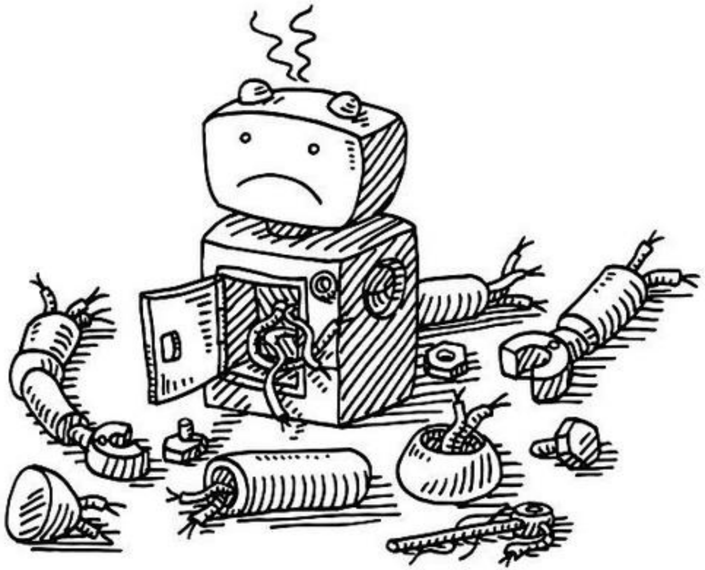
Generally

- **We faced our first attack in 2018:**
  - 2016 — Netflix suggests VMAF,
  - 2017 — we test it,
  - 2018 — we add VMAF into MSU Codec Comparison leaderboard and **immediately** detect attack on VMAF
- **All CVPR NTIRE Challenges face tuning for metrics**
- **Codec developers remove their codecs from subjective comparison** in MSU CC
- **Authors don't want to publish methods in our benchmarks** (leaderboard is by subjective comparison)

3Dvideo 22 ноября в 11:02

Хакинг метрик качества видео или как с приходом ИИ все становится намного сложнее

Программирование*, Сжатие данных*, Машинное обучение*, Научно-популярное, Искусственный интеллект



Сейчас модно писать, что ML пришел туда и все стало отлично, DL пришел сюда и все стало замечательно. А к кому-то пришел сам AI, и там все стало просто сказочно! Возможна ли

We have robust classification/object detection but **no robust quality measurement**

200+ robust models in 3 existing benchmarks on object classification, object detection robustness

RobustML

Only **1 so-called "robust" image/video quality assessment metric**

… which we managed to attack!


Proceedings > AICCC '21 > Hacking VMAF and VMAF NEG: Vulnerability to Different Preprocessing Methods

RESEARCH-ARTICLE

Hacking VMAF and VMAF NEG: Vulnerability to Different Preprocessing Methods

Authors: Maksim Siniukov, Anastasia Antsiferova, Dmitriy Kulikov, Dmitriy Vatolin Authors Info & Claims

https://dl.acm.org/doi/abs/10.1145/3508259.3508272

Our progress (Nov 2022)

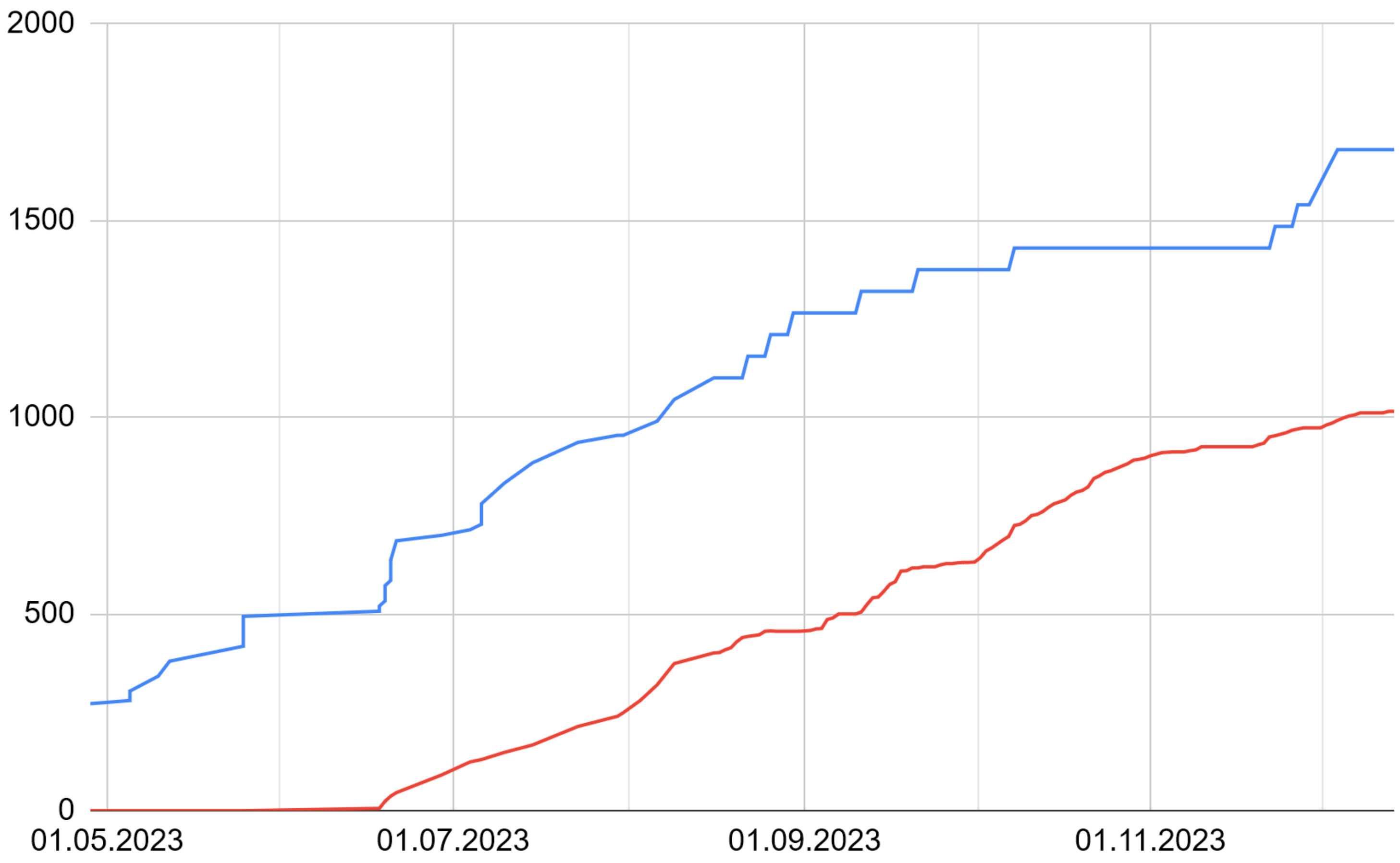
Current hacking rate:

10-20 metrics/method/month

Our goal (Nov 2022)

Hacking of 100-200 metrics/method/month

Метрико-атаки и посчитанные метрико-атаки

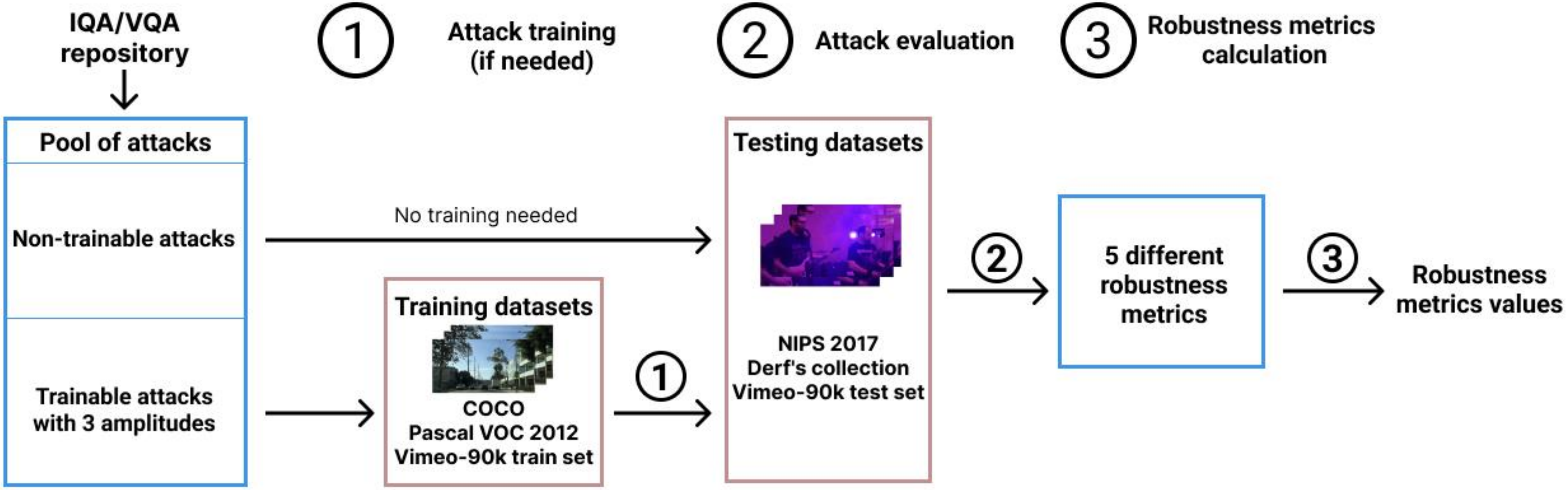


Interim conclusions

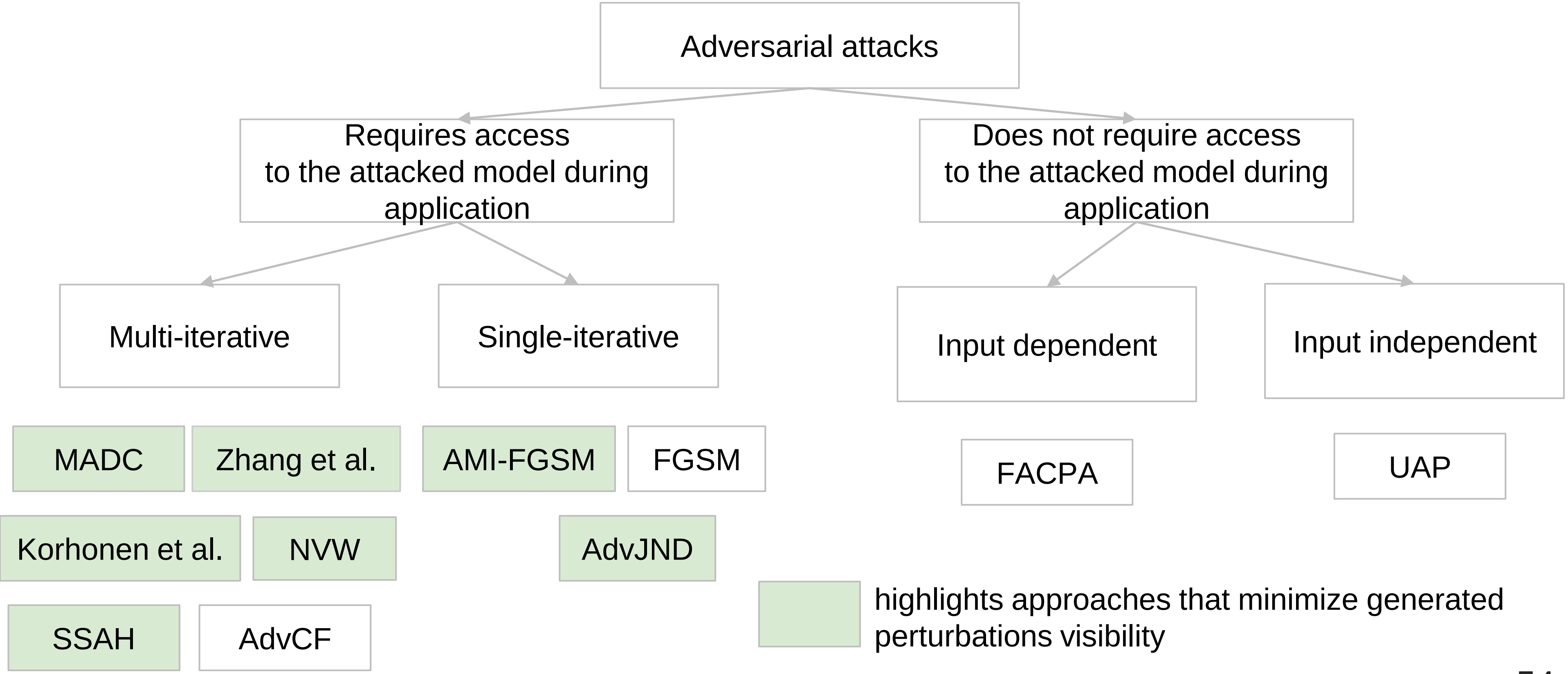
- Up to **<u>28 NVIDIA A100</u>** used now for benchmark
  - It won't be easy to repeat the result
- **Need <u>more computational power</u>**: we're not tuning out attack parameters enough
  - Some of the results will change
- **Work on the <u>defenses has just only begun</u>**
  - There aren't enough people
- **A <u>noticeable improvement in algorithms is clearly possible</u> in this direction** (JPEG AI, SR etc…)

Proposed metrics robustness benchmark

Benchmark pipeline



Pool of attacks

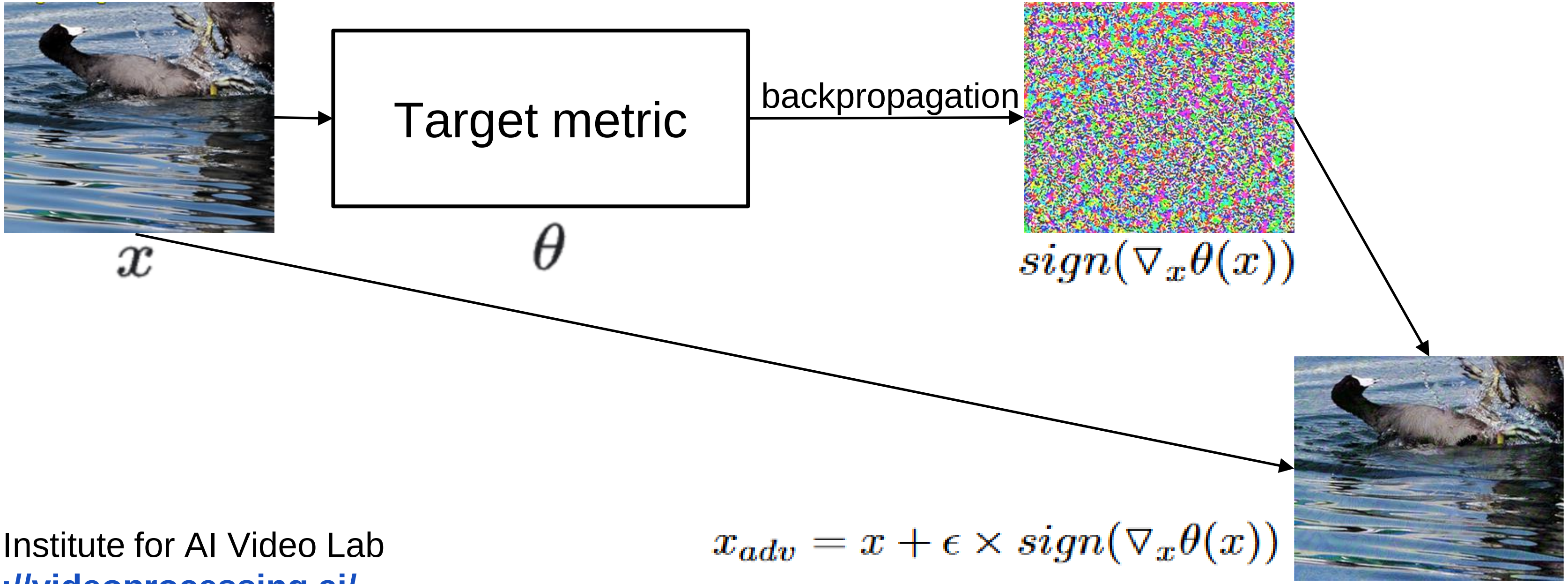


Example: FGSM

Classifiers **vs** **Metrics**

$$loss = -CrossEntropyLoss(target\_class, \theta(x))$$

$$loss = 1 - \frac{\theta(x)}{range(\theta)}$$



Pool of metrics

FR metrics. We implemented 32 metrics, will expand to 50+

MAD	LPIPS	DISTS	AHIQ	SR-SIM	MS-SSIM	PieAPP
VMAF	SWIN-IQA	ST-LPIPS	Brisque	HAAR-PSI	Conformer-BNS	CKDN
NLPD	MS-GMSD	MR-Perceptual	MDSI	ASNA-MACS	VTAMIQ	VIF
GMSD	DSS	IW-SSIM	IQT	FSIM	CW-SSIM	CVRKD-IQA

NR metrics. We implemented 23 metrics, will expand to 50+

MANIQA	PAQ2PIQ	NIMA	Koncept	CLIP-IQA	WSP	RANK-IQA
MUSIQ	DBCNN	TRES	Linearity	HYPER-IQA	VSFA	MDTVSFA
FPR	META-IQA	SPAQ	NIQE	Brisque	Koniq++	LIQE

Robustness evaluation

- Absolute and Relative gain

$$Abs.gain = \frac{1}{n}\sum_{i=1}^{n}(f(x'_i) - f(x_i)), \quad Rel.gain = \frac{1}{n}\sum_{i=1}^{n}\frac{f(x'_i) - f(x_i)}{f(x_i) + 1}$$

- Robustness score*

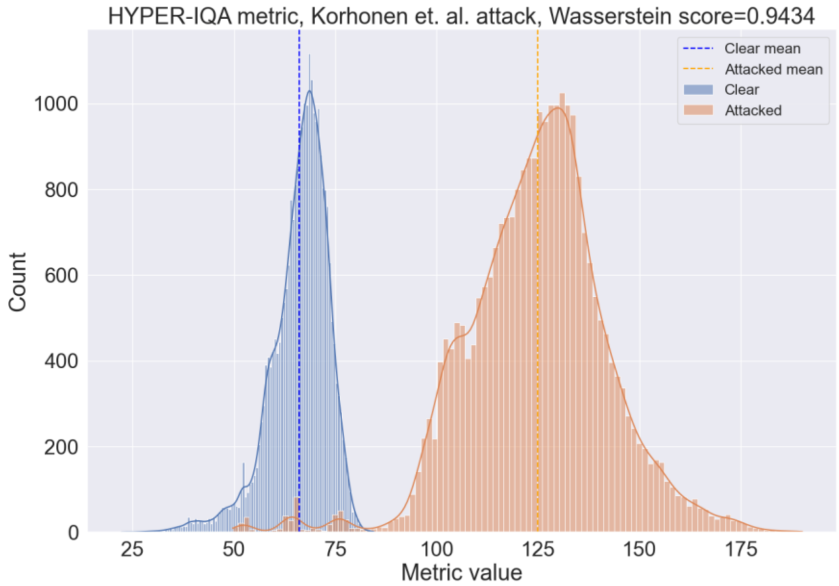
$$R_{score} = \frac{1}{n}\sum_{i=1}^{n} log_{10}\left(\frac{max\{\beta_1 - f(x'_i), f(x_i) - \beta_2\}}{|f(x'_i) - f(x_i)|}\right) \quad \beta_1 = 1 \quad \beta_2 = 0$$

x_i – target image x'_i – attacked image

$f(\cdot)$ – target metric

*Zhang et al., Perceptual attacks of no-reference image quality models with human-in-the-loop, 2022

Robustness evaluation



- Wasserstein score

$$W_{score} = W_1(\hat{P}, \hat{Q}) \cdot sign(\bar{x}_{\hat{P}} - \bar{x}_{\hat{Q}})$$

$$W_1(\hat{P}, \hat{Q}) = \inf_{\gamma \in \Gamma(\hat{P}, \hat{Q})} \int_{\mathbb{R}^2} |x - y| d\gamma(x, y) =$$

$$= \int_{-\infty}^{\infty} |\hat{F}_{\hat{P}}(x) - \hat{F}_{\hat{Q}}(x)| dx$$

- Energy Distance score

$$E_{score} = E(\hat{P}, \hat{Q}) \cdot sign(\bar{x}_{\hat{P}} - \bar{x}_{\hat{Q}})$$

$$E(\hat{P}, \hat{Q}) = (2 \cdot \int_{-\infty}^{\infty} (\hat{F}_{\hat{P}}(x) - \hat{F}_{\hat{Q}}(x))^2 dx)^{\frac{1}{2}}$$

$\hat{P}$ $\hat{Q}$ – empirical distributions of metric values before and after the attack

$\hat{F}_{\hat{P}}(x)$ $\hat{F}_{\hat{Q}}(x)$ – respective empirical Cumulative Distribution Functions

$\bar{x}_{\hat{P}}$ $\bar{x}_{\hat{Q}}$ – respective sample means

Datasets

Dataset	Type	Number of samples	Resolution
Training datasets			
COCO	Images	300,000	640×480
Pascal VOC 2012	Images	11,530	500×333
Vimeo-90k Train set	Triplet of images	2,001	448×256
Testing datasets			
NIPS 2017: Adv.Learning Devel.Set	Images	1,000	299×299
Derf’s collection	Videos	24 (10,000)	1920×1080
Vimeo-90k Test set	Triplet of images	11,346	448×256

Attack propagation in videos

Attacking each frame individually is ineffective:

- Computationally expensive
- Resource-intensive
- No temporal stability

Attack propagation in videos

Proposed solution:

- Attack only several keyframes at equal intervals
- Use keyframe results to generate intermediate perturbations:
  - Linear interpolation
  - Motion estimation + linear interpolation
  - Motion estimation + attacks with a small number of attack iterations

Attack propagation in videos

Linear interpolation, 0.45 s/frame, mean Rel. Gain=0.73

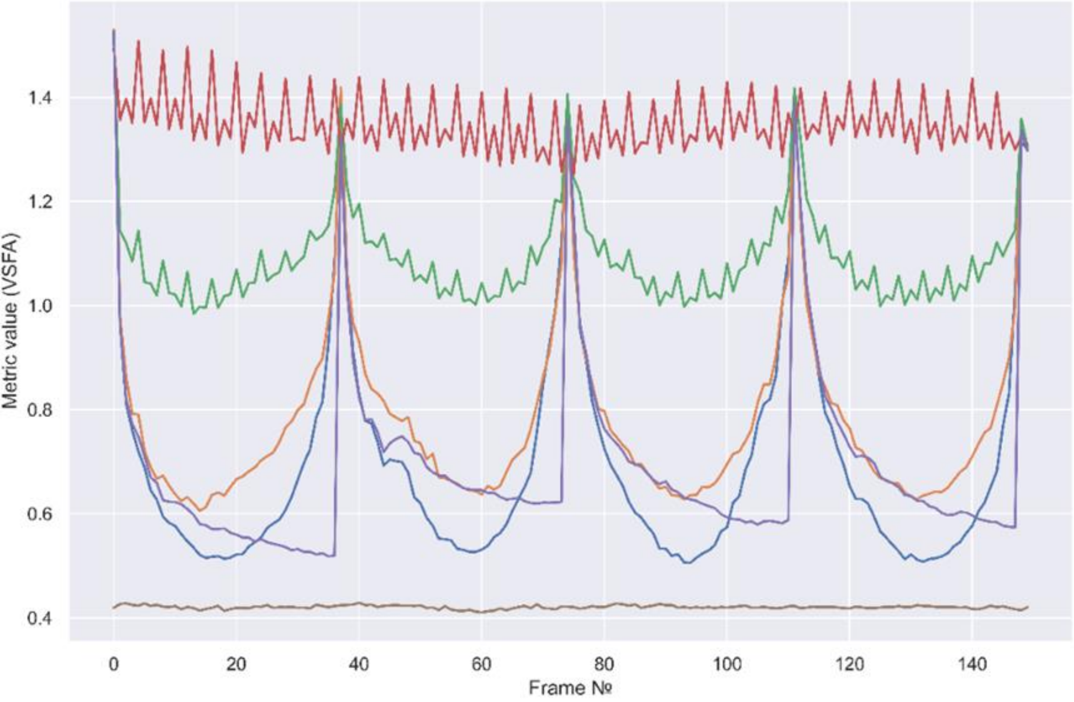
ME interpolation, 0.77 s/frame, mean Rel. Gain=0.92

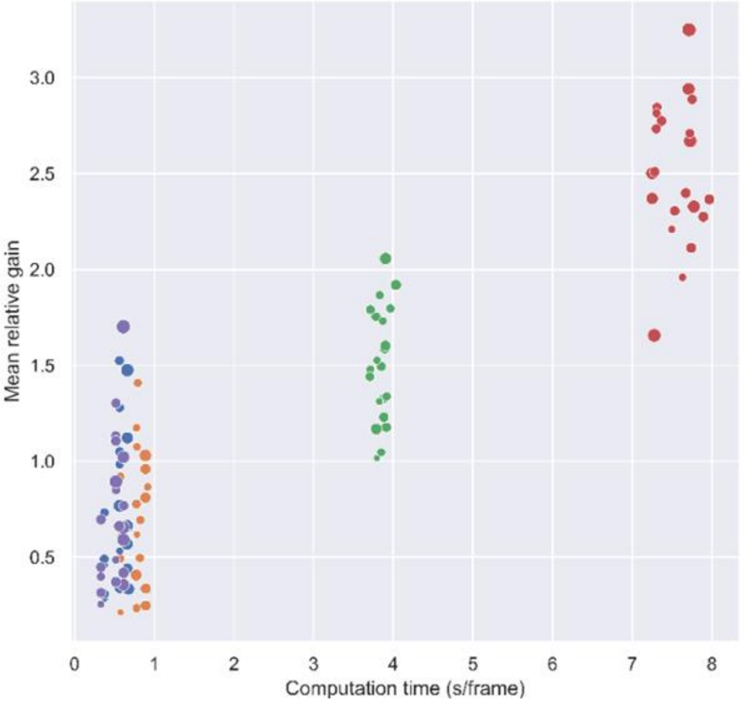
ME+reduced attack, 3.7 s/frame, mean Rel. Gain=1.79

No interpolation, 8.1 s/frame, mean Rel. Gain=2.77

Attack: Korhonen et al. w/ lr=0.5, metric: VSFA, video: Old Town Cross, 1080p

Attack propagation in videos





- Linear interpolation
- ME interpolation
- ME interpolation + reduced attack
- No interpolation
- Repeating from previous frame
- Source video

Characteristics of attacks

- Attack **gain**
- **Applicability** to different kinds of metrics (differentiable, etc)
- **Computational complexity**
- **Visibility** of influence (masking capability)
- **Ability to attack multiple metrics** simultaneously
- Opportunity to **reduce computational complexity for video**
- **Easiness of detecting** (of this attack presence)
- **Resistance** of the attack **to defenses**

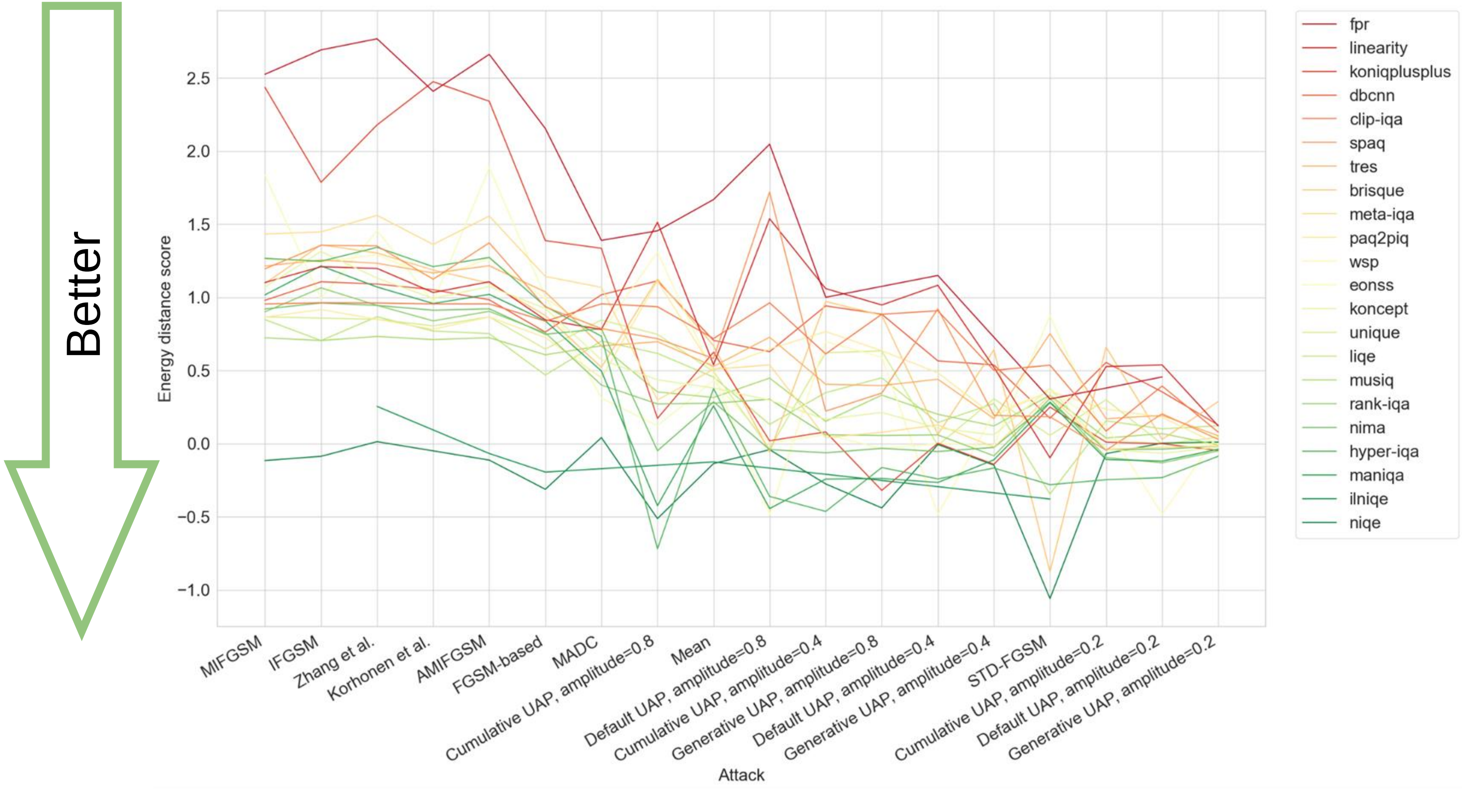
We implemented 30+ types of different attacks for now!

Characteristics of (attacked) metrics

- Metric **correlation**
- Metric **robustness** for different types of attacks
- **Computational complexity**
- **Impact of attacks to subjective score** (mostly negative, content dependent, etc)
- **Easiness of defence** (erase attack for this metric)
- **Uniqueness of contribution** (to participate in combined metrics)

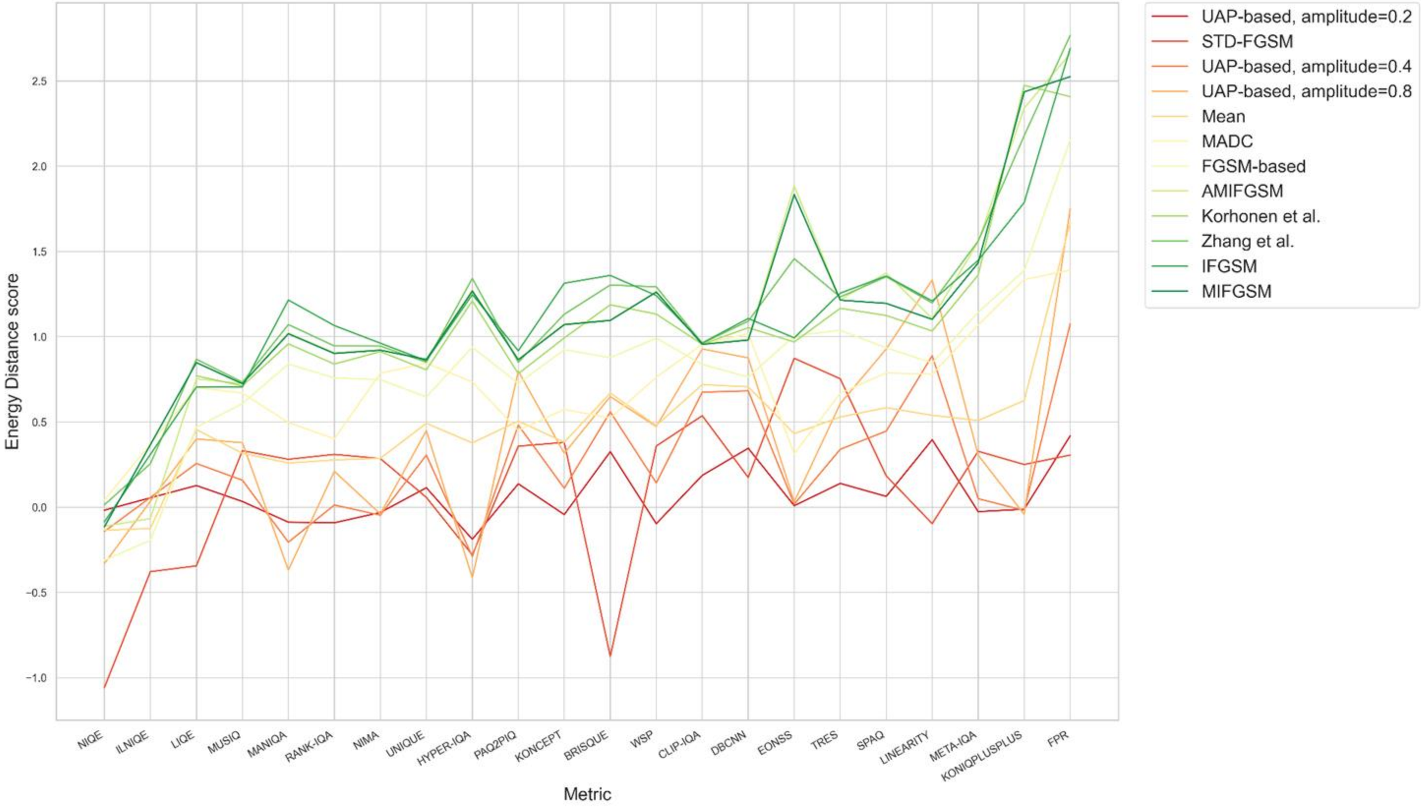
First results of metrics robustness benchmark

Metrics robustness to different attacks



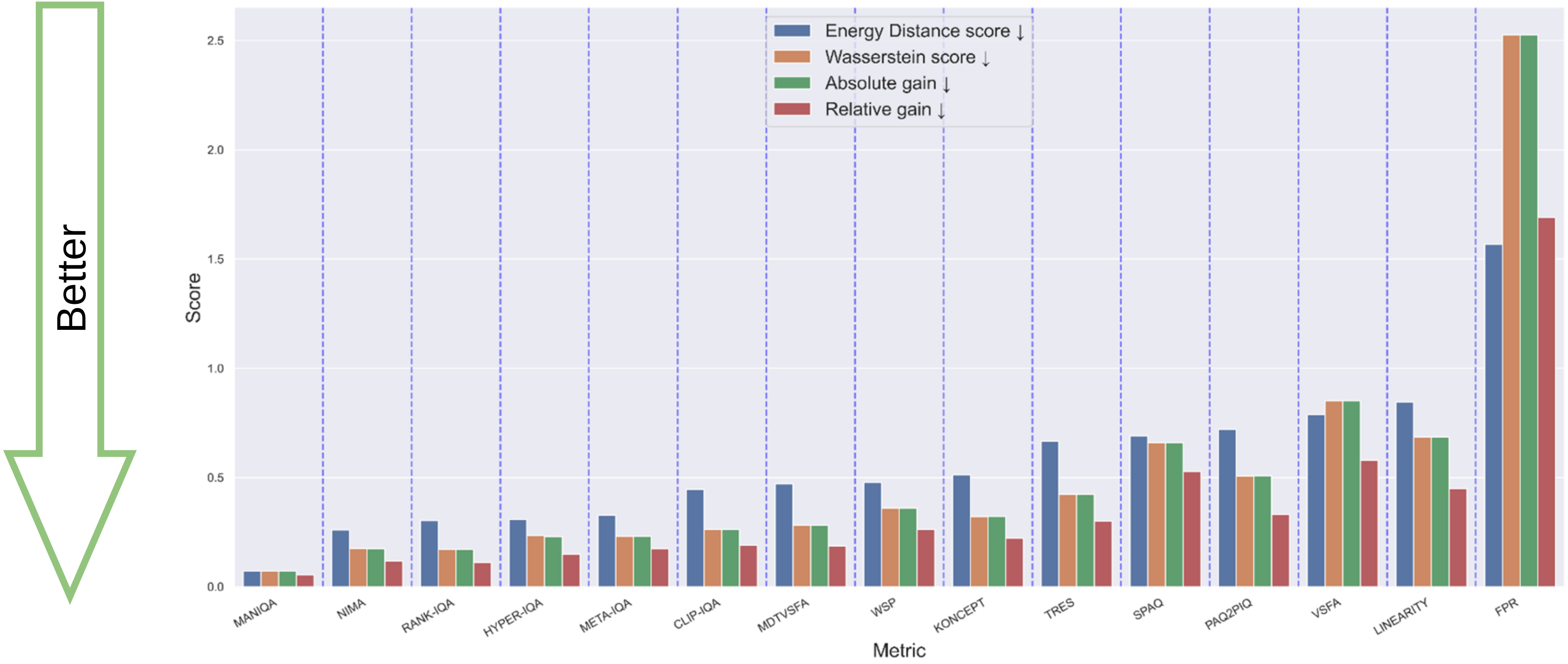
Attacks performance on different metrics

Better

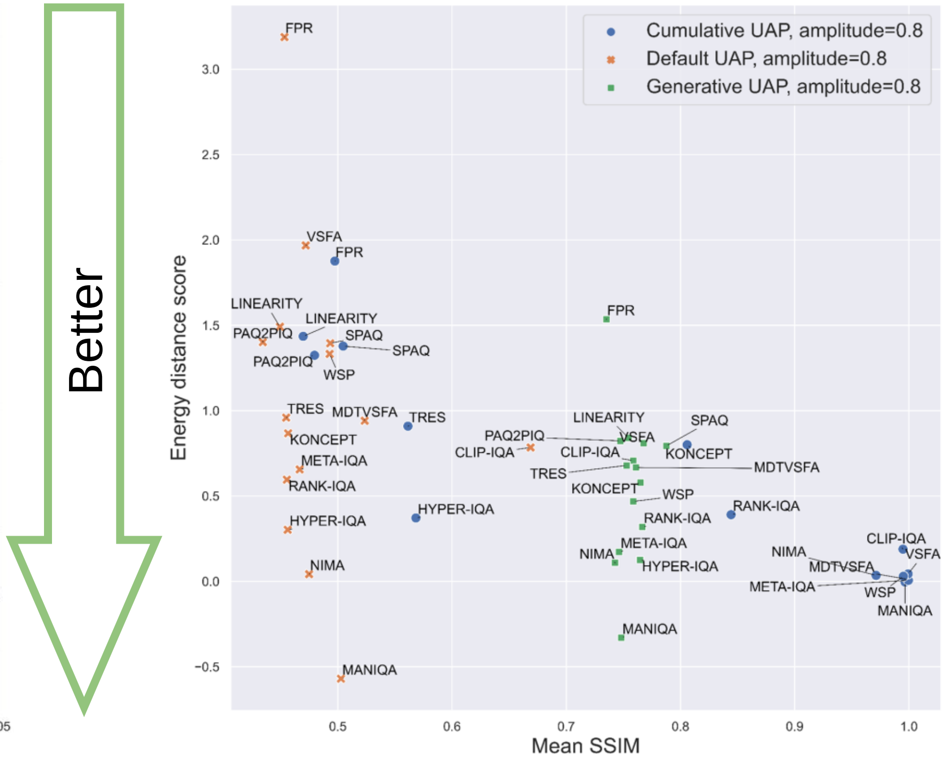


MSU Institute for AI
https://videoproce

Mean metrics robustness

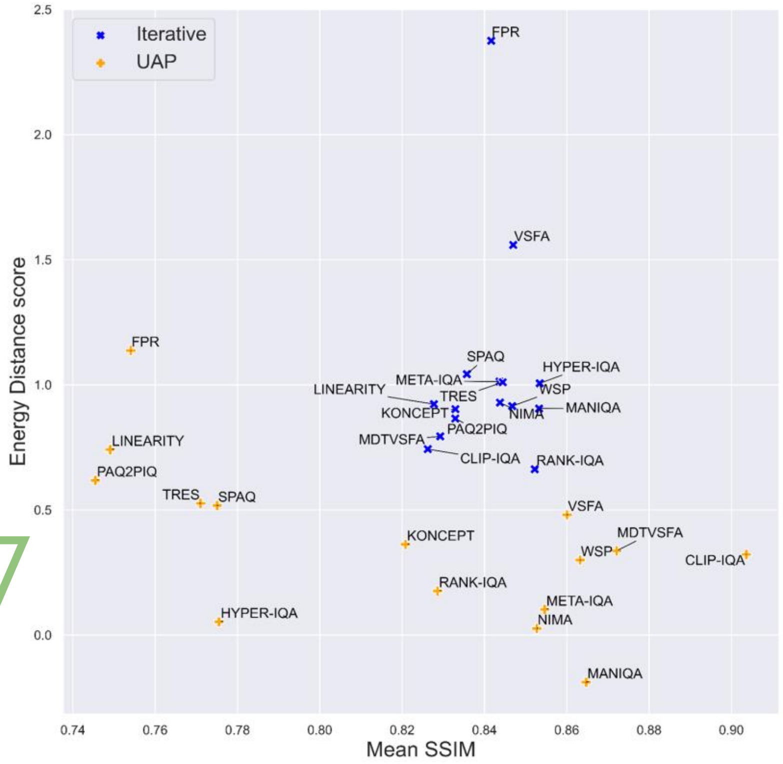


Metrics robustness to different types of attacks

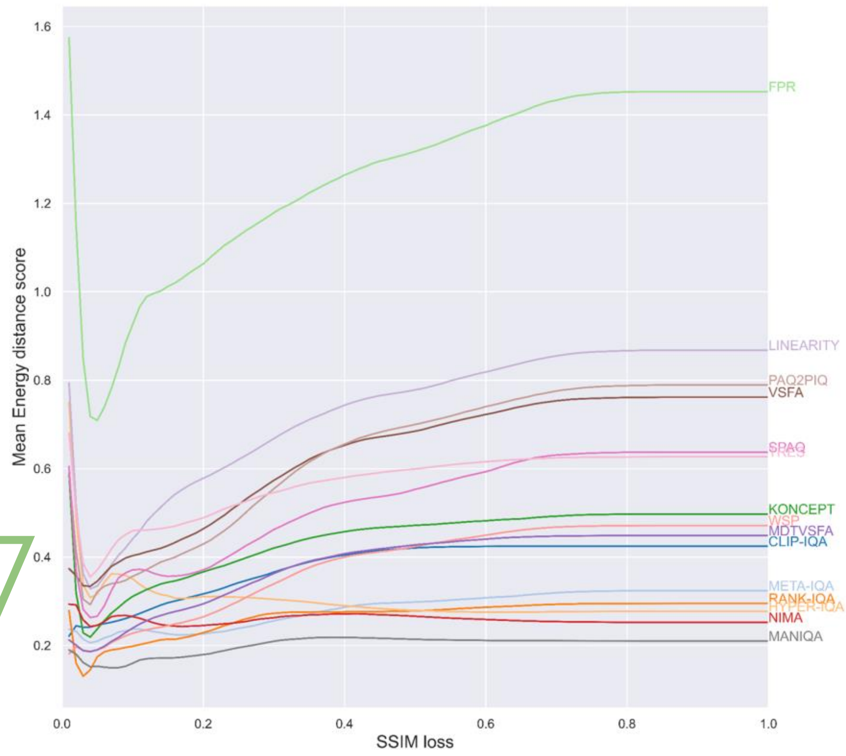


Metrics robustness to different types of attacks

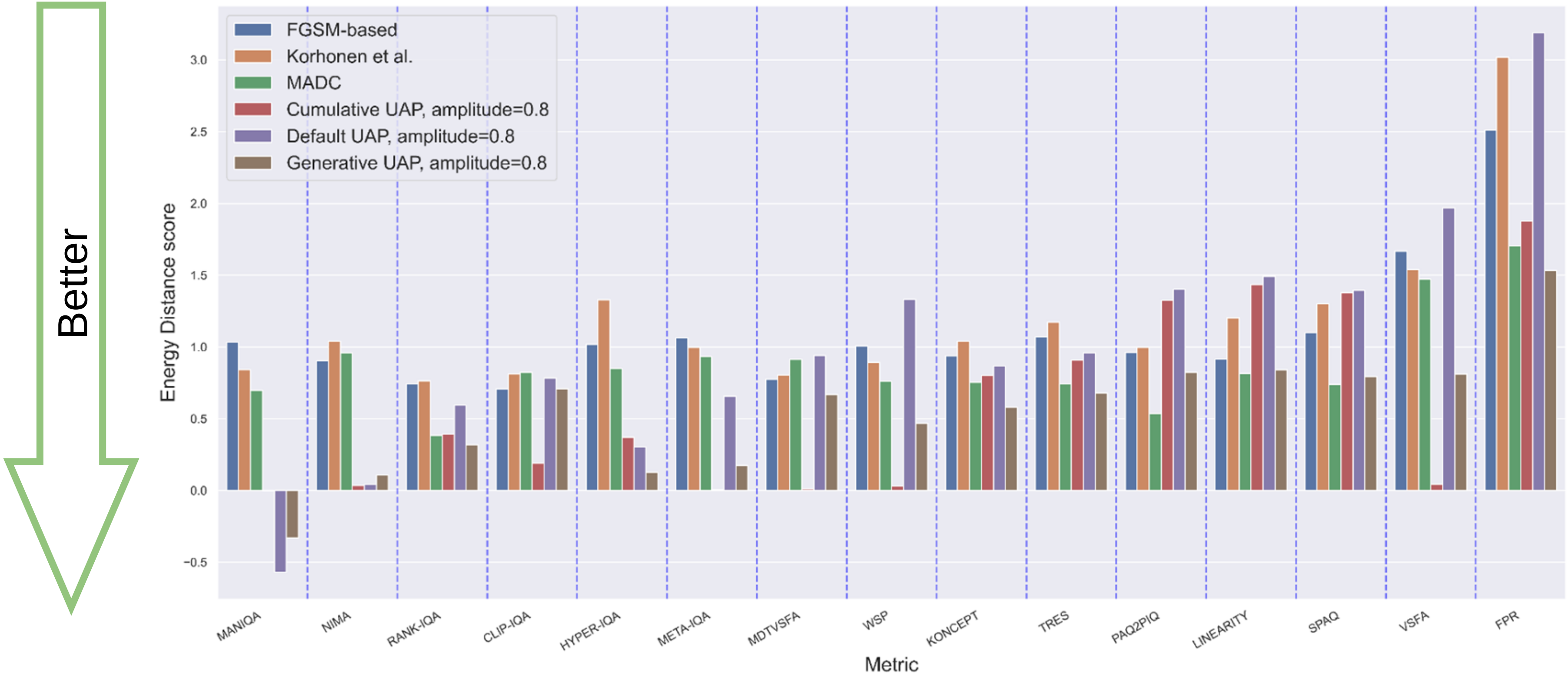
Better



Better



Metrics robustness to different types of attacks



Our papers on ICLR & AAAI

Published as a Tiny Paper at ICLR 2023

FAST ADVERSARIAL CNN-BASED PERTURBATION ATTACK ON NO-REFERENCE IMAGE QUALITY METRICS

Ekaterina Shumitskaya[1], **Anastasia Antsiferova**[2,3] **& Dmitriy Vatolin**[1,2,3]
[1] Lomonosov Moscow State University, Moscow, Russian Federation
[2] ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russian Federation
[3] MSU Institute for Artificial Intelligence, Moscow, Russian Federation
{ekaterina.shumitskaya, aantsiferova, dmitriy}@graphics.cs.msu.ru

ABSTRACT

Modern neural-network-based no-reference image- and video-quality metrics exhibit performance as high as full-reference metrics. These metrics are widely used to improve visual quality in computer vision methods and compare video processing methods. However, these metrics are not stable to traditional adversarial attacks, which can cause incorrect results. Our goal is to investigate the boundaries of no-reference metrics applicability, and in this paper, we propose a fast adversarial perturbation attack on no-reference quality metrics. The proposed attack (FACPA) can be exploited as a preprocessing step in real-time video processing and compression algorithms. This research can yield insights to further aid in designing of stable neural-network-based no-reference quality metrics.

Comparing the robustness of modern no-reference image- and video-quality metrics to adversarial attacks

Anastasia Antsiferova[1,2], **Khaled Abud**[2], **Aleksandr Gushchin**[1,2], **Sergey Lavrushkin**[2], **Ekaterina Shumitskaya**[3], **Maksim Velikanov**[3], **Dmitriy Vatolin**[1,2,3]
ISP RAS Research Center for Trusted Artificial Intelligence[1]
MSU Institute for Artificial Intelligence[2]
Lomonosov Moscow State University[3]
{aantsiferova, khaled.abud, alexander.gushchin, sergey.lavrushkin, ekaterina.shumitskaya, maksim.velikanov, dmitriy}@graphics.cs.msu.ru

Abstract

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

Useful links

- Beta version of benchmark webpage: https://videoprocessing.ai/benchmarks/metrics-robustness.html
- Paper with additional results analysis: https://openreview.net/forum?id=bpsYFVVayV
- GitHub repository for assessing metrics robustness to adversarial attacks and reproducing benchmark results: https://github.com/msu-video-group/MSU_Metrics_Robustness_Benchmark

Reviewer's insight

This paper is sound, interesting, but in my opinion does not innovate enough to be published in high profile journal like IJCV. The paper can be easily compressed into a conference paper. As a side note, I'd say that the ethics of such research is questionable in that it fosters fraud in the evaluation of results, but does not offer a solution. The only deduction one can make from such papers is that NR metrics should be banned from benchmarks and challenges, or that they could no longer be public, so that nobody can train on them. But, perhaps, this deduction is too much hurried up and there might be ways to make any NR metrics robust to such attacks. That would be for sure a valuable contribution.

Review of our paper about adversarial attacks on NR-metrics

This conclusion is applicable for FR- and RR-metrics as well

Are you ready to ban all NN-metrics from all benchmarks, challenges and papers?

Near beautiful future

Video quality metrics
in company reports will soon
mean approximately nothing!

Welcome to the wonderful AI world!

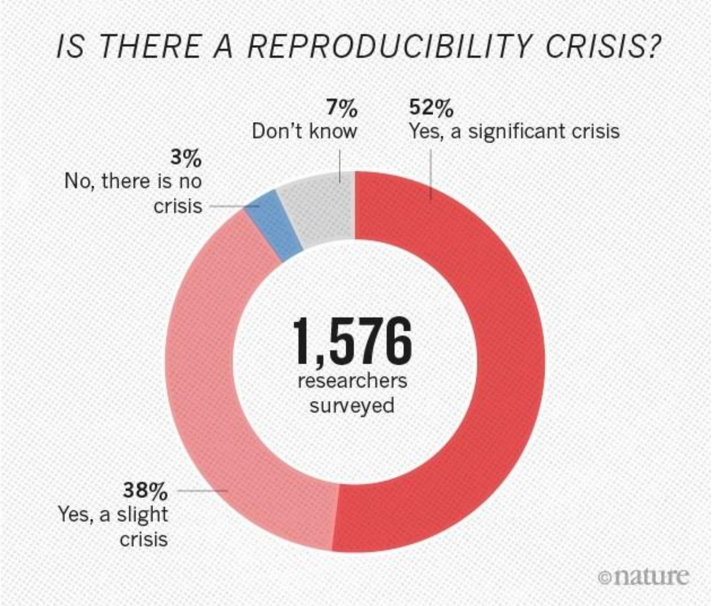
Hackability implications

Without check for adversarial attacks we **<u>can no longer trust</u>** the results in:

- papers
- benchmarks and challenges
- company reports

Reproducibility crisis will become deeper soon!

Adversarial attacks check at least as important as ablation study



Our challenges here

Our challenges:

- Improve **the first hackability (and hack-resistance) benchmark for metrics**
- (more complicated) Create a **methodology to determine the probability of an attempted hack**
- (even more complex) Create a **metric with high correlation and high resistance to hacking**

We are looking for researchers

Our tasks:

- **Implement more attacks**
- **Makes attacks more efficient**
- **Implement more defense**
- **Analyze this multidimentional space efficiently**
- **Suggest new metrics/measurement approaches**
- **Prove new approaches efficiency**

Contacts

Dmitriy Vatolin dmitriy@graphics.cs.msu.ru

- videoprocessing.ai/about
- compression.ru/video
- compression.ru/vqmt
- videocompletion.org
- videomatting.com
- subjectify.us
- evt.guru