





Cross-Modal Transferable Image-to-Video Attack on Video Quality Metrics

Georgii Gotin¹^a, Ekaterina Shumitskaya^{2,3,1}^b, Anastasia Antsiferova^{3,2,4}^c and Dmitriy Vatolin^{1,2,3}^d

¹*Lomonosov Moscow State University, Moscow, Russia*

²*ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia*

³*MSU Institute for Artificial Intelligence, Moscow, Russia*

⁴*Laboratory of Innovative Technologies for Processing Video Content, Innopolis University, Innopolis, Russia*


Keywords: Video Quality Assessment, Video Quality Metric, Adversarial Attack, Cross-Modal, CLIP.


Abstract: Recent studies have revealed that modern image and video quality assessment (IQA/VQA) metrics are vulnerable to adversarial attacks. An attacker can manipulate a video through preprocessing to artificially increase its quality score according to a certain metric, despite no actual improvement in visual quality. Most of the attacks studied in the literature are white-box attacks, while black-box attacks in the context of VQA have received less attention. Moreover, some research indicates a lack of transferability of adversarial examples generated for one model to another when applied to VQA. In this paper, we propose a cross-modal attack method, IC2VQA, aimed at exploring the vulnerabilities of modern VQA models. This approach is motivated by the observation that the low-level feature spaces of images and videos are similar. We investigate the transferability of adversarial perturbations across different modalities; specifically, we analyze how adversarial perturbations generated on a white-box IQA model with an additional CLIP module can effectively target a VQA model. The addition of the CLIP module serves as a valuable aid in increasing transferability, as the CLIP model is known for its effective capture of low-level semantics. Extensive experiments demonstrate that IC2VQA achieves a high success rate in attacking three black-box VQA models. We compare our method with existing black-box attack strategies, highlighting its superiority in terms of attack success within the same number of iterations and levels of attack strength. We believe that the proposed method will contribute to the deeper analysis of robust VQA metrics.


1 INTRODUCTION


Modern No-Reference Video Quality Assessment (NR-VQA) metrics are vulnerable to adversarial attacks (Yang et al., 2024a), (Yang et al., 2024b), (Zhang et al., 2024), (Siniukov et al., 2023), (Shumitskaya et al., 2024a). This raises concerns about the safety of relying on these metrics to automatically assess video quality in real-world scenarios, such as public benchmarks and in more critical situations, such as autonomous driving. Adversarial attacks on VQA metrics can be classified into two categories: white-box and black-box attacks. White-

box attacks operate with complete access to the VQA metric, including its architecture and gradients. In contrast, black-box attacks work without any knowledge of the metric's architecture and can only send queries to receive the metric's response. There is also a subclass of black-box attacks that utilizes a proxy white-box model to generate adversarial perturbations. These generated perturbations can effectively deceive unseen models in black-box settings. However, in (Zhang et al., 2022) the authors demonstrated that VQA metrics exhibit poor transferability across different models. This limitation may appear from the fact that VQA models place significant emphasis on various texture and noise details, which can vary greatly among different models. In contrast, classification tasks typically focus primarily on the semantic content of images, leading to greater consistency in performance across diverse classifi-

^a <https://orcid.org/0009-0007-7176-703X>

^b <https://orcid.org/0000-0002-6453-5616>

^c <https://orcid.org/0000-0002-1272-5135>

^d <https://orcid.org/0000-0002-8893-9340>

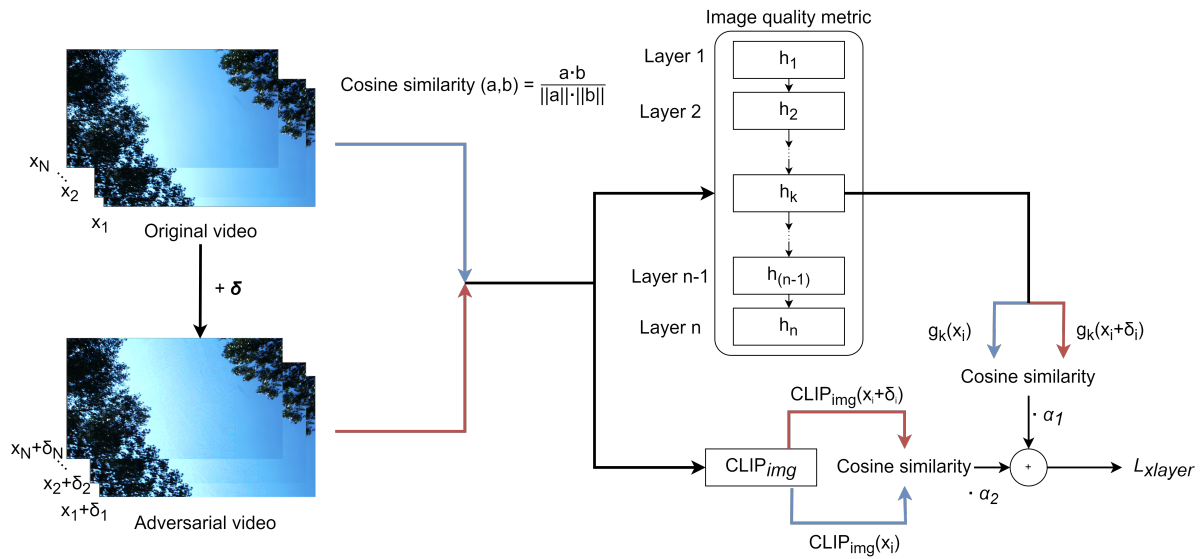


Figure 1: Scheme of the proposed IC2VQA method. Given an original video, each clip runs through image quality metric with saving of output on the k -th layer and through CLIP image model with saving full output. After that attacked video runs same models with saving same outputs. Then cosine similarities of saved outputs are respectively aggregated in cross layer loss.

ication models. In other words, creating a transferable attack for VQA metrics is more challenging than for classification tasks. To address this issue, we propose transferable cross-modal attack to perform white-box attack on Image quality metric and CLIP and transfer it to Video Quality Assessment model (IC2VQA). Figure 1 provides an overview of the proposed IC2VQA method. IC2VQA takes individual frames of the original video and generates adversarial noise for each frame.

Our main contributions are as follows.

- We propose a novel method for transferable cross-modal attacks on NR VQA metrics that utilizes IQA metrics and CLIP model
- We conduct comprehensive experiments using 12 high-resolution videos and 3 target VQA models and show the superiority of the proposed method among existing methods
- We analyze the correlations between features in the deep layers of IQA and VQA metrics
- We made our code available on GitHub: <https://github.com/GeorgeGotin/IC2VQA>.

2 RELATED WORK

2.1 Image- and Video-Quality Metrics

Image and video quality assessment (IQA/VQA) metrics can be divided into full-reference and no-reference (also known as blind IQA/VQA). Full-reference quality metrics compare two images/videos, while no-reference metrics assess the visual quality of a single image/video. These tasks are fundamentally different: full-reference IQA focuses on measuring distances between images in various feature spaces, while no-reference IQA evaluates the quality of an image based solely on the distorted image. No-reference image- and video-quality assessment (NR-VQA) metrics fall into distortion-specific and general-purpose. Distortion-specific approaches predict the quality score for a particular type of distortion, such as compression (Wang et al., 2015) or blurring (Chen and Bovik, 2011). However, these methods have limited real-world applications because it is not always possible to specify the type of distortion. They may not capture the complex mixtures of distortions that often occur in real-world images and videos. However, general-purpose NR-VQA approaches assess the image quality of any distortion. In this work, we focus on the problem of attacking NR-VQA metrics (Li et al., 2019), (Li et al., 2021), (Zhang and Wang, 2022) to find metrics that are ro-

bust to transferable cross-modal attacks.

2.2 Adversarial Attacks on Image- and Video-Quality Metrics

The problem of vulnerability analysis of novel NR IQA models to adversarial attacks was widely discussed in previous works: (Yang et al., 2024a), (Leonkova et al., 2024), (Kashkarov et al., 2024), (Deng et al., 2024), (Konstantinov et al., 2024), (Yang et al., 2024b), (Zhang et al., 2024), (Ran et al., 2025), (Mef-tah et al., 2023), (Siniukov et al., 2023), (Shumitskaya et al., 2024b), (Shumitskaya et al., 2024a). Some works have been conducted as part of the MediaEval task: “Pixel Privacy: Quality Camouflage for Social Images” (MediaEval, 2020), where participants aimed to improve image quality while reducing the predicted quality score. This task is similar to the vanilla adversarial attack on quality metrics, but to decrease the score rather than increase it. In (Bonnet et al., 2020), the authors generated adversarial examples for NR models using PGD attack (Madry et al., 2018). Zhao et al. (Zhao et al., 2023) proposed to attack NR metrics by applying image transformations based on optimizing a human-perceptible color filter. They also demonstrated that this attack is even resistant to JPEG compression. However, these studies are limited to small-scale experiments and lack in-depth analysis. Several comprehensive works have recently been published that systematically investigate adversarial attacks against NR models.

In (Zhang et al., 2022), a two-step perceptual attack was introduced for the NR metrics. The authors established the attack’s goal as a Lagrangian function that utilizes some FR metric, which acts as a “perceptual constraint”, alongside the NR metric representing the target model. By adjusting the Lagrange multiplier, they produced a range of perturbed images that exhibited varying degrees of visibility regarding their distortions. Their extensive experiments demonstrated that the proposed attack effectively deceived four different NR metrics; however, the adversarial examples did not transfer well across various models, indicating specific design vulnerabilities within the NR metrics assessed. In (Shumitskaya et al., 2022), the authors trained the UAP on low-resolution data and then applied it to high-resolution data. This method significantly reduces the time required to attack videos, as it requires only adding perturbations to individual frames. In the study by (Korhonen and You, 2022), the authors create adversarial perturbations for NR metrics by injecting the perturbations into textured areas using the Sobel filter. They also demonstrated that adversarial images gen-

erated for a simple NR metric in white-box settings are transferable and can deceive several NR metrics with more complex architecture in black-box settings. In (Antsiferova et al., 2024), the authors presented a methodology for evaluating the robustness of NR and FR IQA metrics through a wide range of adversarial attacks and released an open benchmark.

To the best of our knowledge, no methods have been designed for transferable cross-modal attacks from NR IQA to NR VQA metrics, which is a subject of this work.

2.3 Transferable Attacks on Image Classification

Adversarial attacks have received significant attention in the domain of machine learning, particularly in image classification tasks. The phenomenon of transferability, where adversarial examples generated on one model can deceive another (potentially different) model, has been investigated in many works. Papernot et al. (Papernot et al., 2016) explored this aspect and demonstrated that transferability is a useful property that could be exploited in black-box settings, where the attacker has limited knowledge of the target model. They also experimentally showed that adversarial examples could be trained on weaker models and successfully deceive more robust classifiers. Various methods have been proposed to enhance the effectiveness of transferable attacks. Some of them (Xie et al., 2019), (Lin et al., 2019), (Dong et al., 2019) apply data-augmentation techniques to enhance the generalization of adversarial examples and reduce the risk of overfitting the white-box model. For example, the translation-invariant attack (Dong et al., 2019) executes slight horizontal and vertical shifts of the input. The second direction to improve transferability is to modify the gradients used to update adversarial perturbations (Dong et al., 2019), (Lin et al., 2019), (Wu et al., 2020a). For example, the momentum iterative attack (Dong et al., 2019) stabilizes the update directions using the addition of momentum in the iterative process. The third approach concentrates on disrupting the shared classification properties among different models (Wu et al., 2020b), (Huang et al., 2019), (Lu et al., 2020). One example is the Attention-guided attack (Wu et al., 2020b), which prioritizes the corruption of critical features that are commonly utilized by various architectures. Recently, innovative cross-modal approaches have been proposed that leverage the correlations between spatial features encoded by different modalities (Wei et al., 2022), (Chen et al., 2023), (Yang et al., 2025). Image2Video attack, proposed in (Wei et al., 2022), is an attack to

successfully transfer from image to video recognition models.

3 PROPOSED METHOD

3.1 Problem Formulation

Let's consider we have a video $x \in X \subset [0, 1]^{N \times C \times H \times W}$, where N — number of frames in video, C — number of channels in video, H, W — height and width of video respectively, X is the set of all possible videos. We define video quality metric as $f : X \rightarrow [0, 1]$, image quality metric as $g : [0, 1]^{C \times H \times W} \rightarrow [0, 1]$. Image quality metric can be expressed in layered form as $g = h_K \circ h_{K-1} \circ \dots \circ h_1$, so

$$g(x_i) = h_K(h_{K-1}(\dots h_1(x_i)\dots)), \quad (1)$$

where each function $h_k : P_{k-1} \rightarrow P_k$ corresponds to a processing layer with $P_0 = [0, 1]^{C \times H \times W}$ being the input feature space and $P_K = [0, 1]$ being the output range of the metric. g_k defines the composition of the first k layers:

$$\begin{aligned} g_k &= h_k \circ \dots \circ h_1 \\ g_k : [0, 1]^{C \times H \times W} &\rightarrow P_k. \end{aligned} \quad (2)$$

Each g_k serves the k -th layer of the quality metric, where P_k represents the feature spaces corresponding to that layer.

3.2 Method

The primary goal of the attack is to make the predicted quality score of video $f(x + \delta)$ on the attacked video deviate from the original score $f(x)$, where δ is the perturbation on the video x . Also, the rank of correlation of predicted score with MOS is important, so our goal is to shrink it as possible. This method was based on method proposed by Zhipeng Wei as I2V(Wei et al., 2022).

The proposed attack is designed to mislead the video quality metric. It creates adversarial frame $\delta_i \in [0, 1]^{C \times H \times W}$ for each i -th frame on the input video. To maintain the imperceptible of this adversarial perturbation, we import a constraint on its magnitude $\|\delta\|_p \leq \epsilon$, where $\|\cdot\|_p$ denotes L_p norm. In our research, we adopt the L_∞ norm due to its computational efficiency compared to other L_p norms.

Based on observation of correlations between layers of video and image quality metrics, we proposed the cross-layer loss, this loss is designed to influence the features of the layers within the image quality metric and enhance it's the effectiveness in black box

settings. The cross-layer loss of the k -th layer defined as follows

$$\mathcal{L}_{xlayer} = \frac{1}{N} \sum_{i=1}^N \frac{g_k(x_i + \delta_i) \cdot g_k(x_i)}{\|g_k(x_i + \delta_i)\| \|g_k(x_i)\|}, \quad (3)$$

where x — the original video with N frames, x_i — i -th frame of the video. We propose multi-modal cross-layer loss for better implementation and generalization across different feature domains. This loss utilizes adversarial perturbation δ to simultaneously optimize an ensemble of image quality metrics $g^{(1)}, \dots, g^{(F)}$ with layers k_f for f -th metric. Consequently, the overall cross-layer loss can be defined as follows:

$$\begin{aligned} \mathcal{L}_{sim} &= \frac{1}{N} \sum_{i=1}^N \sum_{f=1}^F \alpha_f \frac{g_k^{(f)}(x_i + \delta_i) \cdot g_k^{(f)}(x_i)}{\|g_k^{(f)}(x_i + \delta_i)\| \|g_k^{(f)}(x_i)\|} + \\ &+ \frac{1}{F} \sum_{f=1}^F \|1 - \alpha_f\|, \end{aligned} \quad (4)$$

where α_f — constant positive value, initialized with ones.

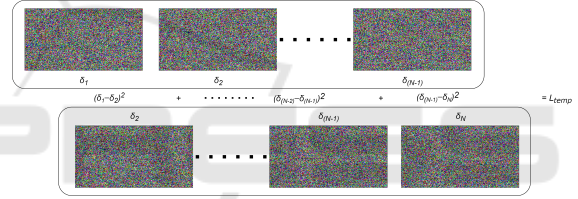


Figure 2: Overview of the temporal loss computing. For each pair of frames from original and attacked videos difference Δ is computed. The temporal loss is computed as square root of sum of all differences.

To enhance temporal stability of the attacked video $x + \delta$ and further ensure that the adversarial perturbation δ is imperceptible, we added a temporal loss component (Figure 2)

$$\mathcal{L}_{temp} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|\delta_{i+1} - \delta_i\|_2. \quad (5)$$

3.3 Algorithm

We construct our attack as presented in Algorithm 1, which is applied to image quality metrics. At each step of the attack, the cross-layer loss for the f -th image quality metric is computed and the adversarial noise is optimized using the Adam optimizer. Subsequently, the noise is clipped to ensure it remains within the bounds of ϵ according to the L_∞ norm. Experiments shown that alternative version of algorithms, where all losses are summed with weights as described in (4), yields lower scores compared to the final algorithm. By applying this algorithm to attacks

Algorithm 1: Algorithm of the consistent attack with multiple image quality metrics.

Data: original video $x \in [0, 1]^{N \times C \times H \times W}$, F image quality metrics
 $g_1, \dots, g_F : [0, 1]^{C \times H \times W} \rightarrow [0, 1]$,
 k_1, \dots, k_F — number of the layer,
perturbation budget ϵ , number of iterations I

Result: $\delta \in [0, 1]^{N \times C \times H \times W}$, *s.t.* $\|\delta\| \leq \epsilon$
 $\delta = (1/255)^{N \times C \times H \times W}$;

for i from 1 to I **do**
 for f from 1 to F **do**
 Calculate \mathcal{L}_{xlayer} as in 3 for $g^{(f)}$;
 Calculate \mathcal{L}_{temp} as in 5;
 $\delta \leftarrow ADAM(\alpha, loss_{xlayer} + loss_{temp})$;
 $\delta \leftarrow clip_{\epsilon}(\delta)$;
 end
end

targeting a single image quality metric, IC2VQA has effectively transformed into a single-metric attack.

4 EXPERIMENTS

4.1 Dataset

We evaluate our attack using a subset of Xiph.org (Derf’s) dataset (Xiph.org, 2001). The subset contains ten videos downsampled from 1080p to 540p and trimmed to 75 frames. The videos have different patterns of image and motions in it such as shooting from a tripod, moving crowd, running water, etc.

4.2 Quality Metrics

4.2.1 Image Quality Metrics

For ensembles of image quality metrics we used NIMA (Talebi and Milanfar, 2018), PaQ-2-PiQ (Ying et al., 2020), SPAQ (Fang et al., 2020) metrics. To further boost the transferability we added additional modalities such as CLIP model (Radford et al., 2021). To get feature vectors, in NIMA model attack utilizes layers after classifier and after global pool, in PaQ-2-PiQ model attack utilizes layers after roi-pool layer and body, in SPAQ model attack utilizes first, second, third and fourth layers. In CLIP model, an output of the CLIP image module was utilized.

4.2.2 Video Quality Metrics

As black-boxed video-metric we used the VSFA (Li et al., 2019), MDTVSFA (Li et al., 2021) and TiVQA

(Zhang and Wang, 2022) trained on the KoNViD-1k (Hosu et al., 2020). These metrics evaluate quality scores by taking into account both the spatial and temporal characteristics of the videos.

4.3 Comparison with Other Methods

Due to the lack of existing black-box image-to-video quality model attacks, we compared our method against one transferable attack, the PGD attack (Madry et al., 2018), adapted for image-to-video scenarios, as well as two black-box attacks: Square Attack (Andriushchenko et al., 2020) and AttackVQA (Zhang et al., 2024). The latter was specifically designed to target VQA metrics. For comparison, we tested all methods using a grid of parameters for ϵ and I to generate attacked videos with varying levels of distortion. Recall that ϵ represents the L_{∞} norm restriction on generated perturbation and I is the number of iterations used for attack. Next, we measured the VQA metric scores of the attacked videos and calculated the correlations between these scores and a corresponding linearly decreasing vector. As the ϵ/I parameters increase while keeping I/ϵ fixed, the quality of the attacked videos tends to degrade in an approximately linear manner. Therefore, an effective VQA metric should exhibit a strong correlation with this vector for the attacked videos. Consequently, if the metric is vulnerable, it will be indicated by a low correlation. Additionally, the most effective attacks will result in lower correlations, so we assess attack success by evaluating their ability to reduce these correlations. In our experiments, we used absolute values Pearson’s (PLCC) and Spearman’s (SRCC) correlations.

4.4 Parameters

We evaluated the proposed and comparison methods using a range of ϵ and I parameters to assess their effectiveness under various conditions. We used the following grids of parameters: $\epsilon = [1/255, 2/255, 5/255, 10/255, 15/255, 20/255, 50/255]$ and $I = [1, 2, 5, 10, 20]$.

5 RESULTS

Results of comparison with other methods shown in the Table 1. The proposed IC2VQA attack method demonstrated promising results across all three VQA models, achieving the reduction in PLCC and SRCC scores up to 0.425 and 0.380 on average, respectively. Additionally, it outperformed competing methods in

Table 1: Comparison of the proposed transferable cross-modal IC2VQA attack with two black-box attacks (Square Attack (Andriushchenko et al., 2020) and AttackVQA (Zhang et al., 2024)) and one transferable PGD attack (Madry et al., 2018) targeting three VQA metrics. The table presents the mean absolute values of PLCC and SROCC correlations across different epsilons between linearly decreasing vectors and attacked VQA scores. For each score Transferable attacks were performed using three different white-box IQA metrics.

Attack	Image quality metric*	Video metric		
		VSFA PLCC↓ / SRCC↓	MDTVSFA PLCC↓ / SRCC↓	TiVQA PLCC↓ / SRCC↓
Square Attack		0.635 / 0.579	0.617 / 0.564	0.570 / 0.521
AttackVQA		0.335 / 0.289	0.429 / 0.384	0.479 / 0.392
PGD	NIMA	0.578 / 0.518	0.546 / 0.470	0.531 / 0.514
	PaQ-2-PiQ	0.619 / 0.571	0.586 / 0.341	0.598 / 0.516
	SPAQ	0.544 / 0.564	0.608 / 0.486	0.480 / 0.492
IC2VQA (ours)	NIMA	0.475 / 0.453	0.369 / 0.348	0.426 / 0.419
	PaQ-2-PiQ	0.450 / 0.404	0.414 / 0.396	0.459 / 0.428
	SPAQ	<u>0.404 / 0.311</u>	<u>0.390 / 0.299</u>	<u>0.439 / 0.366</u>

*Image quality metric used in the proposed method. For the PGD — metric which is attacked. For the IC2VQA — component of cross-layer loss. In the IC2VQA attack, the image quality metric specified in the table was utilized in conjunction with CLIP and temporal losses.

Table 2: Comparison of variations of the IC2VQA attack with different configuration. The table presents the mean absolute values of PLCC and SROCC correlations across different epsilons between linearly decreasing vectors and attacked VQA scores. VSFA was used as VQA.

Loss	PLCC ↓	SRCC ↓
\mathcal{L}_{xlayer}	0.849	0.800
$\mathcal{L}_{xlayer} + \mathcal{L}_{CLIP}$	0.472	0.430
$\mathcal{L}_{xlayer} + \mathcal{L}_{CLIP} + \mathcal{L}_{temp}$	<u>0.515</u>	0.354

two out of the three VQA black-box models. Furthermore, the results demonstrate that methods specifically designed for the VQA task, such as AttackVQA and the proposed IC2VQA, consistently outperform PGD and Square Attack, which are adaptations from classification tasks. This highlights the importance of developing approaches tailored for VQA challenges. Figure 4 presents the example of the proposed attack. We can see that VQA metric fails to accurately assess the quality of the degraded video, assigning it a higher score.

6 ABLATION STUDY

6.1 Loss Configuration

To experimentally demonstrate effectiveness of combination of losses in comparison with single \mathcal{L}_{xlayer} , we evaluated our attack in configuration with only

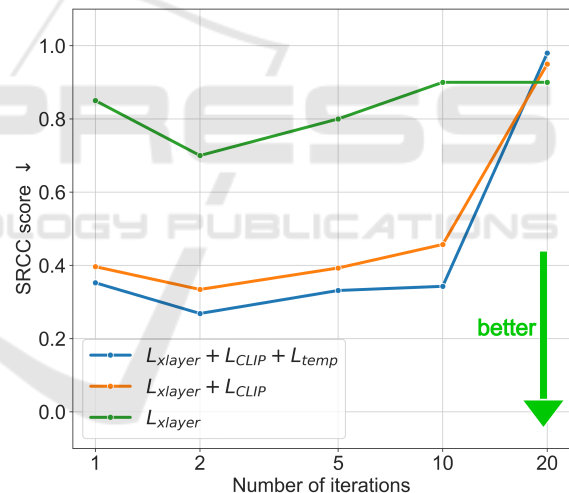


Figure 3: The plot of variations of the IC2VQA attack under different configuration. The plot presents the median value of SRCC score across different epsilon with variation of the number of iterations.

one image quality metric \mathcal{L}_{xlayer} , with one image quality metric and CLIP image model $\mathcal{L}_{xlayer} + \mathcal{L}_{CLIP}$ and with one image quality metric, CLIP image model and temporal regularization $\mathcal{L}_{xlayer} + \mathcal{L}_{CLIP} + \mathcal{L}_{temp}$. In experiment we evaluate IC2VQA configurations on white-box models NIMA, PaQ-2-PiQ and SPAQ and black-box VSFA model and scored them by median absolute value of correlations. The results of the comparison are shown in the Table 2 and Figure 3. From Table 2, we observe that addition of cosine



Figure 4: Example of IC2VQA attack. Cross-layer loss is computed for layer1 of SPAQ, ϵ is set 50/255, number of iterations is set to 20. The visual quality of clean video is obviously higher than that of the attacked video, however, VSFA metric rates the attacked video as having higher quality.

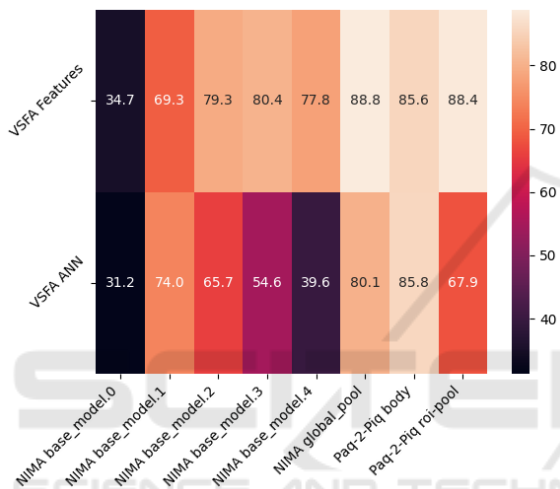


Figure 5: Heatmap of cosine similarity between the features of VSFA layers and those from the NIMA and PaQ-2-PiQ layers. The values represent the cosine similarity scaled by a factor of 100.

similarity between CLIP features (\mathcal{L}_{CLIP}) to the loss function enhances the attack’s success by 1.8 times. The temporal loss increases the attack’s success in terms of SRCC by 1.2 times and slightly decreases PLCC. Figure 3 shows that the combined loss function $\mathcal{L}_{xlayer} + \mathcal{L}_{CLIP} + \mathcal{L}_{temp}$ outperforms others in attack success, as measured by SRCC, across all iteration values.

The results of this experiment show that the addition of all components contributes to the effectiveness of the attack method. Therefore, in the final version of the attack, we use the $\mathcal{L}_{xlayer} + \mathcal{L}_{CLIP} + \mathcal{L}_{temp}$ loss function.

6.2 Feature Correlation

In this section, we analyze the correlations between features in the deep layers of IQA and VQA metrics. Figure 5 presents the heatmap of correlations between

features from the VSFA VQA model and the NIMA and PaQ-2-PiQ IQA models. We observe that these features are often highly correlated, highlighting the fact that addition of IQA modalities to black-box attack on VQA can boost transferability with a high likelihood of success.

7 CONCLUSION

In this paper we propose the novel adversarial attack on VQA metrics that operates as a black-box. The proposed IC2VQA performs a cross-modal transferable attack that utilizes white-box IQA metrics and the CLIP model. The results of extensive experiments showed that IC2VQA generates adversarial perturbations that are more effective compared to previous approaches, significantly reducing the SRCC and PLCC scores of a black-box VQA model. The proposed method can serve as a tool for verifying VQA metrics robustness to black-box attacks. Furthermore, the vulnerabilities identified in this study can contribute to the development of more robust and accurate VQA metrics in the future.

ACKNOWLEDGEMENTS

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

REFERENCES

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

- Antsiferova, A., Abud, K., Gushchin, A., Shumitskaya, E., Lavrushkin, S., and Vatolin, D. (2024). Comparing the robustness of modern no-reference image- and video-quality metrics to adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 700–708.
- Bonnet, B., Furon, T., and Bas, P. (2020). Fooling an automatic image quality estimator. In *MediaEval 2020-MediaEval Benchmarking Initiative for Multimedia Evaluation*, pages 1–4.
- Chen, K., Wei, Z., Chen, J., Wu, Z., and Jiang, Y.-G. (2023). Gcma: Generative cross-modal transferable adversarial attacks from images to videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 698–708.
- Chen, M.-J. and Bovik, A. C. (2011). No-reference image blur assessment using multiscale gradient. *EURASIP Journal on image and video processing*, 2011:1–11.
- Deng, W., Yang, C., Huang, K., Liu, Y., Gui, W., and Luo, J. (2024). Sparse adversarial video attack based on dual-branch neural network on industrial artificial intelligence of things. *IEEE Transactions on Industrial Informatics*.
- Dong, Y., Pang, T., Su, H., and Zhu, J. (2019). Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4312–4321.
- Fang, Y., Zhu, H., Zeng, Y., Ma, K., and Wang, Z. (2020). Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3677–3686.
- Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., Li, S., and Saupé, D. (2020). The konstanz natural video database.
- Huang, Q., Katsman, I., He, H., Gu, Z., Belongie, S., and Lim, S.-N. (2019). Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742.
- Kashkarov, E., Chistov, E., Molodetskikh, I., and Vatolin, D. (2024). Can no-reference quality-assessment methods serve as perceptual losses for super-resolution? *arXiv preprint arXiv:2405.20392*.
- Konstantinov, D., Lavrushkin, S., and Vatolin, D. (2024). Image robustness to adversarial attacks on no-reference image-quality metrics. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 611–615. IEEE.
- Korhonen, J. and You, J. (2022). Adversarial attacks against blind image quality assessment models. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, pages 3–11.
- Leonenkova, V., Shumitskaya, E., Antsiferova, A., and Vatolin, D. (2024). Ti-patch: Tiled physical adversarial patch for no-reference video quality metrics. *arXiv preprint arXiv:2404.09961*.
- Li, D., Jiang, T., and Jiang, M. (2019). Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2351–2359.
- Li, D., Jiang, T., and Jiang, M. (2021). Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129(4):1238–1257.
- Lin, J., Song, C., He, K., Wang, L., and Hopcroft, J. E. (2019). Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*.
- Lu, Y., Jia, Y., Wang, J., Li, B., Chai, W., Carin, L., and Velipasalar, S. (2020). Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 940–949.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.
- MediaEval (2020). Pixel privacy: Quality camouflage for social images. <https://multimediaeval.github.io/editions/2020/tasks/pixelprivacy/>.
- Meftah, H. F. B., Fezza, S. A., Hamidouche, W., and Déforges, O. (2023). Evaluating the vulnerability of deep learning-based image quality assessment methods to adversarial attacks. In *2023 11th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE.
- Papernot, N., McDaniel, P., and Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Ran, Y., Zhang, A.-X., Li, M., Tang, W., and Wang, Y.-G. (2025). Black-box adversarial attacks against image quality assessment models. *Expert Systems with Applications*, 260:125415.
- Shumitskaya, E., Antsiferova, A., and Vatolin, D. (2024a). Towards adversarial robustness verification of no-reference image- and video-quality metrics. *Computer Vision and Image Understanding*, 240:103913.
- Shumitskaya, E., Antsiferova, A., and Vatolin, D. S. (2022). Universal perturbation attack on differentiable no-reference image- and video-quality metrics. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press.
- Shumitskaya, E., Antsiferova, A., and Vatolin, D. S. (2024b). IOI: Invisible one-iteration adversarial attack on no-reference image- and video-quality metrics. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings*

- of *Machine Learning Research*, pages 45329–45352. PMLR.
- Siniukov, M., Kulikov, D., and Vatolin, D. (2023). Unveiling the limitations of novel image quality metrics. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE.
- Talebi, H. and Milanfar, P. (2018). Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011.
- Wang, C., Shen, M., and Yao, C. (2015). No-reference quality assessment for dct-based compressed image. *Journal of Visual Communication and Image Representation*, 28:53–59.
- Wei, Z., Chen, J., Wu, Z., and Jiang, Y.-G. (2022). Cross-modal transferable adversarial attacks from images to videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15064–15073.
- Wu, D., Wang, Y., Xia, S.-T., Bailey, J., and Ma, X. (2020a). Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990*.
- Wu, W., Su, Y., Chen, X., Zhao, S., King, I., Lyu, M. R., and Tai, Y.-W. (2020b). Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1161–1170.
- Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., and Yuille, A. L. (2019). Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739.
- Xiph.org (2001). Xiph.org Video Test Media [derf’s collection]. <https://media.xiph.org/video/derf/>.
- Yang, C., Liu, Y., Li, D., and Jiang, T. (2024a). Exploring vulnerabilities of no-reference image quality assessment models: A query-based black-box method. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Yang, C., Liu, Y., Li, D., Zhong, Y., and Jiang, T. (2024b). Beyond score changes: Adversarial attack on no-reference image quality assessment from two perspectives. *arXiv preprint arXiv:2404.13277*.
- Yang, H., Jeong, J., and Yoon, K.-J. (2025). Prompt-driven contrastive learning for transferable adversarial attacks. In *European Conference on Computer Vision*, pages 36–53. Springer.
- Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., and Bovik, A. (2020). From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3585.
- Zhang, A., Ran, Y., Tang, W., and Wang, Y.-G. (2024). Vulnerabilities in video quality assessment models: The challenge of adversarial attacks. *Advances in Neural Information Processing Systems*, 36.
- Zhang, A.-X. and Wang, Y.-G. (2022). Texture information boosts video quality assessment. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2050–2054. IEEE.
- Zhang, W., Li, D., Min, X., Zhai, G., Guo, G., Yang, X., and Ma, K. (2022). Perceptual attacks of no-reference image quality models with human-in-the-loop. *Advances in Neural Information Processing Systems*, 35:2916–2929.
- Zhao, Z., Liu, Z., and Larson, M. (2023). Adversarial image color transformations in explicit color filter space. *IEEE Transactions on Information Forensics and Security*, 18:3185–3197.