# Black-box Universal Adversarial Perturbations for Image and Video Quality Assessment Methods. Supplementary Materials

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

## I. Related Work

### A. White-box Adversarial Attacks

Goodfellow et al. [1] introduced an early white-box adversarial attack on image classifiers. Their work proposed the fast gradient sign method (FGSM), which generates adversarial perturbations by using the sign of the loss-function gradient with respect to the input image. Kurakin et al. [2] proposed a modified FGSM version, called I-FGSM, that employs an iterative approach to refine adversarial examples. Dong et al. [3] further enhanced I-FGSM by introducing MI-FGSM, which incorporates momentum into the iterative process to make attacks more transferable. Madry et al. [4] suggested projected gradient descent (PGD), which performs gradient-descent steps of fixed size, followed by projection to fit certain constraints. More recently, Sang et al. [5] advanced MI-FGSM by integrating adaptive momentum guided by the NR IQA metric NIQE [6].

Moosavi et al. [7] created DeepFool, a nontargeted attack that employs iterative linearization to approximate a model's decision boundary without requiring direct access to gradients. Unlike gradient-based approaches, DeepFool iteratively computes the minimum perturbation to misclassify an input by estimating the model's decision boundary. Carlini et al. [8] introduced C&W, a highly versatile technique capable of both targeted and nontargeted attacks using $l_1$, $l_2$, and $l_\infty$ norms. Bhattad et al. [9] presented unrestricted attacks, which, to avoid traditional perturbation constraints, manipulate image colors by employing a colorization model and by incorporating style transfer. Additionally, Alaifari et al. [10] and Xiao et al. [11] introduced the ADef and StAdv attacks, respectively, both of which optimize vector fields to deform images and generate adversarial examples.

### B. Black-box Adversarial Attacks

Early black-box attacks, such as those of Papernot et al. [12], train surrogate models to mimic the target model's output and then transfer adversarial examples from the surrogate to the target. Chen et al. [13] pioneered a black-box attack that eschews surrogate models, instead employing coordinate gradient descent with gradient approximation. Ilyas et al. [14] improved this approach by adding a more efficient gradient-estimation scheme based on a natural evolutionary strategy (NES) and used it to update adversarial examples.

Moon et al. [15] developed an effective algorithm that avoids gradients: it solves a discrete surrogate problem using marginal gains to create an attack that is robust against hyper-parameter variations. Square Attack by Andriushchenko et al. [16] also avoids gradient estimation by updating adversarial perturbations with colored square patches from a random-search algorithm.

More recently, Croce et al. [17] and Williams et al. [18] presented effective sparse black-box attacks that focus on a subset of pixels rather than the entire image. These attacks use random-search and evolutionary algorithms, respectively, to craft adversarial examples, demonstrating the continual evolution and specialization of black-box attacks.

### C. IQA Metrics

IQA and VQA methods are generally categorized into full-reference (FR) and no-reference (NR) metrics. FR metrics evaluate quality by comparing both the reference and distorted images, measuring their similarity. In contrast, NR metrics assess quality using only a single image, predicting its subjective quality without the need for a reference.

Classical full-reference (FR) IQA metrics, such as SSIM by Wang et al. [19], its modification MS-SSIM [20], and PSNR, do not rely on machine learning algorithms but remain widely used in practice. While these metrics are robust against adversarial attacks, they generally show lower correlations with subjective quality scores compared to machine learning-based methods. One of the most widely used FR VQA metrics is VMAF, introduced by Netflix [21]. VMAF employs support vector machines (SVM) to predict video quality based on spatial and temporal features. To this day, it is one of the leading metrics for evaluating compression quality [22]. However, VMAF is highly susceptible to specific image processing techniques. To mitigate this vulnerability, Netflix released VMAF NEG [23], an improved version of VMAF with enhanced stability but reduced correlations with subjective scores. For IQA, LPIPS by Zhang et al. [24] has gained popularity. It utilizes AlexNet or VGG as a backbone and computes quality scores based on the distance between features extracted from

reference and distorted images. Both variations of LPIPS were included in our analysis to evaluate their robustness. Ding et al. [25] proposed DISTS, which incorporates feature distance concepts and is specifically designed to be insensitive to resampling of visual textures. AHIQ, introduced by Lao et al. [26], compares image patches, incorporates spatial details, and predicts quality by considering the varied contributions of different patches. PieAPP, developed by Prashnani et al. [27], trains a network using pairwise learning on a dataset labeled with probabilities of human preference between two images. Finally, ASNA-MACS by Ayyoubzadeh et al. [28] employs a Siamese-Difference neural network architecture, incorporating spatial and channel-wise attention mechanisms.

Considering NR metrics, Hosu et al. [29] introduced Koncept512, which employs InceptionResNetV2 as a backbone and trains an IQA model on KonIQ-10k [29]. Li et al. [30] proposed Linearity, which uses the "Norm-in-Norm" loss function to improve convergence and overall performance. Ying et al. [31] proposed PaQ-2-PiQ, which utilizes a ResNet-18 backbone and Region of Interest (RoI) pooling to aggregate quality predictions from image patches. Yang et al. [32] introduced MANIQA, an NR IQA method built on a Vision Transformer (ViT) backbone that uses a Transposed Attention Block and Scale Swin Transformer Block for prediction. Zhang et al. [33] developed DB-CNN, a model that combines two CNNs to address both synthetic and authentic distortions. Another notable metric, UNIQUE, proposed by Zhang et al. [34], is trained on multiple IQA databases and optimizes fidelity loss for robust performance. Chen et al. [35] presented TOPIQ-NR, that employs cross-scale attention within a coarse-to-fine network. This design propagates features in a top-down manner to effectively assess image quality. In the domain of video quality assessment, Li et al. [36] developed VSFA, an NR VQA metric tailored for in-the-wild videos, which accounts for content dependency and temporal memory effects in its predictions. This approach was later refined to MDTVSFA [37], incorporating a pipeline that includes a relative quality assessor, nonlinear mapping, and dataset-specific perceptual scale alignment.

The robustness of several NR metrics has been extensively studied in white-box scenarios. For instance, Antsiferova et al. [38] performed a comprehensive evaluation of 15 NR metrics. However, limited research exists on the robustness of FR metrics. In this paper, we address both of these understudied aspects: black-box adversarial attacks and the robustness of FR metrics.

## II. Attack Parameters

Table I outlines the parameters of our proposed UAP attacks, which we tuned for PaQ-2-PiQ. The experimental evaluation only varied the perturbation strength, $\varepsilon$. We used a fixed value of $\varepsilon = 0.1$, which we previously employed to generate white-box UAPs [39]. All UAPs underwent training over 10,000 iterations using the entire training dataset; the result was 50×10,000 = 500,000 queries total for each IQA metric. Experiments involving UAP sizes greater than 8×8 for DE or 32×32 for NES converged slower and delivered less overall performance under otherwise constant parameters.

Table II summarizes the parameters of the adapted black-box attacks; we similarly tuned these parameters for PaQ-2-PiQ. The experimental evaluation employed $\varepsilon$ values of $0.05$ and $0.1$ for attacks constrained by the $l_\infty$ norm. The value $0.05$ commonly serves in adversarial attacks, whereas $0.1$ facilitates direct comparisons with UAPs trained under the same $\varepsilon$. For attacks constrained by the $l_0$ norm, we selected parameter restrictions in accordance with the original methods. We made adjustments to ensure an equivalent number of perturbed pixels between Patch-RS and $l_0$-RS and to align the number of perturbed pixels in Frame-RS with these approaches. The transformation-parameter bounds for CLAHE, gamma correction, gamma + unsharp, and Drago's tonemap came from the original work [40].

## III. Metric List

Table III details the metrics this paper analyzes, including their type (FR or NR, IQA or VQA), upper and lower bounds, calculated range, original publication, and source-code link. We used the official IQA implementations from GitHub. Our choice of metrics was based on their widespread adoption (e.g., VMAF, DISTS, and LPIPS), strong correlation with subjective-quality scores (e.g., VMAF, DB-CNN, MDTVSFA, and UNIQUE), and robustness to white-box adversarial attacks as demonstrated in [38] (e.g., MANIQA and PaQ-2-PiQ).

To calculate the metric range, we used the DIV2K_valid_HR subsample from DIV2K [41]. For FR metrics, we applied JPEG compression at quality levels of 98 and 10 to generate high and low metric scores, respectively.

All our experiments applied VQA metrics to images as if they were single-frame videos. For Vimeo-90k [43], we computed IQA metrics for each frame and averaged the results over all frames in the video.

## IV. Black-box Attacks on AGIQA-3k Dataset

Table IV shows results of black-box attacks on the AGIQA-3k dataset. Compared with the KonIQ-10k experiments, the results are mostly the same. But the Ran et al. attack delivered generally lower gain than NES. See the main paper for our discussion of the results.

## V. Black-box Attacks for FR Metrics

This section describes black-box adversarial attacks on FR metrics, which measure the difference between a reference image and its distorted counterpart. In this case we added compression to the optimized function:

$$J^{\mathrm{FR}}(x, x^{ref}) = 1 - k_f \frac{f(\mathrm{JPEG}(x), x^{ref})}{\mathrm{range}(f)},$$

where JPEG represents JPEG compression with a quality of 80, which distorts an image but leaves it similar to the original. We employed compression here as a postprocessing step for the adversarial example to obtain different FR-metric values that are far from their lower or upper bounds.

TABLE I
OVERVIEW OF THE UAP BLACK-BOX ADVERSARIAL ATTACKS AND THEIR CORRESPONDING PARAMETER VALUES USED IN OUR EXPERIMENTS.

| Method | Restriction | Bound parameter | Other parameters |
|---|---|---|---|
| DE UAP | $l_\infty$ | $\varepsilon = 0.1$ | popsize = 50, size = (1, 3, 8, 8), mut = 0.3, crossp = 0.5, max_iter = 200 |
| NES UAP | $l_\infty$ | $\varepsilon = 0.1$ | sigma=0.001, N=32, n=20, eta=0.01, max_iters=250 |
| Parsimonious UAP | $l_\infty$ | $\varepsilon = 0.1$ | max_queries=10000, batch_size=64, block_size=32, max_iters=10000 |
| Square Attack UAP | $l_\infty$ | $\varepsilon = 0.1$ | p_init=0.05, max_queries=10000 |
| Ran et al. UAP | $l_\infty$ | $\varepsilon = 0.1$ | p_init=0.05, max_queries=10000, n_sample = 1, n_squares = 1 |

The normalized-gain estimates for FR metrics on the KonIQ-10k subset appear in Table V. These metrics show better resistance to adversarial attacks than NR alternatives do. Some metrics, such as traditional MS-SSIM and the NEG version of VMAF, appear resilient to all our proposed attacks. Several FR models prove vulnerable, however. ASNA-MACS and PieAPP are among the most susceptible. **Square Attack** is the most efficient attack for FR metrics, whereas sparse attacks ($l_0$**-RS**, **Patch-RS**, and **Frame-RS**) are ineffective. Unrestricted attacks, such as **Drago's tonemap**, show promising results, especially against VMAF, for which they were designed. These attacks can also target other FR metrics, demonstrating similar vulnerabilities.

## VI. UAP SCALING

Figure 1 demonstrates the effect of the scaling parameter on UAP performance.

When we train UAP attacks with an $l_\infty$ bound of $\varepsilon = 0.1$, we can apply them to images with a certain amplitude $k \leq 1$ followed by additional clipping $Clip(x + k \cdot P, 0, 1)$. This process mimics the effect of UAPs generated using $\varepsilon = 0.1k$. Such an approach reduces the attack's visibility but also diminishes its effectiveness. Values of $k > 1$ can increase the gain of UAP attacks at the cost of greater visibility. To investigate this phenomenon, we calculated the normalized gains for $k = 0.2, 0.4, 0.8, 1.0, 2.0, 4.0, 8.0$. In general, attack efficiency increases with rising amplitude. But the gain for some metrics, such as TOPIQ-NR, falls as amplitude increases. For ineffective UAPs that yield low metric gains (e.g., MDTVSFA and MANIQA for certain attacks), the metric scores remain relatively unchanged regardless of amplitude. Interestingly, the sign of the metric gain shifts with amplitude in some cases, such as DB-CNN and Linearity. Successful UAPs tend to increase the metric gain with increasing amplitude even for $k > 1$. In most cases, however, the results decline beyond a certain amplitude, as PaQ-2-PiQ and DB-CNN demonstrate. This effect could be due to severe distortion of the base image. UAPs trained on Koncept512 avoid this issue and consistently increase with rising amplitude. Unsuccessful UAPs either leave the metric score unaffected (VSFA and MDTVSFA) or severely decrease it (TOPIQ-NR and Linearity in Parsimonious UAP).

## VII. UAP VISUALIZATIONS

Figure 2 presents the universal perturbations generated by our proposed methods for all evaluated NR IQA metrics. It highlights metrics that exhibit vulnerability to at least one UAP attack. The UAPs appear as images using the formula $5(P + 0.1)$, where $P$ represents a trained UAP. This transformation ensures the resulting pixel values are float numbers in the range $[0, 1]$, given that $P$ is confined to the interval $[–0.1, 0.1]$.

## VIII. BLACK-BOX UAPS FOR FR METRICS

Table VI shows the normalized gains achieved by our UAP attacks on FR metrics. The results indicate most FR metrics are resistant to the generated UAPs, probably because of their strong reliance on the reference image, which aids in accurate predictions. But several UAPs with positive gains correspond to successful attacks, reflecting the original methods' effectiveness in certain cases.

## IX. BLACK-BOX UAPS FOR NR METRICS ON VIMEO-90K AND NIPS 2017 DATASETS

Table VII shows results for black-box attacks on the Vimeo-90k and NIPS 2017 datasets. Relative to the experiments on KonIQ-10k, the results are mostly the same.

## X. VISIBILITY OF ADVERSARIAL ATTACKS

To assess the perceptibility of adversarial distortions, we have conducted a subjective evaluation on the visibility of the tested attacks. Specifically, we selected 5 images from KonIQ-10k and generated adversarial examples optimized for TOPIQ-NR metric, resulting in a total of 85 images. We have conducted the experiment on Subjectify.us and asked 480 participants to point to the image with the highest quality. The evaluation is summarized in Figure 3. The results show that all attacks resulted in a decrease in subjective quality scores. This indicates that each adversarial attack is noticeable and negatively affects subjective quality, even as it increases the metric output.

## XI. DEFENSES AGAINST BLACK-BOX ADVERSARIAL ATTACKS

The adversarial attacks presented in our paper employ various techniques to generate perturbations, necessitating an

extensive experimental evaluation to determine the most suitable defense for each attack type. We tested 2 basic defenses against pre-attacked images, as summarized in Table VIII. The table presents the absolute value of the normalized gains, averaged over seven NR IQA metrics, demonstrating that these defenses significantly reduce attack effectiveness.

## REFERENCES


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

[2] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

[3] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," 2019.

[5] Qingbing Sang, Hongguo Zhang, Lixiong Liu, Xiaojun Wu, and Alan C Bovik, "On the generation of adversarial examples for image quality assessment," *The Visual Computer*, pp. 1–16, 2023.

[6] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[7] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[9] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and DA Forsyth, "Unrestricted adversarial examples via semantic manipulation," in *International Conference on Learning Representations*, 2019.

[10] Rima Alaifari, Giovanni S Alberti, and Tandri Gauksson, "Adef: an iterative algorithm to construct adversarial deformations," in *International Conference on Learning Representations (ICLR 2019)*, 2019.

[11] Chaowei Xiao, Jun Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song, "Spatially transformed adversarial examples," in *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[12] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[13] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

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

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

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

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

[18] Phoenix Neale Williams and Ke Li, "Black-box sparse adversarial attack via multi-objective optimisation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12291–12301.

[19] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Ieee, 2003, vol. 2, pp. 1398–1402.

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

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

[22] Anastasia Antsiferova, Sergey Lavrushkin, Maksim Smirnov, Aleksandr Gushchin, Dmitriy Vatolin, and Dmitriy Kulikov, "Video compression dataset and benchmark of learning-based video-quality metrics," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13814–13825, 2022.

[23] Zhi Li, Kyle Swanson, Christos Bampis, Lukáš Krasula, and Anne Aaron, "Toward a better quality metric for the video community," *The Netflix Tech Blog*, p. 2, 2020.

[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[25] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.

[26] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang, "Attentions help cnns see better: Attention-based hybrid image quality assessment network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1140–1149.

[27] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen, "Pieapp: Perceptual image-error assessment through pairwise preference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1808–1817.

[28] Seyed Mehdi Ayyoubzadeh and Ali Royat, "(asna) an attention-based siamese-difference neural network with surrogate ranking loss function for perceptual image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 388–397.

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

[30] Dingquan Li, Tingting Jiang, and Ming Jiang, "Norm-in-norm loss with faster convergence and better performance for image quality assessment," in *Proceedings of the 28th ACM International conference on multimedia*, 2020, pp. 789–797.

[31] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik, "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*, 2020, pp. 3575–3585.

[32] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1191–1200.

[33] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[34] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

[35] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE Transactions on Image Processing*, 2024.

[36] Dingquan Li, Tingting Jiang, and Ming Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 2351–2359.

[37] Dingquan Li, Tingting Jiang, and Ming Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1238–1257, 2021.

[38] Anastasia Antsiferova, Khaled Abud, Aleksandr Gushchin, Ekaterina Shumitskaya, Sergey Lavrushkin, and Dmitriy Vatolin, "Comparing the

robustness of modern no-reference image-and video-quality metrics to adversarial attacks," in *Proceedings of AAAI*, 2024, vol. 38, pp. 700–708.

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

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

[41] Eirikur Agustsson and Radu Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[42] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock, "Vmaf: The journey continues," *Netflix Technology Blog*, vol. 25, no. 1, 2018.

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

TABLE II

OVERVIEW OF THE ADAPTED BLACK-BOX ADVERSARIAL ATTACKS AND THEIR CORRESPONDING PARAMETER VALUES USED IN OUR EXPERIMENTS.

| Method | Restriction | Bound parameter | Other parameters | Code |
|---|---|---|---|---|
| Frame-RS | Frame range | $\epsilon = 1$ | max_queries=10000, p_init=0.8 rescale_schedule=True | Github |
| NES | $l_\infty$ | $\varepsilon = 0.05$ | sigma=0.001, N=32, n=20, eta=0.1, max_iters=250 | Github |
| $l_0$-RS | $l_0$ | $\epsilon = 169$ | max_queries=10000, p_init=0.8 | Github |
| CLAHE | Unrestr. | — | mu=18, lambda_=24, cxpb=0.5, mutpb=0.49, ngen=12, a0_min=1, a0_max=100, a1_min=0.00001, a1_max=20.0 | |
| Gamma correction | Unrestr. | — | mu=18, lambda_=24, cxpb=0.5, mutpb=0.49, ngen=12, a0_min=1, a0_max=100, a1_min=0.00001, a1_max=20.0 | |
| Gamma + Unsharp | Unrestr. | — | mu=18, lambda_=24, cxpb=0.5, mutpb=0.49, ngen=12, a0_min=30, a0_max=120, a1_min=30, a1_max=120, a2_min=0, a2_max=2 | |
| Drago's tonemap | Unrestr. | — | mu=18, lambda_=24, cxpb=0.5, mutpb=0.49, ngen=12, a0_min=0, a0_max=2.5, a1_min=0, a1_max=3, a2_min=0, a2_max=1 | |
| Parsimonious | $l_\infty$ | $\varepsilon = 0.05$ | max_queries=10000, batch_size=64, block_size=32, max_iters=10000 | Github |
| Patch-RS | $l_0$ | $\epsilon = 169$ | max_queries=10000, p_init=0.8, n_restarts=1 | Github |
| Square Attack | $l_\infty$ | $\varepsilon = 0.05$ | p_init=0.05, max_queries=10000 | Github |
| Ran et al. | $l_\infty$ | $\varepsilon = 0.05$ | p_init=0.05, max_queries=10000, n_sample = 1, n_squares = 1 | Github |

TABLE III

OVERVIEW OF THE NR AND FR IQA METRICS UTILIZED IN OUR STUDY, INCLUDING DETAILS ON METRIC TYPE, UPPER AND LOWER BOUNDS, CALCULATED METRIC RANGE, ORIGINAL PAPER, AND LINKS TO THE ORIGINAL CODE.

| Metric | Metric type | Lower bound $b_l$ | Upper bound $b_u$ | Metric range | $k_f$ | Paper | Code |
|---|---|---|---|---|---|---|---|
| Koncept512 | NR/IQA | 26.403 | 66.869 | 40.466 | 1 | Hosu et al. [29] | Github |
| Linearity | NR/IQA | 25.781 | 83.227 | 57.446 | 1 | Li et al. [30] | Github |
| PaQ-2-PiQ | NR/IQA | 58.38 | 84.171 | 25.791 | 1 | Ying et al. [31] | Github |
| VSFA | NR/VQA | 0 | 1 | 1 | 1 | Li et al. [36] | Github |
| MANIQA | NR/IQA | 0 | 1 | 1 | 1 | Yang et al. [32] | Github |
| MDTVSFA | NR/VQA | 0 | 1 | 1 | 1 | Li et al. [37] | Github |
| DB-CNN | NR/IQA | 26.7 | 83.636 | 56.936 | 1 | Zhang et al. [33] | Github |
| UNIQUE | NR/IQA | -2.657 | 2.196 | 4.853 | 1 | Zhang et al. [34] | Github |
| TOPIQ-NR | NR/IQA | 0.215 | 0.822 | 0.607 | 1 | Chen et al. [35] | Github |
| AHIQ | FR/IQA | -1.497 | 0.786 | 2.283 | 1 | Lao et al. [26] | Github |
| PieAPP | FR/IQA | 4.921 | 0 | 4.921 | -1 | Prashnani et al. [27] | Github |
| LPIPS AlexNet | FR/IQA | 1.586 | 0 | 1.586 | -1 | Zhang et al. [24] | Github |
| LPIPS VGG | FR/IQA | 1.063 | 0 | 1.063 | -1 | Zhang et al. [24] | Github |
| MS-SSIM | FR/IQA | 0 | 1 | 1 | 1 | Wang et al. [20] | Github |
| ASNA-MACS | FR/IQA | -3.066 | 1.932 | 4.998 | 1 | Ayyoubzadeh et al. [28] | Github |
| DISTS | FR/IQA | 0.704 | -0.0 | 0.704 | -1 | Ding et al. [25] | Github |
| VMAF | FR/VQA | -33.136 | 97.986 | 131.123 | 1 | Li et al. [42] | Github |
| VMAF NEG | FR/VQA | -35.898 | 97.428 | 133.326 | 1 | Li et al. [23] | Github |

TABLE IV
THE ROBUSTNESS OF NR METRICS TO THE ADAPTED BLACK-BOX ADVERSARIAL ATTACKS IN TERMS OF THE NORMALIZED GAIN ESTIMATED ON THE SUBSAMPLE OF AGIQA-3K DATASET.

| | Koncept512 | Linearity | PaQ-2-PiQ | VSFA | MANIQA | MDTVSFA | DB-CNN | UNIQUE | TOPIQ-NR |
|---|---|---|---|---|---|---|---|---|---|
| Frame-RS | 0.011 | 0.06 | 0.136 | 0.04 | 0.05 | 0.025 | 0.153 | 0.079 | 0.042 |
| NES | **0.523** | **0.514** | 0.414 | 0.233 | 0.076 | **0.126** | 0.522 | **0.351** | **0.464** |
| CLAHE | 0.105 | 0.023 | 0.172 | 0.009 | 0.041 | 0.007 | 0.054 | 0.023 | 0.014 |
| Gamma correction | 0.019 | 0.029 | 0.017 | 0.0 | 0.034 | 0.003 | 0.062 | 0.065 | 0.059 |
| Gamma + Unsharp | 0.122 | 0.055 | 0.168 | 0.02 | 0.078 | 0.011 | 0.105 | 0.077 | 0.071 |
| Drago's tonemap | 0.192 | 0.087 | 0.218 | 0.038 | 0.069 | 0.018 | 0.146 | 0.088 | 0.11 |
| Parsimonious | 0.302 | 0.268 | 0.511 | **0.28** | **0.242** | 0.121 | 0.371 | 0.25 | 0.203 |
| Patch-RS | 0.269 | 0.109 | 0.259 | 0.046 | 0.069 | 0.026 | 0.134 | 0.126 | 0.069 |
| Square Attack | 0.318 | 0.153 | 0.457 | 0.165 | 0.19 | 0.076 | 0.116 | 0.189 | 0.062 |
| Ran et al. | 0.228 | 0.413 | **0.59** | -0.028 | -0.058 | -0.0 | **0.59** | 0.225 | 0.358 |

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

| | AHIQ | PieAPP | LPIPS AlexNet | LPIPS VGG | MS-SSIM | ASNA-MACS | DISTS | VMAF | VMAF NEG |
|---|---|---|---|---|---|---|---|---|---|
| Frame-RS | -0.003 | 0.014 | -0.022 | -0.045 | -0.001 | 0.192 | -0.019 | 0.0 | -0.001 |
| NES | -0.001 | 0.068 | -0.003 | -0.026 | -0.004 | 0.136 | -0.022 | -0.012 | -0.01 |
| $l_0$-RS | -0.003 | 0.088 | -0.007 | -0.009 | -0.001 | 0.155 | 0.011 | 0.025 | 0.004 |
| CLAHE | 0.011 | -0.006 | 0.001 | -0.0 | -0.003 | 0.046 | -0.0 | 0.304 | -0.079 |
| Gamma correction | -0.001 | -0.0 | -0.0 | -0.0 | **0.0** | 0.002 | 0.0 | 0.0 | 0.0 |
| Gamma + Unsharp | 0.014 | -0.001 | 0.002 | **0.007** | -0.001 | 0.072 | 0.006 | 0.369 | -0.025 |
| Drago's tonemap | **0.028** | **0.393** | -0.002 | -0.0 | -0.01 | 0.041 | -0.001 | **0.392** | -0.007 |
| Parsimonious | -0.057 | 0.284 | -0.015 | -0.023 | -0.015 | **0.485** | 0.009 | 0.013 | -0.036 |
| Patch-RS | -0.003 | 0.032 | -0.0 | -0.0 | -0.0 | 0.021 | 0.003 | 0.056 | 0.0 |
| Square Attack | 0.001 | 0.207 | **0.003** | 0.005 | **0.0** | 0.466 | **0.028** | 0.032 | **0.006** |
| Ran et al. | -0.106 | -0.304 | -0.126 | -0.219 | -0.137 | -0.139 | -0.192 | -0.146 | -0.073 |

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

| | AHIQ | PieAPP | LPIPS AlexNet | LPIPS VGG | MS-SSIM | ASNA-MACS | DISTS | VMAF | VMAF NEG |
|---|---|---|---|---|---|---|---|---|---|
| DE UAP | **0.008** | -0.057 | -0.044 | -0.174 | -0.019 | **0.008** | -0.156 | -0.025 | -0.014 |
| NES UAP | -0.001 | -0.005 | -0.0 | -0.002 | -0.0 | -0.001 | **-0.0** | -0.001 | -0.0 |
| Parsimonious UAP | -0.076 | **0.015** | -0.02 | -0.026 | -0.013 | -0.037 | -0.035 | -0.049 | -0.047 |
| Square Attack UAP | -0.003 | -0.006 | **-0.0** | **-0.0** | **0.0** | -0.064 | -0.021 | **-0.0** | **0.0** |
| Ran et al. UAP | -0.172 | -0.377 | -0.38 | -0.49 | -0.241 | -0.579 | -0.533 | -0.315 | -0.259 |
| NES | -0.002 | 0.068 | -0.004 | -0.03 | -0.004 | 0.149 | -0.026 | -0.014 | -0.011 |
| Parsimonious | -0.099 | 0.409 | -0.04 | -0.062 | -0.042 | 0.604 | -0.037 | -0.048 | -0.105 |
| Square Attack | -0.003 | 0.23 | 0.001 | 0.002 | 0.0 | 0.487 | 0.017 | 0.04 | 0.003 |
| Ran et al. | -0.169 | -0.529 | -0.207 | -0.306 | -0.3 | -0.222 | -0.291 | -0.247 | -0.216 |

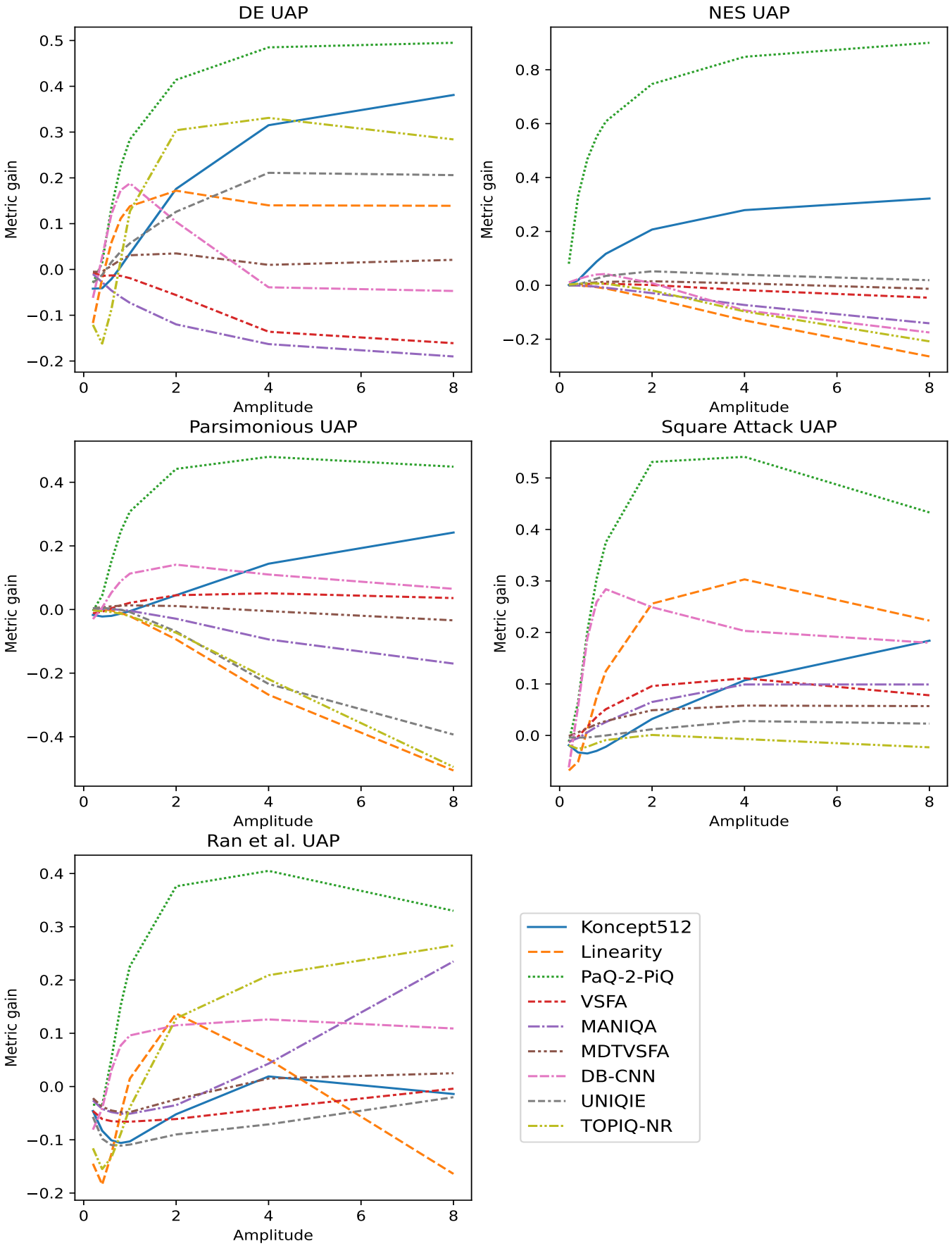


Fig. 1. Variation of the normalized gain for the proposed UAP black-box attacks estimated with different amplitudes and averaged across all testing datasets.

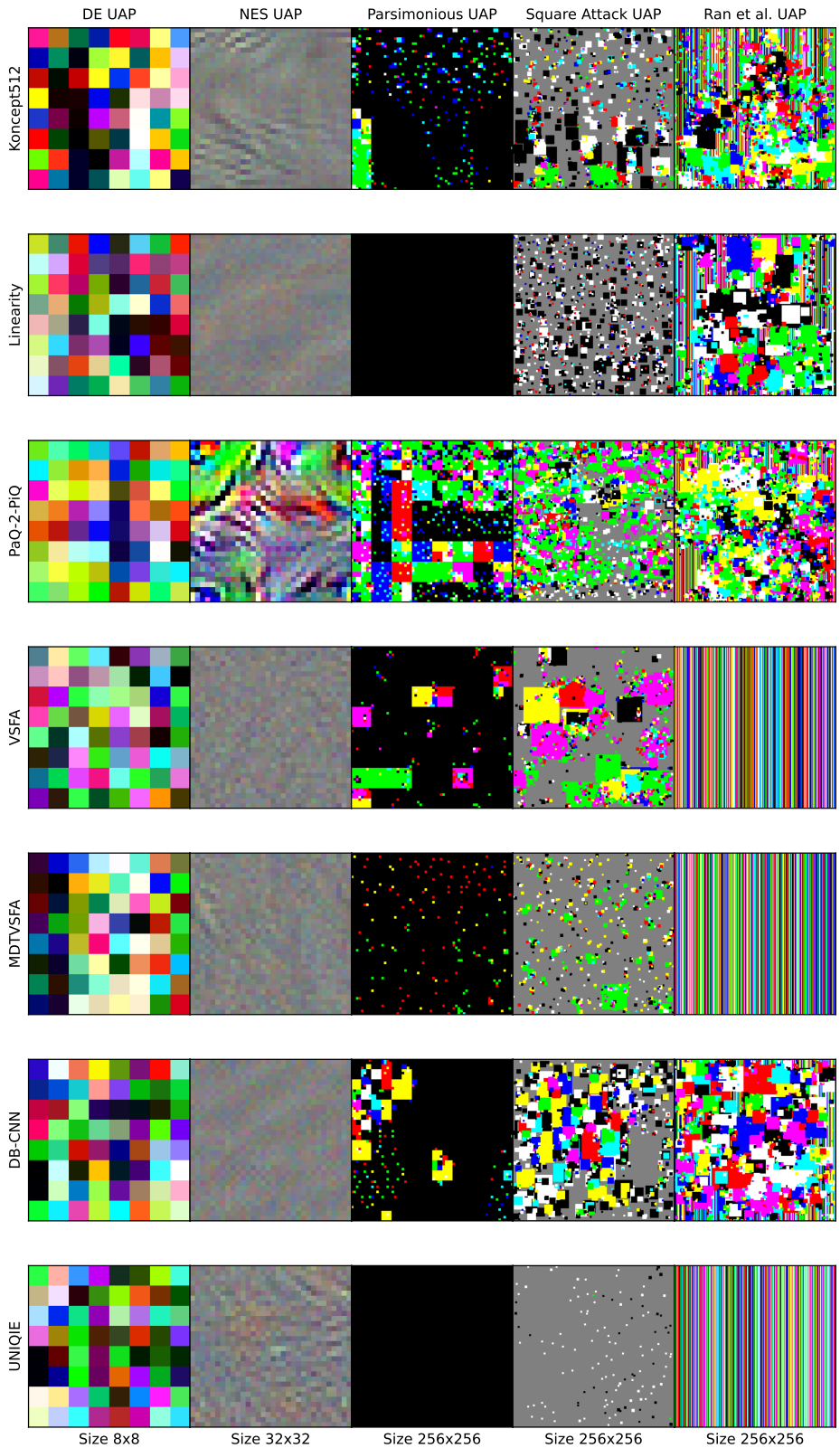


Fig. 2. Examples of UAPs generated for NR IQA metrics using the proposed black-box UAP attack methods.

TABLE VII
THE ROBUSTNESS OF THE NR METRICS TO THE PROPOSED UAP BLACK-BOX ADVERSARIAL ATTACKS IN TERMS OF THE NORMALIZED GAIN ESTIMATED ON VIMEO-90K AND NIPS 2017.

| | Koncept512 | Linearity | PaQ-2-PiQ | VSFA | MANIQA | MDTVSFA | DB-CNN | UNIQUE | TOPIQ-NR |
|---|---|---|---|---|---|---|---|---|---|
| **Vimeo-90k** | | | | | | | | | |
| DE UAP | 0.003 | 0.045 | 0.293 | -0.014 | -0.072 | 0.012 | 0.202 | **0.116** | -0.099 |
| NES UAP | **0.115** | -0.014 | **0.605** | 0.006 | -0.007 | 0.007 | 0.057 | 0.049 | **0.001** |
| Parsimonious UAP | -0.006 | -0.023 | 0.321 | 0.016 | -0.004 | 0.013 | 0.12 | -0.009 | -0.022 |
| Square Attack UAP | -0.021 | **0.121** | 0.389 | **0.073** | **0.018** | **0.032** | **0.303** | 0.0 | -0.004 |
| Ran et al. UAP | -0.1 | 0.025 | 0.236 | -0.066 | -0.049 | -0.048 | 0.11 | -0.104 | -0.034 |
| **NIPS 2017** | | | | | | | | | |
| DE UAP | -0.138 | -0.073 | 0.182 | -0.008 | -0.078 | 0.003 | 0.021 | **0.041** | -0.081 |
| NES UAP | **0.07** | -0.005 | **0.419** | 0.003 | -0.008 | 0.003 | 0.032 | 0.039 | **0.002** |
| Parsimonious UAP | -0.022 | -0.008 | 0.166 | 0.017 | **0.002** | 0.001 | 0.028 | -0.003 | -0.005 |
| Square Attack UAP | -0.057 | **0.014** | 0.221 | **0.047** | -0.005 | **0.008** | **0.078** | -0.008 | -0.004 |
| Ran et al. UAP | -0.137 | -0.102 | 0.108 | -0.07 | -0.071 | -0.054 | -0.059 | -0.169 | -0.1 |

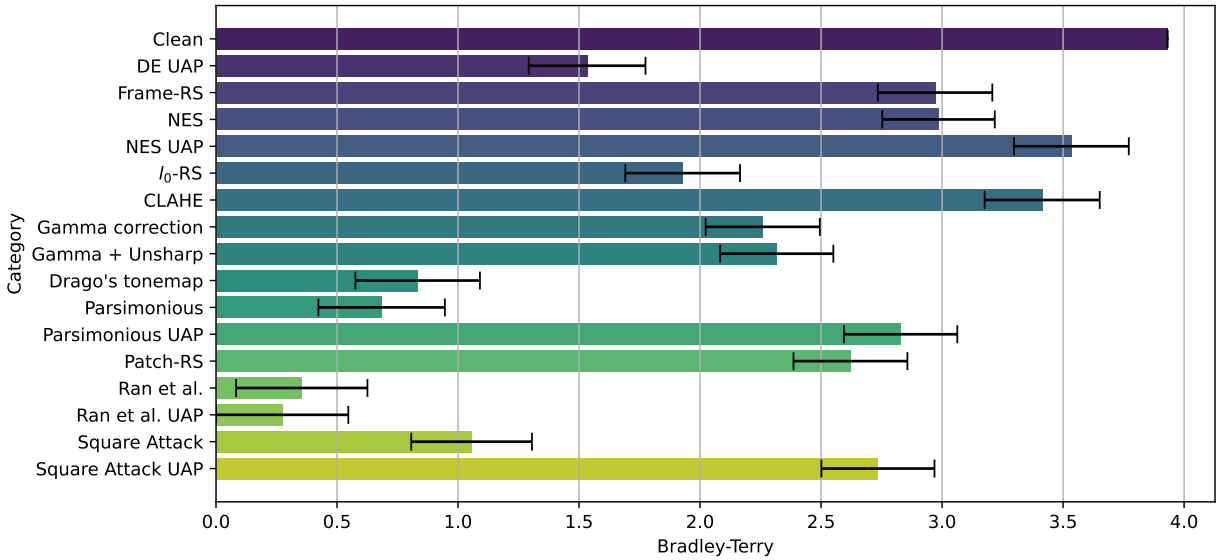


Fig. 3. Subjective comparison of adversarial examples.

TABLE VIII
EVALUATION OF ADVERSARIAL DEFENSES.

| | Patch-RS | Square Attack | Parsimonious | NES |
|---|---|---|---|---|
| W/o Defense | 0.2304 | 0.1859 | 0.2771 | 0.1522 |
| JPEG Compression | 0.1210 | 0.0381 | 0.0996 | **0.0138** |
| Flip | **0.0425** | **0.0236** | **0.0365** | 0.0433 |