PeerJ Computer Science

Assessing the visual realism of deepfake videos: an IQA-based approach

Ahmet Demir[1], Ahmet Emir Dirik[2] and Saffet Vatansever[3]

[1] Department of Electrical and Electronics Engineering, Bursa Uludag University, Bursa, Turkey
[2] Department of Computer Engineering, Bursa Uludag University, Bursa, Turkey
[3] Department of Mechatronics Engineering, Bursa Technical University, Bursa, Turkey

ABSTRACT

As deepfake videos become increasingly realistic, the ability of detection models to generalize across manipulation techniques with varying degrees of visual realism has become critical. However, many existing deepfake datasets contain visually low-realism samples that can hinder effective model training and result in misleading assessments of detection performance. While prior work has focused on dataset-level assessments of visual realism, systematic individual video-level research has received limited attention. To address this gap, this study proposes a video-level visual realism assessment method that quantifies distributional differences in no-reference image quality (NR-IQA) scores between pristine and manipulated videos using Cohen's *d* as an effect-size measure, yielding a per-video score referred to as the Video Fidelity Score (VFS). Experiments conducted on the FaceForensics++, FaceShifter, and Celeb-DF datasets characterize realism variations both within individual datasets and across datasets under a consistent experimental setup. The resulting VFS-based rankings exhibit trends that are consistent with Fréchet Inception Distance (FID) measurements, supporting the reliability of the proposed realism stratification. In addition, a dedicated human evaluation study confirms that VFS aligns well with human visual perception when distinguishing between visually high- and low-realism videos. Overall, the proposed approach offers a practical tool for realism-aware dataset curation, with potential implications for both deepfake generation and detection research.

Submitted 27 October 2025
Accepted 6 April 2026
Published 22 May 2026

Corresponding author
Ahmet Demir,
ahmetdemir@uludag.edu.tr

Academic editor
Paulo Jorge Coelho

Additional Information and Declarations can be found on page 19

DOI **10.7717/peerj-cs.3893**

Copyright
2026 Demir et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Artificial Intelligence, Computer Vision, Data Mining and Machine Learning, Multimedia, Security and Privacy
Keywords Deepfake, FID, Perceptual quality, Visual realism, VRA, Video fidelity, Realism assessment, Face-swap

INTRODUCTION

Recent advances in artificial intelligence (AI) and deep learning have enabled deepfake technology to generate highly realistic content. While this technology can be employed in socially beneficial applications such as filmmaking, advertising, virtual try-on systems, and digital entertainment, it also poses significant risks when used maliciously, including fake news, privacy breaches, and political manipulation (*Ciftci, Demir & Yin, 2020*; *Le & Woo, 2023*; *Tong et al., 2024*). Such malicious activities often rely on manipulated videos. As these threats grow, developing effective and reliable deepfake video detection methods, particularly those based on deep learning, has become increasingly vital (*Seow et al., 2022*; *Mirsky & Lee, 2021*). To support the development and evaluation of such detection models,

several deepfake video datasets have been proposed (*Rossler et al., 2019*; *Jiang et al., 2020*; *Dolhansky et al., 2019*; *Li et al., 2020b*; *Korshunov & Marcel, 2018*; *Yang, Li & Lyu, 2019*).

Despite notable progress in deepfake detection models and the availability of benchmark datasets, detection performance alone does not fully characterize the visual properties of manipulated videos that matter in practice. In particular, two videos that are both correctly classified as fake may differ substantially in their perceived visual realism, and thus in how convincing they appear to human observers. This observation highlights a complementary yet relatively underexplored dimension in deepfake research: the assessment of how visually realistic a manipulation appears, beyond it can be detected by a model. In this context, visual realism emphasizes manipulation-induced perceptual artifacts that influence the perceived plausibility and convincingness of a forged video, rather than generic degradations such as blur or compression alone (*Dragar et al., 2023*; *Peng et al., 2023*). This distinction is important in practice: a forged video may appear highly realistic to human viewers while still being detected by specialized models, whereas other manipulations may be immediately noticeable due to visually implausible artifacts, despite being generated by modern synthesis techniques.

Adopting a realism-oriented perspective has concrete implications for both detection research and dataset construction. From a detection standpoint, training and evaluating models on data with diverse levels of visual realism is important to prevent learning shortcut cues associated with low-quality artifacts and to better reflect real-world scenarios, where manipulations can be subtle. From a dataset design perspective, realism-aware curation helps ensure that datasets contain a representative range of difficulty levels, rather than being dominated by visually obvious forgeries. Although some datasets provide relatively realistic manipulated videos and are therefore more suitable for detection system development, they may still include samples with low perceptual realism that appear unnatural and are easily recognizable to the human eye, even when produced by state-of-the-art methods (*Le & Woo, 2023*; *Concas et al., 2024*). The inclusion of such visually unrealistic samples can introduce noise during training, degrade model generalization, and lead to overly optimistic estimates of detection performance during evaluation (*Lin et al., 2024*; *Le & Woo, 2023*). Consequently, distinguishing between low- and high-realism videos is important for realism-aware dataset curation and for better reflecting the diversity of conditions encountered in downstream detection scenarios.

Although there are several studies in the literature that assess overall visual realism at the dataset level (*Rossler et al., 2019*; *Jiang et al., 2020*; *Cai et al., 2024*; *Kwon et al., 2021*; *Hou et al., 2024*) using either subjective human evaluation or widely adopted generative model metrics such as Fréchet Inception Distance (FID) (*Heusel et al., 2017*) and Inception Score (IS) (*Barratt & Sharma, 2018*), these approaches primarily provide an aggregated view of realism. As a result, they do not directly yield an interpretable per-video realism score. Consequently, substantial within-dataset variability—arising from factors such as source quality, face region size, synthesis conditions, and codec effects—may remain hidden under a single subset label. Therefore, video-level realism characterization is crucial for identifying outliers, forming controlled "*best/worst*" partitions, and enabling fair

comparisons within and across datasets. Overall, principled and scalable per-video realism assessment remains an open and underexplored research problem.

Building on this observation, our central hypothesis is that deepfake generation processes induce not only semantic alterations but also systematic deviations in visual quality characteristics relative to the corresponding pristine videos. As the extent of manipulation increases, these deviations become more pronounced in the distributions of no-reference image quality assessment (NR-IQA) scores. Based on this observation, we propose to quantify such distributional deviations using an effect-size formulation, namely Cohen's d (*Cohen, 2013*), yielding a per-video score that measures how far a manipulated video departs from the visual quality profile of its original counterpart. Importantly, this formulation enables the construction of a lightweight, video-specific reference quality model using only the mean and standard deviation of NR-IQA scores computed from the pristine video.

Related work

A prominent effort in visual realism assessment (VRA) is the DFGC-VRA benchmark introduced by *Peng et al. (2023)*, which defines realism assessment as the subjective counterpart to objective deepfake detection. It provides a standardized evaluation protocol based on human Mean Opinion Scores (MOS) through a dedicated challenge setting. The benchmark is built on the MOS-annotated DFGC-2022 competition dataset (*Peng et al., 2022*), providing a common reference for supervised realism assessment. As the evaluation is tightly coupled to the specific data, domain, and annotation protocol of the competition, the reported results are inherently benchmark-dependent, and broader validation across diverse datasets remains an open issue (*Peng et al., 2023*).

Building on the MOS-annotated DFGC-2022 dataset, *Sun et al. (2023)* present a systematic study on face-swap visual realism assessment by comparing a range of feature representations, from handcrafted descriptors to deep and transferred features, for video-level MOS prediction. Their findings show that deep representations perform well under standard benchmark settings, but performance degrades under more challenging evaluation protocols, pointing to unresolved issues of robustness and generalization. In addition, the reliance on MOS supervision makes such approaches dependent on the availability and consistency of subjective annotations, limiting their applicability for annotation-free realism scoring on unseen datasets (*Sun et al., 2023*).

Along the same MOS-regression paradigm, *Dragar et al. (2023)* propose an ensemble of two convolutional neural network (CNN) backbones trained on the DFGC-2022 dataset to regress video realism scores from frame sequences and report competitive performance in the associated VRA challenge. While this demonstrates the feasibility of learning strong supervised predictors on MOS-labeled data, such models are inherently coupled to the training domain and the subjective labeling protocol. As a result, their portability to different datasets or codec and acquisition conditions typically requires additional validation and, potentially, re-training (*Dragar et al., 2023*). Collectively, these MOS-supervised approaches provide valuable baselines and benchmarks for VRA (*Peng et al., 2023*; *Sun et al., 2023*; *Dragar et al., 2023*). However, because they rely on

subjective MOS annotations, they do not directly offer an annotation-free, scalable way to score per-video realism across unseen datasets and varying codec or acquisition conditions.

Beyond MOS-supervised realism assessment, a recent study *Kim et al. (2024)* investigates the correlation between dataset quality and detection performance across several datasets and models, using no-reference image quality assessment (NR-IQA) (*An et al., 2025*) metrics. While the study mentions video-level visual quality (implicitly visual realism) measurement, this serves as an intermediate step toward aggregated dataset-level analysis; no direct per-video analysis or interpretation is provided. Moreover, it does not explicitly account for codec-related differences across subsets within datasets, where manipulation types differ not only in generative approach but also in compression level and bitrate. Since many NR-IQA metrics are sensitive to compression artifacts, failing to address such factors could introduce bias and affect the validity of conclusions regarding the relationship between perceived visual quality and detection accuracy. In another study, *Song et al. (2024)* propose a curriculum learning framework to improve deepfake detector generalization by estimating video-level forgery quality scores, derived from detection model feedback and identity similarity, which then guide the training process. While the approach provides a structured and adaptive learning path, its reliance on dynamic quality assessment may pose practical challenges. Furthermore, since the method incorporates a face recognition model, ArcFace, to assess identity similarity, it may carry a risk of demographic bias if the similarity metrics are not adequately balanced—that is, not calibrated to account for variations across different demographic groups.

Despite these efforts, there remains a need for an annotation-free, model-independent, and codec-aware per-video visual realism measure that can be applied consistently across datasets generated by different manipulation methods and synthesis pipelines, under varying codec conditions.

Contributions

Motivated by the limitations revealed by prior work, particularly the lack of annotation-free, model-independent, and codec-aware per-video realism measure, this article proposes a video-level visual realism assessment method based on quantifying distributional shifts in NR-IQA scores between pristine and manipulated videos using an effect-size formulation. Unlike prior approaches that rely on detection model outputs or overlook codec-related effects, the proposed method enables consistent and interpretable measurement of the visual realism of individual videos in a systematic and model-independent manner. To demonstrate the practical utility of the proposed method, we first apply it to a cross-dataset visual realism analysis on the distinct video subsets from the FaceForensics++ dataset (https://github.com/ondyari/FaceForensics) (*Rossler et al., 2019*). We then assess its ability to distinguish between high- and low-realism videos within each individual subset.

Following this analysis, the proposed framework is further evaluated on additional datasets to examine its robustness and generalization across different manipulation methods and data characteristics. Specifically, we extend the analysis to the FaceShifter

dataset (*Li et al., 2020a*), which is generated using the same source videos and preprocessing protocol as FaceForensics++ but employs a different face manipulation method, and to the Celeb-DF dataset (*Li et al., 2020b*), which is widely regarded as a high-realism deepfake benchmark. Through these experiments, we analyze variations in visual realism both across datasets and within individual datasets. The observed realism trends are consistent with widely adopted distribution-based measures, particularly the Fréchet Inception Distance (FID) (*Heusel et al., 2017*), suggesting that the proposed Video Fidelity Score (VFS) captures meaningful differences in perceptual realism. In addition, we examine the relationship between VFS and human visual perception in two complementary ways. First, we compare VFS-based realism trends with the human evaluation results reported for FaceForensics++ in *Rossler et al. (2019)*. Second, we conduct a dedicated human evaluation study on FaceForensics++ to directly assess whether VFS aligns with human judgments when distinguishing between visually high- and low-realism videos. Finally, to assess the role of the proposed effect-size-based formulation, we repeat the same analysis using solely the raw NR-IQA scores of the manipulated (deepfake) videos (*i.e.*, without incorporating the effect-size measure) and compare the resulting trends against those obtained using VFS.

The main contributions of this work can be summarized as follows:

- **Video Fidelity Score (VFS) for per-video deepfake visual realism.** We introduce VFS, an annotation-free and detector-agnostic per-video realism measure that quantifies distributional differences in NR-IQA scores using an effect-size formulation (Cohen's d) computed between matched real–fake video pairs. By explicitly measuring distributional shifts relative to the corresponding pristine source, the proposed formulation mitigates bias introduced by source quality and compression variability, yielding an interpretable measure of manipulation-induced perceptual deviation.
- **Realism stratification for dataset analysis and curation.** We demonstrate how VFS can be used to systematically stratify deepfake video datasets into high- and low-realism subsets for controlled analysis and realism-aware curation. The resulting stratifications are supported by converging evidence, including consistency with FID-based distributional measures, a dedicated human realism stratification study, and ablation against raw NR-IQA scoring.

These contributions lead to the following research questions:

RQ1: How can the visual realism of an individual deepfake video be measured in an objective, annotation-free manner that is consistent with human perception?

RQ2: Can the proposed Video Fidelity Score (VFS) reliably stratify deepfake videos into high- and low-realism subsets within and across datasets?

MATERIALS AND METHODS

Data

The proposed approach was evaluated on three deepfake video datasets, namely FaceForensics++, FaceShifter, and Celeb-DF. Table S1 provides an overview of the

datasets, manipulation methods, and the corresponding numbers of real and manipulated videos considered in this study.

FaceForensics++ (*Rossler et al., 2019*) was selected as the primary benchmark due to its comprehensive structure, which comprises four distinct subsets of manipulated videos generated using different deepfake methods. Specifically, the dataset includes 1,000 pristine source videos collected from YouTube, along with their corresponding manipulated versions produced by DeepFakes (*Repository, 2017a*), FaceSwap (*Repository, 2017b*), Face2Face (*Thies et al., 2016*), and NeuralTextures (*Thies, Zollhöfer & Nießner, 2019*). DeepFakes and FaceSwap perform full-face identity replacement, whereas Face2Face and NeuralTextures primarily modify facial expressions. Each video is available under three compression settings: raw (no compression), C23 (medium compression), and C40 (high compression). In this study, we randomly selected 100 original videos under the C23 compression setting and their corresponding manipulated versions for each manipulation method, resulting in a total of 400 fake videos. This selection enables controlled comparison across manipulation types under uniform compression conditions. Throughout the article, each video subset in FaceForensics++ is referred to by the name of the manipulation method used to generate it.

To assess the robustness of the proposed approach under a different face manipulation method while keeping the source content consistent, we additionally evaluated it on the FaceShifter dataset (*Li et al., 2020a*). FaceShifter performs full-face identity swapping and is constructed using the same set of pristine source videos as FaceForensics++, with videos released under multiple compression levels. As summarized in Table S1, the dataset contains 1,000 real videos and 1,000 corresponding manipulated videos. For consistency with the FaceForensics++ experiments, we used 100 real videos and their corresponding 100 FaceShifter-manipulated videos under the C23 compression setting.

To further examine generalization across different data distributions and synthesis pipelines, we employed the Celeb-DF. Celeb-DF contains 590 real videos and 5,639 manipulated videos generated using an improved deepfake synthesis process designed to reduce common visual artifacts. The real videos are collected from publicly available YouTube clips of 59 celebrities with diverse genders, ages, and ethnic backgrounds. To ensure balanced comparisons and consistent experimental conditions across datasets, we randomly selected 100 real videos and 100 manipulated videos from Celeb-DF and used them as released in the dataset.

Preprocessing

From each selected video, 20 frames were extracted by uniformly dividing the video duration into equal temporal segments and selecting the middle frame from each segment. This frame sampling strategy was applied consistently across all datasets and manipulation methods. Facial regions in the extracted frames were detected using the HOG-based face detector provided by the Dlib library (*King, 2009*), after which the detected faces were cropped and resized to a resolution of 300 × 300 pixels. The same face detection and preprocessing pipeline was applied to all datasets to ensure that subsequent analyses reflect differences in visual realism rather than variations in data preparation.

NR-IQA metrics and implementation details

The proposed approach is built upon no-reference image quality assessment (NR-IQA) metrics, which serve as the primary perceptual cues for characterizing manipulation-induced visual artifacts in videos. NR-IQA (also referred to as blind IQA) estimates perceptual image quality without access to a pristine reference image (*Mittal, Moorthy & Bovik, 2012*). This property makes NR-IQA particularly suitable for scenarios involving manipulated visual content, where a distortion-free and frame-aligned reference image is typically unavailable for each manipulated sample. Even when original videos exist, full-reference comparisons at the frame level are often impractical due to differences introduced by manipulation pipelines (*e.g.*, face region replacement), compression, and preprocessing operations. Consequently, NR-IQA provides a practical and scalable mechanism for capturing perceptual quality variations in large-scale benchmarks involving manipulated videos.

In this work, we employ a set of traditional and deep learning–based NR-IQA metrics, namely BRISQUE (*Mittal, Moorthy & Bovik, 2012*), DBCNN (*Zhang et al., 2020*), LIQE (*Zhang et al., 2023*), TReS (*Golestaneh, Dadsetan & Kitani, 2022*), MUSIQ (*Ke et al., 2021*), and MANIQA (*Yang et al., 2022*). Each NR-IQA metric is evaluated independently, and no fusion or aggregation is performed at the NR-IQA level. We select this specific set of metrics because they are widely used in the NR-IQA literature and provide complementary perspectives on perceptual distortions in both authentic and synthetic imagery. BRISQUE (*Mittal, Moorthy & Bovik, 2012*) represents a classic and highly cited natural-scene-statistics-based approach that is sensitive to common distortions in the spatial domain. DBCNN (*Zhang et al., 2020*) extends learning-based assessment by combining deep and bilinear representations, and has been shown to perform competitively across a wide range of authentic and synthetic distortions. To reflect recent progress in perceptual quality modeling, we further include modern deep NR-IQA metrics that have demonstrated strong performance under diverse authentic distortions: LIQE (*Zhang et al., 2023*) and TReS (*Golestaneh, Dadsetan & Kitani, 2022*), as well as transformer-based approaches such as MUSIQ (*Ke et al., 2021*) and MANIQA (*Yang et al., 2022*), which aim to capture perceptual quality cues beyond low-level artifacts.

Importantly, the resulting NR-IQA scores are not used directly for realism assessment. Instead, they are integrated into the proposed Video Fidelity Score (VFS) formulation by computing an effect-size measure (Cohen's d) between the NR-IQA score distributions of matched real and manipulated video pairs, as detailed in the Methodology section.

All experiments were conducted on a desktop computer running Ubuntu 24.04 LTS, equipped with an Intel Core i7-8700K CPU, 32 GB of RAM, and an NVIDIA GeForce GTX 1080 GPU. The software environment consisted of MATLAB R2024a (The MathWorks, Natick, MA, USA) and Python 3.8 (Anaconda environment). The implementation utilized key MATLAB toolboxes, including the Image Processing Toolbox and Computer Vision Toolbox, as well as essential Python libraries including dlib (v19.24.6), opencv-python (v4.10.0), numpy (v1.26), and deep IQA frameworks such as pyiqa and torch.

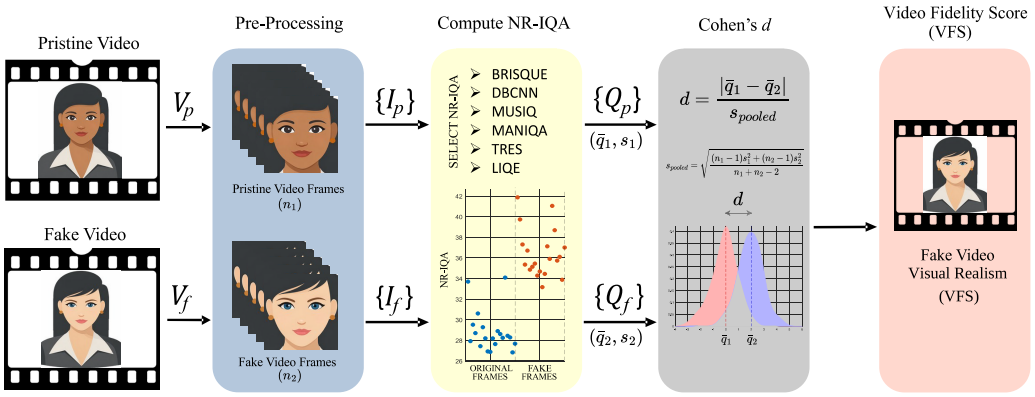


Figure 1 Illustration of the proposed method. The VFS value is calculated by measuring the distributional difference in NR-IQA scores between fake and source videos. Lower VFS values indicate higher visual realism, with VFS = 0 indicating no measurable distributional difference in NR-IQA scores between the fake and pristine videos. Full-size DOI: 10.7717/peerj-cs.3893/fig-1

Proposed approach

This section presents a robust video-level visual realism assessment technique for deepfake datasets. As illustrated in Fig. 1, it comprises three main stages:

(1) pre-processing of frames from both the fake video and its corresponding source (that is, pristine) video;

(2) computation of no-reference image quality assessment (NR-IQA) scores from the facial region of each selected frame;

(3) statistical quantification of the distributional shift between the NR-IQA scores of facial regions in fake and pristine video frames using Cohen's d (*Cohen, 2013*), a measure of standardized mean difference.

In the first stage, a predefined number of frames are selected from both the fake video and its corresponding pristine video. To ensure temporal coverage, these frames should be uniformly spaced across the entire video. For each selected frame, facial regions are detected using Dlib's HOG-based detector (*King, 2009*), though alternative face detection methods can be integrated. The detected face regions are then cropped and resized to a fixed resolution to maintain consistency in the subsequent realism assessment.

In the next stage, a no-reference image quality assessment (NR-IQA) score is computed for each pre-processed facial region using either a traditional or deep learning-based metrics. Figure 2 shows NR-IQA scores, computed separately using each of these metrics, for facial regions extracted from 20 frames of a pristine video (ID: 125) from the FaceForensics++ dataset and its four manipulated counterparts—generated using DeepFakes (*Repository, 2017a*), FaceSwap (*Repository, 2017b*), Face2Face (*Thies et al., 2016*), and NeuralTextures (*Thies, Zollhöfer & Nießner, 2019*)—from the FaceForensics++ dataset (*Rossler et al., 2019*). The results, shown in the figure, demonstrate that these metrics follow different scoring conventions: lower scores indicate higher perceptual quality for BRISQUE (*Mittal, Moorthy & Bovik, 2012*), whereas the opposite is true for the

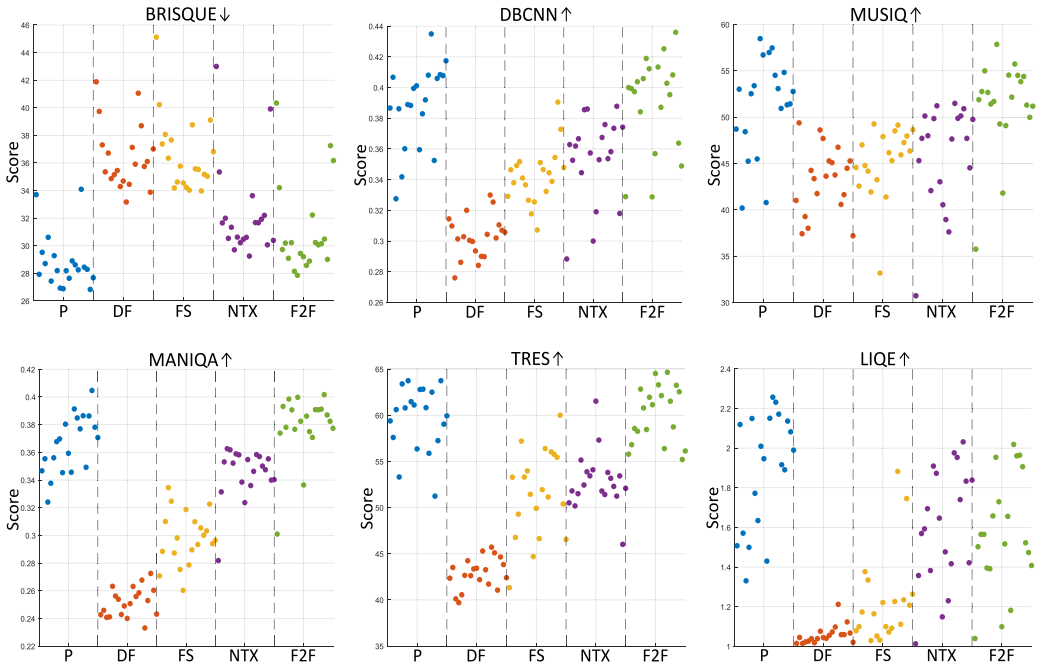


Figure 2 NR-IQA scores computed from facial regions in 20 frames of a pristine video (ID: 125) from the FaceForensics++ dataset and its four manipulated versions using six different NR-IQA metrics. The downward arrow (↓) signifies lower VFS values imply higher visual realism, while the upward arrow (↑) denotes the opposite. P, DF, FS, NTX, and F2F refer to Pristine, DeepFakes, FaceSwap, NeuralTextures, and Face2Face, respectively. Full-size DOI: 10.7717/peerj-cs.3893/fig-2

other methods. The figure also reveals that per-frame quality scores exhibit notable variability within a single video, regardless of the chosen metric. More critically, a consistent distributional divergence of NR-IQA scores between pristine and manipulated videos is observed across most metrics. In particular, the IQA scores resulting from videos generated by DeepFakes and FaceSwap (full-face replacement methods) show greater deviation from pristine scores than NeuralTextures and Face2Face (expression-modifying methods).

Although this final observation from Fig. 2 might suggest that IQA metrics can be directly used to assess manipulation-induced visual realism degradation in deepfake videos, their practical applicability in this context remains questionable for the following reasons. First, NR-IQA metrics are designed to capture a broad range of visual distortions—including noise, blur, compression artifacts, color shifts, unnatural contrast, and geometric deformations—many of which may not directly result from the deepfake generation process itself. Second, in deepfake datasets, the size of the facial region can vary significantly across videos, influencing the amount of visual information and the level of granularity. Third, the original acquisition conditions of source videos—such as resolution, bitrate, or camera specifications—may vary substantially, and these differences often persist even after standardized re-encoding under identical compression settings, causing residual artifacts that vary across videos. For example, C23-type source videos (encoded using constant rate factor 23) from the FaceForensics++ dataset (*Rossler et al., 2019*),

Table 1 Computed MANIQA-based VFS values for pristine video (ID: 125) from the FaceForensics++ dataset across different manipulation methods.

Manipulation method	VFS (MANIQA)	Visual realism
DeepFakes	6.95	Very bad
FaceSwap	3.41	Bad
NeuralTextures	1.04	Good
Face2Face	0.60	Good

evaluated in the "Results and Discussion" section, yielded BRISQUE (*Mittal, Moorthy & Bovik, 2012*) scores ranging from 24 to 58 and TReS (*Golestaneh, Dadsetan & Kitani, 2022*) scores from 27 to 96, even though all of them were encoded using the same compression level. Since deepfake videos often retain compression-related and resolution-dependent artifacts from their source videos, this variability further complicates direct visual realism comparisons between manipulated videos.

In the final stage of the proposed method, the means ($\hat{q}_1,\ \hat{q}_2$) and standard deviations ($s_1,\ s_2$) of the NR-IQA scores are computed separately for the fake and source videos, respectively. These statistics are then used to calculate Cohen's d (*Cohen, 2013*) to quantify manipulation-induced variations that reflect changes in visual realism. Cohen's d is computed as shown in Eq. (1):

$$d = \frac{|\hat{q}_1 - \hat{q}_2|}{s_{pooled}}, \tag{1}$$

where s_{pooled} is the pooled standard deviation, which is computed as:

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}. \tag{2}$$

Here, n_1 and n_2 are the sample sizes, while s_1^2 and s_2^2 are the variances of the two distributions. Cohen's d captures the degree of separation between the two score distributions in a scale-independent manner. The resulting value is termed as the Video Fidelity Score (VFS), which provides an interpretable measure of the visual impact introduced by the manipulation. A lower VFS implies that the fake video more closely resembles its source in terms of perceptual quality, indicating higher realism. Table 1 reports the MANIQA-based VFS values from the same videos used in Fig. 2 across different manipulation methods. To comply with ethical standards concerning identifiable subjects, representative video frames are not included in the manuscript; implementation details and additional experimental materials are available in our publicly accessible project GitHub (https://github.com/aahmetdemir/VFS) repository.

The complete implementation of the Video Fidelity Score (VFS) pipeline—including all MATLAB and Python scripts necessary for reproducing the experimental results—is publicly available on GitHub (https://github.com/aahmetdemir/VFS). The distributed code implements the VFS workflow for a single real–fake video pair. Given such a pair, the pipeline performs face extraction, computes the prescribed NR-IQA metrics, and returns the per-metric Cohen's d-based VFS value corresponding to the fake video.

Human evaluation protocol

To assess whether the proposed VFS is consistent with human visual perception in distinguishing visually high- and low-realism deepfake videos, we conducted a human-driven realism stratification study on FaceForensics++ (*Rossler et al., 2019*). The evaluation was performed using C23-compressed videos from the four manipulation subsets (DeepFakes, FaceSwap, NeuralTextures, and Face2Face).

Candidate videos were manually inspected to identify representative samples exhibiting clear differences in perceptual realism, and two balanced groups were curated: 50 videos judged to have good visual realism and 50 videos judged to have bad visual realism.

For each selected video, we computed (i) the proposed VFS values and (ii) a baseline using the corresponding direct outputs of NR-IQA models, both obtained *via* the same preprocessing and scoring pipeline described in the "Proposed Approach" subsection. The NR-IQA backbones considered were BRISQUE, DBCNN, MUSIQ, MANIQA, TReS, and LIQE. To quantify how well each scoring scheme separates the human-labeled *good* and *bad* groups, we treated the *good* set as class 0 and the *bad* set as class 1, and computed the area under the ROC curve (AUC) for each metric-based variant.

Beyond separability, we tested whether the score distributions of the two groups differ in a statistically meaningful manner. Specifically, we applied Welch's t-test (*Welch, 1947*), which is suitable under potential heteroscedasticity (unequal variances), and the Mann–Whitney U test (*Mann & Whitney, 1947*), a non-parametric alternative that does not assume normality. Statistical significance was assessed at the 5% level; results with $p < 0.05$ were considered statistically significant, suggesting that the observed differences between the two groups are unlikely to be due to random variation.

RESULTS AND DISCUSSION

Cross-dataset visual realism characterization

This section evaluates the proposed method's ability to characterize visual realism variations across different deepfake datasets and manipulation methods. Specifically, we analyze the FaceForensics++ manipulation subsets (DeepFakes, FaceSwap, NeuralTextures, and Face2Face) (*Rossler et al., 2019*), as well as the FaceShifter (*Li et al., 2020a*) and Celeb-DF (*Li et al., 2020b*) datasets. For each dataset, Video Fidelity Score (VFS) values were computed at the video level using six NR-IQA metrics: BRISQUE, DBCNN, MUSIQ, MANIQA, TReS, and LIQE. Table 2 reports the mean VFS values for each FaceForensics++ subset, as well as for the FaceShifter and Celeb-DF datasets. Lower VFS values correspond to higher visual realism. As is evident from the table, among the FaceForensics++ manipulation methods, DeepFakes and FaceSwap consistently exhibit higher mean VFS values than NeuralTextures and Face2Face across most metrics, indicating lower perceived visual realism. This observation aligns with the human evaluation results reported by *Rossler et al. (2019)* and supports the intuition that expression-modifying methods (NeuralTextures and Face2Face) tend to preserve overall visual fidelity better than full-face replacement methods (DeepFakes and FaceSwap). The results further show that Celeb-DF, which is widely regarded as a high-realism deepfake

Table 2 Average VFS values computed using various NR-IQA metrics across different datasets, along with FID scores.

Datasets	$VFS^{(brq)}\downarrow$	$VFS^{(db)}\downarrow$	$VFS^{(msq)}\downarrow$	$VFS^{(mnq)}\downarrow$	$VFS^{(trs)}\downarrow$	$VFS^{(lq)}\downarrow$	FID↓
FF++: DeepFakes	2.81	2.52	2.22	4.03	3.52	2.86	16.27
FF++: FaceSwap	2.12	1.46	1.13	2.09	1.75	1.49	13.10
FaceShifter	1.73	2.39	1.91	2.48	3.36	2.08	12.33
FF++: NeuralTextures	1.23	1.39	0.98	1.08	2.09	1.37	8.94
FF++: Face2Face	0.50	0.94	0.45	0.95	1.11	0.57	8.77
Celeb-DF	0.47	1.00	0.84	1.51	1.39	1.22	8.03
PCC (**VFS**, **FID**)	0.97	0.85	0.86	0.93	0.77	0.87	

Notes:
- $VFS^{(x)}$ indicates VFS values computed using NR-IQA metric x, where $x \in \{brq, db, msq, mnq, trs, lq\}$, corresponding to BRISQUE, DBCNN, MUSIQ, MANIQA, TReS, and LIQE, respectively.
- The downward arrow (↓) shows that lower values denote better case.
- PCC denotes to the Pearson correlation coefficient between **VFS** and **FID** columns.

Table 3 Average runtimes to compute the video fidelity scores (VFS) using different NR-IQA metrics for a single real–fake video pair and for the full setup of 400 pairs (s = second, h = hour).

Test condition	BRISQUE	DBCNN	MUSIQ	MANIQA	TRES	LIQE
Single video pair	7.17 s	165.18 s	148.54 s	281.82 s	240.44 s	253.88 s
400 video pairs	0.80 h	18.35 h	16.50 h	31.31 h	26.72 h	28.21 h

benchmark, achieves comparatively lower mean VFS values across all considered metrics. Notably, its VFS values are consistently lower than those of the FaceForensics++ full-face replacement subsets and the FaceShifter dataset, indicating higher overall visual realism.

To further validate these findings, we compare VFS trends with the Fréchet Inception Distance (FID) (*Heusel et al., 2017*), a widely used distribution-based metric for assessing generative realism that has been shown to correlate with human perceptual judgments (*Jiang et al., 2020*). The FID scores reported in the last column of Table 2 follow trends that are consistent with the corresponding VFS values, with lower FID scores associated with lower VFS values and higher perceived realism. To quantify the relationship between VFS and FID, we computed the Pearson correlation coefficient (PCC) (*Benesty et al., 2009*) between the sequence of mean VFS values—computed using each NR-IQA metric across the considered datasets—and the corresponding FID score sequence, both ordered according to the dataset and subset ranking reported in Table 2. The resulting PCC values are shown in the last row of Table 2. From the table, it is clear that the VFS values based on BRISQUE and MANIQA exhibit the strongest correlation with FID, while those from the other NR-IQA metrics also yield considerably high PCC values.

To evaluate the computational efficiency of the proposed VFS pipeline, the total execution time required to compute the VFS value for a single real–fake video pair was measured using the hardware and software configuration described in the "Data and Preprocessing". Table 3 reports the average runtime for each NR-IQA metric when processing a single video pair. The results indicate that the traditional BRISQUE metric

Table 4 FID scores of the *best* and *worst* 25% videos in each datasets, categorized based on VFS rankings across different NR-IQA metrics, and their corresponding differences.

	FID for VFS$^{(brq)}$			FID for VFS$^{(db)}$			FID for VFS$^{(msq)}$			FID for VFS$^{(mnq)}$			FID for VFS$^{(trs)}$			FID for VFS$^{(lq)}$		
Datasets	**Worst**	**Best**	**Diff.**	**Worst**	**Best**	**Diff.**	**Worst**	**Best**	**Diff.**	**Worst**	**Best**	**Diff.**	**Worst**	**Best**	**Diff.**	**Worst**	**Best**	**Diff.**
FF++: DeepFakes	36.5	29.4	7.1	36.0	26.1	9.9	37.2	26.0	11.2	35.5	25.3	10.2	36.3	24.8	11.5	36.7	26.4	10.3
FF++: FaceSwap	29.3	25.5	3.8	32.1	21.0	11.1	31.4	23.4	8.0	32.2	24.6	7.6	34.7	23.1	11.6	31.3	23.3	8.0
FaceShifter	27.2	21.7	5.5	32.2	19.1	13.1	30.9	18.1	12.8	30.5	20.7	9.8	32.8	18.7	14.1	29.5	19.1	10.4
FF++: NeuralTextures	19.7	18.3	1.4	18.9	18.3	0.6	18.9	18.6	0.3	19.0	18.8	0.2	18.8	17.8	1.0	16.7	19.2	−2.5
FF++: Face2Face	19.8	18.8	1.0	19.1	13.7	5.4	20.3	18.4	1.9	23.8	14.7	9.1	17.3	20.9	−3.6	15.2	17.9	−2.7
Celeb-DF	21.9	15.8	6.1	18.2	16.9	1.3	18.4	13.6	4.8	18.0	14.3	3.7	19.5	15.0	4.5	19.0	16.8	2.2
Average			4.2			6.9			6.5			6.8			6.5			4.3

Note:
• brq, db, msq, mnq, trs, and lq refer to BRISQUE, DBCNN, MUSIQ, MANIQA, TReS, and LIQE, respectively.

executes within a few seconds, highlighting its computational simplicity. In contrast, the deep learning-based metrics, DBCNN, MUSIQ, MANIQA, TRES, and LIQE, incur additional processing overhead stemming from model initialization, GPU inference, and tensor operations.

Finally, to assess the added value of incorporating effect size (*via* Cohen's *d*), we examined the case of using solely the raw NR-IQA scores of the manipulated (deepfake) videos, without applying the proposed VFS formulation. Specifically, the average NR-IQA score of each dataset in Table S2 was correlated with the corresponding FID score sequence, both ordered by dataset ranking, using the PCC. The resulting correlation values were 0.21, −0.14, −0.06, −0.19, −0.05, and −0.31 for BRISQUE, DBCNN, MUSIQ, MANIQA, TReS, and LIQE, respectively (see Table S2). Since the direction of correlation depends on whether a metric assigns higher scores to better or worse visual realism, the sign is not meaningful. However, the relatively lower magnitudes highlight the limitations of relying solely on NR-IQA scores, as discussed in the "Proposed Approach".

Intra-dataset visual realism analysis

While FID (*Heusel et al., 2017*) is useful for measuring overall image quality in deepfake datasets, assessing quality at the level of individual videos remains a challenge for this method. In contrast, the proposed method is well-suited for per-video analysis, as it produces a VFS value independently of other videos. To explore this capability further, this section evaluates the effectiveness of the proposed method in identifying low and high-realism videos within each evaluated dataset (*i.e.*, the FaceForensics++ subsets as well as the FaceShifter and Celeb-DF). In this context, the VFS values of videos within each dataset (as computed in the "Cross-Dataset Visual Realism Characterization") were sorted. For each dataset, the top and bottom 25% of videos, based on their VFS values, were grouped as *best* and *worst*. FID was then used to validate whether these groups differ in terms of perceptual quality. Table 4 presents, for each evaluated dataset, the FID scores of these video groups that were categorized based on the VFS rankings obtained using

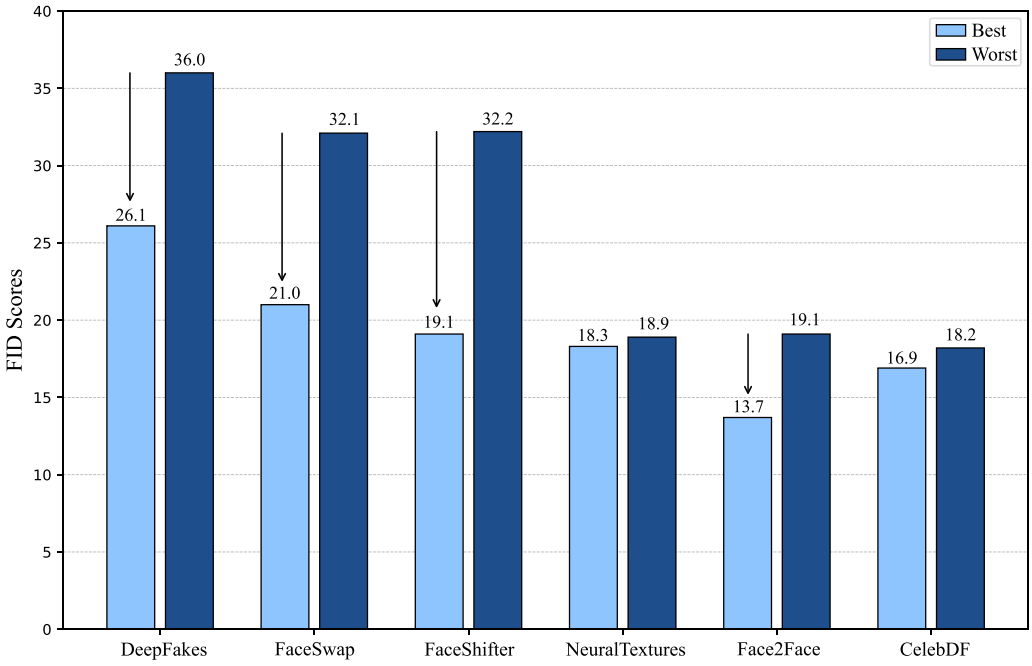


Figure 3 **FID scores for the *best* and *worst* 25% of videos ranked by DBCNN-based VFS values.**
Full-size DOI: 10.7717/peerj-cs.3893/fig-3

different NR-IQA methods: BRISQUE, DBCNN, MUSIQ, MANIQA, TReS, and LIQE. From the table, a considerable distinction between the *best* and *worst* groups can be observed for the DeepFakes and FaceSwap subsets, as well as for the FaceShifter dataset, across all metrics, with larger positive differences reflecting more apparent separation between high- and low-realism video groups. Celeb-DF also shows predominantly positive separations across all metrics, although the magnitude of separation is relatively smaller for most metrics. For the Face2Face subset, the DBCNN- and MANIQA-based VFS yielded substantial separation between the *best* and *worst* groups, whereas TReS- and LIQE-based VFS performed poorly, with negative difference values. For the NeuralTextures subset, the LIQE-based VFS was also ineffective, yielding a negative best–worst difference, whereas the other metrics achieved small but positive separation.

Overall, the average difference in FID scores between the *best* and *worst* video groups, computed across all evaluated datasets in Table 4, was 4.2, 6.9, 6.5, 6.8, 6.5, and 4.3 for BRISQUE, DBCNN, MUSIQ, MANIQA, TReS, and LIQE, respectively, with DBCNN and MANIQA providing the most consistent separation. To make these trends more evident, Figs. 3 and 4 illustrate the separation between the *best* and *worst* video groups for DBCNN- and MANIQA-based VFS. While the corresponding FID scores were already reported in Table 4, the bar plots emphasize the superior discriminative ability of these two metrics.

Here, to examine the added value of incorporating an effect size measure, an ablation study was conducted by ranking fake videos using their NR-IQA scores alone (*i.e.*, without applying the proposed VFS formulation). Based on this ranking, the top 25% and bottom

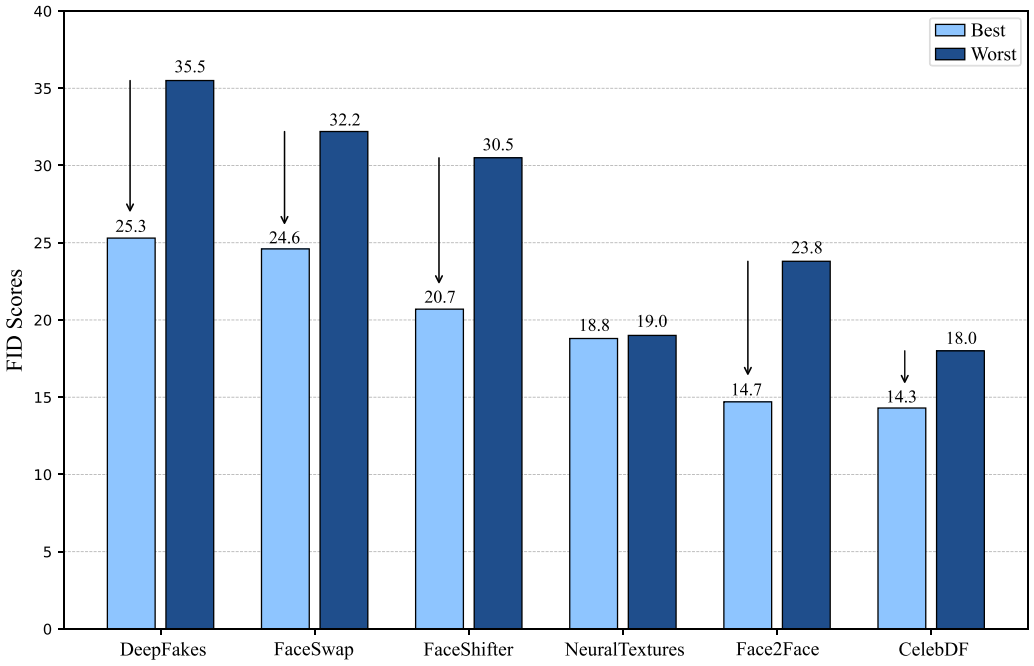


Figure 4 **FID scores for the *best* and *worst* 25% of videos ranked by MANIQA-based VFS values.**
Full-size DOI: 10.7717/peerj-cs.3893/fig-4

Table 5 **FID scores of the *best* and *worst* 25% videos in each FaceForensics++ subset, categorized based on NR-IQA scores rankings across different NR-IQA metrics, and their corresponding differences.**

	FID for BRISQUE			FID for DBCNN			FID for MUSIQ			FID for MANIQA			FID for TRES			FID for LIQE		
Datasets/subsets	**Worst**	**Best**	**Diff.**	**Worst**	**Best**	**Diff.**	**Worst**	**Best**	**Diff.**	**Worst**	**Best**	**Diff.**	**Worst**	**Best**	**Diff.**	**Worst**	**Best**	**Diff.**
DeepFakes	28.8	37.2	−8.4	33.2	33.9	−0.7	30.9	31.9	−1.0	34.1	31.7	2.4	26.9	35.3	−8.4	30.0	33.9	−3.9
FaceSwap	29.2	26.4	2.8	31.2	26.1	5.1	29.2	26.6	2.6	28.1	24.6	3.5	28.1	28.6	−0.5	28.5	28.3	0.2
FaceShifter	19.6	32.4	−12.8	21.5	29.3	−7.8	21.5	28.8	−7.3	20.3	27.1	−6.8	19.3	27.2	−7.9	19.6	29.3	−9.7
NeuralTextures	17.5	17.4	0.1	18.9	22.1	−3.2	19.7	21.6	−1.9	19.6	19.9	−0.3	17.4	21.8	−4.4	20.1	21.4	−1.3
Face2Face	15.9	20.6	−4.7	19.1	21.3	−2.2	17.9	20.6	−2.7	18.3	20.7	−2.4	18.3	22.3	−4.0	18.4	21.2	−2.8
Celeb-DF	22.3	20.8	1.5	16.6	16.3	0.3	15.3	16.5	−1.2	15.9	15.0	0.9	19.8	18.1	1.7	17.5	17.5	0.0
Average			−3.6			−1.4			−1.9			−0.5			−3.9			−2.9

25% were labeled as the *best* and *worst* groups, respectively, and the FID was computed for each group, as in the previous analysis. Table 5 reports the resulting FID scores and best–worst differences across the evaluated datasets for different NR-IQA metrics. Overall, the average FID differences were −3.6, −1.4, −1.9, −0.5, −3.9, and −2.9 for BRISQUE, DBCNN, MUSIQ, MANIQA, TReS, and LIQE, respectively. These results indicate that using raw NR-IQA scores from fake videos fails to reliably distinguish visual realism; in many cases, these scores assign lower FID values (indicating better realism) to the group labeled as *worst*, highlighting the limitations of relying solely on raw NR-IQA scores without the proposed VFS formulation.

Table 6 Average VFS values of *good* and *bad* video set and their AUC, Welch's *t*-test and Mann–Whitney U test results.

	$\text{VFS}^{(brq)}\downarrow$	$\text{VFS}^{(db)}\downarrow$	$\text{VFS}^{(msq)}\downarrow$	$\text{VFS}^{(mnq)}\downarrow$	$\text{VFS}^{(trs)}\downarrow$	$\text{VFS}^{(lq)}\downarrow$
AUC	0.81	0.87	0.83	0.85	0.80	0.83
VFS avg. (good set)	0.89	0.90	0.58	1.03	1.47	0.87
VFS avg. (bad set)	2.42	2.90	2.35	4.48	3.81	2.91
p-value (Welch's *t*-test)	1.06×10^{-7}	1.84×10^{-10}	1.43×10^{-8}	2.46×10^{-9}	2.71×10^{-8}	2.60×10^{-6}
p-value (Mann–Whitney U test)	1.22×10^{-7}	1.69×10^{-10}	1.49×10^{-8}	2.13×10^{-9}	3.31×10^{-7}	1.08×10^{-8}

Notes:
- $\text{VFS}^{(x)}$ indicates VFS values computed using NR-IQA metric x, where $x \in \{brq, db, msq, mnq, trs, lq\}$, corresponding to BRISQUE, DBCNN, MUSIQ, MANIQA, TReS, and LIQE, respectively.
- The downward arrow (↓) shows that lower values denote better case.

The analysis in this subsection reveals that even deepfake datasets widely regarded for their high visual realism, such as Celeb-DF and the FaceForensics++ expression-modifying subsets (NeuralTextures and Face2Face), may still contain visually low-realism videos. The presence of such videos can compromise the reliability of performance evaluations for deepfake detection models. Moreover, if not excluded from training data, these low-realism samples may introduce noise and bias into deep learning pipelines. These findings underscore the importance of quality-aware dataset curation for both evaluation and training, particularly in supporting the robustness and fairness of deepfake detection systems.

Human evaluation study

This subsection evaluates whether the proposed VFS aligns with human visual perception when distinguishing visually high- and low-realism deepfake videos. To this end, we conducted a human-driven realism stratification experiment on FaceForensics++ (*Rossler et al., 2019*) and compared the discriminative capability of VFS against the direct use of raw NR-IQA scores of the fake videos. All videos were manually inspected by the authors in a controlled and consistent manner across the four FaceForensics++ manipulation subsets (DeepFakes, FaceSwap, NeuralTextures, and Face2Face), and two balanced sets were curated: 50 videos judged to have good visual realism (*i.e.*, perceptually convincing with minimal visible artifacts) and 50 videos judged to have bad visual realism (*i.e.*, containing noticeable visual artifacts). Table S3 reports the video IDs and corresponding manipulation methods of the selected bad and good sets. The selection includes samples from all four manipulation subsets, ensuring a balanced and unbiased evaluation across manipulation methods. To comply with ethical standards concerning identifiable subjects, representative video frames are not included in the manuscript. Experimental details and documentation of the selected videos are available in the GitHub (https://github.com/aahmetdemir/VFS) project repository.

Following the protocol described in "Human Evaluation Protocol", we computed VFS scores and raw NR-IQA values for each selected video and quantified how well each method separates the two human-labeled groups. Tables 6 and 7 report the results for VFS and raw NR-IQA, respectively, including the AUC (area under the ROC curve), average scores for the good and bad sets, Welch's *t*-test and the Mann–Whitney U test.

Table 7 Average NR-IQA metric scores of *good* and *bad* video set and their AUC, Welch's *t*-test and Mann–Whitney U test results.

	BRISQUE ↓	DBCNN ↑	MUSIQ ↑	MANIQA ↑	TRES ↑	LIQE ↑
AUC	0.44	0.47	0.53	0.48	0.42	0.58
NR-IQA metric scores avg. (good set)	45.68	0.35	40.78	0.34	50.77	1.55
NR-IQA metric scores avg. (bad set)	44.25	0.36	39.67	0.34	54.45	1.50
p-value (Welch's *t*-test)	0.27	0.58	0.57	0.84	0.20	0.73
p-value (Mann–Whitney U test)	0.29	0.56	0.58	0.78	0.14	0.18

Note:
- The downward arrow (↓) shows that lower values denote better case, while the upward arrow (↑) denotes the opposite.

According to Table 6, VFS yields clear and consistent separability between the *good* and *bad* sets across all six NR-IQA metrics. The AUC values range from 0.80 to 0.87, with the best performance obtained using the DBCNN-based VFS (AUC = 0.87), followed by MANIQA-based VFS (AUC = 0.85).

Importantly, the average VFS values are consistently lower for the *good* set and higher for the *bad* set for all metrics, which matches the intended interpretation of VFS (lower indicates higher visual realism) and is consistent with the earlier FID-based trends (Table 2). For example, the DBCNN-based average VFS increases from 0.90 (*good* set) to 2.90 (*bad* set), while the MANIQA-based average increases from 1.03 to 4.48. The separation is also statistically compelling: for all VFS variants, both Welch's *t*-test and the Mann–Whitney U test yield extremely small *p*-values (*e.g.*, as low as 1.84×10^{-10} for DBCNN-based VFS), providing clear evidence that the two groups differ systematically.

In contrast, Table 7 indicates that directly using raw NR-IQA scores from the fake videos does not reliably distinguish the *good* and *bad* sets. The AUC values remain near chance level, ranging from 0.44 to 0.58, and the corresponding hypothesis tests do not support statistically significant separation. Concretely, Welch's *t*-test and the Mann–Whitney U test yield large *p*-values across all metrics (*e.g.*, $p = 0.84$ for MANIQA and $p = 0.73$ for LIQE under Welch's *t*-test), which are well above the conventional 0.05 threshold. This indicates that, under the human-curated *good/bad* split, the raw NR-IQA score distributions do not exhibit a consistent and statistically meaningful difference. These observations are consistent with our ablation results based on FID validation, where raw NR-IQA rankings produced unreliable *best/worst* groupings.

Figure 5 visualizes the AUC comparison between the proposed VFS (Proposed) and the raw NR-IQA scores, highlighting a consistent and substantial performance gap in favor of VFS across all metrics. Taken together with the cross-dataset FID analyses, these results provide converging evidence from two complementary validation routes: (i) VFS-based rankings form *best/worst* video groups that are supported by FID separation, and (ii) when videos are stratified by human visual judgment, VFS remains highly discriminative with statistically significant group differences, whereas raw NR-IQA scores do not. Overall, these findings indicate that the proposed VFS formulation provides a reliable video-level realism characterization that is closely aligned with human perception.

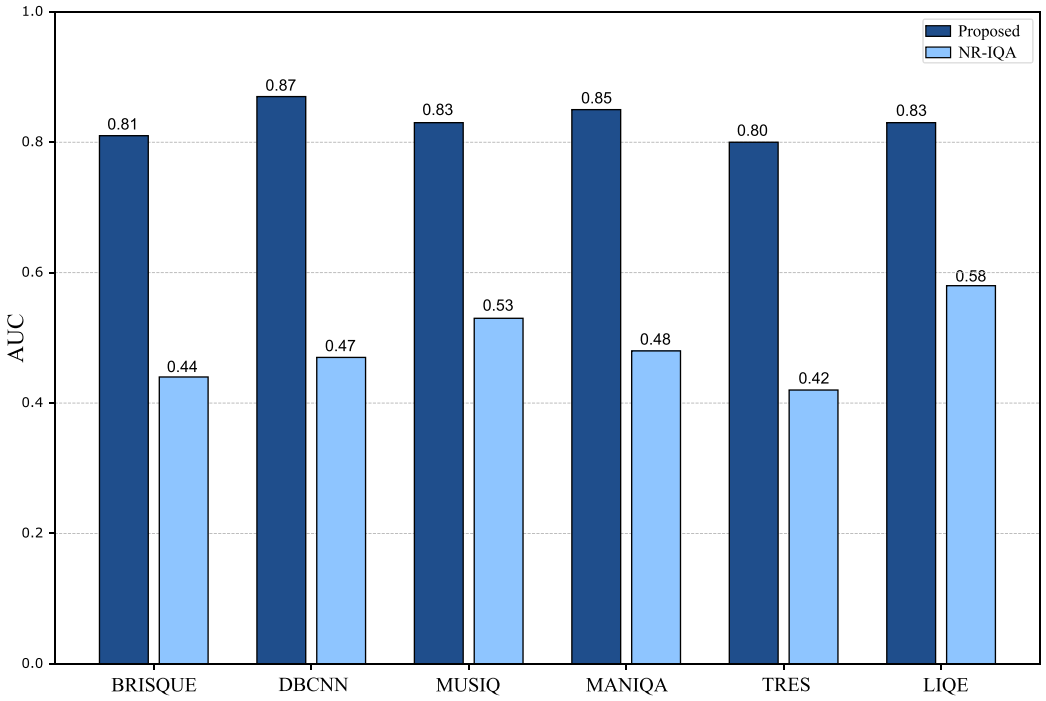


Figure 5 **Per-metric AUC comparison for good/bad video set discrimination: VFS (proposed) *vs.* NR-IQA.** Full-size DOI: 10.7717/peerj-cs.3893/fig-5

CONCLUSION AND FUTURE WORK

This study proposed a video-level visual realism assessment method that quantifies distributional differences in no-reference image quality (NR-IQA) scores between pristine and manipulated videos using Cohen's *d* as an effect-size measure, yielding a per-video score, termed the VFS. Unlike prior work that focuses on aggregate dataset-level analysis or relies on detection-model-based feedback, VFS enables direct and interpretable per-video realism assessment in a scalable and statistically grounded manner. Comprehensive experiments on FaceForensics++, FaceShifter, and Celeb-DF datasets demonstrate that VFS consistently captures variations in visual realism at multiple levels, including across and within manipulation subsets, within individual datasets, and across datasets under a unified evaluation framework. Through FID-based validation, we further showed that VFS-defined *best* and *worst* video groups exhibit clear distributional separation, supporting the effectiveness of VFS as a realism-aware grouping criterion. These results highlight the practical utility of VFS for dataset curation, enabling the identification of low-realism outliers and the construction of controlled *best/worst* partitions that can support realism-aware dataset curation and evaluation, with potential benefits for downstream deepfake detection research.

In addition to distributional validation, we conducted a dedicated human-driven realism stratification study on the FaceForensics++ dataset to directly assess alignment with human perception. The results show that VFS provides clear and consistent discrimination between human-labeled good- and bad-realism video groups across all

evaluated NR-IQA backbones (AUC = 0.80–0.87), with the strongest performance obtained using DBCNN- and MANIQA-based VFS. In contrast, using raw NR-IQA scores without incorporating the effect-size measure yields near-chance separability and fails to produce statistically meaningful group differences, underscoring the importance of the proposed VFS formulation for per-video realism assessment.

Despite these strengths, the proposed method requires paired real and fake videos, which limits its applicability to unpaired or standalone manipulations. Addressing this limitation represents an important direction for future work.

Future work will extend the proposed method to larger and more diverse deepfake datasets and incorporate additional NR-IQA backbones to better capture realism variations under different content and codec conditions. We also plan to scale up human evaluation studies by increasing the number of participants and evaluated samples to further strengthen evidence of alignment with human visual perception. Additionally, investigating the impact of training detection models with or without high-realism samples may provide further insights into improving robustness against realistic deepfake manipulations. These extensions would further strengthen the role of VFS in visual realism-aware dataset analysis and deepfake detection research.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Ahmet Demir conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Ahmet Emir Dirik conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Saffet Vatansever conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available at GitHub and Zenodo:

- https://github.com/aahmetdemir/VFS

- aahmetdemir. (2025). aahmetdemir/VFS: Initial public release of the VFS pipeline (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.17378194

The dataset is available at GitHub:

- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1-11. https://github.com/ondyari/FaceForensics

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3893#supplemental-information.

REFERENCES

An B, Zhou H, Peng P, Zhang L, Ren S, Li J, Xu T. 2025. A survey of learning-based no-reference image quality assessment. In: Yue Y, Leng L, eds. *The International Conference Optoelectronic Information and Optical Engineering (OIOE2024)*. Vol. 13513 International Society for Optics and Photonics, SPIE, 135132S.

Barratt S, Sharma R. 2018. A note on the inception score. ArXiv DOI 10.48550/arXiv.1801.01973.

Benesty J, Chen J, Huang Y, Cohen I. 2009. *Pearson correlation coefficient*. Berlin Heidelberg, Berlin, Heidelberg: Springer.

Cai Z, Ghosh S, Adatia AP, Hayat M, Dhall A, Gedeon T, Stefanov K. 2024. AV-Deepfake1M: a large-scale LLM-driven audio-visual deepfake dataset. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, 7414–7423.

Ciftci UA, Demir I, Yin L. 2020. How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals. In: *2020 IEEE International Joint Conference on Biometrics (IJCB)*. Piscataway: IEEE.

Cohen J. 2013. *Statistical power analysis for the behavioral sciences*. London: Routledge.

Concas S, La Cava SM, Casula R, Orrù G, Puglisi G, Marcialis GL. 2024. Quality-based artifact modeling for facial deepfake detection in videos. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Piscataway: IEEE, 3845–3854.

Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer C. 2019. The deepfake detection challenge (DFDC) preview dataset. ArXiv DOI 10.48550/arxiv.1910.08854.

Dragar L, Peer P, Štruc V, Batagelj B. 2023. Beyond detection: visual realism assessment of deepfakes. ArXiv DOI 10.48550/arXiv.2306.05985.

Golestaneh S, Dadsetan S, Kitani KM. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Piscataway: IEEE, 1220–1230.

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems* **30(1)**:6626–6637 DOI 10.18034/ajase.v8i1.9.

Hou Y, Fu H, Chen C, Li Z, Zhang H, Zhao J. 2024. PolyGlotFake: a novel multilingual and multimodal deepfake dataset. In: *International Conference on Pattern Recognition*. Cham: Springer, 180–193.

Jiang L, Li R, Wu W, Qian C, Loy C. 2020. Deeperforensics-1.0: a large-scale dataset for real-world face forgery detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 2889–2898.

Ke J, Wang Q, Wang Y, Milanfar P, Yang F. 2021. Musiq: multi-scale image quality transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 5148–5157.

Kim H, Lee J, Park L, Kwon T. 2024. On the correlation between deepfake detection performance and image quality metrics. In: *Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes*, 14–19.

King DE. 2009. Dlib-ML: a machine learning toolkit. *The Journal of Machine Learning Research* **10(11)**:1755–1758 DOI 10.1145/1577069.1755843.

Korshunov P, Marcel S. 2018. DeepFakes: a new threat to face recognition? Assessment and detection. ArXiv DOI 10.48550/arXiv.1812.08685.

Kwon P, You J, Nam G, Park S, Chae G. 2021. KoDF: a large-scale Korean deepfake detection dataset. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 10744–10753.

Le BM, Woo SS. 2023. Quality-agnostic deepfake detection with intra-model collaborative learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 22378–22389.

Li L, Bao J, Yang H, Chen D, Wen F. 2020a. Advancing high fidelity identity swapping for forgery detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 5074–5083.

Li Y, Yang X, Sun P, Qi H, Lyu S. 2020b. Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE.

Lin Y, Song W, Li B, Li Y, Ni J, Chen H, Li Q. 2024. Fake it till you make it: curricular dynamic forgery augmentations towards general deepfake detection. In: *European Conference on Computer Vision*. Cham: Springer, 104–122.

Mann HB, Whitney DR. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* **18(1)**:50–60 DOI 10.1214/aoms/1177730491.

Mirsky Y, Lee W. 2021. The creation and detection of deepfakes: a survey. *ACM Computing Surveys (CSUR)* **54(1)**:1–41 DOI 10.1145/3425780.

Mittal A, Moorthy A, Bovik A. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing* **21(12)**:4695–4708 DOI 10.1109/tip.2012.2214050.

Peng B, Sun X, Wang C, Wang W, Dong J, Sun Z, Zhang R, Cong H, Fu L, Wang H, Zhang Y, Zhang H, Zhang X, Liu B, Ling H, Dragar L, Batagelj B, Peer P, Struc V, Zhou X, Liu K, Feng W, Zhang W, Wang H, Diao W. 2023. DFGC-VRA: deepfake game competition on visual realism assessment. In: *2023 IEEE International Joint Conference on Biometrics (IJCB)*. Piscataway: IEEE.

Peng B, Xiang W, Jiang Y, Wang W, Dong J, Sun Z, Lei Z, Lyu S. 2022. DFGC 2022: the second deepfake game competition. In: *2022 IEEE International Joint Conference on Biometrics (IJCB)*. Piscataway: IEEE.

Repository DG. 2017a. Deepfakes GitHub repository. *Available at https://github.com/deepfakes/faceswap* (accessed 2 April 2025).

Repository FG. 2017b. FaceSwap GitHub repository. *Available at https://github.com/MarekKowalski/FaceSwap/* (accessed 2 April 2025).

Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. 2019. Faceforensics++: learning to detect manipulated facial images. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 1–11.

Seow JW, Lim MK, Phan RC, Liu JK. 2022. A comprehensive overview of deepfake: generation, detection, datasets, and opportunities. *Neurocomputing* **513(1)**:351–371 DOI 10.1016/j.neucom.2022.09.135.

Song W, Yan Z, Lin Y, Yao T, Chen C, Chen S, Zhao Y, Ding S, Li B. 2024. A quality-centric framework for generic deepfake detection. ArXiv DOI 10.48550/arXiv.2411.05335.

Sun X, Dong B, Wang C, Peng B, Dong J. 2023. Visual realism assessment for face-swap videos. In: *International Conference on Image and Graphics*. Cham: Springer, 415–426.

Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M. 2016. Face2face: real-time face capture and reenactment of RGB videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 2387–2395.

Thies J, Zollhöfer M, Nießner M. 2019. Deferred neural rendering: image synthesis using neural textures. *ACM Transactions on Graphics (TOG)* **38(4)**:1–12 DOI 10.1145/3306346.3323035.

Tong K, Zhang J, Wang Y, Tohidypour H, Nasiopoulos P. 2024. A high-fidelity partial face manipulation dataset for enhanced deepfake detection. In: *2024 International Conference on Computing, Networking and Communications (ICNC)*. Piscataway: IEEE, 16–20.

Welch B. 1947. The generalization of 'student's' problem when several different population varlances are involved. *Biometrika* **34(1–2)**:28–35 DOI 10.1093/biomet/34.1-2.28.

Yang X, Li Y, Lyu S. 2019. Exposing deep fakes using inconsistent head poses. In: *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*. Piscataway: IEEE, 8261–8265.

Yang S, Wu T, Shi S, Lao S, Gong Y, Cao M, Wang J, Yang Y. 2022. Maniqa: multi-dimension attention network for no-reference image quality assessment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 1191–1200.

Zhang W, Ma K, Yan J, Deng D, Wang Z. 2020. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology* **30(1)**:36–47 DOI 10.1109/tcsvt.2018.2886771.

Zhang W, Zhai G, Wei Y, Yang X, Ma K. 2023. Blind image quality assessment via vision-language correspondence: a multitask learning perspective. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 14071–14081.