

A Comprehensive Framework for Evaluating Deepfake Generators: Dataset, Metrics Performance, and Comparative Analysis

Sahar Husseini^{1,2} and Jean-Luc Dugelay¹

¹Department of Digital Security, Eurecom Research Center, Sophia Antipolis, France

²Docaposte Biometrics Lab, Sophia Antipolis, France

{husseini, dugelay}@eurecom.fr

Abstract

Assessing the realism and accuracy of deepfake generators, especially in cross-reenactment situations, is a major challenge. This challenge is primarily attributed to the absence of ground-truth data, which restricts the application of metrics that rely on explicit ground-truth, such as SSIM and LPIPS. To overcome this challenge, this paper introduces a novel protocol for quantitatively assessing images generated by face-reenactment techniques. To address the scarcity of suitable datasets, two video datasets are generated: the Real Head and the synthesized Metahuman datasets. Furthermore, user studies are conducted to evaluate the efficacy of our proposed protocol. The results demonstrate a strong correlation between subjective evaluations and quantitative metrics obtained within our protocol. Comparative analysis with existing evaluation protocols further validates the effectiveness of our proposed approach. Notably, our protocol exhibits superior performance in analyzing identity preservation, head pose, and facial expression replication. The source code and datasets are made publicly available at https://github.com/SaharHusseini/deepfake_evaluation.git

1. Introduction

The face serves as a highly expressive and complex non-verbal communication channel for humans. The advancements in AI-generated synthetic faces, known as Deepfakes, have brought about significant benefits in various domains, including education, film production, and dubbing.

Among the fundamental techniques in DeepFake face manipulation are face swapping and face-reenactment. Face swapping involves transforming a face from a source image to seamlessly replace the face in a target image, achieving a result where the replacement blends naturally into the tar-

get image. Face-reenactment methods, on the other hand, aim to generate a synthesized video that animates a target face based on the movements captured from a driving video, while preserving the identity conveyed by the source image. This process involves treating the person in the source image as a controllable puppet, with the facial expressions, head pose, and movements from the driving video defining the corresponding actions in the synthesized video.

Recent face-reenactment techniques [27, 10, 24, 20, 18, 25] have leveraged generative models such as Encoder-Decoder (ED) networks [26], Variational Auto-Encoders (VAEs) [15], and Generative Adversarial Networks (GANs) [9] to generate images that push the boundaries of realism, making it increasingly challenging to discern between what is real and what is artificially generated. Despite the progress made in the development and application of face-reenactment methods, evaluating the realism and accuracy of the generated images, particularly in cross-reenactment scenarios where a different individual's face is used to reenact the source face, remains a significant challenge. Directly employing image based quality metrics, such as Structural Similarity Index (SSIM) [13] or facial keypoint errors is impractical due to the absence of ground-truth data.

To address this challenge and quantitatively assess the quality of images generated through cross-reenactment, researchers have investigated the extraction of feature embeddings from both the source and generated faces. Subsequently, they calculate the errors or discrepancies between these extracted features [3, 8, 31]. Although this approach offers partial solutions for cross-reenactment evaluation, it is confined to specific metrics and lacks a comprehensive assessment. Therefore, there is an urgent need to develop a new evaluation protocol that can effectively assess the fidelity of cross-reenactment images.

This paper introduces a novel protocol for the quantitative evaluation of images produced by face-reenactment techniques, particularly in cross-reenact scenarios. The pro-

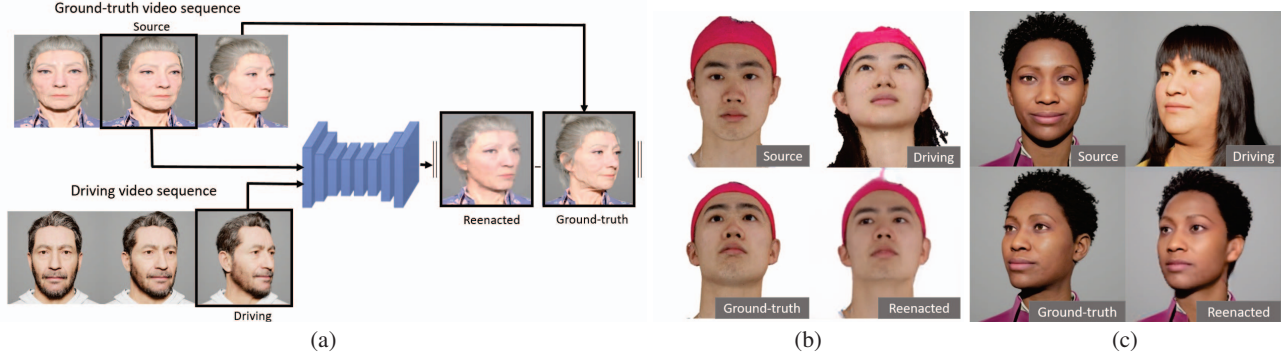


Figure 1: Proposed protocol (a). Examples of the source image, driving video frame, generated frame, and corresponding ground-truth provided by our proposed protocol for both the real (b) and synthesized (c) datasets.

protocol enables assessment of cross-reenactment images using metrics that rely on explicit ground-truth such as SSIM and LPIPS. To overcome the limited availability of appropriate datasets, two video generation approaches are proposed. The first approach involves the utilization of 3D models of real heads acquired using a multi-view system. In the second approach, realistic synthesized head models are employed, encompassing a wide range of human subjects, facial expressions, pose variations, and lighting conditions.

Our proposed protocol is applied using these datasets, along with established metrics such as SSIM [13], Cosine Similarity (CSIM) [6], Learned Perceptual Image Patch Similarity (LPIPS) [32], Average Keypoint Distance (AKD), Fréchet Inception Distance (FID) [11], and Fréchet Video Distance (FVD) [23] to assess the quality of four well known and state-of-the-art reenactment methods: FOMM [20], X2Face [27], LIA [26], and DaGAN [12].

In addition to quantitative evaluation, a series of user studies are conducted to investigate the effectiveness of our proposed protocol. These studies analyze the generated images in terms of identity preservation, head pose and facial expression replication, and overall image similarity, providing further validation of our quantitative results.

2. Related work

Evaluation techniques for face-reenactment can be classified into three categories: self-reenactment evaluation, cross-reenactment evaluation, and subjective test evaluation. The self-reenactment evaluation protocol, as depicted in Figure 2a, involves selecting a single frame from a video as the source image and using the remaining frames from the same video sequence to animate it. Since the source and driver identities originate from the same video sequence, the driver frames serve as a reliable ground-truth reference for comparing the generated images. This ensures a consistent and controlled evaluation of the reenactment process.

To assess the quality of the generated images in self-reenactment studies, image quality metrics such as SSIM

and Peak Signal-to-Noise Ratio (PSNR) [13] are commonly employed [27, 20, 19, 26, 30]. These metrics rely on ground-truth data and provide objective measures of image similarity and fidelity. Additionally, the self-reenactment technique enables the measurement of facial keypoint error such as AKD and Missing Keypoint Rate (MKR) which offers further insights into the accuracy of the reenactment process [26, 20]. To quantitatively evaluate the quality of generated frames, Siarohin et al. [20] utilizes self-reenactment to measure the L1 error, AKD, and Average Euclidean Distance (AED) between the generated frames and the ground-truth frames. Similarly, Gao et al. [8] reports the L1, SSIM, PSNR, FID and AKD error between the generated frames and the corresponding ground-truth frames for the self-reenactment scenario. Wang et al. [26] and Yang et al. [29] utilized the LPIPS to compute the similarity score between generated and ground-truth frames.

To quantitatively evaluate the generated frames in cross-reenactment scenarios and address the absence of ground-truth data, researchers employ a set of metrics that do not rely on explicit ground-truth comparisons. For the evaluation, researchers commonly utilize a cross-reenactment protocol, as illustrated in Figure 2b. In the existing cross-reenactment protocol, a prevalent method involves utilizing a pretrained network to extract identity features from the source and reenacted images [26, 20]. Alternatively, geometric features can be extracted from the driving and reenacted images [1, 4]. These extracted embeddings capture essential characteristics of the face, such as appearance and face pose. The quality of the generated frames can be assessed by computing the distance or dissimilarity between these embeddings. For instance recent face-reenactment methods [8, 10, 31, 3] evaluate the identity preservation by computing CSIM of embedding vectors between the generated frame and the source face [5]. Furthermore, Ha et al. [10] leverage pretrained networks to estimate the head pose angles and Facial Action Units (FAU) of generated image and compare these estimates with the corresponding

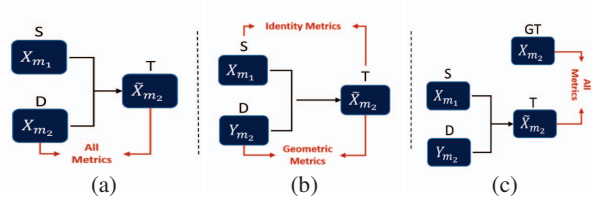


Figure 2: Face-reenactment evaluation protocols: self-reenactment (a), cross-reenactment (b), and our proposed evaluation protocol (c). In this illustration, X and Y represent identities, while S, D, T, and GT correspond to source, driving, target, and ground-truth, respectively. Additionally, m_1 and m_2 indicate the movement of the source and driving, respectively

driver’s head pose and action units, providing insights into the accuracy of the reenactment process.

Subjective test form the third category of evaluation techniques for cross-reenactment. In these evaluations, human observers play a crucial role by providing judgments on various aspects such as the visual quality, realism, and coherence of the generated cross-reenactment frames. For instance, Siarohin et al. [20] and Wang et al. [26] conducted a user study in which participants were presented with a source image, a driving video, and the corresponding results of their method and a competitive method. Participants were asked to select the most realistic image animation. Despite the significant advancements in cross-reenactment evaluation, there is still a need for an automated protocol that can compute errors for metrics relying on explicit ground-truth data. The establishment of such a protocol would contribute to a comprehensive and robust evaluation of cross-reenactment methods, enabling a deeper understanding of their performance and fostering further advancements in the field.

3. Proposed methodology

This section presents our proposed protocol for evaluating the image quality of reenactment methods, with a focus on cross-reenactment scenarios. To fulfill this objective, we generate video sequences comprising different identities with precisely controlled and known head pose and expression for each frame. These video sequences are then employed in conjunction with our proposed protocol and a set of quantitative metrics to measure the fidelity of images generated by various reenactment methods. In the following subsections, we provide a detailed description of the proposed protocol and the process of data generation.

3.1. Protocol

The pipeline of our proposed protocol is depicted in Figure 1. The protocol involves two video sequences, de-

noted as A and B, representing distinct identities. For each frame, the head pose and expression are identical in both sequences. Initially, any frame can be selected from video sequence A as the source image, representing the face to be reenacted. Subsequently, the video frames of identity B are utilized to animate the source image, resulting in frames of identity A that simulate the expressions and movements of identity B. These generated frames, known as deepfake frames, are then compared with the original frames of identity A in the ground-truth video sequence to evaluate the accuracy of the cross-reenactment process. The evaluation protocol can be summarized as follows:

1. Select a frame from video sequence A as the source image. In our experiments, we begin with frames displaying a frontal head pose and a natural expression, gradually introducing extreme variations in head pose and expression.
2. Select a driving video sequence, comprising video frames of identity B, to animate the source image. The head pose and expression in all frames of the driving video correspond to those of the source face.
3. Input the source image and driving video frames into a face-reenactment method to generate a new video sequence representing source identity A. This generated video sequence should accurately reflect the facial expressions and movements that match those of the driving video sequence.
4. Assess the accuracy of the generated frames by comparing them to the ground-truth video using metrics such as SSIM, CSIM, LPIPS, AKD, FID, and FVD.

3.2. Dataset generation

Two video datasets were generated for evaluating face-reenactment techniques: one comprised real face models generated from the Facescape dataset [28], and the other consisted of synthesized MetaHumans [7].

3.3. Real face dataset

To create a dataset comprising real human subjects, we employed the Pyrender 3D environment and utilized FaceScape [28], a well-established 3D face dataset. The FaceScape dataset consists of multi-view RGB images and intrinsic and extrinsic camera parameters, which were captured using 68 DSLR cameras. Leveraging this data, we generate 3D head point clouds with RGB values for various individuals exhibiting a neutral expression. By placing these 3D head models in desired scenes and defining specific camera parameters, we render them in the desired head poses. Figure 3 illustrates the rendering process.

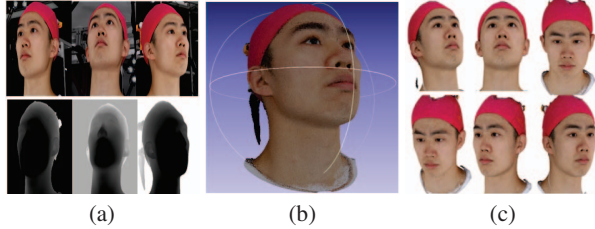


Figure 3: Multiview RGB images and their corresponding depth maps utilized to inverse project pixels into point clouds (a). The resulting reconstructed 3D head model (b). Rendered images of 3D models from desired angles(c).

In our study, we generated a total of 40 video sequences to investigate the impact of head rotations on face-reenactment. These sequences included five unique identities, and for each identity, we incorporated eight specific head rotations. The primary objective was to highlight different head rotation scenarios, namely a rotation around the pitch axis, a rotation around the yaw axis, and a combined rotation involving both pitch and yaw axes. To ensure consistent evaluation, each video began with the frontal head position in the initial frame and gradually transitioned the head towards the desired rotation axis in the final frame. Throughout the duration of the clips, the facial expression of the subjects remained constant. Each video clip consisted of 100 frames with a resolution of 512×512 pixels.

3.4. Synthesized dataset of MetaHumans

Evaluating the performance of face-reenactment methods solely using real data has limitations in assessing their ability to handle different facial expressions, as the individuals in the real dataset maintain a neutral expression throughout all the videos. To establish accurate ground-truth for facial expressions in the context of real datasets, image matching techniques like optical flow can be employed to reconstruct different expressions [2, 14]. However, the potential errors associated with these techniques necessitate an alternative approach. Therefore, we propose utilizing synthesized data, which provides complete control over the scene, allowing precise manipulation of geometry and appearance. This approach ensures data reliability and creates a controlled and accurate evaluation environment.

We utilized the Unreal Engine and the MetaHuman asset from the Quixel Bridge library [22] to generate a realistic synthesized face video dataset. MetaHumans are 3D human models created with advanced scanning, rigging, and animation technology, featuring high-quality photo scans of real skin textures and additional artificial textures for details like light reflection and surface roughness. Their riggability enables precise control over facial expressions and movements. To generate the video dataset, the scene was set up in the Unreal Engine with adjusted lighting conditions and

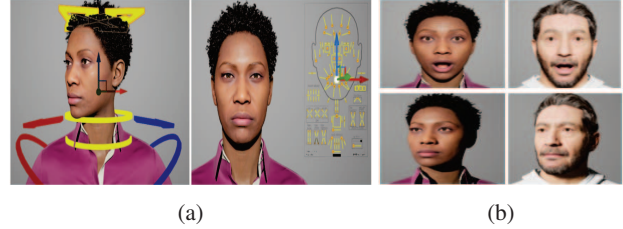


Figure 4: Head (left) and Face (right) Control Rig Boards enabling adjustment of pose and facial expression (a). Two MetaHumans with identical facial expressions and head poses (b).

configured camera properties. MetaHuman characters were placed within the scene and animated using the Control Rig Board as shown in Figure 4a. The resulting animations were rendered, capturing the desired facial expressions and movements.


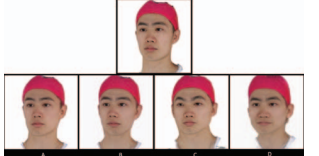
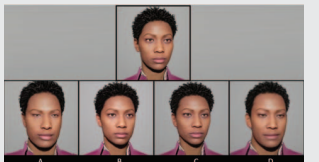
In Unreal Engine, the process of applying animations from one MetaHuman character to another is straightforward. By substituting the model references in the scene, the animations originally designed for the first character can be effortlessly transferred to the second character. This replacement ensures that both characters share the same expression setting, resulting in identical head pose and facial expression. Leveraging this capability, it becomes possible to generate multiple videos, each showcasing a different identity, while preserving consistent head pose and expression across all videos. Figure 4b illustrates two MetaHuman identities with the exact same head pose and expression.

The video sequences were meticulously designed to ensure a structured progression, starting with a frontal head position and neutral expression and culminating in expressive facial expressions or head rotations, or a combination of both. These sequences encompassed a diverse set of facial expressions, including amusement, anger, disgust, laughter, sadness, and surprise. The head rotations in the dataset covered 8 rotations around the yaw axis, pitch axis, and combinations of the pitch and yaw axes, including various directions such as up, down, left, right, and diagonal directions. A total of 20 distinct face movement animations were produced for each of the 10 MetaHuman identities, resulting in a dataset comprising 200 videos (10 identities \times 20 face animation). All videos were rendered at a resolution of 1920×1080 pixels, ensuring a high level of visual quality and detail for the evaluation process.

4. Subjective evaluation

Three subjective evaluations were conducted to assess the proposed protocol and evaluate the strengths and weaknesses of each reenactment method. These evaluations utilized a set-wise ranking method, where participants were presented with a set of videos or frames and tasked with

Table 1: Summary of subjective evaluation methods.

Evaluation name	Objective	Videos/ Images	Number of scenarios	Blind comparison	Subjective Test Example
VR	Assess perceived 'Realism' of generated videos	Videos	46	Test Includes Ground-Truth	
IS	Evaluate users 'Satisfaction' with generated outputs for specific head rotations	Images	132	Explicitly Informed (On Top)	
VI VPE VS	Assess quality focusing on 'Identity' preservation (VI), head 'Pose and Expression' preservation (VPE), and overall 'Satisfaction' (VS)	Videos	20	Explicitly Informed (On Top)	

directly comparing and organizing them based on specific criteria. Table 1 provides a summary of the three evaluation methods along with an example of each test conducted with the participants. The evaluations involved the participation of 23 professionals specializing in computer vision, ensuring their expertise in accurately assessing the fidelity of the generated frames produced by face-reenactment methods. Prior to the evaluation, participants were provided with detailed explanations of each test and completed practice tests to ensure their comprehension of the procedures. To optimize the evaluation time per participant, the test dataset was divided randomly into two batches, allowing participants to complete half of the test. On average, each evaluation session lasted approximately one hour.

In the first evaluation, titled "Realism Assessment," participants were presented with sets of five videos that included one ground-truth video and four reenacted videos. The videos were carefully selected to cover a wide range of facial expressions and head rotations. Participants were asked to rank the videos based on their perceived realism, using a scale from 1 to 5. To minimize bias, the order of the videos within each set was randomized, and participants were unaware of which video was the ground-truth.

The second evaluation, titled "Overall Satisfaction with Head Rotation," aimed to assess users' overall satisfaction with the generated outputs at specific head rotations. Participants were presented with sets of four frames generated by the reenactment methods, along with a ground-truth image depicting a specific head pose. Participants were explicitly informed about the identity of the ground-truth image and instructed to compare each generated image to the ground-truth. They were then asked to assign a rank to each image on a scale of 1 to 4, indicating their overall satisfaction relative to the ground-truth image.

The third evaluation aimed to assess the quality of the

generated videos, focusing on three aspects: 1) identity preservation, 2) head pose and expression preservation, and 3) users' overall satisfaction. Participants were presented with sets of four videos alongside the ground-truth video and were asked to rank each video in relation to the ground-truth. The rankings were reported separately for the preservation of identity, head pose and expression, and overall satisfaction. Participants provided scores ranking from 1 to 4, with 1 indicating the highest satisfaction. The first test consisted of 46 scenarios, the second test had 132 scenarios, and the third test comprised 20 scenarios.

Statistical analysis of subjective evaluation: To assess the distance between reenactment methods through subjective evaluation, each technology is assessed by a group of observers using a set of images and videos. We utilize the outlier detection and scaling method described in the study by Perez et al.[17], which is based on Thurstone's model and its assumptions [21]. This method, given a matrix that includes the results for all participants, measures the probability of observing the data of each observer in comparison to the rest of the observers.

The method uses Maximum Likelihood Estimation (MLE) to compute an inter-quartile-normalized score for each subject. Let's suppose we aim to compare n conditions O_1, \dots, O_n (i.e., n technologies) with unknown underlying true quality scores $q = (q_1, \dots, q_n)$, where $q_i \in R$ represents the quality score for condition O_i . The goal of this analysis is to estimate scores $\hat{q} = (\hat{q}_1, \dots, \hat{q}_n)$ that approximate the true quality q .

The perceived quality of a condition O_i is modeled as a random variable: $r_i \sim N(q_i, \sigma)$, where the mean of the distribution is assumed to be the true quality score q_i . When scaling the data, the focus is primarily on recovering the distance $q_i - q_j$ between underlying quality scores q_i and q_j (as scores are relative). If we know the true probability of

selecting O_i as better than O_j ($P(r_i > r_j)$), the probability that O_i was selected over O_j in exactly c_{ij} trials out of the total number of $n_{ij} = n_{ji} = c_{ij} + c_{ji}$ trials is given by the binomial distribution.

$$L(\hat{q}_i - \hat{q}_j \mid c_{ij}, n_{ij}) = \binom{n_{ij}}{c_{ij}} P(r_i > r_j)^{c_{ij}} (1 - P(r_i > r_j))^{n_{ij} - c_{ij}} = \binom{n_{ij}}{c_{ij}} \Phi\left(\frac{\hat{q}_i - \hat{q}_j}{\sigma_{ij}}\right)^{c_{ij}} \left(1 - \Phi\left(\frac{\hat{q}_i - \hat{q}_j}{\sigma_{ij}}\right)\right)^{n_{ij} - c_{ij}}, \quad (1)$$

Where, c_{ij} represents the count of cases where condition O_i was chosen as better than condition O_j , out of a total number of trials $n_{ij} = n_{ji}$. The true probability of choosing condition O_i over condition O_j can be computed using the cumulative normal distribution Φ , given two Gaussian random variables r_i and r_j .

$$P(r_i > r_j) = P(r_i - r_j > 0) = \Phi\left(\frac{q_i - q_j}{\sigma_{ij}}\right), \quad (2)$$

The parameter σ_{ij} represents the noise parameter in Thurstone’s model [21]. It is typically determined based on the probability p_{ij} of a 1 Just-Objectable-Difference (JOD) unit, as described in Perez et al. [17]. The scaling of the comparisons is then performed by maximizing the products of the likelihoods.

$$\arg \max_{\hat{q}_2, \dots, \hat{q}_n} = \prod_{i,j \in \Omega} L(\hat{q}_i - \hat{q}_j \mid c_{ij}, n_{ij}) \quad (3)$$

where Ω denotes the number of pairs with at least one made comparison. Subjects with an inter-quartile-normalised score above a threshold of 1.5 are tagged as outliers and discarded.

5. Experiment and results

Dataset: Two video datasets were compiled to assess face-reenactment techniques. The first dataset included 40 videos of real face models, featuring five identities with 8 head rotations each. The second dataset comprised 200 synthesized videos of MetaHumans, exhibiting 10 identities with 20 variations of head movement and facial expressions. A systematic approach was employed for both datasets, selecting first frame of one video as the source for each identity and utilizing the remaining videos from the same face animation type but different identities as driving videos. This methodology yielded a total of 1960 scenarios, with 160 scenarios derived from the real dataset and 1800 scenarios from the synthesized dataset. Table 1 provides an overview of the scenario distribution in the three subjective tests, ensuring an equal representation of synthesized and real scenarios in each test. These datasets offer a comprehensive and diverse range of scenarios, providing valuable insights into the performance of face-reenactment methods.

Methods and Metrics: In our evaluation, we compare the performance of four face-reenactment methods: FOMM [20], X2Face [27], LIA [26], and DaGAN [12]. The effectiveness of these methods is evaluated using six widely recognized metrics: SSIM [13], CSIM [6], LPIPS [32], FID [11] and FVD [23]. The CSIM metric utilizes facial embeddings extracted through the ArcFace [5] face recognition model to measure content similarity between generated and ground-truth images. The AKD metric quantifies key-point discrepancies by extracting 468 facial landmarks using the MediaPipe library [16]. To interpret subjective evaluation results, we employ Thurstone’s model assumptions to scale the ranking scores, as detailed in Section 4. The scores are represented on the Just-Objectable-Difference (JOD) scale, where a difference of 1 JOD signifies that 75% of observers favored one condition over another.

Evaluation and Analysis of Protocol Performance: Table 2 presents the performance evaluation results of cross-reenactment methods on real datasets, while Table 3 showcases the results on synthesized Metahuman datasets. The evaluation is conducted using various quantitative metrics, including SSIM, AKD, and LPIPS, which are computed based on 1960 scenarios derived from 240 videos. These metrics are employed to measure the disparities between the reenacted images and the corresponding ground-truth images provided by our protocol design. Additionally, our evaluation protocol incorporates the utilization of CSIM, FID, and FVD, which are commonly employed in existing face-reenactment evaluation.

FID assesses the photo-realism of the generated samples by comparing them to the ground-truth images at a deep feature level, while FVD, a modified version of FID, accounts for temporal coherence by considering spatial-temporal features. Notably, these metrics operate at the data distribution level, rather than focusing on individual frames. The calculation of FID and FVD metrics remains consistent with existing approaches since the ground-truth comprises data distributions of the Metahumans and real head videos.

In our analysis, we also incorporate the calculation of CSIM using the existing protocol depicted in Figure 2b, referred to as $CSIM_{trad}$. This metric evaluates the cosine similarity between the source and reenacted faces. However, the presence of distinct head poses between the source and reenacted faces poses a challenge, resulting in lower CSIM scores in traditional evaluation compared to the measurements obtained through our protocol.

Furthermore, the subjective test results are reported in both Table 2 and Table 3. The subjective evaluation serves multiple objectives in our study: firstly, it allows for the identification of strengths and weaknesses of each face-reenactment method, providing qualitative insights into their performance. Secondly, it enables the assessment of the effectiveness of our proposed protocol compared to ex-

Table 2: Evaluation results for cross-identity reenactment for real dataset.

Method	Quantitative Evaluation using the Proposed Protocol						Subjective Evaluation (JOD)					Traditional CSIM↑
	SSIM↑	LPIPS↓	CSIM↑	AKD↓	FID↓	FVD↓	VR↑	IS↑	VI↑	VPE↑	VS↑	
X2Face [27]	0.749	0.260	0.695	3.892	39.4	224.0	1.065	1	1	1	1	0.52
FOMM [20]	0.788	0.222	0.867	1.983	32.2	202.4	1	1.264	1.244	1.843	2.096	0.71
DaGAN [12]	0.803	0.159	0.833	2.883	34.6	217.1	1.964	2.654	2.139	2.640	2.164	0.66
LIA [26]	0.818	0.133	0.847	2.137	30.9	210.5	3.154	3.989	4.053	4.532	4.165	0.64
Ground-truth							5.071					

Table 3: Evaluation results for cross-identity reenactment for synthesized MetaHuman dataset.

Method	Quantitative Evaluation using the Proposed Protocol						Subjective Evaluation (JOD)					Traditional CSIM↑
	SSIM↑	LPIPS↓	CSIM↑	AKD↓	FID↓	FVD↓	VR↑	IS↑	VI↑	VPE↑	VS↑	
X2Face [27]	0.656	0.190	0.652	4.821	50.6	293.5	1	1	1	1	1	0.61
FOMM [20]	0.687	0.182	0.838	3.971	41.6	257.7	2.159	2.918	1.805	2.187	2.293	0.67
DaGAN [12]	0.821	0.147	0.865	1.902	45.4	271.5	3.075	4.034	2.557	2.789	3.320	0.64
LIA [26]	0.836	0.142	0.874	2.159	43.6	255.2	4.004	5.438	2.996	3.300	3.490	0.68
Ground-truth							5.269					

isting evaluation approaches. Lastly, the subjective results aid in determining the most informative quantitative metrics within our protocol that best describe the quality of reenacted images, thereby facilitating the identification of suitable metrics for future evaluations.

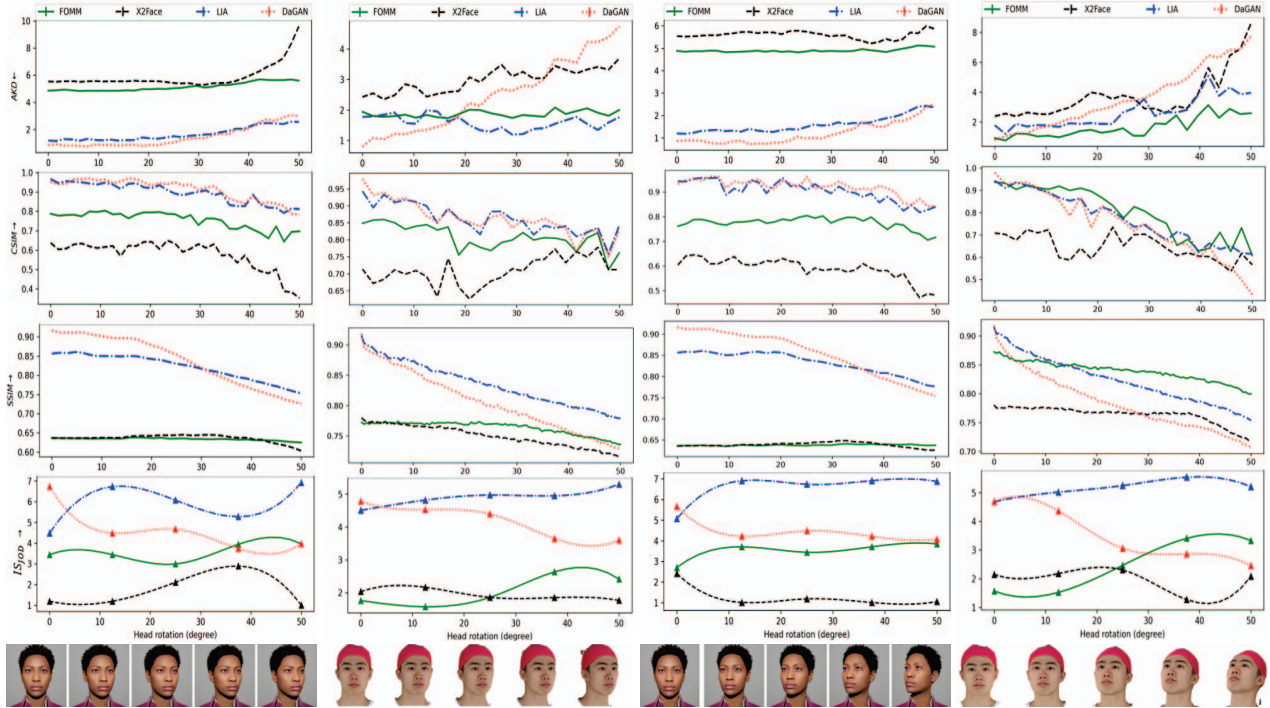
During the subjective tests, the reenactment methods are evaluated based on their performance in generating realistic content (VR_{JOD}), preserving identity (VI_{JOD}), transferring pose and expression (VPE_{JOD}), and overall satisfaction (VS_{JOD}). Statistical analysis reveals that the LIA method consistently achieves the highest scores in all subjective tests, slightly surpassing DaGAN. Both LIA and DaGAN consistently outperform X2Face and FOMM. A significant finding emerges from the blind comparison between the ground-truth and reenacted videos. The VR_{JOD} scores, calculated based on blind comparisons where the ground-truth is questioned alongside the reenacted videos, indicate that all four reenactment methods fail to generate sufficiently realistic content. Human subjects were able to distinguish reenacted content from the ground-truth images. FID and FVD are commonly used metrics to assess image and video realism. It is noteworthy that although FOMM demonstrates a good FID score, it does not align with the qualitative results (VR_{JOD}).

Furthermore, FOMM exhibits good scores in CSIM and AKD, which are considered identity preservation metrics in the literature. For example, its CSIM and $CSIM_{trad}$ scores in real dataset evaluation outperform other methods. It should be noted that FOMM employs relative keypoint locations to address the identity preservation problem, which seemingly increases CSIM, $CSIM_{trad}$, and AKD scores. However, its subjective score VI_{JOD} is lower than both LIA and DaGAN. To determine which quantitative metrics better describe the quality of reenacted images, the Pearson correlation coefficient is presented in Figure 6. The results demonstrate that the frame-based metrics within our

protocol, where the ground-truths are provided, exhibit the strongest correlation with subjective evaluations.

Pose Transferability Evaluation Using Our Proposed Protocol: Supplementing the results in Tables 2 and 3, we conducted a comprehensive analysis encompassing subjective and quantitative results using both the real head dataset and the synthesized dataset. A dedicated subjective test was conducted to assess overall satisfaction with image-based reenactment, with a specific focus on head rotation at different degrees. The driving sequences were incrementally rotated up to 50 degrees while maintaining natural facial expressions. The resulting overall scores, denoted as IS_{JOD} scores, were calculated for various head rotation scenarios, including rotations around the pitch axis, yaw axis, and combinations of pitch and yaw axes. The obtained scores are presented in Table 2 for the real head dataset and in Table 3 for the synthesized MetaHuman dataset.

To further analyze the quality of generated images under specific rotation conditions, we provide results for yaw rotation (right) and yaw-pitch rotation (up and left) in Figure 5. In addition to the subjective evaluations, quantitative scores such as SSIM, CSIM, and AKD were computed using ground-truth data as per our proposed protocol. Based on the findings presented in Tables 2 and 3, both the LIA and DaGAN methods demonstrate comparable performance in generating animated faces. However, based on Figure 5 they exhibit distinguishable sensitivities to head rotation. Through the subjective tests and SSIM evaluation, it is evident that LIA performs better in scenarios with more significant head movement in the driving video. Conversely, DaGAN exhibits superior performance in scenarios involving minimal head rotation, particularly those closer to the frontal head pose. Notably, DaGAN’s quality deteriorates gradually, and beyond a certain threshold (approx. 30°), it becomes comparable to or even worse than FOMM. In contrast, the FOMM method showcases resilience to head



(a) MetaHumans; Yaw

(b) Real dataset; Yaw

(c) MetaHumans; Pitch-Yaw

(d) Real dataset; Pitch-Yaw

Figure 5: Pose transferability evaluation using our proposed protocol. The figure presents the results of the image-based overall satisfaction subjective test scores (I_{SOD}) for different head degrees, along with the corresponding quantitative scores such as SSIM, CSIM, and AKD, computed using ground-truth data following our proposed protocol.

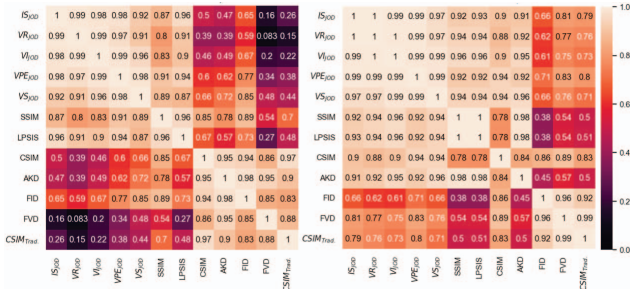


Figure 6: Confusion matrix depicting the correlation of metrics within Real (left) and synthesized (right) datasets

rotation, as the quality of reenacted images remains relatively unaffected and comparable to scenarios with a frontal head pose. When evaluating the CSIM and AKD metrics, FOMM achieves scores on par with those of LIA and DaGAN. However, its SSIM score is notably lower.

6. Future work

The application of our proposed protocol to face swapping methods shows great promise for future research. To implement our protocol for face swapping, we recommend utilizing our MetaHumans dataset and creating a com-

prehensive ground-truth by integrating elements generated from diverse sources. Specifically, the backgrounds, body and hairstyles can be preserved and rendered similarly to the driving videos, while the face identities should be derived from the source images.

7. Conclusion

This paper presents a novel protocol for evaluating the realism and accuracy of face-reenactment generators in cross-reenactment scenarios. Comparative analysis with existing evaluation approaches demonstrates the effectiveness of our protocol, supported by user studies validating its efficacy in analyzing identity preservation, head pose, and facial expression replication. The results reveal a strong correlation between subjective evaluations and frame based metrics (e.g., SSIM and LPIPS) within our protocol.

8. Acknowledgement

This research was financially supported by Docaposte for Biometrics Research and Testing. We extend our sincere gratitude to Fabien Aili and Emmanuel Nars for their invaluable contributions to our project meetings. Their participation has yielded fruitful discussions and insightful comments that have significantly enriched the scope and quality of this work.

References

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019.
- [2] Pouria Babahajiani. Geometric computer vision: Omnidirectional visual and remotely sensed data analysis. 2021.
- [3] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. What comprises a good talking-head video generation? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.
- [4] Marcella Cornia, Matteo Tomei, Lorenzo Baraldi, and Rita Cucchiara. Matching faces and attributes between the artistic and the real domain: the personart approach. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(3):1–23, 2022.
- [5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.
- [6] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pages 14398–14407, 2021.
- [7] Zhixin Fang, Libai Cai, and Gang Wang. Metahuman creator the starting point of the metaverse. In *2021 International Symposium on Computer Technology and Information Science (ISCTIS)*, pages 154–157, 2021.
- [8] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5609–5619, 2023.
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [10] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10893–10900, 2020.
- [11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [12] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022.
- [13] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.
- [14] Sahar Husseini. A survey of optical flow techniques for object tracking. B.S. thesis, 2017.
- [15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [16] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chu-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 2019, 2019.
- [17] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2019.
- [18] Jia-Wen Seow, Mei-Kuan Lim, Raphaël C-W Phan, and Joseph K Liu. A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, 2022.
- [19] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.
- [20] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.
- [21] Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.
- [22] Unreal Engine. Metahumans in quixel bridge. <https://docs.metahuman.unrealengine.com/en-US/metahumans-in-quixel-bridge/>, 2021. Accessed on May 12, 2023.
- [23] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.
- [24] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.
- [25] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: This video does not exist. disentangling motion and appearance for video generation. *arXiv preprint arXiv:1912.05523*, 2019.
- [26] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022.
- [27] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018.
- [28] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face

- prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [29] Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. Face2face ρ : Real-time high-resolution one-shot face reenactment. In *European conference on computer vision*, pages 55–71. Springer, 2022.
- [30] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022.
- [31] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020.
- [32] 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*, pages 586–595, 2018.