MetaHumans help to evaluate deepfake generators

Sahar Husseini
dept. of Digital Security
EURECOM
Sophia Antipolis, France
husseini@eurecom.fr

Jean-Luc Dugelay
dept. of Digital Security
EURECOM
Sophia Antipolis, France
dugelay@eurecom.fr

***Abstract*—The progress achieved in deepfake technology has been remarkable; however, evaluating the resulting videos and comparing different generators remains challenging. A primary concern arises from the lack of ground-truth data, except for self-reenactment scenarios. Additionally, available datasets may have inherent limitations, such as lacking expected animations or demonstrating inadequate subject diversity. Furthermore, there are ethical and privacy concerns when using real individuals' faces in such applications. This paper goes beyond the state-of-the-art dealing with the evaluation of deepfake generators by introducing an innovative dataset featuring MetaHumans. Our dataset ensures the availability of ground-truth data and encompasses diverse facial expressions, variations in pose and illumination conditions, and combinations of these factors. Additionally, we meticulously control and verify the expected animations within the dataset. The proposed dataset enables accurate evaluation of cross-reenactment generated images. By utilizing various established metrics, we demonstrate a high degree of correlation between the generator's scores obtained from deepfake videos of Metahumans and those obtained from deepfake videos of real persons. The synthesized MetaHuman dataset can be accessed at: https://github.com/SaharHusseini/MMSP_2023**

***Index Terms*—Deepfake, Face reenactment, Face Animation, Evaluation, MetaHumans**

I. Introduction

Animating videos using just a single-face image opens up a wide array of applications, ranging from movie production and image editing to dubbing and beyond. Within the realm of face animation techniques, one particular noteworthy approach is face reenactment, posing significant potential for diverse applications.

Face reenactment methods aim to generate a synthesized video animated by the driver's movement while preserving the identity of the source image. More precisely, when a source image is fed to a face reenactment network, the source person in the image will turn into a puppet, and the driving video will define the source's facial expression, head pose, and movement in the targeting video.

The recent face manipulation techniques [1]–[6] utilize generative models such as Encoder-Decoder (ED) networks, Generative Adversarial Networks (GAN) [7], and Variational Auto-Encoders (VAEs) [8] to generate image animation. These recent works based on deep learning have significantly improved the automatic generation of the synthesized videos' quality and realism.

The development of reenactment methods is popular among researchers; however, evaluating the results still poses significant challenges, especially in cross-reenactment scenarios, where a different identity reenacts the source face, commonly referred to as cross-reenactment. This challenge arises from the lack of ground-truth data, which makes it difficult to obtain accurate and objective results. To address this challenge, a subjective test can be conducted; however, it is time-consuming and requires visual inspection.

The second approach for evaluating cross-reenactment is to use feature embeddings. In this approach, the feature embeddings of the source, driving, and generated images are extracted, and depending on the evaluation criteria, the extracted feature of the generated image is compared to the source or driving image. Although this approach/protocol holds promise, it offers only a partial solution to the evaluation problem. This is because only certain existing metrics, such as Cosine Similarity (CSIM) of embeddings can be applied with this protocol to evaluate reenactment results, while metrics that require explicit ground-truth (e.g., Structural Similarity Index (SSIM)) [16] cannot be used with this protocol.

For a more comprehensive evaluation, in our previous work [9] we propose a new protocol [9], depicted in Figure 1, for cross-reenactment evaluation, addressing the lack of ground-truth by utilizing a real Head 3D dataset to assess various deepfake methods. However, our proposed dataset is limited to head rotation and does not include facial expressions. This paper builds on our previous work and proposes a novel approach to create a 3D synthesized dataset that includes both facial expressions and head rotation. The dataset is generated using MetaHumans [13], an advanced platform providing highly realistic synthesized head models with diverse human subjects, facial expressions, variations in pose, and illumination conditions. Leveraging a synthesized dataset ensures compliance with the General Data Protection Regulation (GDPR) by avoiding the use of sensitive personal information. The availability of our proposed dataset enables more robust evaluations of cross-reenactment methods. It overcomes the challenges associated with the absence of ground-truth data, fostering further advancements in this domain.

This paper is structured as follows: Section II provides an overview of related works in face reenactment evaluation. Section III introduces the methodology and the synthetic dataset generated using MetaHumans. In Section IV, we present experiments and results comparing four reenactment methods: FOMM [4], X2Face [1], Fs-vid2vid [3], and ICFace [17]. The

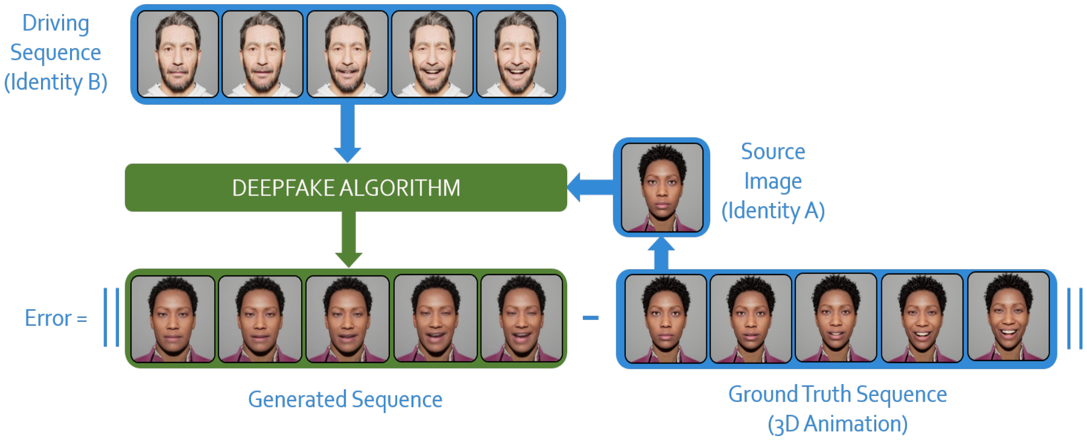


Fig. 1: Cross-reenactment evaluation protocol. The protocol involves two video sequences of different identities (A and B) that share the same rendering conditions (same expression, head pose, and light condition). The first frame of the video sequence, depicting identity A, is used as the source face. The video frames of identity B are then used to animate the source face, thereby generating video frames of identity A that simulate the expressions and movements of identity B. The generated deepfake frames can then be compared to the original frames of the video sequence of identity A to evaluate the accuracy of the cross-reenactment.

results obtained using MetaHumans are compared with those of the real Head dataset discussed in [9] for comparison: Finally, Section V concludes the study and outlines future research directions in this domain.

II. Related Work

Evaluation techniques for face reenactment can be broadly classified into three categories: self-reenactment evaluation, cross-reenactment evaluation, and subjective test evaluation.

Self-reenactment evaluation involves using one video frame as a source image and animating it with the rest of the frames from the same video sequence. As the source and driver identity belong to the same video sequence, the driver frames can serve as ground-truth, allowing a direct comparison with the generated frames. This evaluation protocol enables the assessment of the generated frames using metrics that require explicit ground-truth, such as Peak Signal to-Noise Ratio (PSNR) [16], SSIM, and Facial keypoint error. For instance, Siarohin et al. [4] report the L1 error, Average Euclidean Distance (AED), and Average Keypoint Distance (AKD) between the generated and ground-truth frames for self-reenactment. Similarly, Wiles et al. [1] compute the L1 error between the generated and ground-truth frames.

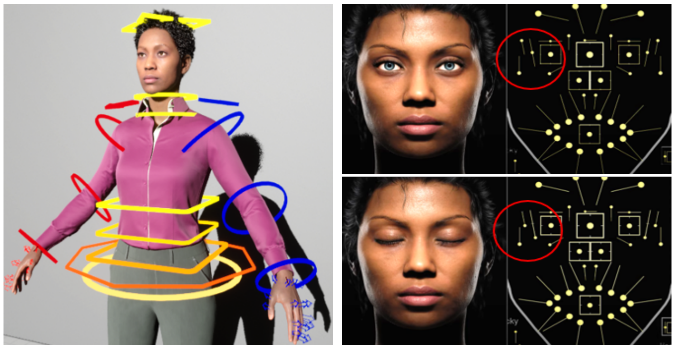
Cross-reenactment evaluation, in contrast, is employed when the source face is reenacted by a different identity. Since the ground-truth data does not exist, evaluating the generated frames using metrics that require explicit ground-truth is challenging. In this protocol, first, a pre-trained network is used to extract some embeddings/features from the generated frames, the driving frames, and the source frame. Depending on the evaluation criteria, the extracted features from the generated frame are then compared with the driving frame or the source frame. For instance, in recent face reenactment methods [2], [10], [11], identity preservation is evaluated by computing the Cosine Similarity (CSIM) of embedding vectors generated by a pre-trained face recognition model [12]. Furthermore, Ha et al. [2] use a pre-trained network to estimate the head pose angles and facial action units of both the generated and driving frames and compute the error between them.

Using the cross-reenactment protocol and pre-trained networks for extracting feature embeddings has partially addressed the challenges of cross-reenactment evaluation. However, it cannot be used to compute metrics that require explicit ground-truth, such as SSIM and PSNR. To enable the evaluation of cross-reenactment generated frames, we recently proposed a protocol in [9] that can be used in conjunction with video sequences that are created using a 3D environment where the animation movement is controlled. Using this protocol and the controlled video sequence dataset one can evaluate cross-reenactment using all existing evaluation metrics.

Subjective tests, on the other hand, involve human observers assessing the quality of the generated frames and can be conducted for both self and cross-reenactment. For example, Siarohin et al. [4] conducted a user study in which participants were presented with a source image, a driving video, and corresponding results generated by different methods. Participants were then asked to select the most realistic image animation. Similarly, Wang et al. [6] employed a pairwise comparison method to evaluate the realness of generated frames by human observers.

III. Proposed Methodology: Synthesized Data Generation

This section presents the overall pipeline for cross-reenactment evaluation, as proposed in [9], and illustrated in Figure 1. The protocol involves two video sequences of



(a) (b)

Fig. 2: Body Control (a) and Face Control Rig Board (b) enabling adjustment of pose and facial expression.

different identities (A and B) that have the exact same head pose and expression in each frame. The first frame of the video sequence, depicting identity A, is used as the source face. The video frames of identity B are then used to animate the source face, thereby generating images of identity A that simulate the expressions and movements of identity B. The generated deepfake images can then be compared to the original images of identity A in video sequence A to evaluate the accuracy of the cross-reenactment.

The protocol is specifically designed to be utilized with video sequences rendered in a 3D environment with controlled animation settings. The 3D environment offers a significant advantage in terms of control over various features in the scene, including lighting conditions and movements [14], [15], [18]. To evaluate the effectiveness of face reenactment methods, we generate synthesized data of MetaHumans. We used the Unreal Engine, a powerful game engine, and content creation tool, in combination with the built-in MetaHuman asset from the Quixel Bridge library [13], to generate a highly realistic synthesized face video dataset. MetaHumans are 3D digital human models created with advanced scanning, rigging, and animation technology. Their textures are a combination of high-quality photo scans of real skins and artificial textures that capture additional details, such as light reflection and surface roughness. The riggability of MetaHumans allows for greater control over their facial expressions and movements.

A rig is a digital framework that defines the movement and behavior of the 3D character. It typically consists of a hierarchy of interconnected bones or joints that can be manipulated to control the deformation and animation of the character. In Unreal Engine, all MetaHuman characters have the same base rig and they interpret the animations in the same way. Therefore, animations created for one MetaHuman can be transferred to other MetaHumans. This method allows us to generate multiple videos, each containing a different identity with the exact same head rotation and expression. Another interesting feature in Unreal engine is the slider associated with each MetaHuman, which allows adjustment of specific parts of the face. Figure 2a illustrates a MetaHuman character with its corresponding body Control Rig, while Figure 2b illustrates the displacement of a specific part of the face (e.i. eyes) using sliders.

To generate the video dataset for cross-reenactment evaluation, we began by defining the scene, lighting conditions, and camera properties in Unreal Engine. We then place a MetaHuman character in the scene and animated it using either the Control Rig Board or pre-made expressions from the Facial Pose Library and we render the animation. Finally, we transfer the same animation to the second character to create several videos with the same head pose and expression. Figure 3 illustrates two MetaHuman identities with the exact same head pose and expression in each column.

We generated a total of 19 distinct video animations featuring the source identity named Ada (depicted in Figure 4). The video sequences are carefully structured, beginning with a frontal head position and neutral expression, and concluding with either an expressive facial expression or head rotation. They include various facial expressions such as amusement, anger, disgust, laughter, sadness, and surprise. Additionally, we incorporate head rotations to assess the Metahumans' ability to accurately reproduce complex head rotations. These rotations involve rotating the head around the yaw axis towards the left side (Head-L) and right side (Head-R), as well as rotating the head around a combination of the pitch and yaw axes towards a combination of down and right (Head-DR), down and left (Head-DL) and up and left (Head-UL).

Furthermore, we created a set of videos that combine both head rotations and expressions. In these videos, the head is rotated towards the left, right, or a combination of down and right, while simultaneously transitioning the expression from neutral to amusement (Head-L-Amusement, Head-R-Amusement, and Head-DR-Amusement).

To explore the impact of lighting on reenactment performance, we also produced videos with different lighting directions. These videos maintained the same head rotation scenarios as the previous ones but introduced a left-sided light source (Head-DR-LightL, Head-L-LightL, Head-R-LightL, Head-UL-LightL).

By incorporating this diverse set of video animations, we aimed to conduct a comprehensive evaluation of cross-reenactment by Metahumans. This evaluation allowed us to assess their ability to accurately reproduce a wide range of facial expressions and head rotations, as well as their performance under varying lighting conditions.

To create the driving video sequence, we transfer Ada's animation to five other MetaHuman characters/identities, namely Emory, Gavin, Maria, Nasim, and Robin resulting in a total of 95 videos (5 identities × 19 animations). All videos are of resolution 256 × 256.

IV. Experiment and Results

This section presents the results using the cross-reenactment protocol and the synthesized dataset of MetaHumans as introduced in Section III. We conduct a comparative evaluation of four reenactment methods (FOMM, X2Face, Fs-vid2vid, and

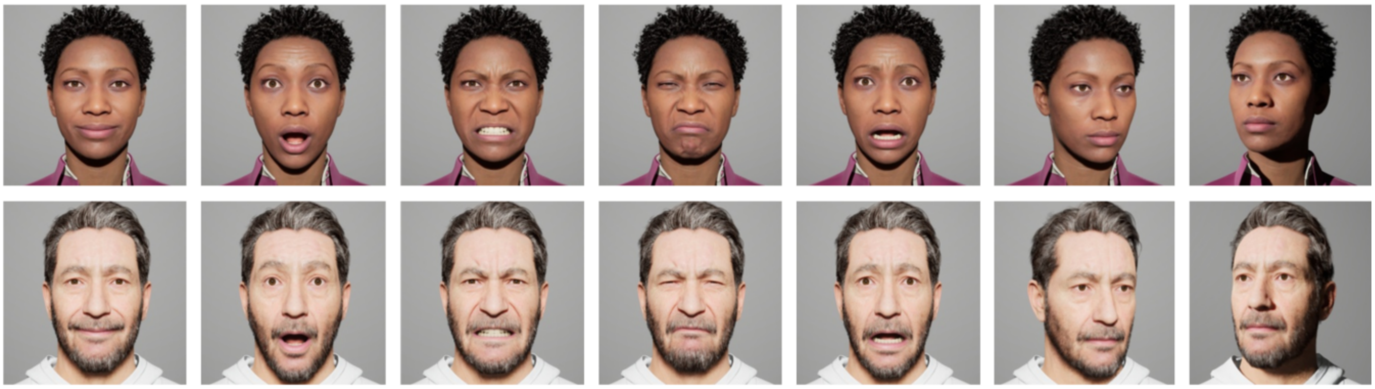


Fig. 3: Two MetaHumans sharing the same expression. Expressions from left to right: amusement, surprise, anger, disgust, fear, head rotated to right-hand side and to left-up by 30 degrees.

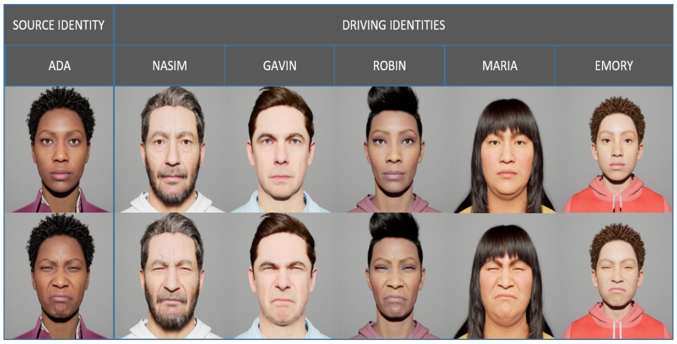


Fig. 4: The source and driving identities for cross-reenactment evaluation. Ada, Emory, Gavin, Maria, Nasim, and Robin are generated using MetaHumans. Identity Ada is the source face, and the rest of the identities are the driving identities.

ICFace) using our dataset, which is characterized by a broad range of head rotations, expressions, and lighting variations. By utilizing our dataset, we investigate the effect of driving identity on the generated images. Additionally, we analyze the sensitivity of each method to different head rotations and expressions.

We employed AKD and SSIM metrics to evaluate the quality of the generated images using the synthesized MetaHuman dataset. The results are presented in Tables I. Furthermore, for comparison purposes, we present the results on the Real Head dataset [9] in Table II, which includes different head rotations with static expressions. Each row in the tables represents the metric value for different variations in the datasets, while each column corresponds to a specific reenactment method. The last row in the tables presents the average error of the method for the specific metric across the entire dataset.

For the evaluation, we used a total of 12 video clips containing 1,200 frames from the Real Head Dataset, and 95 video clips containing 5,600 frames from the synthesized dataset. All experiments start from the initial frontal head pose and neutral expression, enabling a consistent comparison across methods.

The synthesized dataset results, as illustrated in Table I, indicate that the average AKD error for FOMM and X2Face is statistically similar, while the Fs-vid2vid and ICFace methods have higher AKD error values of 7.85 and 11.03, respectively. Upon closer examination of the errors, it becomes evident that X2Face generates significantly larger errors for certain head rotations, such as head rotation toward the left (Head L), when compared to the FOMM method. However, the error for videos with only expression is comparable to FOMM's performance. This finding suggests that while X2Face can maintain similar landmark accuracy to FOMM, it struggles with more pronounced head rotations, leading to a notable increase in error.

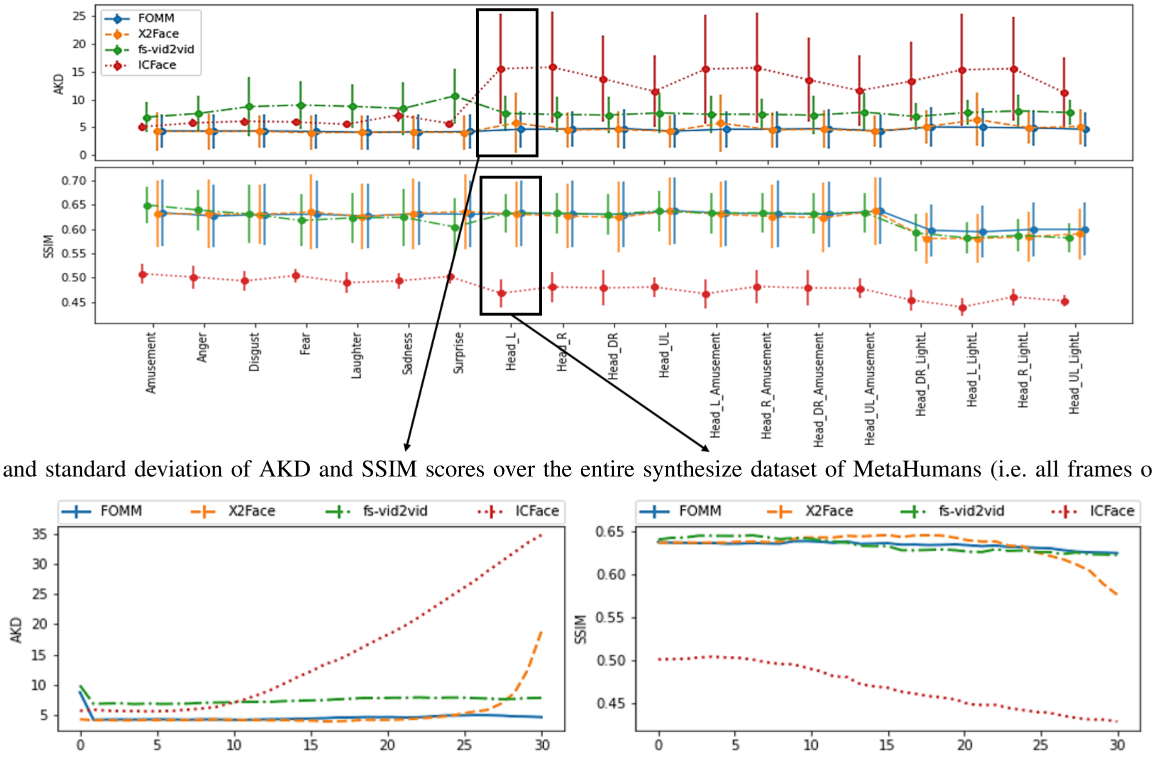
Moreover, Figure 5a shows that FOMM and X2Face exhibit similar AKD values for most cases with minimal head rotation. However, upon closer analysis, as depicted in Figure 5b, the error per frame/head degree for the specific case of Head-L video shows that the results for both methods are nearly identical up to 25 degrees of head movement. Beyond this threshold, the error for the X2Face method increases significantly. Furthermore, the SSIM scores suggest that all techniques struggle to preserve image quality when confronted with substantial head rotations.

Furthermore, we compared the results obtained in Table I with the results from the Real Head Dataset in Table II. As shown in Table II, the FOMM method achieved the best performance among the four reenactment methods with an AKD error of 1.99, followed by X2Face with an error of 3.36. The ICFace method exhibited the highest average keypoint error of 11.5, indicating difficulties in accurately reproducing the target face's keypoints.

Regarding SSIM scores, FOMM outperformed the other methods with a score of 0.73, indicating its effectiveness in generating images with similar structural content to the ground-truth images. The X2Face method had a lower SSIM score of 0.68 but still relatively high compared to the other two techniques. Meanwhile, the Fs-vid2vid and ICFace methods had lower SSIM scores of 0.54 and 0.45, respectively, suggesting difficulties in accurately generating images with

	AKD↓				SSIM↑			
VARIATION	**FOMM**	**X2FACE**	**FS-VID2VID**	**ICFACE**	**FOMM**	**X2FACE**	**FS-VID2VID**	**ICFACE**
Amusement	4.32	4.25	6.83	5.07	0.63	0.63	0.64	0.50
Anger	4.29	4.22	7.47	5.78	0.62	0.63	0.63	0.50
Disgust	4.31	4.26	8.72	6.08	0.63	0.63	0.63	0.49
Fear	4.24	4.012	8.99	5.96	0.63	0.63	0.61	0.50
Laughter	4.07	4.05	8.77	5.54	0.62	0.62	0.62	0.48
Sadness	4.17	4.203	8.38	7.19	0.63	0.63	0.62	0.49
Surprise	4.22	4.022	10.64	5.68	0.63	0.63	0.60	0.50
Head_L	4.65	5.77	7.50	15.56	0.63	0.63	0.63	0.46
Head_R	4.68	4.50	7.30	15.82	0.63	0.62	0.63	0.48
Head_DR	4.76	4.60	7.19	13.71	0.63	0.62	0.62	0.47
Head_UL	4.30	4.27	7.58	11.47	0.63	0.63	0.63	0.48
Head_L_Amusement	4.65	5.73	7.33	15.49	0.63	0.63	0.63	0.46
Head_R_Amusement	4.65	4.50	7.33	15.72	0.63	0.62	0.63	0.48
Head_DR_Amusement	4.74	4.62	7.18	13.57	0.63	0.62	0.63	0.47
Head_UL_Amusement	4.33	4.31	7.71	11.59	0.63	0.63	0.63	0.47
Head_DR_LightL	5.06	5.05	2.30	13.28	0.59	0.58	0.59	0.45
Head_L_LightL	4.99	6.38	7.66	15.37	0.59	0.58	0.58	0.43
Head_R_LightL	4.88	4.91	7.94	15.55	0.59	0.58	0.58	0.46
Head_UL_LightL	4.64	5.14	7.65	11.21	0.59	0.59	0.58	0.45
Average	**4.52**	4.67	7.85	11.03	**0.62**	**0.62**	**0.62**	0.47

TABLE I: The evaluation results for cross-reenactment using synthesized dataset. The arrows pointing upward and downward correspond to metrics that show better results with higher and lower values, respectively. The best values are highlighted in bold.



(a) Mean and standard deviation of AKD and SSIM scores over the entire synthesize dataset of MetaHumans (i.e. all frames of all videos).

(b) AKD and SSIM scores frame by frame for the head rotated toward the left video sequence (Head-L).

Fig. 5: The mean and starndard deviation over different facial expressions and head rotations. Notably, there is a substantial increase in error for head rotations compared to facial expressions (a). Focus on the specific scenario of head rotation toward left . These curves provide a more detailed analysis by zooming in and reporting the error degree by degree for entire video frames, rather than average value.

similar structural content to the ground-truth images.

Our analysis of the MetaHuman dataset revealed consistent findings with those from the Real Head Dataset with animations limited to head pose motion. We observed that the FOMM method outperformed the other methods in terms of maintaining facial keypoints. Conversely, the ICFace method

	AKD↓				SSIM↑			
VARIATION	**FOMM**	**X2FACE**	**FS-VID2VID**	**ICFACE**	**FOMM**	**X2FACE**	**FS-VID2VID**	**ICFACE**
Head_U	1.90	2.76	4.07	8.45	0.74	0.71	0.53	0.48
Head_L	1.82	3.03	4.07	10.89	0.74	0.69	0.61	0.47
Head_UL	2.27	4.29	9.18	15.16	0.70	0.64	0.48	0.41
Average	**1.99**	3.36	5.77	11.5	**0.73**	0.68	0.54	0.45

TABLE II: The evaluation results for cross-reenactment using the real Head dataset [9], which includes three types of head rotations: rotation around the pitch, yaw, and combination of pitch and yaw. The facial expression of this dataset is constant.

exhibited the highest error, indicating potential difficulties with accurately reproducing the target face's keypoints.

V. Conclusion

This paper highlights the challenges associated with accurately evaluating and comparing cross-reenactment technologies. These challenges primarily arise from the absence of ground-truth data. Additionally, existing datasets have limitations in effectively representing diverse subjects, head movements, expressions, and adhering to General Data Privacy Regulation (GDPR) [19]. To overcome these issues, we introduce a novel dataset of MetaHumans that addresses the aforementioned limitations by ensuring diversity in ethnicity, age, and gender. This dataset encompasses a wide range of facial expressions, pose variations, and illumination changes.

To ensure that the synthesized dataset accurately represents real-world data, we utilize various established metrics and compare the results obtained from MetaHuman synthesized dataset with those from the Real Head dataset. Through this analysis, we show a correlation between the generator's scores obtained from deepfake videos of MetaHumans and those obtained from deepfake videos of real individuals. Our current findings are of significant interest, and there are several promising avenues for future research to build upon our work.

Firstly, an extension of this study could involve computing the error for additional metrics, such as CSIM and FID, thereby providing a more comprehensive evaluation of the methods. Furthermore, conducting a subjective test can be beneficial to evaluate the effectiveness of the proposed protocol. By considering these aspects, further advancements can be made in the domain of face reenactment evaluation.

Acknowledgment

This work was funded by Docaposte for Biometrics Research and Testing.

References

[1] Wiles, Olivia, A. Koepke, and Andrew Zisserman. ”X2face: A network for controlling face generation using images, audio, and pose codes.” Proceedings of the European conference on computer vision (ECCV). 2018.

[2] Ha, Sungjoo, et al. ”Marionette: Few-shot face reenactment preserving identity of unseen targets.” Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.

[3] Wang, Ting-Chun, et al. ”Few-shot Video-to-Video Synthesis.” Advances in Neural Information Processing Systems 32 (2019).

[4] Siarohin, Aliaksandr, et al. ”First order motion model for image animation.” Advances in neural information processing systems 32 (2019).

[5] Hong, Fa-Ting, et al. ”Depth-aware generative adversarial network for talking head video generation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[6] Wang, Yaohui, et al. ”Latent Image Animator: Learning to Animate Images via Latent Space Navigation.” ICLR 2022-The International Conference on Learning Representations. 2022.

[7] Goodfellow, Ian, et al. ”Generative adversarial networks.” Communications of the ACM 63.11 (2020): 139-144.

[8] Kingma, Diederik P., and Max Welling. ”Auto-Encoding Variational Bayes.” stat 1050 (2014): 1.

[9] Husseini, Sahar, et al. ”A 3D-Assisted Framework to Evaluate the Quality of Head Motion Replication by Reenactment DEEPFAKE Generators.” ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

[10] Zakharov, Egor, et al. ”Fast bi-layer neural synthesis of one-shot realistic head avatars.” Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer International Publishing, 2020.

[11] Chen, Lele, et al. ”What comprises a good talking-head video generation?.” IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.

[12] Deng, Jiankang, et al. ”Arcface: Additive angular margin loss for deep face recognition.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[13] Unreal Engine. Metahumans in quixel bridge. https://docs.metahuman.unrealengine. com/en-US/metahumans-in-quixel-bridge/, 2021. Accessed on May 12, 2023

[14] Pouria Babahajiani, “Geometric computer vision: Omnidirectional visual and remotely sensed data analysis,” 2021.

[15] Husseini, Sahar, Pouria Babahajiani, and Moncef Gabbouj. ”Color constancy model optimization with small dataset via pruning of CNN filters.” 2021 9th European Workshop on Visual Information Processing (EUVIP). IEEE, 2021.

[16] Hore, Alain, and Djemel Ziou. ”Image quality metrics: PSNR vs. SSIM.” 2010 20th international conference on pattern recognition. IEEE, 2010.

[17] Tripathy, Soumya, Juho Kannala, and Esa Rahtu. ”Icface: Interpretable and controllable face reenactment using gans.” Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020.

[18] Husseini, Sahar. A survey of optical flow techniques for object tracking. BS thesis. 2017.

[19] Zaeem, Razieh Nokhbeh, and K. Suzanne Barber. ”The effect of the GDPR on privacy policies: Recent progress and future promise.” ACM Transactions on Management Information Systems (TMIS) 12.1 (2020): 1-20.