Avatar Fingerprinting for Authorized Use of Synthetic Talking-Head Videos: Supplementary Document

Ekta Prashnani, Koki Nagano, Shalini De Mello, David Luebke, and Orazio Gallo

NVIDIA
{eprashnani, knagano, shalinig, dluebke, ogallo}@nvidia.com

1 Future Work

In this work, we present the novel task of avatar fingerprinting, that aims to verify the authorized use of synthetic talking-head avatars, along with the first method to solve for this task as well as a novel dataset that caters precisely to the needs of avatar fingerprinting. While talking-head avatars are becoming ubiquitous in across several applications, the avatar generation technology itself is also rapidly advancing towards generating full-body and 3D avatars [2,16,17,23], generating realistic synthetic audio [12], controlling talking-head avatars with audio only [19], as well as controlling upper-body avatars including hands with audio only [3]. This makes our proposed first step towards verifying authorized use of synthetic talking-head avatars all the more pertinent, and paves the way for future research directions that verify the "dynamic identity signatures" from newer forms of avatars ranging from full-body avatars, to 3D, to audio-driven ones.

2 Dataset

We now provide additional details of our proposed dataset, including details of the question prompts and sentences spoken in both stages, instructions to the participants, demographics of the dataset, and other relevant statistics.

General Instructions to Subjects. The subjects were asked to join pre-assigned Google Meet video calls using a laptop or a desktop. For the recorded video call, the subjects were also asked to position themselves so that their face was centered and parallel to the screen. However, in some cases with specific video-conferencing setups, this constraint was only approximately satisfied. Additionally, subjects were instructed to avoid hand motion since it can occlude their face, and also excessive body motion that might impair the visibility of their face. Before beginning each monologue, subjects were asked to speak "start topic" in a loud, clear voice, and, similarly, the end of each monologue was marked by the subjects speaking "end topic". These keywords allowed for effectively using the time-stamped transcription to automatically isolate relevant portions of the Google Meet recordings. Right after a subject said "start topic", they were instructed to pause for a few seconds and look directly at the camera with a frontal head pose, while holding a neutral expression. These frames with neutral expressions are crucial for a successful generation of synthetic talking-head videos using face-vid2vid [21], LIA [22], and TPS [8]. These generators work by transferring expression changes from a driving video to the target image. Therefore, it is important that the expression of the target image and that of the first frame of the driving video match. Asking subjects to provide a neutral expression before commencing with their monologues proves to be an effective way to achieve this: these neutral frames serve as good target images, while driving videos that start with these neutral frames allow for effectively animating the target image showing a similar expression. During the second stage of the data capture, where we record scripted monologues, subjects were instructed to memorize and speak the sentences to their recording partner, without referring back to the printed text from which they memorized the sentences. In case the subject forgot a sentence, they were instructed to start from the beginning of the sentence set. The whole recording session with both subjects in a video call typically lasted an hour, which also included miscellaneous interactions in between the monologues. The current dataset release excludes such interactions and only focuses on data captured for the two stages (Free-Form Monologues and Scripted Monologues).

Stage I: Free-Form Monologues. Subjects were asked to alternate between speaker and prompter roles. The prompter's task was to ask each of the following questions to the speaker, and the speaker was instructed to answer these questions in their natural manner.

1. Describe a day when you had to rush to an appointment.
2. Talk about an important milestone you have missed in the past and your feelings about it.
3. What is your favorite family holiday?
4. How is the weather in your area typically?
5. Is there a household chore you don't like doing?
6. Tell me about an incident that really surprised you.
7. Tell me about an incident that really scared you.

Stage II: Scripted Monologues. The following sentence sets were memorized and recited by each subject (alternating with their recording partner) in the second stage of the data capture. We did not ask subjects to explicitly demonstrate specific emotions for any sentence set. Rather, we chose to allow subjects to perform these memorized sentences in a manner natural to them.

1. My friend has a very cute dog. But, he can be scary when he barks.
2. Will you please answer the darn phone? The constant ringing is driving me insane!
3. My aunt was in the hospital for a week. Unfortunately, she passed away yesterday and I will need some time to grieve.
4. I hate rushing to get to the airport. The stress is too much for me to handle.
5. A slice of cake is the perfect ending to a meal. Wouldn't you agree?
6. It is going to be great working with you! I am surprised we didn't connect sooner!
7. You need to take the trash out right now! Your whole apartment smells like rotten eggs!
8. My internet connection is unreliable today. I hope it gets better before my meeting or I will have to call in!
9. I know the deadline is around the corner, but I just don't have any updates yet, I'm sorry.
10. Why can't the banker figure out what's going on? I should have got my money last night!
11. It's really nice out today. I might go for a walk if I get off work early and the kids aren't back from school.

12. There is a famous coffee shop around the corner that also serves snacks. Would you like to go tonight?
13. My dog almost got run over by a car today! Thank God he is safe!
14. It is getting very cold outside. I feel like having some hot chocolate. Would you like some?
15. I have been exercising so much lately. But I am not getting any stronger!
16. I have an old tie that I can wear to the interview. My grandfather gave it to me last year.
17. I had fun last night - we had quite a few drinks. But I have a really bad hangover this morning and I am considering calling in sick.
18. Please don't interrupt me when I am talking! Now I have forgotten what I wanted to tell you.
19. It was such a pleasure talking to you. I hope we stay in touch.
20. I can't believe I misplaced my keys yet again! I have to leave for the airport right now.
21. Gosh! the boy jumped right off the cliff into the ocean. He is lucky he didn't hit a rock.
22. The baby just spit up on my brand new clothes. I am going to be late for our dinner tonight.
23. The food smells disgusting but tastes delicious. How strange is that!
24. I was about to park when I saw a person with a gun. I kept driving and called the police right away.
25. I decided to take a nap during my lunch break. I am so glad I did! I feel very refreshed.
26. The food didn't get delivered on time. We had to keep our guests waiting while we searched for options.
27. I was walking down an alley the other night. I had the strange feeling that someone was following me.
28. She twisted her ankle while ice-skating. It was her final performance for the season.
29. Who moved my boxes from this room? I need to find my shoes before I can head out.
30. We miss our old home in the mountains quite a bit. This new place just doesn't feel as cozy.

Subject demographics. Out of the total pool of subjects that volunteered data for our 2-stage data capture, 50% are female, 47.8% are male, and the remaining chose "a gender not listed here". Amongst different age groups, 37% of the participants are 25-34 years old, 32.6% are $35-44$ years old, 17.4% are 45-54 years old, 6.5% are 18-24 years old, and 6.5% are 55-64 years old. In terms of race and ethnicity, 41.3% are Caucasian, 47.8% are Asian (including South Asian, East Asian, South-east Asian), 6.5% are African, 2.2% are Hispanic / Latino, 2.2% are Pacific islander, and others remained unspecified.

Synthetic Talking-Head Videos. As mentioned briefly in the main paper, we pool together videos for the 46 identities from our own 2-stage data capture, along with videos from 24 identities of RAVDESS (scripted monologues only) [15], and 91 from CREMA-D (short scripted monologues only) [1], resulting in a total of 161 unique identities. For each of these 161 identities, the remaining 160 are used to drive cross-reenactments, with 8 driving videos randomly selected from the total set of videos for each driving identity. For any given target identity, we incorporate synthetic videos driven by *every* remaining identity. During training, such a large variety of cross-reenactments enable effectively learning an appearance-agnostic dynamic facial identity feature space.

Privacy Considerations. Face videos are sensitive data, since a person's face is a key identifier. We took on this task with care to ensure good data governance. Our proposal for the data capture protocol was approved by an Institutional Review Board (IRB). Our goal was to provide exhaustive and transparent information to participants about our data capture procedure, future plans with the dataset (including our intent to create synthetic data samples), and conditions under which future research would be conducted—by us and interested third parties. The participants were also asked to confirm whether their data can be used for research beyond avatar fingerprinting, and whether it could be shown in public disclosures. Each file in our dataset is annotated with their responses.

2.1 Importance of the proposed dataset

As discussed in Section 4 of the main paper, the existing datasets of facial talking-head videos satisfy a subset of the requirements of avatar fingerprinting. This has motivated us to design our own dataset. Here we elaborate further on additional existing datasets (over and above the ones already discussed in the main paper), to re-emphasize the relevance of our proposed dataset.

CelebDFv2 [14]: This dataset features in-the-wild youtube videos of celebrities along with synthetic videos as well. However, these do not contain any face reenactment generators—only FaceSwap generators—and do not contain any self-reenactments. Furthermore, the real videos only contain free-form monologues for celebrities: to verify whether our model latches on to specific choice of words, we need scripted monologues in the dataset as well. Overall, the lack of scripted monologues in real videos, and missing self-reenacted and face-reenactment generators makes it tough to use CelebDFv2 dataset for our work.

ForgeryNet [7] is a diverse dataset that features a large collection of multiple types of synthetic generators. However, the construction of the dataset caters more to traditional forensics tasks like deepfake detection. Specifically, it lacks multiple cross-reenactments driven by all driving identities in the dataset, for each target identity. Moreover, there are no self-reenactments in the dataset. This limits the applicability of ForgeryNet to avatar fingerprinting.

MEAD [20] contains a large collection of real videos of individuals. Upon inspection, however, we observed that often subjects do not look directly at the camera (which is by design to capture different viewpoints of the face), and do not have a naturally-interactive facial behavior. For effectively training a model for avatar fingerprinting, we require driving videos for synthetic video generation process to feature the natural facial dynamics for the driving identity. A natural interactiveness in the facial behavior and voice intonation is a pre-requisite to capturing the dynamic identity signatures needed for avatar fingerprinting. This makes MEAD less applicable to avatar fingerprinting, since capturing natural facial behavior of individuals is a pre-requisite from the set driving videos. In contrast, in RAVDESS [15] and CREMA-D [1], subjects look directly at the camera and are talking in a more natural conversational manner. This is also aligned with the natural interactive manner in which our own video-conferencing dataset is captured.

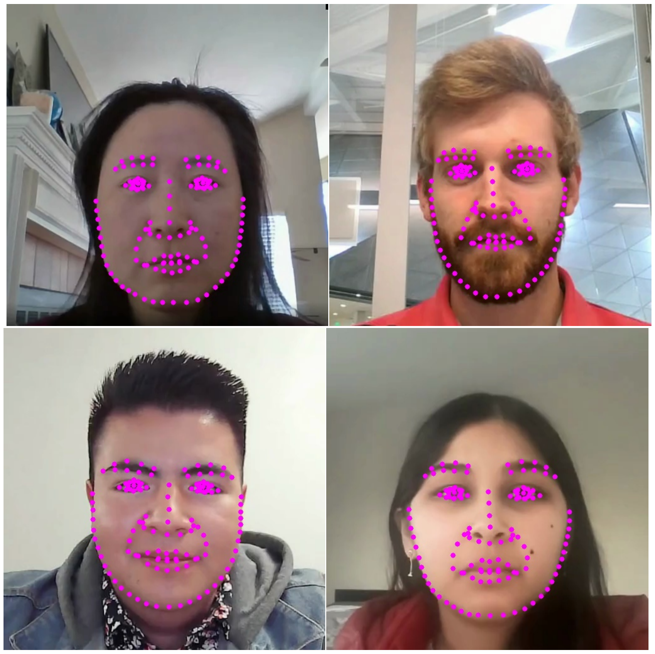

Fig. 1: Visualization of all landmarks.

Apart from the above-mentioned datasets, and the datasets already discussed in the main paper (CREMA-D [15], VFHQ [6], FaceForensics++ [18], KoDF [13]), there are many existing datasets catering specifically to deepfake detection, such as DFor [9, 10], DFDC [5], that share the above-mentioned limitations (lack of self-reenactments, not enough cross-reenactments, scripted videos missing, etc.). This motivated us to design our own dataset for avatar fingerprinting.

3 Implementation Details

Figure 1 visualizes the full set of 126 landmarks that we use to compute the input features for our model. Our model is based on the temporal ID net [4]. It is trained using our input features derived from the pairwise landmark distances, after appropriately modifying the number of input channels to match our feature dimension. To adjust the receptive field of the temporal ID net so that it predicts an embedding vector for longer or shorter input clips, we modify the number of layers of the network, and the dilation factor for the layers. Specifically, here are the kernel sizes and dilation factors for each of the layers in the temporal ID net, depending on the choice of input clip duration F:

1. F = 31 frames: $(1, 1, 1, 1, 2, 2, 2, 2, 4)$
2. F = 51 frames: $(1, 1, 1, 1, 2, 2, 2, 4, 4, 4, 4)$
3. F = 71 frames: $(1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4)$

The kernel sizes are all set to 3 apart from the first layer, which is 1. All other details of the temporal ID network are adapted from the existing implementation [4]. To implement the push and pull terms in Equations 1, 2, 3 from the main paper, n and t

Clip duration F	AUC
31	0.820
51	0.852
71	0.868

Table 1: Ablation study for different values of F.

compression level	AUC
Low (CRF 5)	0.868
Medium (CRF 25)	0.859
High (CRF 40)	0.849

Table 2: Robustness to video distortions.

experiment name	positive set	negative set	AUC
reference-driven cross reenactments in +ve set	driver = ref. ID; target = all IDs	driver = other IDs; target = ref. ID	0.83
2*scripted-set analysis	others utterances; driver = ref. ID	same utterance as test vid.; driver = other IDs	0.79
	others utterances; driver = ref. ID	other utterances; driver = other IDs	0.80

Table 3: Analysis with scripted subset.

span over 5 consecutive F-frame clips in a video. That is, during training, the temporal ID net receives as input (F+4)-frame videos, and outputs 5 embedding vectors, one for each of the 5 F-frame clips in the video. The max operation in Equations 1, 2, 3 from the main paper is performed over the 5 different clips (therefore, 5 different values of n), and the overall loss term in Equation 5 accumulates over 5 values of t. So, when a batch of videos is loaded for a training iteration, it comprises of (F+4)-frame "videos", which are split into 5 clips. These (F+4) frames are randomly selected from the entire video.

Additional training details. In each batch, we include 8 unique identities. For each identity ID_i, the pull term (Equation 1) comprises of 16 clips: 8 are self-reenactments, randomly sampled from the full set, and the remaining are cross-reenactments with ID_i as the *driving* identity. These cross-reenactments can potentially show the same words being spoken by different target identities. This is crucial: it allows the neural network to learn to to pull together videos based purely on the facial motion, regardless of the appearance of the video. The push term (Equation 2) for ID_i is composed of clips with the remaining 7 identities in the batch serving as driving identities (8 clips per driving identity). Therefore, for each identity, 72 clips are included in a batch. The training is performed for 100,000 iterations, with Adam optimizer [11] and a learning rate of $1e^{-4}$. During evaluation, one clip is sampled from each video and AUC is reported based on comparisons of each clip against the positive and the negative set.

4 Evaluation

We report additional ablation studies and robustness analyses in this section. Specifially, we evaluate the impact of reducing the clip duration (Table 1), the robustess of our trained model (trained on facevid2vid videos, at F = 71) to varying levels of compression (Table 2), and controlled analyses on scripted subset of our dataset (Table 3).

Impact of varying clip duration. In Table 1, we report the the results of our experiment with varying values of F, which is the number of frames provided to the network to make a prediction about the dynamic facial temporal identity signature. The performance gained with increasing values of F diminishes at high values. We choose 71 frames as the default for most of our experiments. For cases where shorter clips are desirable, such as for efficiency or for doing frequent verification, we observe that F=31 is plausible—with an AUC of 0.820. Note that, as also discussed in Section 4 we include a large set of scripted monologues in our dataset, which are crucial for a complete evaluation of avatar fingerprinting. These video clips tend to be of a shorter duration (since subjects can recite only a few sentences at a time). Therefore, we are constrained by the shortest-duration video clips in the dataset (98% of the video clips are at least 71 frames long) and the largest value of F we experiment with is 71.

Robustness to distortions. We vary the quality of videos by compressing the test videos to three different levels of compression (color perturbations are observed with higher compression) and test the performance of our trained model (trained on face-vid2vid, f = 71). We observe a negligible performance drop—see Table 2.

Analysis with scripted subset (Table 3). In Table 3, we analyze the performance of our model over the scripted subset in the test set. We want to assess the following:

1. Are cross-reenactments driven by a given identity (“reference identity”) are closer to each other in the embedding space, as compared to those cross-reenactments where reference is the target for other drivers?
2. In the scripted set, do embeddings for video driven by reference lie closer than the cross-reenactment videos where reference is driven by other driver speaking the same utterance and different utterances?

The evaluation in such a controlled setting, which is only possible using our dataset, allows us to assess the clustering of the embedding space independent of the words being spoken. We report the average AUC over the scripted test subset for when test-video utterance matches the cross-driven set (row 3 Table 3), and when it does not (row 4 Table 3). The result is slightly worse when the utterance of the cross-driven set (“negative class”) matches the test video, given the similar lip movements.

Additional results. In Figure 2, we show more samples similar to the ones shown in Figure 4 from the main paper, for face-vid2vid generator (since this generator shows the best quality). For each row of results in Figure 2, we choose a reference identity, and a held-out set of reference self-reenacted videos for each of these identities. Then, we report the average Euclidean distance of the following videos with respect to the held-out self-reenacted videos for the reference:

1. a new self-reenacted video by the reference identity (not included in the held-out reference set) – highlighted with a green border,
2. a cross-reenacted video where the reference identity is the driver – highlighted with a green border, and
3. two cross-reenacted videos where the reference identity is the target, driven by some other identity – highlighted with a red and a blue border.

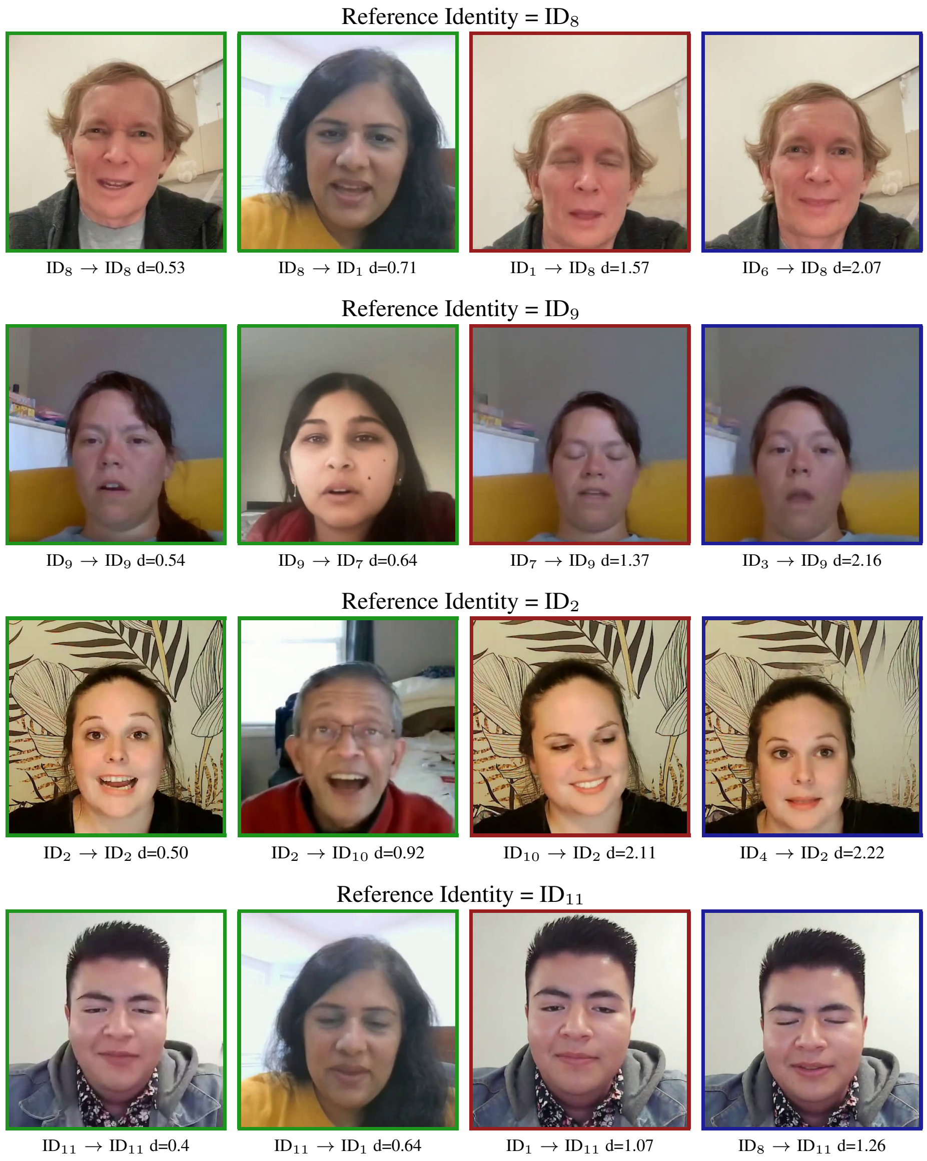
Reference Identity = ID_8
$\mathrm{ID}_8 \rightarrow \mathrm{ID}_8$ d=0.53
$\mathrm{ID}_8 \rightarrow \mathrm{ID}_1$ d=0.71
$\mathrm{ID}_1 \rightarrow \mathrm{ID}_8$ d=1.57
$\mathrm{ID}_6 \rightarrow \mathrm{ID}_8$ d=2.07
Reference Identity = ID_9
$\mathrm{ID}_9 \rightarrow \mathrm{ID}_9$ d=0.54
$\mathrm{ID}_9 \rightarrow \mathrm{ID}_7$ d=0.64
$\mathrm{ID}_7 \rightarrow \mathrm{ID}_9$ d=1.37
$\mathrm{ID}_3 \rightarrow \mathrm{ID}_9$ d=2.16
Reference Identity = ID_2
$\mathrm{ID}_2 \rightarrow \mathrm{ID}_2$ d=0.50
$\mathrm{ID}_2 \rightarrow \mathrm{ID}_{10}$ d=0.92
$\mathrm{ID}_{10} \rightarrow \mathrm{ID}_2$ d=2.11
$\mathrm{ID}_4 \rightarrow \mathrm{ID}_2$ d=2.22
Reference Identity = ID_{11}
$\mathrm{ID}_{11} \rightarrow \mathrm{ID}_{11}$ d=0.4
$\mathrm{ID}_{11} \rightarrow \mathrm{ID}_1$ d=0.64
$\mathrm{ID}_1 \rightarrow \mathrm{ID}_{11}$ d=1.07
$\mathrm{ID}_8 \rightarrow \mathrm{ID}_{11}$ d=1.26

Fig. 2: Animated figure. Open in a media-enabled viewer like Adobe Reader and click on the inset. Continuing Figure 4 from the main paper, we show more visual results to demonstrate that our method indeed predicts embedding vectors that lie close together when the clips have the same driving identity. Here, we show synthetic videos generated by face-vid2vid generator and the embedding vectors are predicted by the model trained on videos by the same generator. As a reminder, for each row, we pick a reference identity. The green box indicates reenactments driven by the reference identity, the red and blue are cross-reenactments of the reference identity. We compute the average distance of each clip shown here against all other clips driven by the reference identity. The average distance to the other clips of the reference identity is consistent for a given motion, and lower (better) when the reference identity is driving as compared to the cross-reenactments that use the reference identity as target.

Based on the reported distance values, we observe that videos where the reference identity is the driver are closer to the set of other self-reenacted videos driven by the reference identity and far from those where reference identity is the target to be driven by other identities. This further confirms the ability of our model to fingerprint synthetic avatars based purely on facial motion, independent of the appearance of a synthetic talking-head video.

References

1. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing (2014)
2. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16123–16133 (2022)
3. Corona, E., Zanfir, A., Gabriel Bazavan, E., Kolotouros, N., Alldieck, T., Sminchisescu, C.: Vlogger: Multimodal diffusion for embodied avatar synthesis. In: arXiv (2024)
4. Cozzolino, D., Rössler, A., Thies, J., Nießner, M., Verdoliva, L.: ID-Reveal: Identity-aware DeepFake video detection. In: IEEE International Conference on Computer Vision (ICCV) (2021)
5. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)
6. Fox, G., Liu, W., Kim, H., Seidel, H.P., Elgharib, M., Theobalt, C.: VideoForensicsHQ: Detecting high-quality manipulated face videos. In: IEEE International Conference on Multimedia and Expo (2021)
7. He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., Liu, Z.: Forgerynet: A versatile benchmark for comprehensive forgery analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4360–4369 (2021)
8. Jian Zhao, H.Z.: Thin-plate spline motion model for image animation (2022)
9. Jiang, L., Guo, Z., Wu, W., Liu, Z., Liu, Z., Loy, C.C., Yang, S., Xiong, Y., Xia, W., Chen, B., Zhuang, P., Li, S., Chen, S., Yao, T., Ding, S., Li, J., Huang, F., Cao, L., Ji, R., Lu, C., Tan, G.: DeeperForensics Challenge 2020 on real-world face forgery detection: Methods and results. arXiv preprint arXiv:2102.09471 (2021)
10. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In: CVPR (2020)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv (2014)
12. Kovela, S., Valle, R., Dantrey, A., Catanzaro, B.: Any-to-any voice conversion with f 0 and timbre disentanglement and novel timbre conditioning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2023)
13. Kwon, P., You, J., Nam, G., Park, S., Chae, G.: Kodf: A large-scale korean deepfake detection dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10744–10753 (2021)
14. Li, Y., Sun, P., Qi, H., Lyu, S.: Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
15. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one (2018)
16. Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Peng, X.B., Rempe, D.: Multi-track timeline control for text-driven 3d human motion generation. arXiv preprint arXiv:2401.08559 (2024)

17. Remelli, E., Bagautdinov, T., Saito, S., Wu, C., Simon, T., Wei, S.E., Guo, K., Cao, Z., Prada, F., Saragih, J., et al.: Drivable volumetric avatars using texel-aligned features. In: ACM SIGGRAPH 2022 Conference Proceedings (2022)
18. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to detect manipulated facial images. In: IEEE International Conference on Computer Vision (ICCV) (2019)
19. Tian, L., Wang, Q., Zhang, B., Bo, L.: Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. arXiv preprint arXiv:2402.17485 (2024)
20. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: European Conference on Computer Vision (ECCV) (2020)
21. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
22. Wang, Y., Yang, D., Bremond, F., Dantcheva, A.: Latent image animator: Learning to animate images via latent space navigation. In: International Conference on Learning Representations (ICLR) (2022)
23. Yuan, Y., Li, X., Huang, Y., De Mello, S., Nagano, K., Kautz, J., Iqbal, U.: Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. arXiv preprint arXiv:2312.11461 (2023)