Avatar Fingerprinting for Authorized Use of Synthetic Talking-Head Videos

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

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

Abstract. Modern avatar generators allow anyone to synthesize photorealistic real-time talking avatars, ushering in a new era of avatar-based human communication, such as with immersive AR/VR interactions or videoconferencing with limited bandwidths. Their safe adoption, however, requires a mechanism to verify if the rendered avatar is trustworthy: does it use the appearance of an individual without their consent? We term this task *avatar fingerprinting*. To tackle it, we first introduce a large-scale dataset of real and synthetic videos of people interacting on a video call, where the synthetic videos are generated using the facial appearance of one person and the expressions of another. We verify the identity driving the expressions in a synthetic video, by learning motion signatures that are independent of the facial appearance shown. Our solution, the first in this space, achieves an average AUC of 0.85. Critical to its practical use, it also generalizes to new generators never seen in training (average AUC of 0.83). The proposed dataset and other resources can be found at: `https://research.nvidia.com/labs/nxp/avatar-fingerprinting/`.

Keywords: Synthetic Media Verification · Avatar Generators · Avatar Attribution

1 Introduction

Recent digital avatar generators have fueled a myriad of computer vision and graphics applications, allowing anyone to synthesize real-time photorealistic personas. Major companies are now supporting avatar-driven remote interactions over immersive AR/VR (*e.g.* Meta's Pixel Codec Avatar [48], Apple's Vision Pro Persona [1]) or video conferencing (*e.g.* NVIDIA's MAXINE [5], Microsoft's Teams [3]), and selfie filters for altering and enhancing appearance (*e.g.* by Snap and Tiktok). While the avatar generation technology today is still young, the legitimate use of synthetic avatars will be ubiquitous in the future. Without proper guardrails, this poses a real risk of unauthorized use and large-scale spread of visual disinformation. To ensure the safe use of such a technology, the relevant question is no longer whether the content is "real" or not—since, by design, the videos and avatars are all *synthetic*—but rather, whether the synthetically-generated videos and avatars are "trustworthy" or not.

When video conferencing, for instance, a synthetic video portrait generator can be used to save valuable bandwidth by reconstructing a synthetic avatar of the sender at the receiver's end, using only a frame of the target identity and a compact representation of the speaker's facial motion. To ensure the authorized use of such synthetically-generated videos, we want to verify if the driving identity behind a synthetic video (ID_1 or ID_2 in Figure 1) is authorized to control the likeness, or the appearance, of the

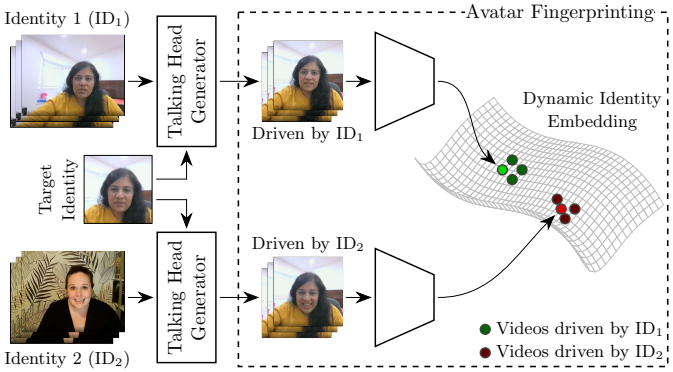

Fig. 1: Talking-head avatar generators can synthesize realistic videos of a target identity from driving videos of different identities. Our method extracts appearance-agnostic temporal facial features and learns an embedding in which the synthetic videos driven by one identity fall close to each other and far from those driven by other identities, regardless of the appearance of the synthetic video. By comparing distances in the embedding space, we evaluate whether an avatar is driven by an authorized identity or not. During evaluation, we only rely on the synthetic videos as input, without requiring any prior knowledge about the driving identity.

synthetic video (target identity in Figure 1). Crucially, we want to only leverage the synthetic video avatars to do so. We call this novel task *avatar fingerprinting*.

We leverage a simple but fundamental observation: facial motions tend to be idiosyncratic, that is, individuals talk and emote in unique ways. For instance, someone may raise one of her eyebrows more than the other, or smile more while talking. These "dynamic identity signatures" [51] have been shown to carry enough information for humans to recognize other individuals, *even when the physical appearance of their face is altered* [31, 41, 51]. This makes them attractive for our task, as they can be derived solely from the driving identity of a talking-head video, regardless of the appearance.

Fortunately, modern avatar generators are becoming increasingly accurate at capturing the facial motion of a person and rendering it onto a target identity. As a first solution to avatar fingerprinting, we then propose to estimate a dynamic identity signature for an identity from the synthesized videos they drove—regardless of the target identities shown. Specifically, we extract facial landmarks and their temporal dynamics from the video. We then introduce a novel contrastive loss to learn a *dynamic identity embedding*, a space where the dynamic identity signatures of a driving identity across multiple videos and target identities are close to each other, and far from those of other driving identities (see Figure 1). We show that this approach, albeit straightforward, is robust and generalizes to generators not seen in training.

Avatar fingerprinting is a new task, and no existing datasets serve its training and validation requirements. There are two key requirements from a dataset. First, we need videos of multiple subjects delivering both scripted and free-form monologues, captured under realistic conditions, such as varying video quality and gaze direction. This allows us to assess if a model leverages talking styles, and not the specific choice of words. Second, to evaluate if the model learns to extract dynamic identity signatures effectively from synthetic videos, we need synthetic talking-head videos for the case in which the driving and target identities are different (cross-reenactment), and that in which they match (self-reenactment). Unfortunately, existing datasets of real videos

only fulfill a subset of these requirements (Table 1 and Section 4). Further, no existing dataset of synthetic videos contain *both* self- and cross-reenactments per subject or use the state-of-the-art talking-head generation technology. To foster research in this new domain, then, we introduce the NVIDIA Facial Reenactment (NVFAIR) dataset, containing real and synthetic face reenactment talking-head videos (Figure 2). Our dataset, which includes ethnically diverse subjects, provides $10\times$ more synthetic facial reenactments than the second largest dataset, for a total of over $650,000$ synthetic videos. It is also the only one using multiple state-of-the-art face reenactment generators [34,60,62], or to provide cross-reenactments driven by all identities, which is critical for training and evaluating avatar fingerprinting algorithms.

In summary:

- We introduce the novel task of avatar fingerprinting, which focuses on verifying the driving identity of synthetic talking-head avatars, rather than classifying them as real or synthetic (by design, all inputs to our model are synthetic).
- We release the first large-scale dataset of subjects delivering scripted and natural monologues, complete with self- and cross-reenactment videos synthesized with multiple state-of-the-art generators.
- We propose a solution for this novel task in the context of video conferencing by extracting person-specific motion signatures, and demonstrate its robustness to various distortions and avatar generators not seen in training.

2 Related Work

Our proposed avatar fingerprinting task aims to verify authorized use of synthetic avatars: this is fundamentally a different problem than traditional forensics research where one aims to detect synthetic media (*e.g*. deepfake detection) or actively mark synthetic content. In our case, by design, the content being evaluated is *always synthetic*: we aim to evaluate its authorized use. Since this is a novel task, no methods exist to directly address it. No methods currently exist to directly address this novel task. Here we discuss the related areas of research.

Learning-based Attribution of Synthetic Media. Learning-based approaches have been used to identify the origin of synthetic media, or to determine if it has been manipulated or altered in some way. Previous work used a pre-trained GAN generator to attribute a synthesized image to its generator via GAN inversion by leveraging the fact that a real image is less invertible [11, 36, 63]. Yet other works focus on attributing other forms of synthetic media, such as text [50]. In contrast, our focus is to attribute a talking-head avatar to the identity driving it, regardless of the appearance of the avatars. Some existing works learn fingerprints associated with cameras to determine whether a video is manipulated [19], or embed watermarks into images and videos [12,25,47,58], which are also shown to transfer to GAN-generated images [64]. Subsequent research introduced a watermark-based conditional GANs for scalable fingerprinting [65]. Our method, in contrast, is a *passive* technique that does not rely on active watermarks.

Deepfake Detection Based on Identity-Specific Features. Deepfake detection ("is a video synthetic?") and avatar fingerprinting ("whose identity is used to generate this synthetic video?") are fundamentally different tasks. Most existing solutions for deepfake detection train a real-vs-synthetic classifier [26, 28, 43, 52, 57, 68], and therefore

cannot be adapted to avatar fingerprinting (where all inputs are synthetic). However, a specific class of detectors leverage identity-specific features to detect synthetic videos by posing the detection problem as an identity-recognition problem. In our experiments, we evaluate some of these methods as baselines. Specifically, Agarwal *et al*. exploit person-specific patterns in facial expressions to detect fake videos [9]. ID-Reveal used facial shapes and motions encoded in a low-dimensional space of a 3D morphable model [13] to handle both face-swapping and face-reenactment deepfakes [20]. Other works explored soft-biometric approaches such as leveraging vocal mannerisms [15], phoneme-viseme consistencies [8], word-facial expression consistencies [6,10], and dynamics of ears [7]. While many of these works need person-specific training, previous works [6, 18, 20, 45] extended this idea to train a CNN-based detector using a large-scale in-the-wild video data [17] and variants of contrastive learning [27, 38, 55, 61]. Agarwal *et al*. combined static facial appearance using a facial recognition model and dynamic facial behaviors using a CNN, and showed that this approach is effective for detecting face-swap deepfakes [6]. Yet another line of research has explored temporal inconsistencies of face identities within a video [45], and identity inconsistencies of inner and outer face regions [22]. Our experiments show that for avatar fingerprinting, such features that are designed to distinguish real from synthetic videos are not reliable, since we only have access to synthetic videos.

Dynamic Facial Identity Signatures. Cognitive scientists have studied the impact of "dynamic facial identity signatures" (*i.e*., characteristic or identity-specific movements of a face) for identity recognition for humans [51]. In one experiment, scientists projected facial animations generated by human actors onto an average head and found that subjects discriminated between individuals based solely on facial motion [31]. In another, subjects correctly attributed animations of synthetic faces to their morphed versions [41]. While these studies point to the existence of "dynamic facial identity signatures" that humans rely on, ours is the first method that isolates these from videos.

Talking-Head Datasets and Generators. Existing talking-head datasets include those that contain only real videos showing a variety of emotions [16, 46, 59], and others that also contain synthetic videos [21, 24, 30, 35, 42, 44, 53]. These datasets cater to traditional forensics and facial analysis tasks. Therefore, they do not contain self-*and* cross-reenactments driven by multiple identities, as well as scripted and free-form monologues across diverse capture settings. There novel requirements posed by the avatar fingerprinting task motivate us to design our own dataset. We focus specifically on face-reenactment talking-head avatar generators for synthetic video generation—this class of generators are the most relevant to AR/VR interactions, video conferencing, and several other applications [2–5, 48]—and combine various modes of human expression for capturing real videos. Given a target facial image and a driving video, these generators reenact the target image using the facial expressions and head pose from the driving video [23, 33, 34, 37, 49, 56, 60, 62, 66, 67]. Another class of talking-head generators use person-specific models [40] and some models aim to preserve the style of the target identity in the synthesized video [39]. However, these models require person-specific training, making them difficult to scale.

3 Terminology

We seek to verify the trustworthiness of a synthesized talking-head video, termed *target video*. We assume that an avatar-generation tool (*e.g*., [60]) created it by animating

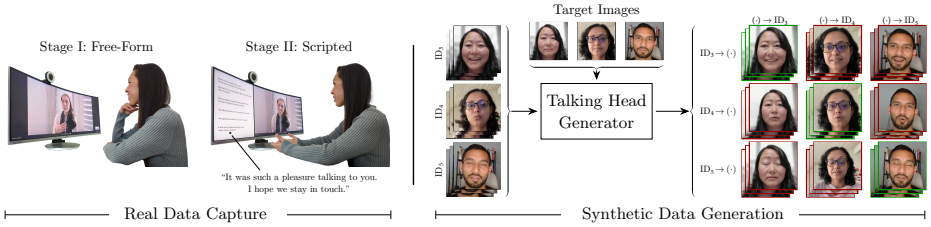

Fig. 2: We introduce the NVFAIR dataset, containing real and synthetic talking-head videos. We capture subjects talking in both scripted and free-form settings. To encourage natural performance, we record the subjects while videoconferencing with each other (left). We then synthesize more than $650,000$ talking-head videos—the largest collection till date—using three state-of-the-art face-reenactment talking-head generators. On the right, each row corresponds to a driving identity ($\text{ID}_i \to (\cdot)$) and each column corresponds to a different target identity ($(\cdot) \to \text{ID}_i$). The videos in which driving and target identity match are self-reenactments, the rest are cross-reenactments.

an image (*target image*) using the expressions and head poses obtained from another video, the *driving video*. We call *driving identity* the identity of the person in the driving video, and *target identity* the identity of the person in the target image. When driving and target identities match, the target video is a *self-reenactment*, while the case of a driving identity used to animate a different target identity is *cross-reenactment*. In both cases, the appearance of the synthesized video is derived from the target identity. This terminology allows us to formally state our goal: we want to verify that a target video is a self-reenactment. With this terminology in mind, we introduce our dataset, which includes real videos as well as self- and cross-reenactment videos.

4 The NVIDIA Facial Reenactment (NVFAIR) Dataset

Recall that avatar fingerprinting is not about detection of synthetic media. Rather, we already know a video to be synthesized, and seek to verify that the driving identity is authorized. This new task dictates a set of requirements for the dataset to be effective for training and evaluation. Specifically, we need a dataset that contains

1. multiple real videos per identity, with scripted and free-form conversations, and with both natural and prescribed emotions,
2. self- *and* cross-reenactments of target identities, with cross-reenactments driven by all subjects to allow for a variety of driving facial structures, and
3. multiple face-reenactment generators.

All relevant existing datasets only capture a subset of these requirements (see Table 1; the Supplement contains further discussion). Moreover they do not use the state-of-the-art talking-head generation technology to synthesize the self- and cross-reenactment videos. We introduce the NVFAIR dataset that features *all* of the above properties. With over $650,000$ synthetic videos, it provides $10\times$ as many videos as the next largest dataset, and uses three different state-of-the-art generators for the self- and cross-reenactment videos. Figure 2 shows an overview of data capture and synthesis.

4.1 Real Data Capture

Capturing videos of monologues delivered by different subjects for the purpose of identity verification introduces two conflicting goals. On the one hand a controlled evalua-

Dataset	# Subjects (source)	(F)ree / (S)cripted?	Emotion: (N)atural / (P)rescribed	(R)eal / (S)elf- / (C)ross-reenact.?	#Face Reenact.	Avg. Videos per Subject	# Face-reenact. Generators	Ethnic Diversity
RAVDESS [46]	24 (new)	(S)	(P)	(R) only	N/A	120 (R)	N/A	✓
MEAD [59]	60 (new)	(S)	(P)	(R) only	N/A	720 (R)	N/A	✓
CREMA-D [16]	91 (new)	(S)	(P)	(R) only	N/A	81 (R)	N/A	✓
VFHQ [24]	36 ([46])	(S)	(P)+(N)	(R)+(S)	1,737	120	1	✓
FF++ [54]	1000 (YT)	(F)	(N)	(R)+(C)	2,000	1(R) + 2(C)	2	✓
KoDF [42]	**403 (new)**	**(F)+(S)**	(N)	(R)+(C)	61,000	150(R) + 151(C)	1	✗
NVFAIR (Ours)	161 (46 new + [46], [16])	**(F)+(S)**	**(P)+(N)**	**(R)+(S)+(C)**	**654,726**	**76(R) + 228(S) + 3,840(C)**	**3**	✓

Table 1: The existing talking-head video datasets were designed for tasks such as deepfake detection or facial emotion analysis. Avatar fingerprinting is a fundamentally different task. As a result, no existing dataset satisfies the requirements for training and evaluating models for avatar fingerprinting. To overcome this limitation, we introduce the NVFAIR dataset, which is the first dataset that offers the complete set of monologue modalities, and features the largest collection of facial reenactments to date. Specifically, it provides scripted and free-form monologues, with natural and prescribed emotions, and self- and cross-reenactments (driven by *all* remaining subjects) synthesized using three generators, alongside original videos for newly-recorded subjects.

tion of the trained models requires predictability of what is spoken to prevent identification algorithms from latching onto the spoken content itself. On the other, we want the subjects to act as they would in a casual conversation, rather than reciting memorized text, to capture their uniquely identifying mannerisms. We address this trade-off by recording the subjects while videoconferencing in pairs, which creates the impression of being in a natural conversation. This differs from existing datasets, in which the subjects look at the camera, but are not interacting with another person during the recording [16,42,46]. We also design two distinct recording strategies: a free-form stage where the subjects are given only general guidance on the topics, and a more controlled scripted stage in which subjects speak short, memorized monologues of 2-3 sentences each, see Figure 2(a). To capture the variability of real-life scenarios, we provided minimal instructions on how to setup the video call, allowing for diverse face, scale, and lightning, bandwidth stability, and background scene clutter. In total we record 46 subjects of diverse genders, ages, and ethnicities, while strictly abiding by a pre-approved IRB protocol (see Supplement for details and privacy considerations).

Stage I: Free-Form Monologues. In this first stage, the two subjects on the call alternate between asking and answering seven pre-defined questions. The questions are designed to avoid sensitive or potentially inflammatory topics. This is critical because we later use sentences spoken by one individual to animate the video of a second individual, quite literally putting words in their mouths. The complete list of questions is in the Supplement. To further create a natural interaction, the subject listening is encouraged to actively but silently engage with the one speaking (*e.g.*, by nodding or smiling).

Stage II: Scripted Monologues. For this stage, we prepared thirty short utterances consisting of two or three sentences each. We chose this length to allow for memorization, while still providing enough content to trigger facial expressions. However, to avoid inducing unnatural expressions, we do not prescribe specific emotions for each utterance. For instance, we do not ask to express anger for a sentence, but we do choose sentences that may naturally evoke it, and used punctuation to encourage it, *e.g.*, "Will you please

answer the darn phone? The constant ringing is driving me insane!" We instruct the subjects to split their screens to show both this list and video call and encourage them to speak to their recording partner when reciting, see Stage II in Figure 2(a). More details, including the full list of utterances can be found in the Supplement.

4.2 Synthetic Talking-Head Videos

Using the videos described in Section 4.1, as well as the original videos from the CREMA-D [16] and RAVDESS [46] datasets, we generate synthetic talking-head videos to train and evaluate our avatar fingerprinting algorithm. Specifically, we pool the 91 identities from the original videos of CREMA-D [16], the 24 identities from those of RAVDESS [46], and the 46 from our own video-conferencing data capture, for a total of 161 unique identities $\mathcal{I}$. Recall that we have several real videos for each identity $\text{ID}_i \in \mathcal{I}$. To avoid a combinatorial explosion of synthetic videos, for all pairs of identities ID_i and ID_j, we use ID_j as the target identity and we randomly select 8 of the videos of ID_i to generate 8 cross-reenactment videos, $\{\text{ID}_i^k \rightarrow \text{ID}_j\}_{k=\{1,..,8\}}$ (all 8 share the same target image). We also generate self-reenactment videos for each of the target identities, by animating their neutral-face images derived from captured videos with each of their real videos.

We use three different generators for synthesizing the videos for all the 161 identities: face-vid2vid [60], LIA [62], and TPS [34]. This allows us to test if our model generalizes across generators. We chose these talking-head generators because they are the state of the art and they preserve the identity-specific facial motion dynamics well. Nevertheless, the reconstruction is not perfect; for instance, in the third row of Figure 2(b) the person in the driving video (ID_5) is squinting, but the eyes are shut in all the synthetic videos, including the self-reenactment video. In total we generate more than $650,000$ synthetic videos, which required more than $2,500$ RTX 3090 GPU hours. More details are in the Supplement.

5 Method

Overview. We seek to verify the driving identity of a synthetic video, independently of the target identity. We leverage the finding from cognitive science research that each person emotes in unique ways when communicating, and that this signal is sufficient for recognition, even when the actual appearance is artificially corrupted [31, 41, 51]. We note that our method does not latch onto artifacts introduced by the generators—a property that we demonstrate by showing generalization to new generators not seen during training. Rather, our features capture the dynamics of the expressions, like the way a person frowns, or the way she smiles. Notably, they are distinct from the temporal artifacts introduced by the generator, and that existing algorithms use to detect whether a video is synthetic or real [29].

An overview of our algorithm for avatar fingerprinting is shown in Figure 3. To capture expressions, we extract the relative position of facial landmarks over time from the input video, as shown in Figure 3 (Section 5.1). We learn to project these temporal signatures onto a *dynamic identity embedding* in which features belonging to the same driving identity are close to each other regardless of the target identity, *i.e.*, independently of appearance (Figure 1). To learn this embedding we train a 3DCNN (where the

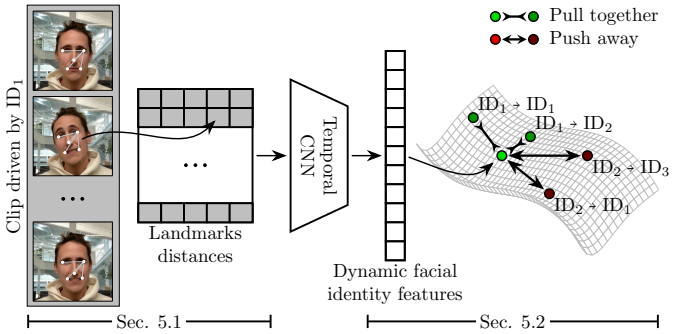

Fig. 3: We extract landmarks from the frames of a talking-head clip, compute their normalized pairwise distances, and concatenate the frame-wise features. We then learn an identity embedding using a loss that pulls closer features of videos driven by the same identity and pushes away those driven by others. $\text{ID}_i \rightarrow \text{ID}_j$ indicates a video that looks like identity j (the "target" identity), and is driven by identity i.

third dimension spans video frames—a *temporal* CNN) with a novel contrastive loss that pulls together all embedding vectors of synthetic videos driven by an individual, while pushing away the embedding vectors of videos driven by all others individuals (Section 5.2). More implementation details are in Section 5.3.

5.1 Dynamic Facial Identity Features

Our first step is to extract temporal features that summarize short segments of the video we wish to fingerprint. We identify the following guiding principles for the extracted features. We would like features that:

1. have minimal dependency on the appearance of the face in the video (that is, the target identity),
2. reflect the dynamics of the expressions, and
3. capture subtle expressions.

One choice could be per-frame 3DMM features [14]: a strategy also used by Cozzolino *et al*. to detect synthetic videos [20]. However, we empirically observe that 3DMM features are not sufficiently expressive, and do not satisfy desideratum 3 (see ablation experiments in Section 6). We observe a similar behavior for action units [9]. Facial landmarks [32] address this issue, but are sensitive to the shape of the face in the video, and thus to the target identity.

To leverage the expressiveness of facial landmarks while abstracting from the underlying facial shape, we compute the pairwise normalized Euclidean distances between each pair of landmarks of a frame. We concatenate these distances into a single vector for the frame, $\mathbf{d}_f$. A subset of the facial landmarks and distances are shown in Figure 3.

We then break the input video into *clips*, which are sequences consisting of F frames and offset by one frame (*e.g*., [1,F], [2,F+1], *etc*.), and concatenate the vectors from all the frames in each clip. Using the change in the relative position of the landmarks over a short period of time (the length of a clip) allows us to capture temporal dynamics with minimal dependence on the absolute position of each landmark, *i.e*., independently of the shape of the face.

We show empirically that our features are a good representation for our task, by comparing against alternative choices for input features such as 3DMM (Section 6).

5.2 Dynamic Identity Embedding Contrastive Loss

While the features described in Section 5.1 are designed to extract low-level motion dynamics, they cannot be used directly to disambiguate two target videos based on their driving identities. We tackle this problem by learning a dynamic identity embedding, a space where videos driven by one subject map to points that are close to each other and far from the videos driven by anybody else.

Specifically, we use a temporal 3DCNN to extract an embedding vector from a clip, which, as described before, is a short segment of an input video. To train the network we use a dataset of synthetic videos driven by different identities. We denote as $\mathcal{C}^k_{\mathrm{ID}_1 \to \mathrm{ID}_2}(t)$ the embedding produced by the network for the clip starting at time t in the k-th video, of a target identity ID_2 driven by identity ID_1. As stated above, we have two main objectives, which we capture with the following terms in our proposed loss function.

We Want to Pull Together All the Videos Driven by ID$_1$. To achieve this, we define the following term:

$$\mathrm{N}_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t) = \sum_{\mathrm{ID}_l,k} \max_n s(\mathcal{C}^j_{\mathrm{ID}_1 \to \mathrm{ID}_2}(t), \mathcal{C}^k_{\mathrm{ID}_1 \to \mathrm{ID}_l}(n)), \tag{1}$$

where $s(\cdot,\cdot) = e^{-\|\cdot,\cdot\|^2}$ is a similarity metric. Intuitively, Equation 1 takes two videos, j and k, both driven by ID_1. Given a clip starting at time t in the first video, it looks for the most similar clip in the second video. Since the driving identity is the same for both videos, Equation 1 encourages an embedding where clips that capture a similar expression are closer to each other. Equation 1 is high even if only one clip from video k has a similar temporal signature to $\mathcal{C}^j_{\mathrm{ID}_1 \to \mathrm{ID}_2}(t)$. That is because even just one occurrence of the same expression is evidence that the driving identity may be the same. Of course, other driving identities may use similar expressions and we address that with the loss term described below. Additionally, we note that k spans the set of *all* videos driven by ID_1, and ID_l spans *all* identities, including $\mathrm{ID}_l = \mathrm{ID}_1$ and $\mathrm{ID}_l = \mathrm{ID}_2$.

We Want to Push Away Videos not Driven by ID$_1$. We define the following term:

$$\mathrm{Q}_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t) = \sum_{\mathrm{ID}_l \neq \mathrm{ID}_1,k} \max_n s(\mathcal{C}^j_{\mathrm{ID}_1 \to \mathrm{ID}_2}(t), \mathcal{C}^k_{\mathrm{ID}_l \to \mathrm{ID}_2}(n)), \tag{2}$$

where, similarly to Equation 1, we take a clip from video j, and look for the most similar clip in video k. This time the two videos share the same target identity, but are driven by different identities: we want all the videos driven by identities different from ID_1 to be pushed away from those driven by ID_1, including videos where ID_1 is the target identity. Note that ID_2 spans *all* identities, including $\mathrm{ID}_2 = \mathrm{ID}_1$ and $\mathrm{ID}_2 = \mathrm{ID}_l$.

We Want to Rely on the Temporal Dynamics of the Videos Driven by ID$_1$. Although we use a temporal CNN, the model could still learn to rely on static expressions, such as a snapshot of the person smiling, rather than the *temporal* progression of expression leading to, or following, the smile. To further encourage the model to learn from the temporal dynamics, we introduce an additional term:

$$\mathrm{R}_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t) = \sum_{\mathrm{ID}_l,k} \max_n s(\mathcal{C}^j_{\mathrm{ID}_1 \to \mathrm{ID}_2}(t), \overset{\sim}{\mathcal{C}}{}^k_{\mathrm{ID}_1 \to \mathrm{ID}_l}(n)), \tag{3}$$

where $\overset{\sim}{\mathcal{C}}^{k}_{\mathrm{ID}_1 \rightarrow \mathrm{ID}_l}$ denotes a version of the clip $\mathcal{C}^{k}_{\mathrm{ID}_1 \rightarrow \mathrm{ID}_l}$ from Equation 1 with randomly shuffled frame ordering. We want such time-shuffled versions of the clips driven by ID_1 to be pushed away from the pristine clips driven of ID_1. Effectively, this means that the driving identity of the time-shuffled clips is regarded as different from ID_1. In other words, we want to pull together video clips in the learned embedding space only when the temporal facial dynamics are characteristic of ID_1. We further show the importance of this term in Section 6.3.

Combining Equations 1, 2, and 3, we write the probability that the embedding vector $\mathcal{C}^{j}_{\mathrm{ID}_1 \rightarrow \mathrm{ID}_2}(t)$ lies close to the embedding vectors for *all* video clips driven by ID_1 and far from *all* the videos driven by others as

$$p_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t) = \frac{\mathrm{N}_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t)}{\mathrm{N}_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t) + \mathrm{Q}_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t) + \mathrm{R}_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t)}, \tag{4}$$

and the complete loss term as

$$\mathcal{L} = \sum_{j,\mathrm{ID}_1,\mathrm{ID}_2,t} -\log(p_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t)). \tag{5}$$

5.3 Implementation

Parameter Choices. To extract the per-frame dynamic facial identity features $\mathbf{d}_f$, we detect 126 facial landmarks for each frame [32], and compute the per-frame normalized pairwise Euclidean distances between these landmarks. The clip duration is set to 71 frames. We find that this is sufficient to capture the facial dynamics that are meaningful for avatar fingerprinting, while also maintaining a good trade-off between speed and accuracy. We also experiment with shorter-duration video clips (see Supplement). The input tensor to the temporal CNN is obtained by concatenating $\mathbf{d}_f$ across 71 frames. In each batch, we include 8 unique identities. For each identity ID_i, the pull term (Equation 1) comprises 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. This allows the neural network 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 consists of clips with the remaining 7 identities in the batch serving as driving identities (8 clips per driving identity), as well as the time-shuffled self-reenactments of ID_i (Equation 3). Additional training details can be found in the Supplement.

Training, Validation, and Testing Datasets. Of the 161 total identities (pooling together the identities from our dataset, RAVDESS, and CREMA-D, see Section 4.2), we reserve 35 for testing, 14 for validation, and 112 for training. We ensure that there are no cross-set cross-reenactments: that is, identities in the training set only drive other training-set identities (and similarly for the validation and test sets). This allows us to test the models on facial appearances and expressions that were not seen during training. To evaluate the generalization of trained models to new generators, we train our network on videos generated for the training set identities using one generator, and evaluate on the synthetic videos of test-set identities synthesized using remaining two generators.

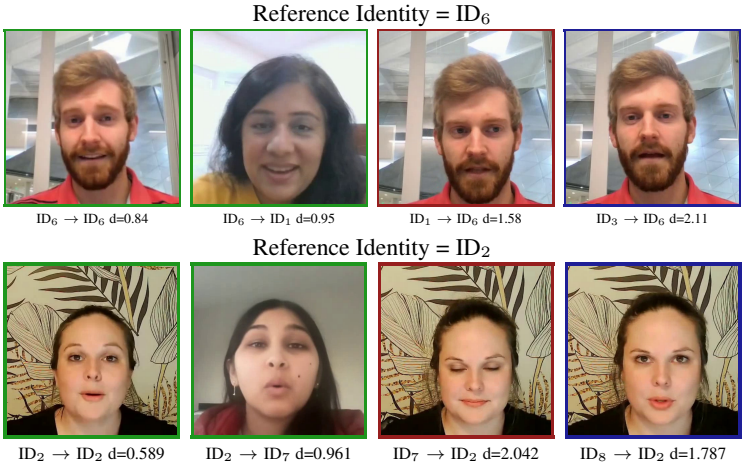

Fig. 4: Animated figure. Open in a media-enabled viewer like Adobe Reader and click on the inset. Our embeddings capture the dynamics of an expression, rather than the appearance of the face. For each row, we pick a reference identity. The green box indicates reenactments driven by the reference identity, 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. Here, we show videos generated by face-vid2vid [60], and use the embedding vectors predicted by the model trained on the same generator (see Figure 6 generalization to new generators not seen during training).

6 Evaluation

We thoroughly evaluate our algorithm both qualitatively and quantitatively. Our algorithm outperforms reasonable baselines (Section 6.1), it generalizes to generators not seen in training (Section 6.2), and it is robust to video compression (Supplement). We also perform a number of ablation studies to analyze our design choices: our input features, the importance of the time-shuffling term in the loss function (Section 6.3), and the impact of clip duration (Supplement).

We begin by evaluating qualitatively our method's ability to extract embedding vectors based on the driving identity. Figure 4 shows a set of self- and cross-reenacted clips (please view the animation in a media-enabled viewer, such as Adobe Acrobat). For each row, we take one identity as reference and we compute the embedding vectors of clips that use it both as the driving and the target identity. We then compute the average Euclidean distance of the resulting embedding vectors against those of *the self-reenactments by the same reference identity*. We note that the average distance d is lower when the driving identity matches the reference identity (first two columns). We also note that the distances between the clips in the first two columns are similar: this confirms that the distance is a function of the facial motion, rather than the facial appearance. When the driving identity changes, the average distance increases, even if the target identity matches the reference identity, which is precisely our goal. More results are in the Supplement.

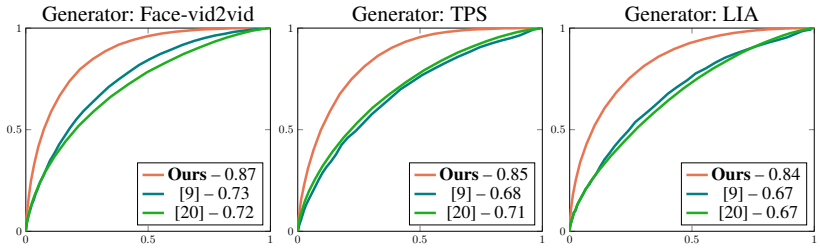

Fig. 5: ROC curves and AUC values for our method and two baselines: Agarwal *et al*. [9] and ID-Reveal [20]. Each sub-plot shows the results on our test set for each of the three talking-head generators: face-vid2vid [60], LIA [62], and TPS [34].

To evaluate our approach more formally, we use the 35 unique test-set identities that are not used as driving or target identities in the training set (Section 5.3). One at a time, we treat each identity ID_i as target and synthesize cross-reenactments using all the remaining identities as drivers. This is the set of "unauthorized" synthetic videos for ID_i. The self-reenacted samples for ID_i form the "authorized" set. Note that there are several self-reenacted videos of ID_i, one per original video of ID_i.

For each target identity ID_i, we extract the dynamic identity embedding vector of all the clips in the pool of its self- and cross-reenacted videos, and compute their Euclidean distances. That is, for clip k we compute

$$\begin{aligned} &d(\mathcal{C}^k_{\text{ID}_i\rightarrow\text{ID}_i}, \mathcal{C}^l_{\text{ID}_i\rightarrow\text{ID}_i}), \quad \forall l \neq k, \quad \text{and} \\ &d(\mathcal{C}^k_{\text{ID}_i\rightarrow\text{ID}_i}, \mathcal{C}^l_{\text{ID}_j\rightarrow\text{ID}_i}), \quad \forall l \neq k, \quad \forall i \neq j. \end{aligned} \tag{6}$$

We threshold these distances for each target identity to get an ROC curve, and average across the target identities to get the overall area under the curve (AUC). We note that this AUC measures one model's ability to classify a synthetic video as self-reenactment or as cross-reenacted. We conduct further analysis of our model's ability to classify other categories of videos—such as, evaluating AUC on same-utterance self- vs. cross-reenactments or on scripted vs. free-form monologues—in the Supplement.

6.1 Comparisons with Existing Methods

Avatar fingerprinting is a novel task, and no existing methods directly address it. The closest related works aim at detecting real versus synthetic media. As discussed in Section 2, some of these detectors learn identity-specific features such as facial expressions and head poses [9], or facial shapes and motion [20] and can serve as baselines for the task of avatar fingerprinting with some adaptation. The work by Agarwal *et al*. trains a model to detect synthetic videos of a *specific* identity [9]. To adapt it to our task, we train 35 different models, one for each identity in the evaluation, by splitting the corresponding original videos into two subsets. We then test each model on the self- and cross-reenactment videos of the corresponding identity. ID-Reveal, trained on a large-scale dataset, learns an embedding space where real videos of a specific identity are grouped together [20]. Since it shows good generalization to new identities, for the task of synthetic media detection, we directly use the pre-trained model on our data to detect, once again, self- versus cross-reenactment. Figure 5 shows the ROC curves for our method compared to these baselines, on three face-reenactment generators (face-vid2vid [60],

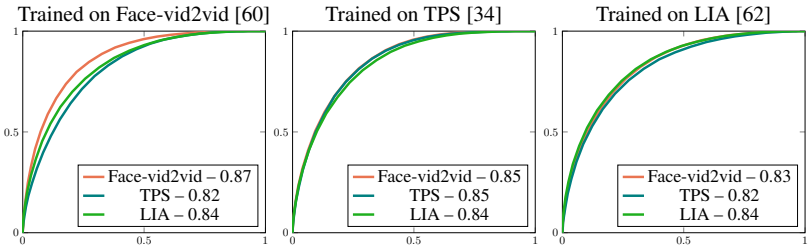

Fig. 6: Generalization to new generators. To study the robustness to new talking-head generators, we train three version of our model on three different generators and test on all three.

LIA [62], and TPS [34]). Our method (AUC=0.868 on face-vid2vid) outperforms by a wide margin both ID-Reveal (AUC=0.720 on face-vid2vid), and the method by Agarwal *et al*. (AUC=0.726 on face-vid2vid). We also note that, unlike ID-Reveal and our method, Agarwal *et al*. uses a different model per identity.

6.2 Generalization to New Generators

For an avatar fingerprinting model algorithm to be broadly applicable, generalization to new talking-head generators that are not seen in training is crucial. Since our dataset contains videos synthesized by three different generators, we can train three models, one with each generator, and test these models on all three generators. Figure 6 shows the resulting ROC curves and AUC values: the overlap of the curves and similar AUC values in each subplot confirms that our method generalizes well to new generators.

6.3 Ablation Study

Our method outperforms existing baselines by introducing two novel components: the dynamic facial identity features, which capture the facial dynamics in a compact and expressive way, and the loss function, which defines the shape of the identity embedding. Here we study the contribution of each, using the face-vid2vid generator for training and testing. We evaluate the contribution of our dynamic facial identity features by swapping them with 3DMM features [13], a popular choice to capture facial dynamics. Since we use a temporal CNN backbone similar to the one from ID-Reveal, for this ablation we use the loss function proposed in their original paper [20]. We re-train the same network using our features and observe a jump from 0.718 to 0.754 in terms of AUC. Upon inspection we notice that the 3DMM features tend to over-smooth the facial motion, and are unable to capture subtle dynamics that prove critical to avatar fingerprinting, and which our features capture. We also evaluate the contribution of our dynamic identity embedding loss and observe a further improvement (AUC 0.868). With our loss formulation, the advantage of $\mathrm{R}_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t)$ in Eq. 5 is also evident when compared against a model trained without this term (AUC 0.850). Table 2 summarizes this ablation study. Additional experiments in the Supplement show the impact of F, performance on scripted vs. free-from monologues, and robustness to video distortions.

6.4 Limitations

Our algorithm is less discriminative of subjects that are less emotive and more neutral. In the future, relying on more granular dynamic signatures that can extract micro-

Input Features	Loss	AUC
3DMM	ID-Reveal rec. loss [20]	0.718
Landmark distances	ID-Reveal rec. loss [20]	0.754
Landmark distances	Our loss without $\mathrm{R}_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t)$	0.850
Landmark distances	Our loss with $\mathrm{R}_{j,\mathrm{ID}_1,\mathrm{ID}_2}(t)$	0.868

Table 2: Ablation study showing the importance of our input features and loss function design.

expressions can help alleviate this. The performance of our method degrades when expressions that are critical to verifying the driving identity are not captured by the synthetic portrait generator. Lastly, our dataset currently features only one style of interaction: one-on-one conversations. We plan to expand to other conversation styles, such as one-way speeches, in future.

7 Societal Impact

We acknowledge the societal importance of introducing guardrails when it comes to the use of talking-head generation technology. We present this work as a step towards trustworthy use of such technologies. Nevertheless, our work could be misconstrued as having solved the problem and inadvertently accelerate the unhindered adoption of synthetic talking-head technology. We do not advocate for this. Instead, we stress that this is the first work on this topic and underscore the importance of further research.

8 Conclusions

Highly photo-real synthetic talking-head generators are becoming increasingly beneficial to applications such as video conferencing and AR/VR-based remote interactions. This trend raises the important new research question of how best to also ensure their safe use in such scenarios. To this end, we investigate the new problem of avatar fingerprinting, to authenticate legitimate talking-heads created by authorized users. We leverage the fact that driving individuals have uniquely identifying dynamic motion signatures, which are also preserved in the synthetic videos that they drive. Since none exists, we contribute a new large-scale dataset carefully designed for avatar fingerprinting and related tasks. We hope that our work lays the foundation for further research.

Acknowledgements. We would like to thank the data-capture participants, and Desiree Luong, Woody Luong, and Josh Holland for their help with Figure 2. We acknowledge David Taubenheim for the voiceover in the demo video, and Abhishek Badki for help with the training infrastructure. We thank Joohwan Kim, Rachel Brown, Anjul Patney, Ben Boudaoud, Josef Spjut, Saori Kaji, Nikki Pope, and Kai Pong for their help with putting together the data capture protocols, informed consent form, photo release form, and agreements for data governance and third-party data sharing. Koki Nagano, Ekta Prashnani, and David Luebke were partially supported by DARPA's Semantic Forensics (SemaFor) contract (HR0011-20-3-0005). This research was funded, in part, by DARPA's Semantic Forensics (SemaFor) contract HR0011-20-3-0005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

References

1. Apple's vision pro. https://www.apple.com/apple-vision-pro/, accessed: 2024-03-06
2. Heygen. https://www.heygen.com, accessed: 2023-11-16
3. Microsoft teams mesh. https://www.microsoft.com/en-us/microsoft-teams/microsoft-mesh, accessed: 2024-03-06
4. Myheritage. https://www.myheritage.com, accessed: 2023-11-16
5. Nvidia's maxine. https://developer.nvidia.com/maxine, accessed: 2024-03-06
6. Agarwal, S., El-Gaaly, T., Farid, H., Lim, S.N.: Detecting deep-fake videos from appearance and behavior. 2020 IEEE International Workshop on Information Forensics and Security (WIFS) pp. 1–6 (2020)
7. Agarwal, S., Farid, H.: Detecting deep-fake videos from aural and oral dynamics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2021)
8. Agarwal, S., Farid, H., Fried, O., Agrawala, M.: Detecting deep-fake videos from phoneme-viseme mismatches. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2020)
9. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting world leaders against deep fakes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
10. Agarwal, S., Hu, L., Ng, E., Darrell, T., Li, H., Rohrbach, A.: Watch those words: Video falsification detection using word-conditioned facial motion. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2023)
11. Albright, M., McCloskey, S.: Source generator attribution via inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
12. Baluja, S.: Hiding images in plain sight: Deep steganography. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
13. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of SIGGRAPH (1999)
14. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2003)
15. Boháček, M., Farid, H.: Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. Proceedings of the national academy of Sciences (2022)
16. 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)
17. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
18. Cozzolino, D., Nießner, M., Verdoliva, L.: Audio-visual person-of-interest deepfake detection (2022)
19. Cozzolino, D., Poggi, G., Verdoliva, L.: Extracting camera-based fingerprints for video forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
20. 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)
21. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Canton Ferrer, C.: The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397 (2020)

22. Dong, X., Bao, J., Chen, D., Zhang, T., Zhang, W., Yu, N., Chen, D., Wen, F., Guo, B.: Protecting celebrities from deepfake with identity consistency transformer. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
23. Drobyshev, N., Chelishev, J., Khakhulin, T., Ivakhnenko, A., Lempitsky, V., Zakharov, E.: MegaPortraits: One-shot megapixel neural head avatars (2022)
24. 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)
25. Fridrich, J.: Steganography in Digital Media: Principles, Algorithms, and Applications. Cambridge University Press (2009)
26. Ge, S., Lin, F., Li, C., Zhang, D., Wang, W., Zeng, D.: Deepfake video detection via predictive representation learning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) (2022)
27. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2006)
28. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
29. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
30. 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)
31. Hill, H., Johnston, A.: Categorizing sex and identity from the biological motion of faces. Current Biology (2001)
32. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
33. Hong, F.T., Zhang, L., Shen, L., Xu, D.: Depth-aware generative adversarial network for talking head video generation (2022)
34. Jian Zhao, H.Z.: Thin-plate spline motion model for image animation (2022)
35. 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
36. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
37. Khakhulin, T., Sklyarova, V., Lempitsky, V., Zakharov, E.: Realistic one-shot mesh-based head avatars. In: European Conference on Computer Vision (ECCV) (2022)
38. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
39. Kim, H., Elgharib, M., Zollhöfer, M., Seidel, H.P., Beeler, T., Richardt, C., Theobalt, C.: Neural style-preserving visual dubbing. ACM Transactions on Graphics (ToG) (2019)
40. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics (ToG) (2018)
41. Knappmeyer, B., Thornton, I., Bülthoff, H.: Facial motion can determine facial identity. Journal of Vision (2001)
42. 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)

43. Li, J., Xie, H., Yu, L., Zhang, Y.: Wavelet-enhanced weakly supervised local feature learning for face forgery detection. In: Proceedings of the 30th ACM International Conference on Multimedia (2022)
44. 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)
45. Liu, B., Liu, B., Ding, M., Zhu, T., Yu, X.: Ti2net: Temporal identity inconsistency network for deepfake detection. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2023)
46. 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)
47. Luo, X., Zhan, R., Chang, H., Yang, F., Milanfar, P.: Distortion agnostic deep watermarking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
48. Ma, S., Simon, T., Saragih, J., Wang, D., Li, Y., De La Torre, F., Sheikh, Y.: Pixel codec avatars. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
49. Mallya, A., Wang, T.C., Liu, M.Y.: Implicit Warping for Animation with Image Sets. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
50. Munir, S., Batool, B., Shafiq, Z., Srinivasan, P., Zaffar, F.: Through the looking glass: Learning to attribute synthetic text generated by language models. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (2021)
51. O'Toole, A.J., Roark, D.A., Abdi, H.: Recognizing moving faces: A psychological and neural synthesis. Trends in Cognitive Sciences (2002)
52. Passos, L.A., Jodas, D., da Costa, K.A., Júnior, L.A.S., Rodrigues, D., Del Ser, J., Camacho, D., Papa, J.P.: A review of deep learning-based approaches for deepfake content detection. arXiv preprint arXiv:2202.06095 (2022)
53. 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)
54. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1–11 (2019)
55. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
56. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
57. Sun, K., Yao, T., Chen, S., Ding, S., Li, J., Ji, R.: Dual contrastive learning for general face forgery detection (2022)
58. Tancik, M., Mildenhall, B., Ng, R.: Stegastamp: Invisible hyperlinks in physical photographs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
59. 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)
60. 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)
61. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
62. 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)

63. Yacoob, Y.: Gan-scanner: A detector for faces of stylegan+ (2021), https://github.com/yaseryacoob/GAN-Scanner
64. Yu, N., Skripniuk, V., Abdelnabi, S., Fritz, M.: Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In: IEEE International Conference on Computer Vision (ICCV) (2021)
65. Yu, N., Skripniuk, V., Chen, D., Davis, L., Fritz, M.: Responsible disclosure of generative models using scalable fingerprinting. In: International Conference on Learning Representations (ICLR) (2022)
66. Zakharov, E., Ivakhnenko, A., Shysheya, A., Lempitsky, V.: Fast bi-layer neural synthesis of one-shot realistic head avatars. In: European Conference on Computer Vision (ECCV) (2020)
67. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: IEEE International Conference on Computer Vision (ICCV) (2019)
68. Zheng, Y., Bao, J., Chen, D., Zeng, M., Wen, F.: Exploring temporal coherence for more general video face forgery detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)