IdentityMask: Deep Motion Flow Guided Reversible Face Video De-identification

Yunqian Wen, Bo Liu, *Senior Member, IEEE,* Jingyi Cao, Rong Xie, *Member, IEEE,* Li Song, *Senior Member, IEEE,* and Zhu Li, *Senior Member, IEEE,*

***Abstract*—Unprecedented video collection and sharing have exacerbated privacy concerns and led to increasing interest in privacy-preserving tools. A satisfactory video de-identification tool should be able to remove sensitive identity information from face videos while maintaining useful information for other identity-agnostic tasks. Meanwhile, it is necessary to allow the authority to inspect real identity when abnormal events are detected. Existing methods only focus on the study of de-identification, and lack the desired recovery ability when granting permissions. Furthermore, they all process the videos frame by frame, which hardly benefit from motion and inter-frame information. In this paper, we propose a modular architecture for reversible face video de-identification, called IdentityMask, which leverages deep motion flow to avoid per-frame evaluation. Our framework consists of two processes: the de-identification process provides a protective mask for identity information, while the recovery process can remove the protective mask if and only if the right key is provided. To this end, a *Protection Module* and a *Recovery Module* are built as two major functional modules, both based on an identity disentanglement network and guided by a crucial *Motion Flow Module*. An *Affine Transformation Module* provides simple but reliable assistance. Extensive experiments on a diverse natural video dataset (gender, ethnicity, age, etc.) demonstrate the effectiveness of the proposed framework for reversible face video de-identification.**

***Index Terms*—Reversible video de-identification, privacy protection, security and forensics.**

I. Introduction

The proliferation of smartphones and short-video platforms has changed the way people create and consume video. Ordinary individuals have become the primary producers and consumers of video activities [1]. With the surge in the number of online videos, the sensitive information (such as human faces) contained in these videos has caused unprecedented violations in the field of personal privacy protection [2]. New privacy laws and regulations begin to forbid the public disclosure of personal sensitive information. However, since the access and utilization of such videos are neither easy to monitor nor to prevent, it is essential to grant users the option to obfuscate themselves out of these videos.

Advanced computer vision technology and blooming online social networks have greatly facilitated both daily social interactions and face videos sharing [3]. While the media users are willing to guard their personal privacy, they are also eager to enjoy the convenience of advanced identity-agnostic computer vision applications. These applications do not need to identify the people in the videos, for instance, face detection, face reenactment, emotion analysis, action recognition and so on. Therefore, maintaining the utility of identity-protected videos to support existing identity-agnostic tasks and normal online social use becomes a new and appealing topic. In addition, the Internet is not an extrajudicial land. When an incident such as a crime occurs, authorities should be able to examine the original videos for forensics purposes.

Reversible face video de-identification is an effective solution to the aforementioned issues. But it is very challenging to design a satisfactory technique to achieve this target. On the one hand, it requires obfuscating the sensitive identity information of the subject while minimizing distortion or changes in other non-identity features [4]–[6], i.e., ensuring visual similarity including appearance, posture, expression, background information, etc. On the other hand, in the case of "after-the-fact forensics", it allows the authorized party to fully restore the anonymous videos [7].

Existing face video privacy-preserving methods [4]–[6], [8]–[12] only focus on the former aspect, and lack the restoration ability, which does not meet the privacy requirements of keeping pace with the times. Furthermore, these methods process video frame by frame without considering the temporal relationship between frames. This can easily make the de-identified video flicker due to temporal inconsistency and cause excessive computational overhead.

In order to overcome the above problems, in this paper, we present a novel and effective reversible face video de-identification modular framework guided by deep motion flow, called IdentityMask. Our framework contains two main functional modules (*Protection Module* and *Recovery Module*), both of which are guided by the crucial *Motion Flow Module*, while an *Affine Transformation Module* provides simple but reliable assistance. Instead of per-frame processing, it lets only the first affined frame go through the *Protection*/*Recovery Module*, and calculates the deep motion flow between every two adjacent frames via a motion flow generator. Then the subsequent de-identified/recovered frames can be generated based on the first protected/recovered frame by the guide of the relative motion representation. All the synthesized videos can be visually pleasing without flickering. Also, we design a discrete key space where keys condition identity changes to securely enable the recovery transformation only for the authorized parties. Specifically, any video that the user wants to obfuscate will be transformed into the de-identified one

Y. Wen, J. Cao, R. Xie and L. Song are with The Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (email: wenyunqian@sjtu.edu.cn; cjycaojingyi; xierong@sjtu.edu.cn; song_li@sjtu.edu.cn).

B. Liu is with School of Computer Science, University of Technology Sydney, NSW 2007, Australia (email:bo.liu@uts.edu.au).

Z. Li is with the Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City, Kansas City, MO 64110 USA (e-mail: lizhu@umkc.edu).

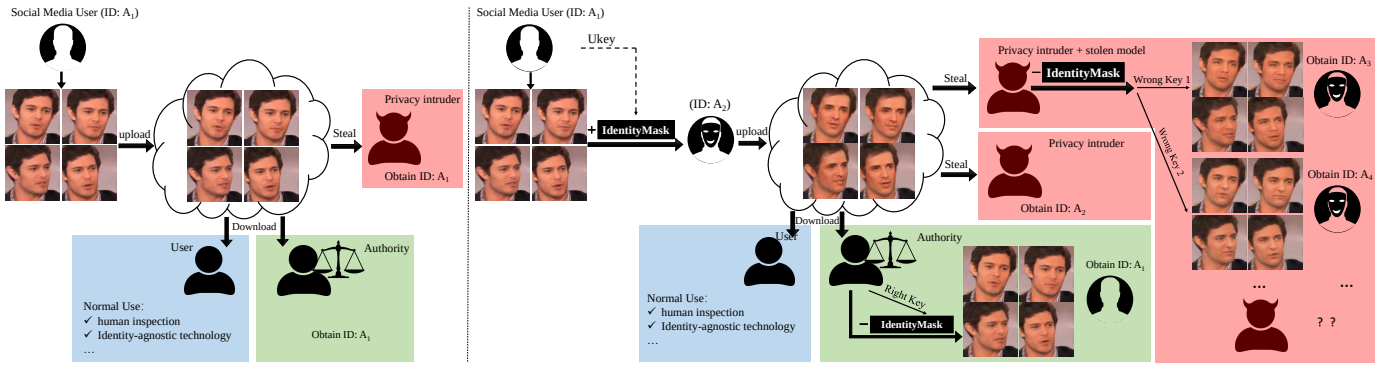


Fig. 1. Comparison between a vulnerable social media platform (left panel) and a IdentityMask protected social media platform (right panel) in maintaining normal use, safeguarding legal supervision as well as handling malicious privacy intruders for stealing personal identity information.

with an assigned Ukey (a number that matches the user's UID). Then, given a de-identified video, the original video can only be recovered if the correct key and the trained recovery pipeline are provided. We further increase security as follows: given an anonymized video, even the trained recovery pipeline is stolen, if a wrong key is provided, it changes to a new identity that is still different from the original one (Fig. 1, "Wrong Key"), with a natural appearance. When the framework is used in practical applications, the Ukey can be a number defined according to the specific situation. For example, specified by the user, distributed by the video platform, and so on.

This paper is built upon our prior work [13] and [14] with multiple improvements. Compared with [13], we add a recovery process and achieve reversible face video de-identification, which is more conducive to the establishment of orderly online social networks. We also add Ukey to enable users to control the de-identification process. The method in [14] works on still images and cannot be directly applied to videos. While in this paper, we use two specifically designed modules to process videos. In addition, the reference identity in [14] is obtained by randomly selecting k (during the experiment, we set k=3000) different identities from the training set, which is inconvenient and may cause legal disputes. We solve this shortcoming by leveraging random seeds conditioned on Ukeys to generate reference identities, which is more flexible and gets rid of the need for auxiliary faces.

In summary, the main contributions of this paper are described as follows:

- To the best of our knowledge, the proposed IdentityMask is the first method that can conduct reversible de-identification for face videos. Our proactive defence technique well addresses the growing concerns about personal privacy protection during online video sharing. On the one hand, the users can protect individual identity with a certain Ukey (equivalent to a password) before sharing. On the other hand, when the identity-protected video is released, the authorized party can still obtain the recovered video with the original identity through the correct Ukey, while it is difficult for unauthorized parties to infer the true identity.
- We introduce deep motion flow into video de-identification tasks to avoid per-frame processing. We show that the *Motion Flow Module* can provide important guidance for IdentityMask pipeline to generate identity-protected/identity-recovered videos, resulting in significantly improved synthesis quality and reduced computational overhead.
- Experimental results on a diverse face video dataset (gender, ethnicity, age, etc.) have demonstrated the effectiveness of our proposed IdentityMask. In addition, we introduce evaluation metrics designed for videos, which are lacking in existing literature.

II. Related Work

To our best knowledge, our work is unique and there is no previous similar work to directly compare with. Nevertheless, it is closely related to previous video de-identification work, which can be classified into two categories according to the application scenarios, as described below.

A. Face video de-identification

The face videos, such as vlogs, live-streaming sales, speeches and interviews, are shot with human head and part of the upper body as the main subject, and have become a popularity in social media in recent years [15]–[17]. Therefore, the corresponding de-identification research is emerging. We classify these approaches into two categories.

Identity swapping-based methods. Replacing the identity in a face video with someone else is a straightforward but effective idea of de-identification. The "someone" here can be either a real identity provider or a somehow synthesized identity that doesn't exist in reality. Generally, the latter is a more thorough way of privacy protection.

Zhu *et al.* [12] applied deepfake technology to de-identify medical examination videos by explicitly swapping the patients' faces with open-source characters. However, such simple operation will lead to an extreme deterioration in visual similarity, thus more skillful identity swapping-based methods are proposed. With several pre-trained active appearance models (AAMs), Samarzija *et al.* [9] found the best fit model of the original face, and swapped the face region with another face taken from the training dataset. Meden *et al.* [10] replaced the

TABLE I
A COMPARISON TO THE EXISTING FACE VIDEO DE-IDENTIFICATION METHODS

	Gross, [8]	Samarzija, [9]	Meden, [10]	Li, [11]	Zhu, [12]	Ren, [4]	Gafni, [5]	Maximov, [6]	ours
Without auxiliary faces	yes	no	no	no	no	yes	no	no	yes
Demonstrated on a diverse video dataset (gender, ethnicity, age, etc.)	no	no	yes	no	no	yes	yes	yes	yes
Demonstrated on a diverse face video dataset (gender, ethnicity, age, etc.)	no	no	no	no	no	no	yes	no	yes
Without per-frame processing	no	no	no	no	no	no	no	no	yes
Recover original face	no	no	no	no	no	no	no	no	yes
Reference to a comparison with ours						Fig. 5	Fig. 6	Fig. 5	

original faces with surrogates generated from a small number of identities. Instead of synthesizing the surrogate faces through simple pixel averaging, they used a convolutional neural network (CNN) to generate artificial surrogates. Li *et al.* [11] used a trained facial attribute transfer model (FATM) to map the non-identity related facial attributes to the face of donors, who were a small number (usually 2 to 3) of consented subjects. Gafni *et al.* [5] utilized a multi-level face descriptor to convert the identity of the original face to that of the target face. Specifically, the removal of identity was done via distancing the face descriptors of the output video from those of the original image. Maximov *et al.* [6] removed the identification characteristics of input people in the bottleneck of the generator via a one-hot label which encoded the desired identity, meanwhile they leveraged the input landmark images with some original identity information left to preserve the pose, thus the generated identity was a composition of both the landmark identity and the desired identity.

Identity disentanglement-based methods. Although the former kind of methods have evolved to a stage with amazing results, its reliance on auxiliary identities can make it difficult to apply under increasingly stringent regulations. For example, consent from the target identity provider should be obtained regularly, which is kind of inconvenient.

Consequently, another pattern that deals with face video de-identification through certain face models by training to extract facial feature representations begins to rise. Once the representations have been disentangled, a de-identified face video can then be generated based on the new representations originated in which the protected identity information has been eliminated, reduced, or obfuscated. During this time, a new virtual identity will generate. Our method follows this pattern.

Gross *et al.* [8] factorized input images into identity and non-identity factors using a generative multi-factor model, and then applied a de-identification algorithm on the combined factorized data before using the bases of the multi-factor model to reconstruct de-identified images. With the development of deep neural network, deep face models can better undertake the task of disentanglement. Ren *et al.* [4] employed a multi-task extension of the generative adversarial network (GAN), where a face anonymizer tried to minimize the identification accuracy and an activity detector tried to maximize spatial action detection performance.

We provide a comprehensive comparison between the previous face video de-identification methods and ours in Table I.

B. Surveillance video de-identification

Video surveillance systems have been omnipresent for a considerable time, with large systems being deployed in strategic places such as public transportation, airports, city centers, or residential areas. In order to address the never-ending concerns about personal privacy protection, a large amount of targeted de-identification technologies have been proposed. As the surveillance videos typically contain multiple people (with full body) and complex surrounding environment, these methods always attach great importance to efficient face detection and tracking, and apply anonymization on the segmented origins. Here we classify them into three categories.

Obfuscation-based methods. These methods achieve video de-identification by obfuscating each frame's privacy sensitive region in some way. Specifically, Dufaux *et al.* [18] used domain scrambling methods to achieve distortion. Schiff *et al.* [19] employed solid ellipsoidal overlays, while minimized the overlay area to maximize the remaining observable region of the scene. Chen *et al.* [20] implemented an EMHI approach to obscure the entire body. Agrawal *et al.* [21] applied the exponential blur of pixels in the voxel or line integral convolution. Mrityunjay *et al.* [22] obscured the segmented bounding box region by using Gaussian Blur of the pixels and binarizing the intensity values. Ivasic-Kos *et al.* [23] applied 2D Gaussian filtering to automatically obfuscate the human body shape information. Blažević *et al.* [24] replaced humans with rendered 3D human models. Ryoo *et al.* [25] presented an inverse super resolution (ISR) paradigm that used extreme low-resolution (e.g., 16×12) videos to achieve de-identification and benefit activity recognition. Flouty *et al.* [26] introduced a sliding window smoother for temporal smoothing on the detections. [27] obfuscated the privacy-sensitive parts at multiple privacy levels by using a random corruption matrix. Kim *et al.* [28] fundamentally protected privacy by blurring unwanted blocks in images, yet ensured that the robots could understand the video for their perception. Wang *et al.* [29] used a lensless coded aperture (CA) camera, which placed only a coded aperture in front of an image sensor, the resulting CA images would be visually unrecognizable and were difficult to restore with high fidelity. Zhou *et al.* [1] proposed a novel PsOP framework which was extendable to any potential privacy-sensitive objects pixelation after leveraging pre-trained detection networks as the backbone. Tu *et al.* [30] generated bounding boxes to cover the regions of interest, then the pixels inside bounding boxes could be modified to achieve a certain

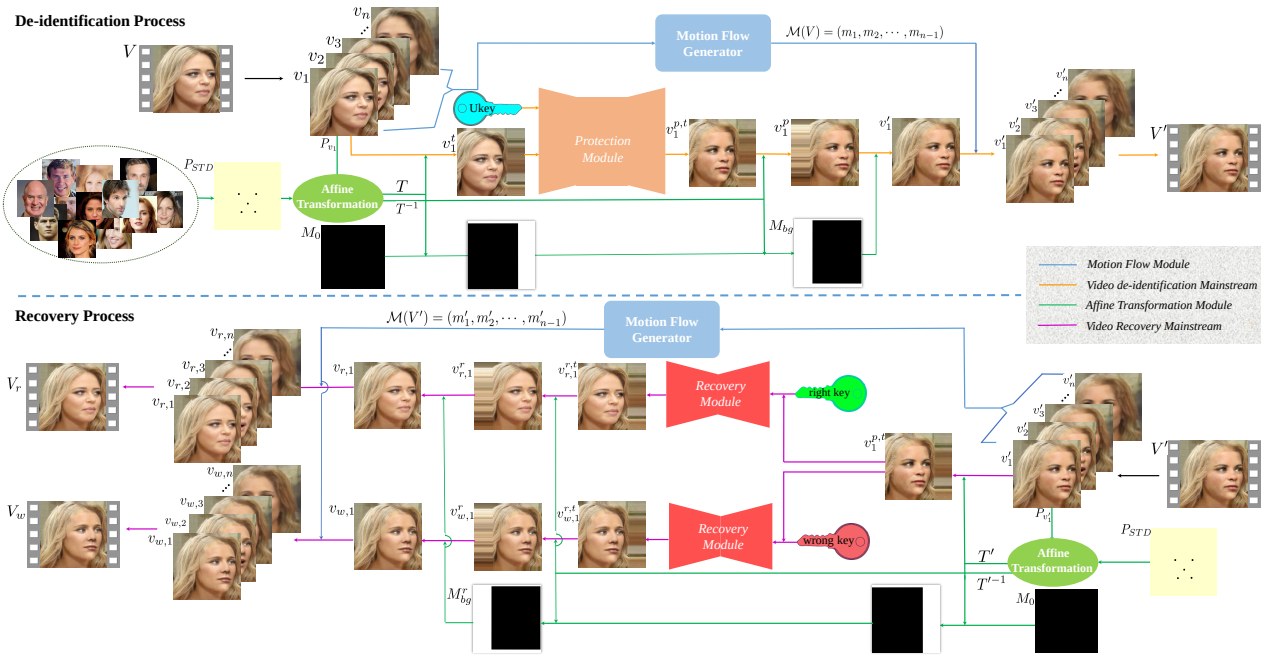


Fig. 2. The overall architecture of the proposed reversible face video de-identification method, IdentityMask. Our framework consists of two processes: the de-identification process provides a protective mask for identity information, while the recovery process removes the protective mask if and only if the right key is provided. The former relies on the *Protection Module* and the latter relies on the *Recovery Module*, both of which are guided by the crucial *Motion Flow Module*, and are assisted by the simple but reliable *Affine Transformation Module*.

degree of content-obscuring to obscure the person-identifiable contents.

Style transfer-based methods. Style transfer has also been used to do de-identification. Winkler *et al.* [31] generated an abstracted version of the security regions that showing only outlines of persons. Erdélyi *et al.* [32] presented a resource-aware cartooning privacy protection filter which converted raw images into abstracted frames where the privacy revealing details were removed. Brkić *et al.* [33] altered the appearance of the segmented pedestrians through a neural art algorithm that used the responses of a deep neural network to render the pedestrian images in a different style. [34] proposed two privacy protection schemes by using false colors on entire images. PECAM [7] converted the real-world images (domain-X) into the privacy-enhanced ones (domain-Y) through cartoon style rendering.

Identity disentanglement-based methods. Recently, the development of deep CNNs has also inspired new methods that based on identity disentanglement [35]. Li *et al.* [36] developed an encoder-decoder network architecture which could separately disentangled the facial feature representation into an appearance code and an identification code. The anonymous face was synthesized by recombining the original identity code and another appearance code from the target set to protect the individual privacy. Proença *et al.* [37] used a binary vector labelling ID, gender, ethnicity, age and hairstyle predicted by an attribute classifier to keep full control over the appearance of the anonymous faces.

Especially, among all these video surveillance de-identification methods, there exist five methods [7], [24], [27], [34], [37] that are reversible and can recover the original scene. Therefore, it is imperative to develop similar reversible de-identification technology for face videos. These five technologies focus on the accurate recording of events in supervised scenes, while little attention is paid to the generation of subtle details due to the original low resolution. In contrast, we strive to generate visual-pleasing facial details and maintain accurate facial motion.

III. Preliminaries of Problem Formulation

A reversible face video de-identification model generally can be viewed as a combination of a complex function δ and its inverse function δ^{-1}. To be more specific, the function δ maps a given face video $V = (v_1, v_2, \cdots, v_n)$ (v_i represents the i^{th} frame) to a de-identified video $V' = (v'_1, v'_2, \cdots, v'_n)$, aiming to conceal the real identity, and can be formulated as:

$$\delta(V) = V' \tag{1}$$
$$s.t. : 1 \leq i \leq n, \mathrm{ID}\{v_i\} \neq \mathrm{ID}\{v'_i\}.$$

After this, video V' can still be used normally, and when given the right key, the function δ^{-1} can restore a video $V_r = (v_{r,1}, v_{r,2}, \cdots, v_{r,n})$ with the original identity, but if the key is wrong, the function δ^{-1} restores a visual-pleasing video $V_w = (v_{w,1}, v_{w,2}, \cdots, v_{w,n})$ whose identity is different from the original video's identity. It can be formulated as follows: when the right key is given:

$$\delta^{-1}(V') = V_r \tag{2}$$
$$s.t. : 1 \leq i \leq n, \mathrm{ID}\{v_{r,i}\} = \mathrm{ID}\{v_i\};$$

and when the wrong key is given:

$$\delta^{-1}(V') = V_w \tag{3}$$
$$s.t. : 1 \leq i \leq n, \mathrm{ID}\{v_{w,i}\} \neq \mathrm{ID}\{v_i\}.$$

TABLE II
NOTATIONS

superscript	t	affined frame
	p	ID-protected frame
	′	ID-protected frame with background
	r	ID-recovered frame
subscript	r	ID right recovered frame
	w	ID wrong recovered frame
model	$\mathcal{M}$	motion flow generator
	$\mathcal{F}$	fusion network

IV. Deep Motion Flow Guided Reversible Face Video De-identification

In this section, we propose a modular architecture, called IdentityMask, to address the reversible face video de-identification problem. From the perspective of realized function, IdentityMask includes two-directional mappings: a de-identification process and a recovery process. When given an original non-protected face video V, the de-identification process aims to transform it into an identity-protected one (V'), whose identity change conditioned on the Ukey, and the recovery process aims to transform the de-identified video V' into an identity-recovered one (V_r with right key or V_w with wrong key). Fig. 2 illustrates the whole pipeline.

From the perspective of framework structure, IdentityMask consists of two main functional modules: the identity protection module *Protection Module* and the identity restoration module *Recovery Module*, both of which are guided by the vital *Motion Flow Module*. With the simple but reliable assistance of the *Affine Transformation Module*, IdentityMask efficiently achieves reversible de-identification. In the following subsections, we first introduce the four modules respectively, and then describe the entire IdentityMask pipeline.

A. Protection Module

We achieve de-identification by following the identity disentanglement pattern. As shown in Fig. 4, when given an original clean face frame v_1^t, we apply an identity encoder and an attribute encoder to extract two disentangled representations of the latent space, denoted as $r_{id}(v_1^t)$ and $r_{attr}(v_1^t)$. Among them, the identity representation r_{id} contains all the information relevant to the identity which affects face verification systems to judge whether it is the same person, and the attribute representation r_{attr} contains the rest of information carried by the image which guarantees the visual similarity (e.g. pose, expression, overall structure, background and so on). Based on this, we firstly use the Ukey as a randomness seed to generate a reference identity vector r_{refer} whose size equals to $r_{id}(v_1^t)$, which is formulated as:

$$r_{refer} = \mathcal{R}_{Ukey}, \tag{4}$$

Here the Ukey is a number that uniquely represents the user's identity. Then, a component vector $r_{\perp}(v_1^t)$ that is orthogonal to $r_{id}(v_1^t)$ in r_{refer} can be decomposed as follows:

$$r_{\perp}(v_1^t) = r_{refer} - (r_{id}(v_1^t) \cdot r_{refer}) \cdot r_{id}(v_1^t), \tag{5}$$

It allows us to create a new identity $r_{new}(v_1^t)$ by rotating $r_{id}(v_1^t)$ with a controllable parameter θ, and we denote it as:

$$r_{new} = r_{id}(v_1^t) \cdot \cos\theta + r_{\perp}(v_1^t) \cdot \sin\theta. \tag{6}$$

Finally we synthesis the de-identified face $v_1^{p,t}$ with new identity representation r_{new} and original attribute representation $r_{attr}(v_1^t)$ through a well-trained fusion network as follows:

$$v_1^{p,t} = \mathcal{F}(r_{new}, r_{attr}(v_1^t)), \tag{7}$$

B. Recovery Module

Given a de-identified face frame $v_1^{p,t}$, our *Recovery Module*, which is based on the same identity disentanglement network structure as the *Protection Module*, can restore the original frame with real identity if and only if the right key is provided. To be more specific, we firstly imply the aforementioned identity and attribute encoders to extract its identity representation $r_{id}(v_1^{p,t})$ and attribute representation $r_{attr}(v_1^{p,t})$, which has the relationship as:

$$\begin{aligned} r_{id}(v_1^{p,t}) &= r_{new}, \\ r_{attr}(v_1^{p,t}) &= r_{attr}(v_1). \end{aligned} \tag{8}$$

Then when given the right key (i.e., R_key, which we define to be equal to the Ukey), the recovered identity embedding r_{rid} can be calculated as:

$$r_{rid} = \frac{r_{id}(v_1^{p,t}) - \mathcal{R}_{R\_key} \cdot \sin\theta}{\cos\theta - A \cdot \sin\theta}, \tag{9}$$

where

$$A = \frac{\cos^2\theta - (r_{id}(v_{1,p}) - \mathcal{R}_{R\_key} \cdot \sin\theta) \cdot r_{id}(v_1^{p,t})}{\sin\theta \cdot \cos\theta}, \tag{10}$$

In fact, middle parameter A equals $r'_{id} \cdot r_{refer}$. Finally, the right recovered image with original real identity can be obtained as:

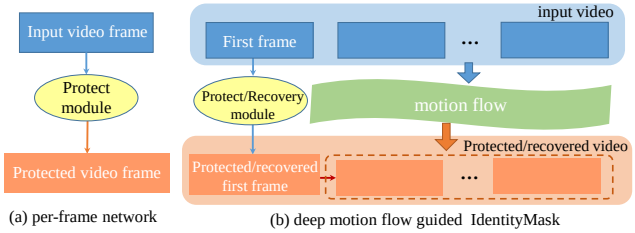
$$v_{r,1}^{r,t} = \mathcal{F}(r'_{id}, r_{attr}(v_{1,p})), \tag{11}$$

C. Motion Flow Module

The above *Protection* and *Recovery Modules* work well for images, and it is straightforward to directly apply them to videos in a frame by frame way. However, since both modules rely on the disentanglement of latent convolutional features, direct per-frame processing is time-consuming. Typically, the flow estimation and feature propagation are much faster than the computation of convolutional features [38], and consecutive face video frames are highly similar, so we exploit the similarity to reduce computational cost and achieve speedup. Specifically, either the *Protection Module* or the *Recovery Module* only processes the first frame, then we use a motion flow generator to calculate the relative motion flow of every two adjacent frames (see Fig. 3), which is denoted as:

$$\mathcal{M} = (m_1, m_2, \cdots, m_{n-1}), \tag{12}$$

where m_i ($i \leq n$) denotes the relative motion flow that can warp the processed (i.e., de-identified or recovered) i^{th} frame to the next $(i+1)^{th}$ frame.



(a) per-frame network (b) deep motion flow guided IdentityMask

Fig. 3. Illustration of (a) existing face video de-identification technologies using per-frame network generation and (b) the proposed deep motion flow guided reversible face video de-identification.

D. Affine Transformation Module

The position and pose of faces in online sharing videos vary widely, which usually differ from the "standard" frontal alignment that commonly used in large face datasets. However, it is well-known that computing deep representations by using a pre-trained CNN does have a restriction: the test image needs to lie close to the image distribution trained by the CNN. Otherwise, the latent optimization may fail to reproduce on the test image, leading to poor feature maps. Therefore, directly applying *Protection Module* or *Recovery Module* on the first frame is invalid, and we design an affine transformation, which can standardize and restore the distribution of the first frame.

To be specific, We calculate several keypoints of all faces in the training datasets of *Protection* and *Recovery Module*, compute the average and set it as the standard pattern (denoted as P_{STD}). Every time before the first frame is input to the *Protection* or *Recovery Module*, its keypoints (denoted as P_{v_1} or $P_{v_1'}$) are firstly computed in the same way. Then these keypoints are matched to the standard keypoint pattern P_{STD} with an affine transformation, which is obtained by minimizing the distortion between the two sets of points. Using this affine transformation, we warp every pixel of the input first frame face to the corresponding position of the average face. We then copy the edge color to fill the warped image into the same dimension as the input. More formally, we denote

$$T = P_{v_1}\vec{\upsilon}P_{STD}, \quad T' = P_{v_1'}\vec{\upsilon}P_{STD}, \tag{13}$$

where T represents the affine transform matrix and $\vec{\upsilon}$ denotes the affine transformation between two point patterns. Besides, we also need the inverse affine transformation to restore the original face position, and it is formulated as:

$$T^{-1} = P_{STD}\vec{\upsilon}P_{v_1}, \quad T'^{-1} = P_{STD}\vec{\upsilon}P_{v_1'}, \tag{14}$$

where T^{-1} represents the inverse transform matrix.

E. The Entire IdentityMask Pipeline

Our pipeline consists of a de-identification process and a recovery process (see Fig. 2).

The de-identification process takes the original clean video $V = (v_1, v_2, \cdots, v_n)$ as input. First of all, it is sent into the *Motion Flow Module*, where the motion flow generator generates the relative motion flow between every two adjacent frames, which is formulated as:

$$\mathcal{M}(V) = (m_1, m_2, \cdots, m_{n-1}). \tag{15}$$

Based on this, the first frame v_1 firstly enters the *Affine Transformation Module* to generate the affine transform matrix T and the inverse affine transform matrix T^{-1} (see Equ. (13) and (14)). Then the image v_1^t that lies in the "standard" distribution is obtained via:

$$v_1^t = v_1 \cdot T. \tag{16}$$

This warped first frame v_1^t is sent to the *Protection Module*, through which the real identity is concealed and a new identity r_{new} conditioned on the Ukey is generated. We denote

$$v_1^{p,t} = \mathcal{F}(r_{new}, r_{attr}(v_1^t)). \tag{17}$$

Next, the de-identified frame $v_1^{p,t}$ restores to the same layout as the original input v_1 through an inverse affine transformation:

$$v_1^p = v_1^{p,t} \cdot T^{-1}. \tag{18}$$

In order to preserve the original background, a background mask M_{bg} is generated by applying a black image M_0 whose dimension is the same as the input v_1 through two affine transformations, which is denoted as

$$M_{bg} = M_0 \cdot T \cdot T^{-1}. \tag{19}$$

With the help of M_{bg}, we can get the de-identified first frame:

$$v_1' = v_{p,1} \cdot (1 - M_{bg}) + v_1 \cdot M_{bg}. \tag{20}$$

Finally, we can obtain the entire identity-protected video $V' = (v_1', v_2', \cdots, v_n')$ on the basis of the successfully de-identified first frame v_1' and the relative motion flow $\mathcal{M}(V)$. Specifically, for $1 < i \leq n$:

$$v_i' = v_{i-1}' \circledast m_{i-1}, \tag{21}$$

where $\circledast$ denotes the inference of v_i' with the former de-identified frame v_{i-1}' and the relative motion flow m_{i-1}.

The de-identification process is summarized in Algorithm 1. The recovery process is similar except that the *Protection Module* is replaced by the *Recovery Module*, and is summarized in Algorithm 2.

V. Implementation

In this section, we introduce the promising module instantiations and their training process in more detail.

A. Identity Disentanglement Network Configuration

As mentioned in Sec. IV, both the *Protection* and the *Recovery Module* are established on the condition of identity disentanglement. Our identity disentanglement network contains an identity encoder E_{id}, an attribute encoder E_{attr} and a fusion network $\mathcal{F}$, which are pre-trained as a whole on the CelebA-HQ dataset [39].

Identity Encoder. As existing studies on face verification and recognition have made arduous efforts in finding discriminative face features for face identification, we employ a pre-trained state-of-the-art face recognition model [40] as our identity encoder. It can provide highly discriminative features for identity verification to avoid training from scratch, and has a clear geometric interpretation due to the exact correspondence to the geodesic distance on the hypersphere. Given an original face image X, the identity representation r_{id}

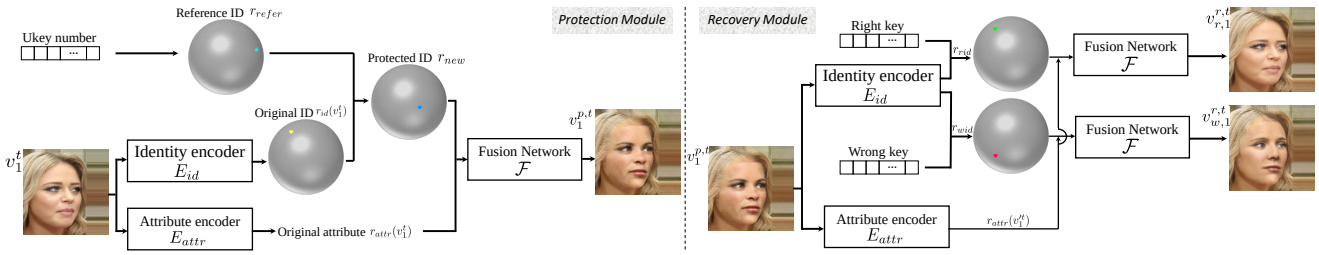


Fig. 4. The detailed architecture of the identity disentanglement network in the proposed *Protection Module* and *Recovery Module* with geometrical interpretation of identity changes. Each point on the sphere represents one normalized feature. Different colors denote different identities.

Algorithm 1 De-identification process.

Input: Original non-protected video $V = \{v_i\}_{i=1}^n$, Ukey.
Output: De-identified video $V' = \{v'_i\}_{i=1}^n$.

Generate the relative motion flow $\mathcal{M}(V) = \{m_i\}_{i=1}^{n-1}$
Generate affine transform matrixes T in Equ. (13) and T^{-1} in Equ. (14).
Generate background mask M_{bg} with black image M_0:
$$M_{bg} = M_0 \cdot T \cdot T^{-1}.$$
while $i = 1$ **do**
  $v_1^t = v_1 \cdot T$.
  Generate new identity embedding r_{new} with Ukey in Equ. (6) and de-identify the affined first frame:
  $$v_{p,1}^t = \mathcal{F}(r_{new}, r_{attr}(v_1^t)).$$
  Restore the original layout: $v_1^p = v_1^{p,t} \cdot T^{-1}$.
  Generate de-identified first frame with preserved background:
  $$v'_1 = v_1^p \cdot (1 - M_{bg}) + v_1 \cdot M_{bg}.$$
  i = i + 1.
end while
for $1 < i \leq n$ **do**
  $v'_i = v'_{i-1} \circledast m_{i-1}$.
  i = i + 1.
end for

is defined to be the last normalized feature vector before the final FC layer, which is denoted as:

$$r_{id}(X) = E_{id}(X). \tag{22}$$

It is believed that all the embedding features r_{id} are distributed around each feature centre on a normalized 512-D hypersphere [40]. Fig. 4 shows the feature distribution visualization of identity changes. Each point on the sphere represents one normalized feature. Different colors denote different identities.

Attribute Encoder. Attribute representation, which determines pose, expression, overall structure, background and so on, intuitively carries more spatial information than identity. Therefore, in order to preserve different level details, we construct a U-Net-like structure with a depth of 8, and then use the 8 feature maps generated from the U-Net decoder as the attributes representations r_{attr}. More formally, we denote

$$r_{attr}(X) = E_{attr}(X) = \left\{r_{attr}^1(X), r_{attr}^2(X), \cdots, r_{attr}^8(X)\right\}, \tag{23}$$

where $r_{att}^k(X)$ represents the *k*-th level feature map from the U-Net decoder.

Algorithm 2 Recovery process.

Input: De-identified video $V' = \{v'_i\}_{i=1}^n$, key.
Output: Right recovered video $V_r = \{v_{r,i}\}_{i=1}^n$ or wrong recovered video $V_w = \{v_{w,i}\}_{i=1}^n$.

Generate the relative motion flow $\mathcal{M}(V') = \{m'_i\}_{i=1}^{n-1}$
Generate affine transform matrixes T' in Equ. (13) and T'^{-1} in Equ. (14).
Generate background mask M_{bg}^r with black image M_0:
$$M_{bg}^r = M_0 \cdot T' \cdot T'^{-1}.$$
while $i = 1$ **do**
  $v_1^{p,t} = v'_1 \cdot T$.
  if key_is_right **then**
    Recover the right identity embedding r_{rid} with key in Equ. (9) and correctly restore the affined first frame:
    $$v_{r,1}^{r,t} = \mathcal{F}(r_{rid}, r_{attr}(v_1^{p,t})).$$
    Restore the original layout: $v_{r,1}^r = v_{r,1}^{r,t} \cdot T'^{-1}$.
    Generate right recovered first frame with preserved background:
    $$v_{r,1} = v_{r,1}^r \cdot (1 - M_{bg}^r) + v'_1 \cdot M_{bg}^r.$$
    i = i + 1.
  else
    Recover a wrong identity embedding r_{wid} with key in Equ. (9) and wrongly restore the affined first frame:
    $$v_{w,1}^{r,t} = \mathcal{F}(r_{wid}, r_{attr}(v_1^{p,t})).$$
    Restore the original layout: $v_{w,1}^r = v_{w,1}^{r,t} \cdot T'^{-1}$.
    Generate wrong recovered first frame with preserved background:
    $$v_{w,1} = v_{w,1}^r \cdot (1 - M_{bg}^r) + v'_1 \cdot M_{bg}^r.$$
    i = i + 1.
  end if
end while
for $1 < i \leq n$ **do**
  $v_{x,i} = v_{x,i-1} \circledast m'_{i-1}, x \in \{r, w\}$.
  i = i + 1.
end for

Fusion Network. The fusion network F is required to implement face reconstruction based on r_{id} and r_{attr}. Previous research [41] has verified that direct feature concatenation can easily lead to blurry results and is not expected to be used. To solve this problem, the novel Adaptive Attentional Denormalization (AAD) ResBlk [42] has been proposed to improve feature integration in multiple levels. We integrate 8 cascaded AAD ResBlks to the body of our fusion network,

TABLE III
NETWORK STRUCTURES OF IDENTITY ENCODER, ATTRIBUTE ENCODER AND FUSION NETWORK

Identity Encoder	Attribute Encoder			Fusion Network
			BilinearUpsample ×2	ConvTranspose 4×4,2,1 BN+LeakyRELU
	Conv 4×4,2,1 BN+LeakyRELU	CON	ConvTranspose 4×4,2,1 BN+LeakyRELU	AAD ResBlk(1024,1024) BilinearUpsample ×2
	Conv 4×4,2,1 BN+LeakyRELU	CON	ConvTranspose 4×4,2,1 BN+LeakyRELU	AAD ResBlk(1024,1024) BilinearUpsample ×2
	Conv 4×4,2,1 BN+LeakyRELU	CON	ConvTranspose 4×4,2,1 BN+LeakyRELU	AAD ResBlk(1024,1024) BilinearUpsample ×2
model [40]	Conv 4×4,2,1 BN+LeakyRELU	CON	ConvTranspose 4×4,2,1 BN+LeakyRELU	AAD ResBlk(1024,512) BilinearUpsample ×2
	Conv 4×4,2,1 BN+LeakyRELU	CON	ConvTranspose 4×4,2,1 BN+LeakyRELU	AAD ResBlk(512,256) BilinearUpsample ×2
	Conv 4×4,2,1 BN+LeakyRELU	CON	ConvTranspose 4×4,2,1 BN+LeakyRELU	AAD ResBlk(256,128) BilinearUpsample ×2
	Conv 4×4,2,1 BN+LeakyRELU	CON	ConvTranspose 4×4,2,1 BN+LeakyRELU	AAD ResBlk(128,64) BilinearUpsample ×2
				AAD ResBlk(64,3) BilinearUpsample ×2

Conv 4×4,2,1 represents a Convolutional Layer with kernel size 4, stride 2 and padding 1. ConvTranspose 4×4,2,1 represents a Transposed Convolutional Layer with kernel size 4, stride 2 and padding 1. CON represents feature map concatenating. AAD ResBlk (c_{in}, c_{out}) represents an AAD ResBlk with input and output channels of c_{in} and c_{out}. All LeakyRELUs have $\alpha = 0.1$.

in order to adjust the attention regions of r_{id} and r_{attr}, so that they can harmoniously participate in synthesizing different facial parts. And we can get the reconstructed face X' as:

$$X' = F(r_{id}(X), r_{attr}(X)). \tag{24}$$

The whole training process is discussed in the following.

Training Process. We use the identity consistency loss $\mathcal{L}_{id}$ to make sure the identity of the reconstructed face $\hat{X}$ still keeps the same:

$$\mathcal{L}_{id} = 1 - \frac{r_{id}(X') \cdot r_{id}(X)}{\|r_{id}(X')\|_2 \cdot \|r_{id}(X)\|_2}. \tag{25}$$

Here cosine similarity is chosen because it best fits our angular margin based Identity Encoder [40].

We also define the attributes consistency loss $\mathcal{L}_{attr}$, which can be formulated as

$$\mathcal{L}_{attr} = \frac{1}{2}\sum_{k=1}^{n} \left\| r_{attr}^{k}(X') - r_{attr}^{k}(X) \right\|_2^2. \tag{26}$$

This loss function has been proved to encourage the generated images to be perceptually similar (but not identical) to the target image [43]. We tried other methods ($\mathcal{L}_1$ distance, Huber loss, and cosine similarity) to measure attributes similarity, however, $\mathcal{L}_2$ distance performs best.

If the restored result X' is generated with the same r_{id} and r_{attr}, it should be as similar to the original image as possible. We set pixel-level $\mathcal{L}_2$ distance as the reconstruction loss:

$$\mathcal{L}_{rec} = \frac{1}{2} \left\| X' - X \right\|_2^2. \tag{27}$$

We take advantage of adversarial learning to train the framework and introduce the adversarial loss $\mathcal{L}_{adv}$ to constrain the generated results indistinguishable from real images. To promote the image quality, it is necessary to expand the perception range of the discriminator, so we adopt m multi-scale discriminators [44] with hinge loss functions for different resolution versions of the generated image.

$$\mathcal{L}_{adv}(X'_m, X_m) = log(D(X_m)) + log(1 - D(X'_m)). \tag{28}$$

where X_m indicates the low-resolution image after m-th downsampling.

The total loss function is the weighted sum of the above losses, which can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda_{attr}\mathcal{L}_{attr} + \lambda_{id}\mathcal{L}_{id} + \lambda_{rec}\mathcal{L}_{rec}, \tag{29}$$

where λ_{att}, λ_{id} and λ_{rec} are the weight parameters for balancing different terms.

In the training process, we use the Adam optimizer [45] with momentum parameters $\beta_1 = 0, \beta_2 = 0.999$. The learning rate is set to 4×10^{-4}. The parameters in Eq. (29) are set to $\lambda_{att} = \lambda_{rec} = 10, \lambda_{id} = 5$.

B. Other Implementation Details

In the *Motion Flow Module*, we employ a pre-trained CNN [46] as our motion flow generator to model the relative dense motion flow. In the *Affine Transformation Module*, we calculate the 5 keypoints (left/right eye, leftmost/rightmost tip of the mouth, and nose) of all faces in CelebA-HQ dataset by [47], and compute the average as the standard point pattern. Then the Umeyama algorithm [48] is utilized to calculate the affine transform matrixes between two point patterns.

VI. EXPERIMENTS

A. Experimental Setup

Dataset. We choose the VoxCeleb dataset [49], which contains 22496 videos extracted from YouTube, to demonstrate the effectiveness of our reversible face video de-identification method. After preprocessing like [46], we obtain 12775 videos with lengths varying from 64 to 1024 frames, which are resized to 256×256 preserving the aspect ratio. For simplicity, we use the ID number annotated in the dataset as Ukey and define the right key as a number equal to the Ukey, while a random number other than the Ukey is generated as the wrong key.

Comparison methods. To validate the effectiveness of the proposed IdentityMask, we compare to three state-of-the-art methods: ACTION [4], LIVE [5] and CIAGAN [6].

Evaluation Metrics. We evaluate the proposed IdentityMask in terms of two metrics, as described below.

(1) Privacy metrics. We measure the cosine similarity of embedding vectors from the generated and original face extracted by pre-trained face recognition model, denoted as **CSIM**, to evaluate the quality of identity protection and restoration. For a fair comparison, we employ the well-known FaceNet identification model [50], which is excluded from our training model and pre-trained on two public datasets (CASIA-Webface [51] and VGGFace2 [52]) respectively.

(2) Utility metrics. With today's advanced technology, ensuring that the faces in a synthesised video can still be detected is

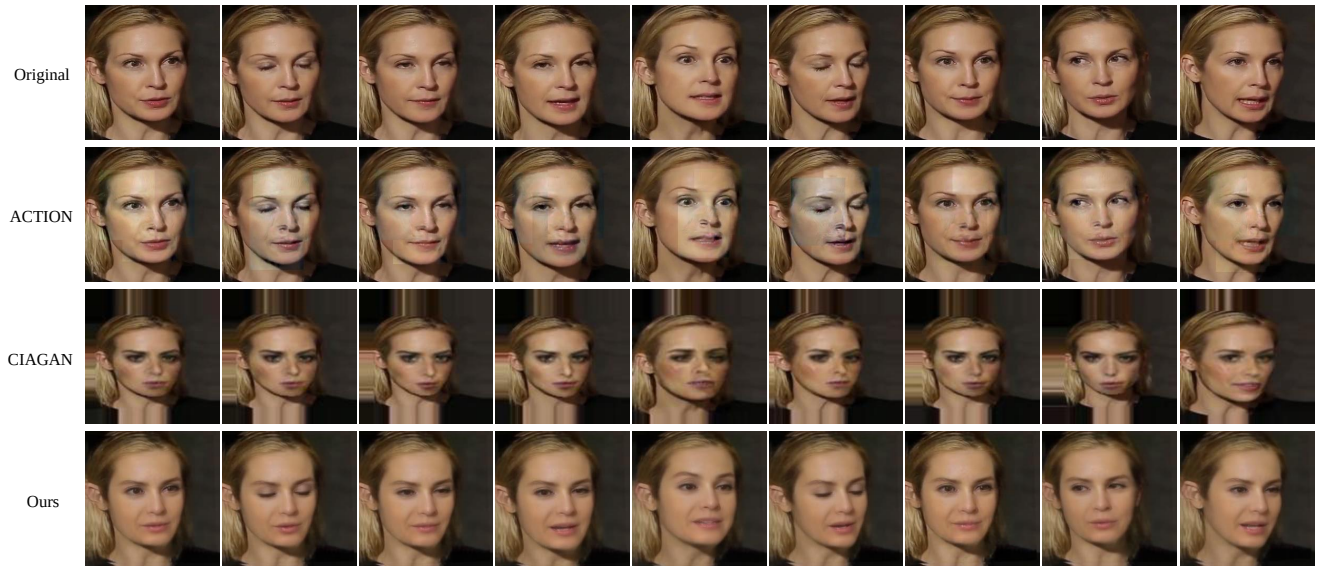


Fig. 5. Comparison of the de-identified faces between ACTION [4], CIAGAN [6] and our method on the VoxCeleb dataset.Videos are available at supplementary material.

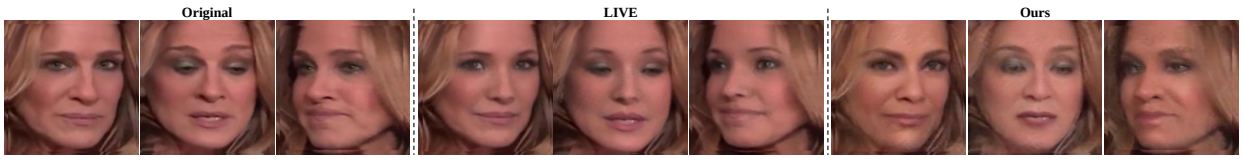


Fig. 6. Comparison of the de-identified faces between LIVE [5] and our method.

very trivial [53]. Therefore, instead of using the face detection rate, we borrow several metrics which have been commonly used in face swapping and face reenactment tasks. They are designed exactly for videos to evaluate the utility performance. Specifically, the $\mathcal{L}_2$ distances between pose and expression vectors from the generated and original face extracted by an open-sourced pose estimator [54] and a 3D facial model [55] are calculated as pose (denoted as **POSE**) and expression (denoted as **EXP**) similarity. The **FID** score is chosen to evaluate the generation quality as it can measure the distance between the generated distribution and the real distribution. In addition, we evaluate whether the motion of the input video is preserved by computing the average distance of facial landmark keypoints [56] from the generated and original face, which is denoted as **AKD**.

Unless otherwise specified, each metric is calculated independently for each frame.

B. Comparison in De-identification

In this subsection, we compare our IdentityMask with state-of-the-art face de-identification methods.

The qualitative comparison with ACTION [4] and CIAGAN [6] is shown in Fig. 5, while the quantitative results are shown in Table IV. It can be seen that the faces generated by ACTION is too visually similar to the original faces, which makes it easy for people to think that they are still the same person, thus does not realize the identity protection from human beings. Besides, incomprehensible artifacts and blurs with light or dark bounding boxes often occur, resulting in obvious video jitter. This makes it difficult to share the generated videos online. In addition, the crucial CSIM value is high, which implies that ACTION is vulnerable to the identification of advanced face verification model.

TABLE IV
QUANTITATIVE COMPARISONS OF IDENTITY PROTECTION ON VOXCELEB. THE BEST RESULTS ARE IN BOLD. ↑ MEANS HIGHER IS BETTER, AND ↓ MEANS LOWER IS BETTER

Method	CSIM↓		POSE↓	EXP↓	FID↓	AKD↓
	CASIA	VGGFace2				
ACTION	0.904	0.869	**2.45**	2.69	19.34	1.60
CIAGAN	0.520	0.507	14.69	8.23	31.50	4.16
Ours	**0.518**	**0.503**	**2.45**	**2.64**	**18.70**	**1.58**

We can see that the frames generated by CIAGAN can maintain some basic face attributes as well as the rough head orientation, but most of which are not visually similar to the original. These de-identified faces can effectively hide the true identity information from both human eyes and machines. However, distortions and artifacts often occur. When the characters perform large poses or expressions, there will even be large deformation. These are very unfavorable to the video. As can be seen from Table IV, its utility metrics deteriorate significantly, which will make the synthesised videos hard to meet the requirement for online sharing.

In contrast, our method produces more natural looking images which achieve a great advantage in visual similarity to the input frame with visually perceptible changes, and enables

Fig. 7. Qualitative results of our method about identity protection and identity recovery on the VoxCeleb dataset. Videos are available at supplementary material.

Fig. 8. Qualitative results of identity recovery when given multiple wrong keys. The black background indicates the original videos, and the red background indicates the wrong-recovered videos.

TABLE V
QUANTITATIVE EVALUATION WITH STATE-OF-THE-ART METHODS ON LFW DATASETS

Method	True Positive Rate↓	
	CASIA	VGGFace2
Original	0.965 ± 0.016	0.986 ± 0.010
ACTION	0.696 ± 0.015	0.714 ± 0.014
LIVE	0.035 ± 0.011	0.038 ± 0.015
CIAGAN	0.019 ± 0.008	0.034 ± 0.015
Ours	**0.017 ± 0.011**	**0.026 ± 0.014**

TABLE VI
QUANTITATIVE RESULTS OF IDENTITY RECOVERY ON VOXCELEB.

Method	CSIM		POSE↓	EXP↓	FID↓	AKD↓
	CASIA	VGGFace2				
R_key	0.961	0.959	1.62	1.59	8.50	1.34
W_key	0.475	0.461	2.75	2.96	23.18	1.61

de-identification for both human beings and machines. Furthermore, from Table IV, the lowest CSIM value indicates that our method is superior to the compared method in protecting the real identity. Meanwhile, the best performance under utility metrics shows that our method also well preserves the non-identity aspects of the original frame, i.e., pose, expression, facial motion, and overall structure. So we can best ensure the subsequent normal use of the de-identified faces.

A comparison with the work of LIVE [5] is given in Fig. 6. Our results are at least visually as good as the original ones, despite having to run on the cropped faces extracted from the paper PDF.

To make the comparison more convincing and fairer, we follow the evaluation protocol that has been used in [5] and [6], which is conducted on the LFW benchmark. Specifically, two FaceNet identification models (pre-trained on CASIA-Webface and VGGFace2 respectively) are employed and the main evaluation metric is the true acceptance rate. Table V presents the results on de-identified LFW image pairs for a given person, while the de-identification method is applied to the second image of each pair. It can be seen that all methods can significantly reduce the true positive rate. In particular, our method achieves the best privacy protection.

C. Analysis in Identity Recovery

In this subsection, we evaluate our performance in identity restoration. The effect of one original video being de-identified and recovered respectively with the right and wrong key (denoted as "R_key" and "W_key") is presented in Fig. 7. It can be seen that the identity-protected frames obtain a new identity, while still maintain a high visual similarity (i.e., appearance, pose, expression, and facial motion), which ensures the rationality of subsequent use. Then the right key can restore a video which is exactly similar to the original video with the real identity, while the wrong key can restore a realistic video with another new identity different from the original identity. Moreover, each wrong key maps to a unique identity. In this way, we provide security via ambiguity: even if a privacy intruder guesses the correct key, it is extremely difficult to know that without having access to any other identity revealing meta-data, since each key—regardless of whether it is correct or not—always leads to a different realistic identity. In particular, the effect of being recovered by multiple wrong keys is shown in Fig. 8.

The quantitative results of identity recovery are shown in Table VI. It shows that after de-identification: 1) the original

Fig. 9. Quantitative comparisons of the influence of different *Motion Flow Module*. The first row is the original frames. The rest rows demonstrate results when using *STD* [46], *AVD* [57] and *RLT* [46]. Videos are available at supplementary material.

TABLE VII
QUANTITATIVE EXPERIMENTAL RESULTS OF RIGHT RECOVERY QUALITY ON VOXCELEB

	LPIPS↓	PSNR↑	SSIM↑	MAE↓
R_key	0.077	25.492	0.875	0.036

TABLE VIII
QUANTITATIVE EVALUATION OF THE MOTION FLOW MODULE

Method	CSIM↓		POSE↓	EXP↓	FID↓	AKD↓
	CASIA	VGGFace2				
STD	0.522	0.513	**2.45**	**2.62**	20.96	1.59
AVD	0.520	0.507	2.53	2.78	22.11	1.61
RLT(ours)	**0.518**	**0.503**	**2.45**	2.64	**18.70**	**1.58**

TABLE IX
ABLATION STUDY OF THE PROPOSED IDENTITYMASK PIPELINE

Method	CSIM↓		POSE↓	EXP↓	FID↓	AKD↓
	CASIA	VGGFace2				
w/o AT	**0.376**	**0.368**	22.58	10.27	43.04	3.96
w/o MF	0.513	0.492	2.50	2.71	20.47	2.24
Ours	0.518	0.503	**2.45**	**2.64**	**18.70**	**1.58**

identity can be recovered excellently with the correct key, which is conducive to the supervision of network abnormal events; 2) when given the wrong key, it is almost impossible to restore the original identity; 3) whether the video is recovered by the right or wrong key, its utility is always impressive.

To better evaluate the right recovery quality, we apply LPIPS (Learned perceptual image patch similarity) distance [58] to measure perceptual similarity, PSNR (Peak signal-to-noise ratio) [59] and MAE (Mean absolute error) to measure distortion at the pixel level, and SSIM (Structural similarity) [60] to measure the structure similarity. The results in Table VII demonstrate that the right recovered frames are extremely similar to the original frames, which is consistent with the intuitive expectation. To the best of our knowledge, IdentityMask is the first work to achieve de-identified face video restoration, so the above results are summarized as the baseline for future research.

D. Model Analysis and Discussions

In this subsection, considering the de-identification process $V \rightarrow V'$ and the recovery process $V' \rightarrow V$ are symmetrical, while previous comparison methods can only do de-identification, we take the former as the example.

1) *Motion Flow Module Selection*. We pick and compare three state-of-the-art motion flow modeling methods: *STD* [46], *RLT* [46] and *AVD* [57]. *STD* computes the deep motion flow between input and output video frame by frame. *RLT* calculates the deep motion flow between every two adjacent frames of the input video first and then applies this relative dense motion flow to the first frame of the output video. *AVD* also computes the deep motion flow between input and output video frame by frame, except it disentangles the shape and pose of objects in the region space and forces decoupling of foreground from background. Different motion flow modeling methods are suitable for different application scenarios. *STD* directly transfers object shape from the input video into the generated video, while the *RLT* requires that objects be in the same pose in the first frame of the input and output video, and the *AVD* is designed specifically for videos of articulated objects. Since the face is exactly in the same pose in the first frame of the input and synthesized video in either de-identification process or recovery process, the *RLT* is theoretically the best motion flow modeling method for IdentityMask. Fig. 9 and Table VIII reveal the qualitative and quantitative results of the influence of the *Motion Flow Module*. We can see that if *STD* is used, the synthesized face can retain a slightly more similar expression to the original face than using *RLT*. However, the transfer of the original face shape leads to a decline of privacy protection ability, while lower FID and AKD values also indicate poorer generation quality and motion flow modeling. In addition, the background of the generated face has severe distortion. If *AVD* is used, not only the eyes and mouth have obvious distortion, but also all the quantitative metrics are worse than employing *RLT*. Therefore, *RLT* is the best choice of our *Motion Flow Module*.

2) *Ablation Study*. We take two variants of the proposed

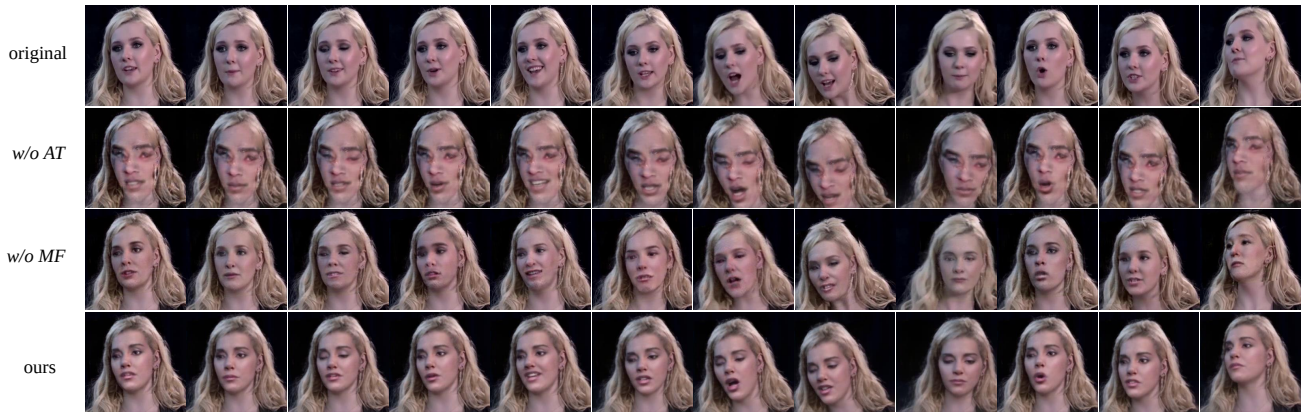


Fig. 10. Ablation study of our method. The first row is the original frames, the second row to the fourth row shows the corresponding de-identified results of *w/o AT* (the model without the *Affine Transformation Module*), *w/o MF* (the model without the *Motion Flow Module*) and the full model.

Fig. 11. De-identified results with variant parameter θ values.

IdentityMask pipeline for ablation study in order to validate effectiveness of *Affine Transformation Module* and *Motion Flow Module*. Specifically, *w/o AT* indicates the variant without the *Affine Transformation Module*, and *w/o MF* indicates the variant without the *Motion Flow Module*, which means that the input videos have to be processed frame by frame. We let "ours" indicate the full model. Fig. 10 shows the qualitative results and Table IX presents the quantitative comparison. It can be seen that *w/o AT* generates a very casual face contour and results in a substantial decline in data utility. This is unacceptable for media users. As for *w/o MF*, although it can protect identity slightly better than the full model, its pose and expression are less similar to the original video. Also, its image quality is poor, especially the preservation of facial movements. These will render the synthesized video unfavorable for subsequent identity-agnostic use. Therefore, each module in our method is indispensable and only the full model can achieve the most wonderful de-identified effects without affecting the subsequent identity-agnostic use.

3) *Parameter Selection*. In this subsection, the performance variation of de-identification with respect to the controllable parameter θ is studied. We conduct a group of identity protection experiments with respect to the parameter θ. During the test, we randomly select 1000 videos from the VoxCeleb dataset, and change θ from 0 to 90 for each video to synthesise the corresponding de-identified videos. Fig. 11 shows the qualitative results. It can be observed that with the increase

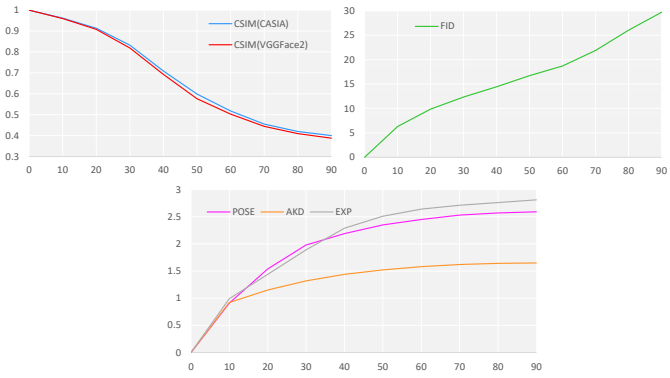


Fig. 12. The performance variation of de-identification with respect to the parameter θ. The x-axis indicates θ value and the y-axis indicates the metric values.

of θ, the visual identity difference between the synthetic faces and the original faces expands, while the identity-independent attributes are still maintained. Here, both the privacy metrics and the utility metrics are used to evaluate the overall identity protection effect, and are shown in Fig. 12. It can be seen that the degree of identity protection can be adjusted, accompanied by utility variations. Considering the identity protection effect and the utility performance comprehensively, we set θ to 60 for all other experiments.

4) *Computational Overhead Analysis*. We explore the contribution of the *Motion Flow Module* to saving computa-

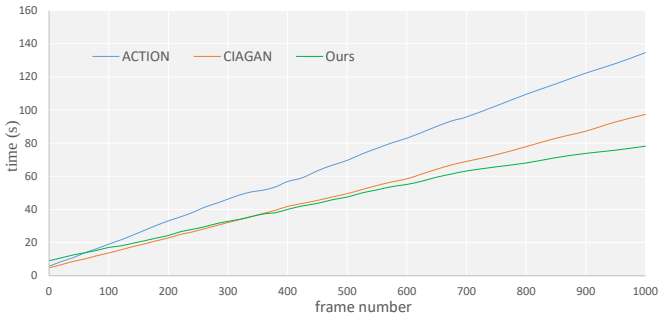


Fig. 13. Comparison of computational overheads between ACTION [4], CIAGAN [6] and our method on the VoxCeleb dataset.

tional overheads in de-identification tasks. We compare with ACTION and CIAGAN on an NVIDIA GTX 1080 Ti, and the results are shown in Fig. 13. We observe that when the number of video frames is greater than 80, our method has a lower computational complexity than ACTION, and when the number of video frames is greater than 360, our method is less computationally complex than CIAGAN. Since the complexity of per frame processing is almost linear with the number of frames, this advantage becomes more obvious as the number of video frames increases. It demonstrates the superiority of motion flow guided evaluation over per frame processing.

TABLE X
SECURITY ANALYSIS OF THE PROPOSED IDENTITYMASK

Method	CViT [61]	LRNet [62]
Fake video detection rate	74.8%	63.1%

5) *Security Analysis*. Previous experimental results have shown that IdentityMask can generate realistic identity-protected videos. However, we are worried about the potential misuse. Once abused, even if the authority can obtain the real identity through the recovery process, unpleasant effects (such as fraud) in the dissemination process may have occurred. Therefore, we apply two advanced deepfake detection models, CViT [61] and LRNet [62], to examine the security of IdentityMask. We calculate the proportion of de-identified videos that are judged as fake, and name it as the **fake video detection rate**. As shown in Table X, the probability of the synthetic videos being judged as "deepfake" is relatively high, which proves that our identity protection technology has good security despite its state-of-the-art utility performance.

VII. CONCLUSIONS

In this paper, we have proposed a reversible face video de-identification framework, IdentityMask, guided by deep motion flow. Our framework consists of a de-identification process and a recovery process. The former is able to conceal the real identity with a visually similar appearance in a seamless way, and the latter aims to recover the original identity only when given the right key. The proposed framework is the first one suitable for reversible face video de-identification. It presents a quality that surpasses the literature methods in the de-identification task, and is impressive in the identity recovery process. Besides, instead of existing per-frame processing, we take advantage of motion flow to guide consecutive frames generation, which alleviates the computational overhead and improves the synthesis effect. Extensive experimental results on a standard diverse dataset verify the effectiveness and efficiency of our framework.

While our reversible face video de-identification results are visibly convincing, additional improvements are possible. As part of our future work, we plan to elaborate the mapping function between Ukey and right key to further enhance the security of identity protection.

ACKNOWLEDGMENTS

This work was supported by MoE-China Mobile Research Fund Project (MCM20180702), the 111 Project (B07022 and Sheitc No.150633) and the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

REFERENCES

[1] J. Zhou, C.-M. Pun, and Y. Tong, "Privacy-sensitive objects pixelation for live video streaming," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3025–3033, 2020.

[2] S. Jia, X. Li, C. Hu, G. Guo, and Z. Xu, "3d face anti-spoofing with factorized bilinear coding," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[3] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–36, 2021.

[4] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," in *Proceedings of the european conference on computer vision (ECCV)*, pp. 620–636, 2018.

[5] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9378–9387, 2019.

[6] M. Maximov, I. Elezi, and L. Leal-Taixé, "Ciagan: Conditional identity anonymization generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5447–5456, 2020.

[7] H. Wu, X. Tian, M. Li, Y. Liu, G. Ananthanarayanan, F. Xu, and S. Zhong, "Pecam: privacy-enhanced video streaming and analytics via securely-reversible transformation," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pp. 229–241, 2021.

[8] R. Gross, L. Sweeney, J. Cohn, F. De la Torre, and S. Baker, "Face de-identification," in *Protecting privacy in video surveillance*, pp. 129–146, Springer, 2009.

[9] B. Samarzija and S. Ribaric, "An approach to the de-identification of faces in different poses," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1246–1251, IEEE, 2014.

[10] B. Meden, R. C. Mallı, S. Fabijan, H. K. Ekenel, V. Štruc, and P. Peer, "Face deidentification with generative deep neural networks," *IET Signal Processing*, vol. 11, no. 9, pp. 1046–1054, 2017.

[11] Y. Li and S. Lyu, "De-identification without losing faces," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, pp. 83–88, 2019.

[12] B. Zhu, H. Fang, Y. Sui, and L. Li, "Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 414–420, 2020.

[13] Y. Wen, B. Liu, R. Xie, J. Cao, and L. Song, "Deep motion flow aided face video de-identification," in *2021 IEEE International Conference on Visual Communications and Image Processing (VCIP)*.

[14] J. Cao, B. Liu, Y. Wen, R. Xie, and L. Song, "Personalized and invertible face de-identification by disentangled identity information manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3334–3342, 2021.

[15] W. Sun, J. Zhou, Y. Li, M. Cheung, and J. She, "Robust high-capacity watermarking over online social network shared images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1208–1221, 2020.

[16] L. Wu, Y. Wang, H. Yin, M. Wang, and L. Shao, "Few-shot deep adversarial learning for video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 1233–1245, 2019.

[17] L. Wu, R. Hong, Y. Wang, and M. Wang, "Cross-entropy adversarial view adaptation for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2081–2092, 2019.

[18] F. Dufaux and T. Ebrahimi, "Scrambling for privacy protection in video surveillance systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1168–1174, 2008.

[19] J. Schiff, M. Meingast, D. K. Mulligan, S. Sastry, and K. Goldberg, "Respectful cameras: Detecting visual markers in real-time to address privacy concerns," in *Protecting Privacy in Video Surveillance*, pp. 65–89, Springer, 2009.

[20] D. Chen, Y. Chang, R. Yan, and J. Yang, "Protecting personal identification in video," in *Protecting Privacy in Video Surveillance*, pp. 115–128, Springer, 2009.

[21] P. Agrawal and P. Narayanan, "Person de-identification in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 299–310, 2011.

[22] M. Mrityunjay and P. Narayanan, "The de-identification camera," in *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2011 Third National Conference on*, pp. 192–195, 2011.

[23] M. Ivasic-Kos, A. Iosifidis, A. Tefas, and I. Pitas, "Person de-identification in activity videos," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1294–1299, IEEE, 2014.

[24] M. Blažević, K. Brkić, and T. Hrkać, "Towards reversible de-identification in video sequences using 3d avatars and steganography," *arXiv preprint arXiv:1510.04861*, 2015.

[25] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[26] E. Flouty, O. Zisimopoulos, and D. Stoyanov, "Faceoff: anonymizing videos in the operating rooms," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pp. 30–38, Springer, 2018.

[27] M. Yamaç, M. Ahishali, N. Passalis, J. Raitoharju, B. Sankur, and M. Gabbouj, "Reversible privacy preservation using multi-level encryption and compressive sensing," in *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, IEEE, 2019.

[28] M. U. Kim, H. Lee, H. J. Yang, and M. S. Ryoo, "Privacy-preserving robot vision with anonymized faces by extreme low resolution," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 462–467, IEEE, 2019.

[29] Z. W. Wang, V. Vineet, F. Pittaluga, S. N. Sinha, O. Cossairt, and S. Bing Kang, "Privacy-preserving action recognition using coded aperture videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[30] N. A. Tu, K.-S. Wong, M. F. Demirci, Y.-K. Lee, *et al.*, "Toward efficient and intelligent video analytics with visual privacy protection for large-scale surveillance," *The Journal of Supercomputing*, pp. 1–31, 2021.

[31] T. Winkler and B. Rinner, "Trustcam: Security and privacy-protection for an embedded smart camera based on trusted computing," in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 593–600, IEEE, 2010.

[32] A. Erdélyi, T. Barát, P. Valet, T. Winkler, and B. Rinner, "Adaptive cartooning for privacy protection in camera networks," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 44–49, IEEE, 2014.

[33] K. Brkić, T. Hrkać, and Z. Kalafatić, "Protecting the privacy of humans in video sequences using a computer vision-based de-identification pipeline," *Expert Systems with Applications*, vol. 87, pp. 41–55, 2017.

[34] S. Çiftçi, A. O. Akyüz, and T. Ebrahimi, "A reliable and reversible image privacy protection based on false colors," *IEEE transactions on Multimedia*, vol. 20, no. 1, pp. 68–81, 2017.

[35] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng, "Coupled bilinear discriminant projection for cross-view gait recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 734–747, 2019.

[36] J. Li, L. Han, H. Zhang, X. Han, J. Ge, and X. Cao, "Learning disentangled representations for identity preserving surveillance face camouflage," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 9748–9755, IEEE, 2021.

[37] H. Proença, "The uu-net: Reversible face de-identification for visual surveillance video footage," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 496–509, 2021.

[38] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2349–2358, 2017.

[39] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[40] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

[41] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards open-set identity preserving face synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6713–6722, 2018.

[42] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.

[43] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, pp. 694–711, Springer, 2016.

[44] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems*, vol. 32, pp. 7137–7147, 2019.

[47] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[48] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Computer Architecture Letters*, vol. 13, no. 04, pp. 376–380, 1991.

[49] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[50] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[51] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[52] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74, IEEE, 2018.

[53] H. Wu, G. Liu, Y. Yao, and X. Zhang, "Watermarking neural networks with watermarked images," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[54] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 2074–2083, 2018.

[55] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[56] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1021–1030, 2017.

[57] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13653–13662, 2021.

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[59] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[60] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[61] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126*, 2021.

[62] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia, "Improving the efficiency and robustness of deepfakes detection through precise geometric features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3609–3618, 2021.