The UU-Net: Reversible Face De-Identification for Visual Surveillance Video Footage

Hugo Proença, *Senior Member, IEEE*

***Abstract*—We propose a reversible face de-identification method for video surveillance data, where landmark-based techniques cannot be reliably used. Our solution generates a photorealistic de-identified stream that meets the data protection regulations and can be publicly released under minimal privacy concerns. Notably, such stream still encapsulates the information required to later reconstruct the original scene, which is useful for scenarios, such as crime investigation, where subjects identification is of most importance. Our learning process jointly optimizes two main components: 1) a *public* module, that receives the raw data and generates the de-identified stream; and 2) a *private* module, designed for security authorities, that receives the public stream and reconstructs the original data, disclosing the actual IDs of the subjects in a scene. The proposed solution is landmarks-free and uses a conditional generative adversarial network to obtain synthetic faces that preserve pose, lighting, background information and even facial expressions. Also, we keep full control over the set of soft facial attributes to be preserved/changed between the raw/de-identified data, which extends the range of applications for the proposed solution. Our experiments were conducted in three visual surveillance datasets (BIODI, MARS and P-DESTRE) plus one video face data set (YouTube Faces), showing highly encouraging results. The source code is available at https://github.com/hugomcp/uu-net.**

***Index Terms*—Visual Surveillance, Video Processing, Anonymization, Privacy, Security and Forensics.**

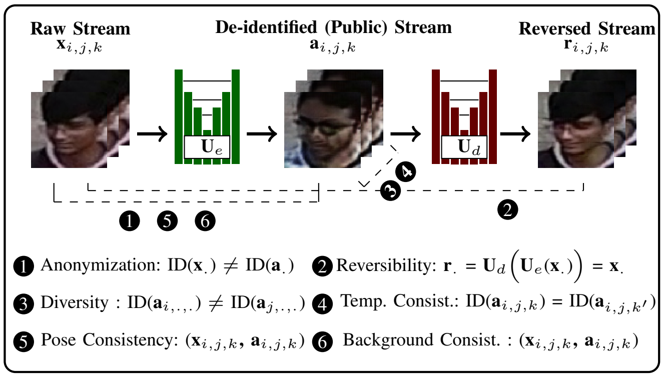
Raw Stream
$\mathbf{x}_{i,j,k}$
De-identified (Public) Stream
$\mathbf{a}_{i,j,k}$
Reversed Stream
$\mathbf{r}_{i,j,k}$
$\mathbf{U}_e$
$\mathbf{U}_d$
1
5
6
3
4
2
1 Anonymization: $\mathrm{ID}(\mathbf{x}_.) \neq \mathrm{ID}(\mathbf{a}_.)$
2 Reversibility: $\mathbf{r}_. = \mathbf{U}_d\Big(\mathbf{U}_e(\mathbf{x}_.)\Big) = \mathbf{x}_.$
3 Diversity : $\mathrm{ID}(\mathbf{a}_{i,.,.}) \neq \mathrm{ID}(\mathbf{a}_{j,.,.})$
4 Temp. Consist.: $\mathrm{ID}(\mathbf{a}_{i,j,k}) = \mathrm{ID}(\mathbf{a}_{i,j,k'})$
5 Pose Consistency: $(\mathbf{x}_{i,j,k}, \mathbf{a}_{i,j,k})$
6 Background Consist. : $(\mathbf{x}_{i,j,k}, \mathbf{a}_{i,j,k})$

Fig. 1. Key properties of a reversible video de-identifier: the identity of every face in the public stream should be obfuscated (1), while keeping pose (5) and background (6) information, to assure seamless transitions and photorealism. Also, the de-identified faces should be diverse among identities (3) and consistent (4) across the frames of a sequence. Finally, it should be possible to reconstruct the original data (2) exclusively based in the public stream.

I. Introduction

Video-based surveillance regards *the act of watching a person or a place, esp. a person believed to be involved with criminal activity or a place where criminals gather*[1]. While this kind of technologies has been sustaining the growth of social monitoring and control tools, it also hosts crime prevention measures throughout the world, raising debates about proper solutions that balance security/privacy issues [50].

Human re-identification has been extensively studied and solutions are becoming more effective. Zhu *et al.*[71] addressed the problem of matching images of different quality, proposing a classification-verification-classification strategy, where models are trained iteratively: at first for multi-class classification and then in a verification setting. Qi [39] described a loss-sensitive GAN model, training a model to distinguish between real/fake samples, while also learning a generator that produces realistic samples. Upon Lipschitz regularisation, this technique is claimed to better generalize than classic GANs. Qi *et al.* [38] described a localised GAN that uses local coordinates to create manifold geometry features. The locality nature of these models enables generators to directly access the local manifold geometry, alleviating the possibility of mode collapse. Observing that handcrafted-based methods tend to lose spatial information during the encoding phase, Tao *et al.* [55] considered images as two-order tensors from where a low-dimensional tensor-subspace is obtained, keeping information of the image structure. Recently, Wu *et al.* [57] evaluated empirically the effectiveness of metric learning solutions in handcrafted/deep-learning representations, which have been closely tied to re-id problems. Assuming the difficulties to obtain ground-truth labels in this kind of problems, Tao *et al.* [56] proposed a weighted majority voting procedure for crowdsource annotations, based in the expertise of annotators and domain adaptation concepts.

As a response to the increasing effectiveness of re-id methods, anonymising publicly-recorded video streams is seen as a solution to privacy concerns, to comply with data protection regulations like GDPR[2] and CCPA[3]. In this context, the earliest de-identification techniques obfuscated privacy-sensitive information by low-level image processing operations, such as downsampling, blurring or masking (e.g., [5] and [35]), but also destroying privacy-insensitive information and decreasing photorealism. Recently, more sophisticated techniques were proposed, based in active appearance models (AAMs) and

H. Proença is with the IT: Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, Portugal, E-mail: hugomcp@di.ubi.pt

Manuscript received ? ?, 2020; revised ? ?, ?.

[1] https://dictionary.cambridge.org/dictionary/english/surveillance

[2] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[3] https://privacyrights.org/resources/california-consumer-privacy-act-basics

facial landmarks (e.g., [16], [31] and [47]). In particular, conditional Generative Adversarial Networks (cGANs) are a popular choice to control the appearance of synthesised data, being used in cross-domain image synthesis, text-to-image translation and fashion synthesis (e.g., [25], [41], [65] and [69]).

As illustrated in Fig. 1, reversible video de-identification is a challenging task. Not only the original stream needs to be seamless modified, keeping concerns about distortions and other artefacts, but also the IDs must be obfuscated in a visually pleasant way, while considering constraints such as background information, pose and lighting conditions.

In this work, we consider the *face* as the most sensitive identifier in public data. We propose a reversible face de-identification solution for video data based in a two-phase adversarial learning process. In inference time, our model is decomposed into two disjoint parts: 1) an *encoder* $\mathbf{U}_e$ that receives the raw data and generates their de-identified version, ensuring IDs obfuscation and temporal consistency, while preserving pose, lighting, background information and facial expressions. This yields the *public* stream, usable for analytics tools and social media; and 2) a *decoder* $\mathbf{U}_d$, available only to authorities, that reconstructs the original scene exclusively based in the public stream. Note that neither the original stream nor any sensitive meta-information are stored or transmitted over the network, assuring the individuals' right to privacy (for public exposure purposes), while still enabling to disclosure the actual IDs in a crime scene.

The root of our solution is a cGAN composed of two sequential U-shaped models [44] (hence UU-Net), used respectively for de-identification/reconstruction purposes. At first, a multi-label CNN classifier is inferred, to estimate the agreeing/disagreeing labels between image pairs (ID, *gender*, *ethnicity*, *hairstyle* and *age*). This model is used in the second learning phase, which works under the adversarial paradigm: a generator attempts to fool a *PatchGAN* [25] discriminator, that is responsible to distinguish between the raw, anonymised and reconstructed faces. Depending of the weights given to each component of the pairwise discriminator responses, we keep full control over the appearance of the anonymised faces, and determine the labels that should agree/disagree between the raw/de-identified faces.

Considering that our method was designed to work in surveillance data of relatively poor resolution, we kept it *landmarks-free* and independent of any face alignment step based in fiducial points. Instead, its unique pre-requisite is a face detector (e.g., [42] or [13]). In inference, once the generator $\mathbf{U}_e$ creates the anonymised faces, we use image steganography [14] to seamlessly hide information of the bounding boxes in the public stream. This is important for reconstruction purposes, to define the regions-of-interest reversed by the decoder model $\mathbf{U}_d$. In summary, we provide the following contributions:

- we propose a two-stage learning process and an architecture to de-identify sequences of facial images in visual surveillance video streams;
- based on the responses provided by an image pairwise analyzer, we offer full control over the labels that should agree/disagree between the original/de-identified data;
- using image steganography, we encapsulate the anonymised faces and the corresponding regions-of-interests in the public video stream, which can be released without compromising the individuals' right to privacy;
- using exclusively the publicly available data, the second part of our model reconstructs the original scenes and disclosures the actual ID of the subjects in a scene. This module is designed for security authorities, to be used in crime scenes investigation.

The remainder of this paper is organized as follows: Section II summarizes the most relevant research in the scope of the paper. Section III provides a detailed description of the proposed method. Section IV discusses the results of our empirical evaluation, and the conclusions are given in Section V.

II. Related Work

This section summarizes the existing works in the image/video-based face de-identification context.

A. Image-Based Face De-identification

The earliest methods used simple image processing operations, such as blacking-out, pixelation or blurring (e.g., [5] and [35]), yielding poor realistic anonymised data. Later, Blanz *et al.* [3] estimated shape, pose and illumination in pairs of faces, and fitted morphable 3-D models to each one, rendering new faces by transferring parameters between source/target models. [36] proposed an eigenvector-based solution in which faces are reconstructed by a fraction of the *eigenface* vectors, such that ID information is lost. Similarly, Seo *et al.* [49] were based on watermarking, hashing and PCA representations of data. Bitouk *et al.* [4] proposed a method that replaces a target by a gallery element, selected according to its similarity to the query. Gross *et al.* [19] used multi-factor models that unify linear, bilinear and quadratic data fitting solutions, but requiring a AAM to provide landmarks information.

The *k*-Same algorithm [34] provided the rationale for various techniques. Considering the *k*-anonymity model [51], linear combinations of the gallery elements were obtained per probe, creating realistic anonymised data that depend on the alignment between the gallery/probe elements. Du *et al.* [15] used gallery samples to change each probe, obtaining "average" faces that lack in terms of photorealism. There are various recent methods still based in this concept, such as the k-Same-Net [32] and the attributes preserving approaches due to Jourabloo *et al.* [26] and Yan *et al.* [62].

Upon the deep learning breakthrough, Korshunova *et al.* [28] learned one generative model per identity. However, by restricting the output patches to gallery elements of the same identity, this solution limits the variability of the results. Using segmented silhouettes, Brkic *et al.* [6] proposed a model where obfuscation depends on the masked input data. *DeepPrivacy* [24] anonymises facial images while retaining the original data distribution, for photorealism purposes. A cGAN is the central component of the face/pose encoding process. Li and Lyu [30] used a face attribute transfer model to preserve the consistency of non-identity attributes between input/anonymised data. Sun *et al*. [53] combined parametric face synthesis techniques and GANs, keeping control over the facial parameters while adding fine details and realism into the resulting images. This approach depends of a computationally demanding face alignment step. Yamac *et al*. [61] introduced a reversible privacy-preserving compression method, that combines multi-level encryption with compressive sensing. Finally, Gu *et al*. [20] described a generative adversarial learning scheme based in image data and passwords that feed the models using additional input channels. The idea is to train a generative model that reconstructs the original input only when the right password is also given.

B. Face De-identification in Videos

The earliest approach was due to Dufaux *et al*. [16], which scrambled the quantized transform coefficients of facial blocks by random flipping/permutations, allowing reversibility but completely failing in terms of photorealism. Agrawal and Narayanan [1] performed 3D segmentation (in space and time) and blurred data in both domains to prevent reversal. Dale *et al*. [11] used 3D multilinear models to track the facial appearance of source/target videos. Using 3D geometry, they warped the source to the target face, keeping concerns about the temporal consistency of the result. Samarzija and Ribaric [47] grouped the gallery faces according to ID/pose information, representing each cluster by an AAM. Queries were matched to each AAM and the best match used as anonymised data.

Ren *et al*. [43] proposed a GAN-based video face anonymizer where the de-identified data preserve action information. Gafni *et al*. [18] proposed a feed-forward encoder-decoder architecture that fuses the input to masked outputs of a U-net backbone. This method requires facial landmarks to produce visually acceptable high resolution data. Sun *et al*. [54] proposed a GAN-based solution that partially changes the face texture, according to landmarks and head pose information given as input. Bao *et al*. [2] proposed a GAN-based framework to synthesise faces from two input images, one used for identity and the other for style attributes preservation. Similarly, Shen and Liu [52] and He *et al*. [23] proposed two models based in similar insights. However, both methods fail essentially in terms of the temporal consistency across different frames. Maximov *et al*. [31] proposed the CIAGAN, also based on conditional GANs, to obtain de-identified versions of the input, while keeping control of the soft biometric features of the output. This method produces highly photorealistic images, yet it requires the availability of facial landmarks for proper alignment.

There are also various examples of real-time video privacy protection techniques, where anonymity is assured by face masking [48], cryptographic obscuration [10], encryption [58] or blurring [33], but photorealism is not a concern.

C. Face Attributes Transfer

Face de-identification is closely related to transferring (swapping) attributes in human faces, which has been motivating several works. Zong *et al*. [68] used mid-level features of a CNN as disentangled representations of facial features, while Cao *et al*. [8] proposed a Multi-task CNN with partially shared layers that learn facial attributes. Xiao *et al*. [60] proposed the ELEGANT framework, that infers a disentangled representation in a latent space, where the various components refer different facial attributes in a quasi-independent way. The Attribute-GAN [23] and Style-GAN [27] are other good examples of generative models that enable to change facial features, while retaining all the other kind of information. In this context, the work due to Qi *et al*. [38] should also be mentioned, which uses local coordinate charts to parameterize the local geometry of data transformations across different locations on the manifold, preventing vanishing data variations and mode collapse.

III. The UU-Net: Reversible Face De-Identification in Visual Surveillance Video Data

A global perspective of our method is shown in Fig. 2. We divide the learning process into two phases, using insights from previous published architectures, such as [23], [27] and [70]: 1) a pairwise attribute matcher $\mathbf{D}_a$ is inferred, predicting the agreeing/disagreeing labels (ID, gender, ethnicity, age and hairstyle) between image pairs; and 2) the $\mathbf{D}_a$ responses are used to constraint the properties of the de-identified elements in the adversarial learning phase, along with an adversarial discriminator $\mathbf{D}_f$ that distinguishes between the input images and the generator ($\mathbf{U}_e$/$\mathbf{U}_d$) outputs.

A. Learning I: 'Same'/'Different' *Pairwise Attributes Classifier*

Let $\mathbf{x}^{\mathbf{l}}_{i,j,k}$ denote a face ROI in the kth frame of the jth sequence of the ith subject in the learning set. Also, let $\mathbf{a}^{\mathbf{l}}_{i,j,k}$ be the corresponding de-identified data and $\mathbf{r}^{\mathbf{l}}_{i,j,k}$ its reconstructed version. $\mathbf{l} \in \mathbb{N}^t$ is a column-vector containing the ground-truth attributes of $\mathbf{x}$. We consider $t = 4$ labels: $\{$*ID*, *gender*, *ethnicity*, *hairstyle*$\}$. For every pair of images $\mathbf{x}^{\mathbf{l}}$/$\mathbf{x}'^{\mathbf{l}'}$, we define a binary column vector $\mathbf{b}$, zeroed in the positions where labels between **l**/**l**' disagree:

$$\mathbf{b} = \left[\mathbb{1}_{\{l^1 == l'^1\}}, \ldots, \mathbb{1}_{\{l^t == l'^t\}}\right]^T, \tag{1}$$

being "==" the equality test operator and $\mathbb{1}$ the characteristic function. The attribute classifier $\mathbf{D}_a$: $\mathbb{N}^n \times \mathbb{N}^n \rightarrow \mathbb{N}^t$ receives a pair of images (each of length n) and predicts their common labels:

$$\hat{\mathbf{b}} = \mathbf{D}_a(\mathbf{x}^{\mathbf{l}}, \mathbf{x}'^{\mathbf{l}'}). \tag{2}$$

We use a cross-entropy loss for $\mathbf{D}_a$, which is optimized using the ground-truth $\mathbf{b}$ and predicted $\hat{\mathbf{b}}$ attributes:

$$\mathcal{L}_{\text{ce}} = \mathbb{E}_{\mathbf{b},\hat{\mathbf{b}}} \; -\mathbf{b}^T \log(\hat{\mathbf{b}}) - (\overrightarrow{\mathbf{1}}^T - \mathbf{b}^T)\log(1 - \hat{\mathbf{b}}), \tag{3}$$

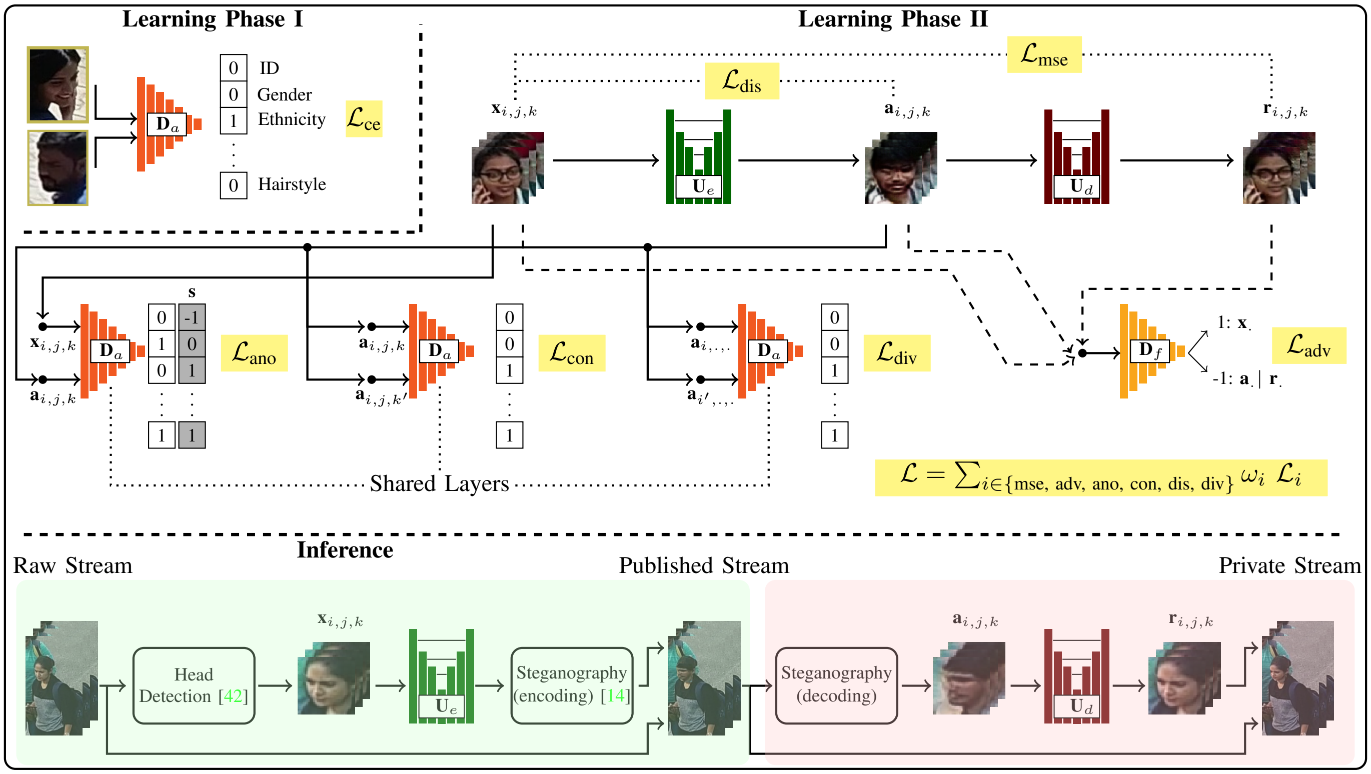

Fig. 2. Cohesive perspective of the proposed method. At first, we learn a pairwise attributes matcher $\mathbf{D}_a$ that infers the agreeing/disagreeing labels between image pairs. Next, a double-U sequential architecture is proposed, where the first part $\mathbf{U}_e$ receives the raw data $\mathbf{x}$ and creates the de-identified versions $\mathbf{a}$. The reverser $\mathbf{U}_d$ exclusively analyzes the de-identified data and reconstructs the original samples $\mathbf{r}$. The pairwise matcher is the basis of the anonymization, temporal consistency and diversity losses, along with an adversarial discriminator $\mathbf{D}_a$ that enforces the *facial appearance* of the generated data. In inference time, we crop the head regions and feed the $\mathbf{U}_e$ network. Then, using image steganography, the anonymized regions-of-interest are hidden in the published stream. Such streams are used by the reverser network $\mathbf{U}_d$ (available to security authorities), that reconstructs the original scenes.

with $\overrightarrow{\mathbf{1}}$ representing an all-ones column vector of t components, and the logarithmic/subtraction operations being applied component-wise. The classifier inferred is given by:

$$\mathbf{D}_a^* = \arg\min_{\mathbf{D}_a} \mathcal{L}_{\text{ce}}. \tag{4}$$

B. Learning II: Reversible De-Identification

We start by defining the **reconstruction loss** $\mathcal{L}_{\text{mse}}$, to guarantee the fidelity of elements reconstructed by $\mathbf{U}_d$ with respect to $\mathbf{x}$. This way, $\mathbf{U}_d$ will attempt to reconstruct $\mathbf{x}$, while at the same time $\mathbf{U}_e$ encapsulates hidden features in $\mathbf{a}_. = \mathbf{U}_e(\mathbf{x})$ that enable such reconstruction:

$$\mathcal{L}_{\text{mse}} = ||\mathbf{x}_. - \mathbf{U}_d\big(\mathbf{U}_e(\mathbf{x}_.)\big)||_2. \tag{5}$$

The **adversarial loss** is based in a *face plausibility* discriminator $\mathbf{D}_f$, that distinguishes between the input elements $\mathbf{x}$ and their encoded counterparts, either in the de-identified $\mathbf{a}_. = \mathbf{U}_e(\mathbf{x}_.)$ or in the reconstructed domain $\mathbf{r}_. = \mathbf{U}_d(\mathbf{a}_.) = \mathbf{U}_d\Big(\mathbf{U}_e(\mathbf{x}_.)\Big)$ This loss forces the encoded data to have *facial appearance*, as both generators $\mathbf{U}_e$ and $\mathbf{U}_d$ will attempt to fool $\mathbf{D}_f$ during the adversarial learning process. From the discriminator perspective, the loss is formulated as:

$$\mathcal{L}_{\text{adv}_1} = -2.\mathbb{E}_{\mathbf{x}} \; \mathbf{D}_f(\mathbf{x}_.) + \mathbb{E}_{\mathbf{a}_.} \; \mathbf{D}_f(\mathbf{a}_.) + \mathbb{E}_{\mathbf{r}_.} \; \mathbf{D}_f(\mathbf{r}_.), \quad \text{s.t. } ||\mathbf{D}_f||_\infty \leq \delta_{\text{gp}}, \tag{6}$$

where $||.||_\infty$ denotes the maximum gradient δ_{gp} allowed to avoid mode collapse and enhance training stability, as in WGAN-GP [21]. The optimal discriminator is formulated as:

$$\mathbf{D}_f^* = \arg\min_{\mathbf{D}_f} \mathcal{L}_{\text{adv}_1}. \tag{7}$$

From the encoder loss perspective, the previously used terms have opposite sign:

$$\mathcal{L}_{\text{adv}_2} = -\mathbb{E}_{\mathbf{a}} \; \mathbf{D}_f(\mathbf{a}_.) - \mathbb{E}_{\mathbf{r}} \; \mathbf{D}_f(\mathbf{r}_.). \tag{8}$$

All the remaining terms evolved in the generator use the pairwise discriminator $\mathbf{D}_a(.,.)$ obtained in the previous phase.

Let $\mathbf{s} \in \{-1, 0, 1\}^t$ be a column vector where '1' values denote labels that should agree between image pairs, '-1' values denote labels disagreement and '0' determines the independence between the raw/de-identified labels (Fig. 3). During optimization, in a minimization problem context, the inner product between $\mathbf{s}$ and $\mathbf{D}_a(\mathbf{x}, \mathbf{a})$ determines that high $\mathbf{D}_a()$ responses will be privileged for positions where s_i=-1. Similarly, low responses will be privileged for positions where s_i=1 (if s_i=0, the corresponding value in $\mathbf{D}_a()$ is ignored, as it is cancelled by the inner product operation). This strategy is used to keep control over the facial attributes that should be kept/changed, depending of the properties desired for the de-identified data. In every case, the ID position of **s** is always set to -1, guaranteeing the de-identification property of the model.

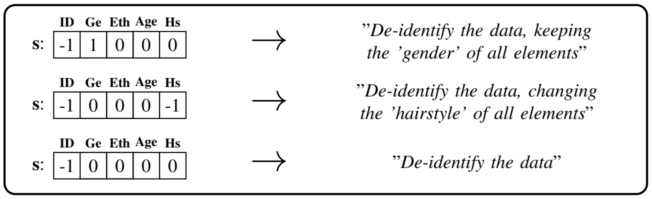
ID Ge Eth Age Hs
s: -1 1 0 0 0
"De-identify the data, keeping the 'gender' of all elements"
ID Ge Eth Age Hs
s: -1 0 0 0 -1
"De-identify the data, changing the 'hairstyle' of all elements"
ID Ge Eth Age Hs
s: -1 0 0 0 0
"De-identify the data"

Fig. 3. Illustration of the role of **s** parameterizations in the de-identified data. We keep control of the soft labels in the de-identified faces, imposing agreement (1), independence (0) or disagreement (-1) between the corresponding labels in **x/r** elements.

The **anonymization loss** forces that the corresponding $\mathbf{x}_. / \mathbf{a}_.$ elements follow the attribute configuration determined by **s**:

$$\mathcal{L}_{\text{ano}} = \mathbb{E}_{\mathbf{x}_{i,j,k},\mathbf{a}_{i,j,k}} \mathbf{s} \odot \Big(2.\mathbf{D}_a(\mathbf{x}_{i,j,k}, \mathbf{a}_{i,j,k}) - 1\Big), \tag{9}$$

where $\odot$ denotes the inner product and the $\big(2.\mathbf{D}_a(.,.) - 1\big)$ term maps the output of $\mathbf{D}_a$ into the [-1, 1] interval.

The **temporal consistency** loss guarantees that all samples of one sequence $\mathbf{a}_{i,j,k}/\mathbf{a}_{i,j,k'}, \forall k, k' : k \neq k'$, have the same soft attributes, for photorealism purposes:

$$\mathcal{L}_{\text{con}} = -\mathbb{E}_{\mathbf{a}_{i,j,k},\mathbf{a}_{i,j,k'}} \overrightarrow{\mathbf{1}}^T \odot \mathbf{D}_a(\mathbf{a}_{i,j,k}, \mathbf{a}_{i,j,k'}), \tag{10}$$

with $\overrightarrow{\mathbf{1}} \in \mathbb{N}^t$ denoting an all-ones column vector of t components.

The **diversity loss** assures that different sequences of one subject $\mathbf{a}_{i,j,.}/\mathbf{a}_{i,j',.}, \forall j, j' : j \neq j'$ are mapped to different virtual IDs, to avoid any malicious inference of subjects and scenes patterns:

$$\mathcal{L}_{\text{div}} = \mathbb{E}_{\mathbf{a}_{i,j,.},\mathbf{a}_{i,j',.}} \overrightarrow{\mathbf{1}} \odot \mathbf{D}_a(\mathbf{a}_{i,j,.}, \mathbf{a}_{i,j',.}). \tag{11}$$

Finally, the **distribution loss** assures that the color distributions of the corresponding **x/a** elements are similar, again for photorealism purposes:

$$\mathcal{L}_{\text{dis}} = \mathbb{E}_{\mathbf{x}_{i,j,k},\mathbf{a}_{i,j,k}} \chi^2_{h(\mathbf{x}),h(\mathbf{a})}, \tag{12}$$

where $h(.)$ is the histogram operator and $\chi^2_{\mathbf{v}^{(1)},\mathbf{v}^{(2)}}$ denotes the Chi-square distance between the distributions of $(\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$: $\sum_i \frac{(\mathbf{v}_i^{(1)} - \mathbf{v}_i^{(2)})^2}{\mathbf{v}_i^{(1)} + \mathbf{v}_i^{(2)}}$, with $\mathbf{v}_i^{(.)}$ denoting the ith bin density.

Overall, the full loss function is the weighted sum of the above described terms:

$$\mathbf{U}_e^*, \mathbf{U}_d^* = \arg \min_{\mathbf{U}_e,\mathbf{U}_d} \omega_{\text{mse}}\mathcal{L}_{\text{mse}} + \omega_{\text{adv}}\mathcal{L}_{\text{adv}_2} + \omega_{\text{ano}}\mathcal{L}_{\text{ano}} + \omega_{\text{con}}\mathcal{L}_{\text{con}} + \omega_{\text{div}}\mathcal{L}_{\text{div}} + \omega_{\text{dis}}\mathcal{L}_{\text{dis}}, \tag{13}$$

where $\omega_.$ are the hyper-parameters that weight each term in the learning process (details about the values used are given in Section IV).

C. Image Steganography

Steganography is used in this work for two purposes: 1) to avoid that the head/face detector has to be used both in the de-identification and reconstruction phases; and 2) to assure that the regions-of-interest (ROIs) used to reconstruct $\mathbf{a}_.$ elements are the same from where $\mathbf{x}_.$ were cropped. This is a sensitive point, as we observed a decrease in the fidelity of the reconstructed data, in case of misalignments between the ROIs cropped for corresponding $\mathbf{x}_./\mathbf{a}_.$ elements. The output of the head detector [42] is incapsulated in the public stream, using the protocol:

$$\begin{aligned} \text{message} &:= \text{n} + \text{","} + [\text{ROI}]_n \\ \text{ROI} &:= \text{x} + \text{","} + \text{y} + \text{","} + \text{w} + \text{","} + \text{h} + \text{","} \\ \text{n} &:= \{\mathbb{N}\} \\ \text{x} &:= \{\mathbb{N}\} \\ \text{y} &:= \{\mathbb{N}\} \\ \text{w} &:= \{\mathbb{N}\} \\ \text{h} &:= \{\mathbb{N}\}, \end{aligned}$$

where n denotes the number of bounding boxes in the frame, $[.]_n$ denotes n occurrences of one element, $\{\mathbb{N}\}$ stands for "*one natural number*" and (x, y, w, h) provide the top left corner (x, y), plus the width w and height h of the ROI bounding box.

This way, every frame in the public stream encapsulates (using [14]) a message containing the number and position of the head regions in the frame. As an example, the message "*2,10,16,9,15,25,45,8,14,*" informs about two ROIs, one starting at position $(10, 16)$, with dimensions $(9, 15)$ and another starting at position $(25, 45)$ with dimensions $(8, 14)$.

D. Inference

For security purposes, the data are de-identified before being published or transmitted through the network (bottom row of Fig. 2). This can be done *in situ*, embedded in the camera hardware and starts by head detection [42] in each frame. Next, the detected $\mathbf{x}_.$ feed the $\mathbf{U}_e$ generator, that returns their de-identified versions $\mathbf{a}_.$. Such virtual IDs are then overlapped in each frame, with image steganography [14] being used to encapsulate ROIs information.

The second part of the generator $\mathbf{U}_d$ is supposed to be available only to authorities. Upon a security incident, the de-identified stream contains all the information required to reconstruct the original scene and disclosure the actual ID of the subjects there. Using [14], we get the bounding boxes of the $\mathbf{a}_.$ elements in each frame and feed $\mathbf{U}_d$, that reconstructs the corresponding **r** representations.

IV. Experiments and Results

A. Datasets and Empirical Protocol

Our experiments were mostly conducted in one proprietary (BIODI) and two freely available visual surveillance datasets (MARS and P-DESTRE).

The BIODI[4] dataset is proprietary of *Tomiworld*[5], and is composed of 849,932 images/13,876 sequences, taken from 216 indoor/outdoor video surveillance sequences. All images are manually annotated for 14 labels: {'gender', 'age', 'height', 'body volume', 'ethnicity', 'hair color' and 'hairstyle', 'beard', 'moustache', 'glasses' and 'clothing' (x4)}. As this set is not annotated for ID, the face recognition experiments were exclusively performed in the remaining sets. MARS [67] contains 1,261 IDs from around 20,000 tracklets, automatically extracted by the Deformable Part Model [17] detector and the GMMCP [12] tracker. In this set, the soft labels {'gender', 'ethnicity"} were automatically inferred by the Matlab SDK for *Face++*[6] system, and the *hairstyle* was manually annotated. Finally, the P-DESTRE [29] provides video sequences of pedestrians in outdoor environments (taken from UAVs), and is fully annotated at the frame level, for ID and 16 soft labels: 'gender', 'age', 'height', 'body volume', 'ethnicity', 'hair colour', 'hairstyle', 'beard', 'moustache', 'glasses', 'head accessories', 'body accessories', 'action' and 'clothing information' (x3). It contains 253 identities and over 14.8M bounding boxes.

Complementary, the YouTube Faces [59] dataset was used to perceive the performance of the proposed solution in case of learning/test sets with very different features (the tracklets and the soft labels were obtain as in the MARS set). When compared to the visual surveillance sets used, images in this set have substantially higher resolution. Hence, the idea was to perceive if such amount of additional information used in the learning phase will enable to obtain sharpen de-identified/reconstructed samples, regardless of the notoriously different learning/test domains. The obtained results are discussed in Section IV-I.

For the image pairwise matcher $\mathbf{D}_a$, we used a classic VGG-like architecture, detailed in Table I. A different model was inferred independently for each label (*ID*, *gender*, *ethnicity* and *hairstyle*). Then, during inference, the pairs of RGB images to be matched were resized and concatenated along the depth axis, resulting in 64 × 64 × 6 inputs, from where the 4-dimensional output vectors were inferred.

For the second learning phase, both the encoder $\mathbf{U}_e$ and decoder $\mathbf{U}_d$ models shared the well known *U-Net* architecture [44], with a minor adaptation to receive 64 × 64 × 4 (encoder) and $64 \times 64 \times 3$ (decoder) data. The encoder receives the raw facial images represented in the RGB space (scaled to the unit interval) and a forth channel of random values drew from a standard uniform distribution $\mathcal{U}(0, 1)$. The adversarial discriminative model $\mathbf{D}_f$ uses the *PatchGAN* [25] architecture. Upon empirical optimization and grounded on the human perception of the $\mathbf{a}_.$ elements generated, we set δ_{gp}=0.01 to assure the stability of the adversarial learning process and the weight parameters: ω_{mse}=50, ω_{adv}=1, ω_{ano}=1, ω_{con}=1, ω_{dis}=1 and ω_{div}=1.

To our knowledge, there are no prior works to perform reversible de-identification in video surveillance video data.

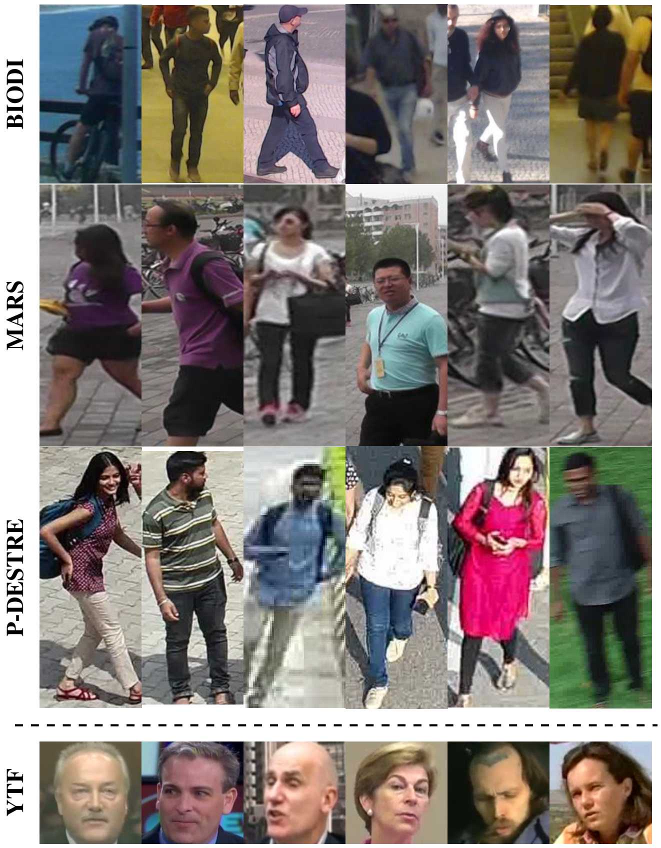

Fig. 4. Top rows: surveillance datasets used in the empirical validation of the method proposed in this paper (BIODI, MARS and P-DESTRE are shown). The bottom row provides some examples of the YouTube Faces set, used to perceive the performance of the proposed solution in case of *substantially different* features between the learning/test sets.

TABLE I
ARCHITECTURE OF THE CNN MODELS USED IN OUR EXPERIMENTS. ('NK': NUMBER OF KERNELS; 'KS': KERNEL SIZE; 'ST': STRIDE; 'MM': MOMENTUM).

$\mathbf{D}_a$ model	$\mathbf{U}_e$ model
Input: $(64 \times 64 \times 6)$ → Convolution (nk: 16, ks: 3 × 3, st: 2) → Batch Normalization (mm: 0.8) → LeakyReLU → Dropout (0.25) → Convolution (nk: 64, ks: 3 × 3, st: 1) → Batch Normalization (mm: 0.8) → LeakyReLU → Dropout (0.25) → Convolution (nk: 128, ks: 3 × 3, st: 1) → Batch Normalization (mm: 0.8) → LeakyReLU → Dropout (0.25) → [Convolution (nk: 64, ks: 3 × 3, st: 2) → BN (mm: 0.8) → LeakyReLU → Dropout (0.25)] × 2 → Flatten → Dense (128) → ReLU → Dense (t) → Sigmoid	Input: $(64 \times 64 \times 4)$ → *U-Net* [44]
	$\mathbf{U}_d$ model
	Input: $(64 \times 64 \times 3)$ →*U-Net* [44]
	$\mathbf{D}_f$ model
	Input: $(64 \times 64 \times 3)$ → *PatchGAN* [25]

Though, we considered two baselines to compare the face detection effectiveness in de-identified data: the Super-pixel [7] method, that replaces each pixel by the average value of the corresponding super-pixel and the Blur-based method [45], where images are downsampled to low-resolution and then upsampled back.

B. Face Detection

The photorealism of the de-identified data was evaluated by comparing the face detection scores obtained for the raw

[4] http://di.ubi.pt/~hugomcp/BIODI/
[5] https://tomiworld.com/
[6] http://www.faceplusplus.com/

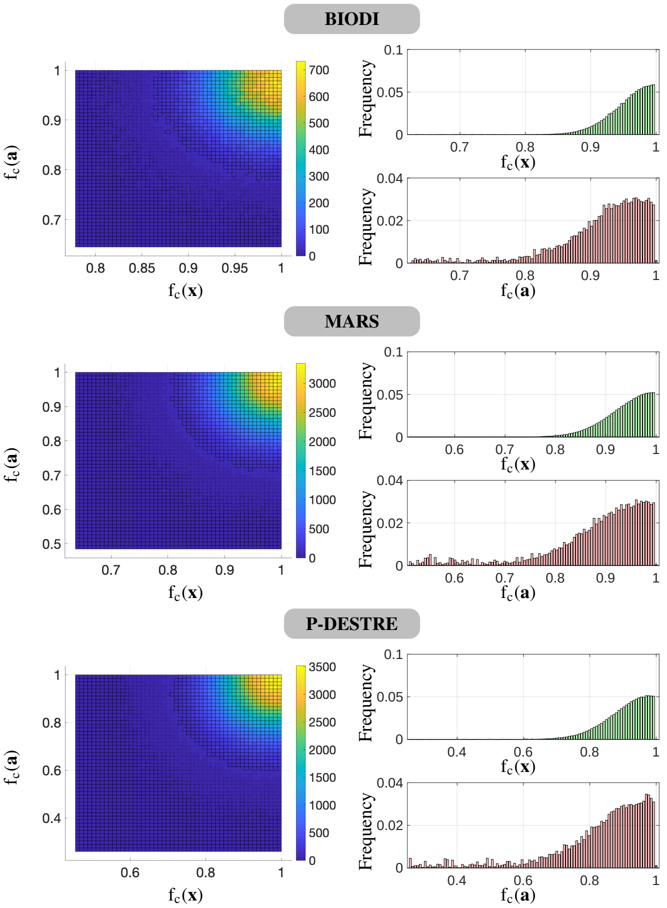

Fig. 5. At left: 3D histograms showing the correlation between the MTCNN [64] face detection confidence scores $f_c(.)$ of $\mathbf{x}_{.}/\mathbf{a}_{.}$ elements, for the BIODI, MARS and P-DESTRE sets. The corresponding unidimensional distributions are given at the right side.

$\mathbf{x}_{.}$ and the de-identified $\mathbf{a}_{.}$ elements, according to [64]. This method provides a confidence score $f_c(.)$ for having a face at a given position. The results for the three data sets are shown in Fig. 5, where the 3D histograms at the left column show the poor correlation between the f_c values for **x** (horizontal axis) and **a** elements (vertical axis). The linear correlation values (Pearson's coefficients) were of 0.011 (BIODI), 0.023 (MARS) and 0.025 (P-DESTRE). The unidimensional histograms at the right side provide the distributions for $f_c(\mathbf{x})/f_c(\mathbf{a})$ values. Overall, the average values for $f_c(\mathbf{a})$ elements decreased about 2.11% (BIODI), 3.22% (MARS) and 4.40% (P-DESTRE) with respect to $f_c(\mathbf{x})$ (BIODI: $f_c(\mathbf{x})= 0.954 \rightarrow f_c(\mathbf{a})= 0.931$, MARS: $f_c(\mathbf{x})= 0.968 \rightarrow f_c(\mathbf{a})= 0.930$ and P-DESTRE: $f_c(\mathbf{x})= 0.918 \rightarrow f_c(\mathbf{a})= 0.877$).

Table II summarizes the face detection effectiveness [64] on **a** data, with respect to the performance in **x** elements. Also, as baselines, we provide the results obtained by two simple de-identification techniques (due to Butler *et al.* [7] and Ryoo *et al.* [45]). For these experiments, random samples composed of 90% of the test data were created (drew with repetition) and the mean Average Precision (mAP) taken in each split, from where the mean and standard deviation values were taken. The proposed solution attained mAP values for

TABLE II
COMPARISON BETWEEN THE FACE DETECTION [64] PERFORMANCE IN THE DE-IDENTIFIED DATA, WITH RESPECT TO THE BASELINE DETECTION VALUES. AVERAGE ± STANDARD DEVIATION 'MEAN AVERAGE PRECISION' (MAP) VALUES ARE GIVEN.

Method	**Params.**	**BIODI**	**MARS**	**P-DESTRE**
Baseline Detection mAP [64] ($\mathbf{x}_{.}$)		0.82 ± 0.02	0.84 ± 0.03	0.63 ± 0.01
Proposed	ω_{mse}=50, ω_{adv}=1, ω_{ano}=1, ω_{con}=1, ω_{div}=1, ω_{dis}=1	0.73 ± 0.06	0.75 ± 0.06	0.59 ± 0.10
Butler *et al.* [7]	8 superpixels	0.23 ± 0.06	0.21 ± 0.04	0.20 ± 0.04
Butler *et al.* [7]	16 superpixels	0.36 ± 0.05	0.32 ± 0.04	0.22 ± 0.05
Ryoo *et al.* [45]	resolution 5x3	0.20 ± 0.03	0.19 ± 0.03	0.17 ± 0.02
Ryoo *et al.* [45]	resolution 7x4	0.31 ± 0.04	0.27 ± 0.04	0.16 ± 0.06

a elements that were about 89% (BIODI), 89% (MARS) and 93% (P-DESTRE) of the values obtained for **x**. Errors occurred typically for the poorest resolution samples, where the detection method was still able to find the original face but not its de-identified version. Also, we observed that the de-identified elements tend to have less details (i.e., lower entropy) than the original samples, which might justify this gap in performance. The remaining techniques got far worse performance, even stressing that both not aim at reversibility.

C. Face Recognition

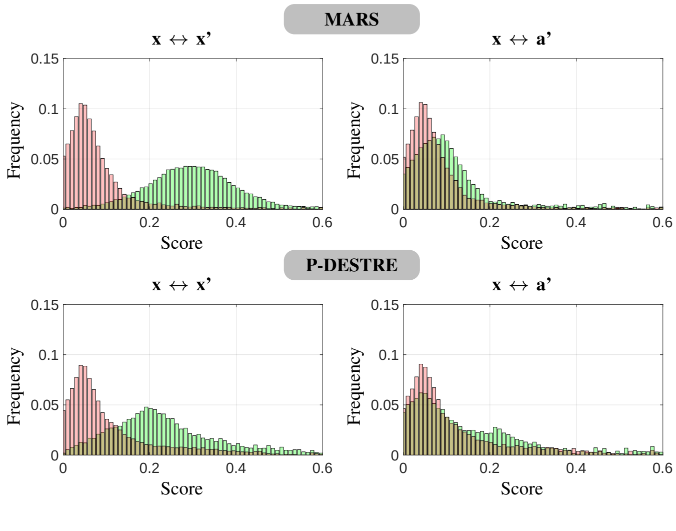

Fig. 6. Comparison between the decision environments resulting from the VGG-Face2 [9] face recognition method, working in pairs of images of the MARS and P-DESTRE datasets ($\mathbf{x} \leftrightarrow \mathbf{x}'$, left column). The right column provides the corresponding values when the second image in each pairwise comparison was de-identified ($\mathbf{x} \leftrightarrow \mathbf{a}'$).

A second question to address is the possibility of models being simply swapping faces between identities, rather than creating virtual IDs. To verify this hypothesis, we learned two

VGG-Face2 [9] recognizers (MARS and P-DESTRE sets), including 80% of the IDs in the learning set ($\mathbf{x}_{.}$ elements). Then, in inference time, we sampled the remaining 20% IDs and created 50K *impostors* + 10K *genuine* pairwise comparisons ($\mathbf{x} \leftrightarrow \mathbf{x}$'), where the VGG-face2 scores were obtained. This experiment was repeated when de-identifying the second image in each pair (i.e., $\mathbf{x} \leftrightarrow \mathbf{a}$'). Results are given in Fig. 6, that shows the decision environments for MARS (top plots) and P-DESTRE (bottom plots). The green bars correspond to the distributions of the *genuine* scores, while the red bars denote the *impostors* distributions. The decidability values of the decision environments were also obtained:

$$d' = \frac{\mu_G - \mu_I}{\sqrt{\sigma_G^2 + \sigma_I^2}}), \qquad (14)$$

where (μ, σ) denote the mean/standard deviation statistics and 'G'/'I' stand for the *genuine* and *impostors* pairwise comparisons. Values decreased from 1.885 (MARS) and 0.839 (P-DESTRE) ($\mathbf{x} \leftrightarrow \mathbf{x}$') to 0.162 (MARS) and 0.155 (P-DESTRE)($\mathbf{x} \leftrightarrow \mathbf{a}$'), with an evident movement of the *genuines* distributions toward the *impostors* region. The corresponding decreases in the AUC values were of MARS: 0.962 → 0.570 and P-DESTRE: 0.820 → 0.568, in both cases turning the identification based in $\mathbf{a}_{.}$ elements almost equivalent to a random choice. Even though, a slight difference between the right tails of both distributions was observed in MARS/P-DESTRE sets, which was justified by potential overfitting, i.e., lack of sufficient learning data to sustain enough variability in the de-identification space.

D. Temporal Consistency

The temporal consistency of the de-identified samples is of most importance for photorealism purposes. In a subjective evaluation perspective, Fig. 7 provides several examples of ($\mathbf{x}$, $\mathbf{a}$, $\mathbf{r}$) elements, obtained for frames of a sequence at time t and $t+i$. In all cases, the consistency between the overall appearance of $\mathbf{a}_t$ and $\mathbf{a}_{t+i}$ is evident.

To obtain a quantitative measure of temporal consistency, we compared the decision environments for MARS and P-DESTRE sets when the *genuine* pairwise comparisons were exclusively composed of $\mathbf{a}_{i,j,t}$/$\mathbf{a}_{i,j,t+k}$ elements, with $k \in \{1, \dots, s_l\}$ (s_l is the sequence length). The *impostors* distributions were obtained as in the previous experiments. The bottom row in Fig. 7 provides the results, with an evident separation between the *impostors*/*genuine* scores. When comparing both empirical data distributions against the *null* hypothesis ('*both samples come from the same distribution*'), the Kolmogorov-Smirnov test enabled to reject the *null* hypothesis with asymptotic p-values lower than $1e^{-8}$ in both sets. With respect to the values given in Fig. 6, note that here only pairs of the same session were considered *genuine*, which justifies the large separability between the *genuine*/*impostors* distributions. In both cases, the genuine scores spread homogeneously along the unit interval, yet there is a fraction of cases ($< 15\%$) where consecutive de-identified elements suddenly changed their appearance and soft labels.

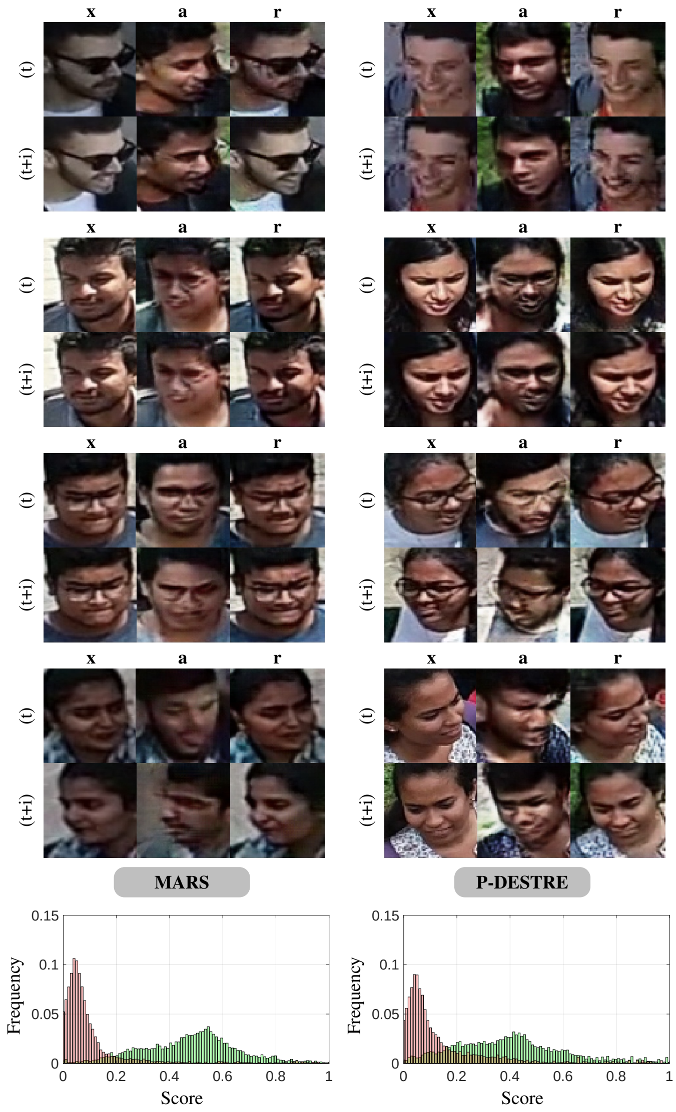

Fig. 7. Top rows: Examples illustrating the temporal consistency of our solution, with each group providing two within-subject examples of a sequence $\mathbf{x}$ taken at times t and $t+i, i \geq 1$. The central image in each group regard the de-identified images $\mathbf{a}$, and the reconstructed samples are given at the rightmost column $\mathbf{r}$. The bottom row shows the decision environments when the *genuine* pairwise comparisons were exclusively composed of $\mathbf{a}_{i,j,t}$/$\mathbf{a}_{i,j,t+k}$ elements.

E. Soft Labels Consistency/Inter-Session Diversity

The consistency of the soft labels generated and the diversity of the virtual IDs per subject were evaluated according to the responses provided by the pairwise labels discriminator $\mathbf{D}_a$. Regarding the soft labels, we were interested in confirm that the {'gender', 'ethnicity', 'hairstyle'} labels inferred for $\mathbf{a}_{.}$ meet the constraints determined by $\mathbf{s}$ in the learning phase. For each $\mathbf{a}_i$ element, we obtained the $\mathbf{D}_a(\mathbf{a}_i, \mathbf{x}_i)$ values and measured their Pearson correlation values with respect to the ground-truth labels, also taking into account the configuration of $\mathbf{s}$. Having drew 50 random samples composed of 90% of the test data (with repetition), the linear correlation values are given in Table III, which were regarded as good indicators of the consistency between $\mathbf{x}_{.}$/$\mathbf{a}_{.}$ soft labels. Some examples of $\mathbf{a}_{.}$ elements generated when

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3066054, IEEE Transactions on Circuits and Systems for Video Technology

$\mathbf{s}$ = ['ID', 'Gender', 'Ethnicity', 'Hairstyle'] = $[-1, 1, 1, 1]$ ('*All Equal*' labels) and $\mathbf{s} = [-1, -1, -1, -1]$ ('*All Different*' labels) are shown in Fig. 9, enabling to perceive the varying appearance of $\mathbf{a}$ elements according to the $\mathbf{s}$ configuration used in learning. Also, we observed that the '*All Equal*' labels configuration reduces the variability of the synthesised virtual IDs, while also increasing the similarity in appearance between the $\mathbf{x}_{.}/\mathbf{a}_{.}$ elements.

TABLE III
PEARSON CORRELATION BETWEEN THE SOFT LABELS INFERRED FOR THE DE-IDENTIFIED ELEMENTS $\mathbf{A}$ WITH RESPECT TO THE SOFT LABEL CONFIGURATION DETERMINED BY $\mathbf{S}$.

Soft Label Consistency	**BIODI**	**MARS**	**P-DESTRE**
Gender	0.818 ± 0.096	0.890 ± 0.081	0.803 ± 0.107
Ethnicity	0.702 ± 0.112	0.750 ± 0.099	0.622 ± 0.144
Hairstyle	0.647 ± 0.106	0.663 ± 0.102	0.594 ± 0.118

To illustrate the labels consistency and the diversity of the virtual IDs generated for each subject, Fig. 8 provides two embeddings for the projection of the 4,096 coefficients of the VGG-19 '*fc7*' layer[7] for **x/a** elements into a 2D space, according to the Neighbourhood Component Analysis [63] algorithm. A VGG classification model was inferred, using images $\mathbf{x}$ of 50 subjects plus the corresponding 50 virtual identities $\mathbf{a}$ (one sequence per subject) of the P-DESTRE set, in a classification task (100 classes). Then, six subjects of a (disjoint) test set were used, with 10 images per subject/sequence considered. The $\mathbf{x}$ elements are denoted by black borders and the corresponding $\mathbf{a}$ de-identified elements appear borderless. The left embedding corresponds to the *'All Equal'* soft labels configuration, with a relative proximity between the corresponding **x/a** elements seen in all cases. In opposition, when the *'All Different'* soft labels configuration is enforced (right embedding), the IDs and their counterpart virtual IDs appear in the antipodes of the embedding. In both types of embeddings, the virtual IDs $\mathbf{a}$ were observed to spread more than the original elements $\mathbf{x}$, which was justified by the intrinsic variability of $\mathbf{x}$ features plus the stochastic nature of the de-identification model.

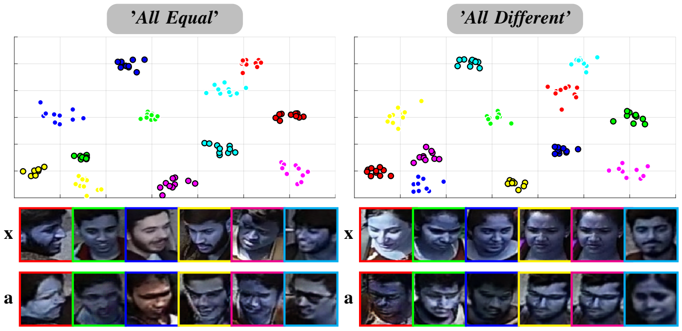

Fig. 8. Comparison between the 2D embeddings yielding from the $\mathbf{x}$ (black borders)/$\mathbf{a}$ (borderless) VGG-19 '*fc7*' representations of six subjects (corresponding to different colors), using 10 images per sequence. Results are shown for the *'All Equal'*/*'All Different'* soft labels configurations.

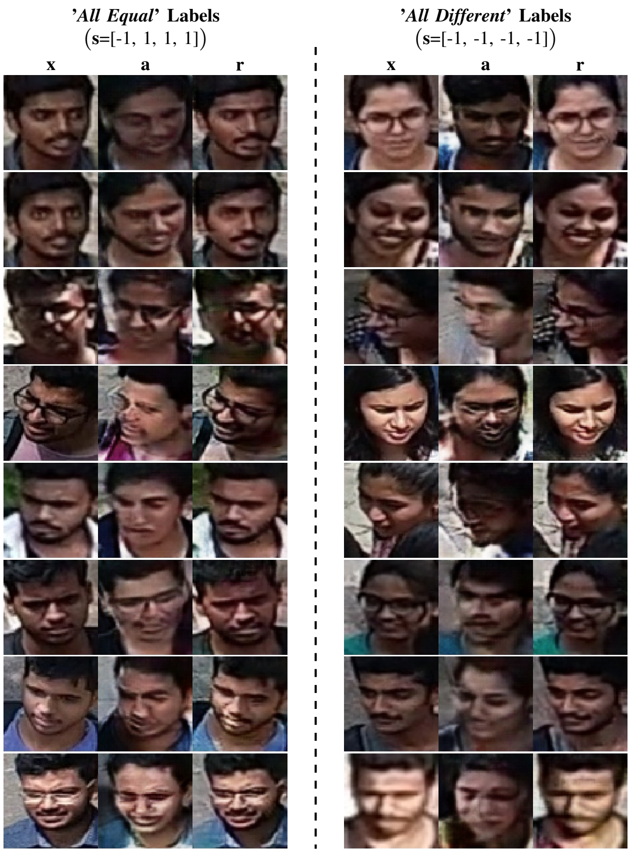

Fig. 9. Comparison between the results typically attained when the soft labels coherence/discrepancy is enforced. The left column provides examples for soft labels agreement $\mathbf{x}_{.}/\mathbf{a}_{.}$ ('*All Equal*'), while the right column illustrates the disagreement labels case ('*All Different*').

The diversity of the IDs generated per session is illustrated in the top rows of Fig. 10, with the bottom row providing the decision environments when the *genuine* pairs were exclusively composed of $\mathbf{a}_{i,j,.}$/ $\mathbf{a}_{i,k,.}$ elements. Here, even though there was an evident separation between both distributions (d'=0.491 for MARS and 0.329 for P-DESTRE), genuine distributions were skewed toward the higher values region (i.e., corresponding to the typical *genuine* region), which suggests that the IDs generated for different sessions might still share some undesirable patterns.

F. Reversibility and Input/Output Variability

The models proposed in this paper are stochastic. In inference time, the encoder $\mathbf{U}_e$ receives (apart the RGB data) one random input channel that has an effect in the system response. This way, when *repeating* (excluding the random channel) the input, there will be some variations in the corresponding outputs. We measured such variability, in terms of the point-by-point differences in the de-identified **a/a**' and in the reconstructed **r/r**' domains. More importantly, we perceived how likely such variations imply a change in the ID inferred for the output image. Results are shown in Fig. 11. The histograms in the top row provide the point-by-point residuals between outputs yielding from the same input, in the de-identified (at

[7] https://www.mathworks.com/help/deeplearning/ref/vgg19.html

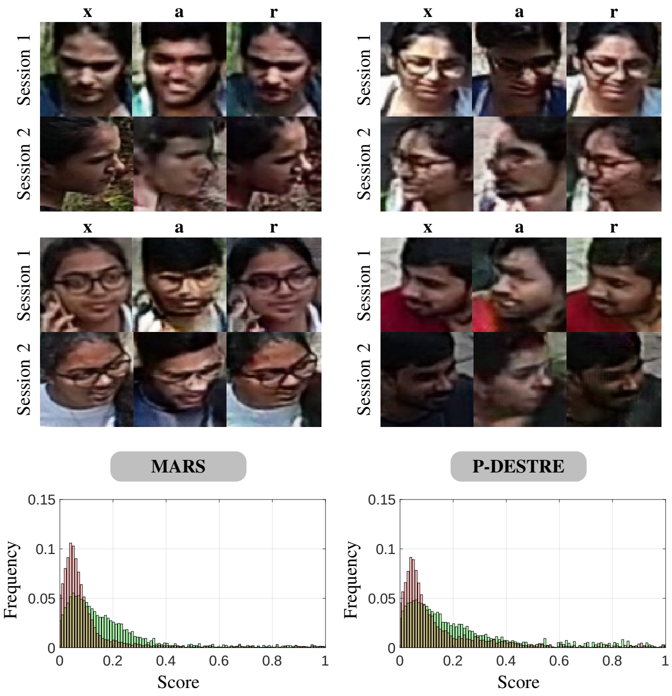

Fig. 10. Top rows: diversity of the de-identified samples for different sequences of a subject. The bottom row provides the decision environments obtained for the MARS (at left) and P-DESTRE (at right) sets, when the *genuine* pairs were exclusively composed of $\mathbf{a}_{i,j,.}$/ $\mathbf{a}_{i,k,.}$ (j $\neq$ k) elements.

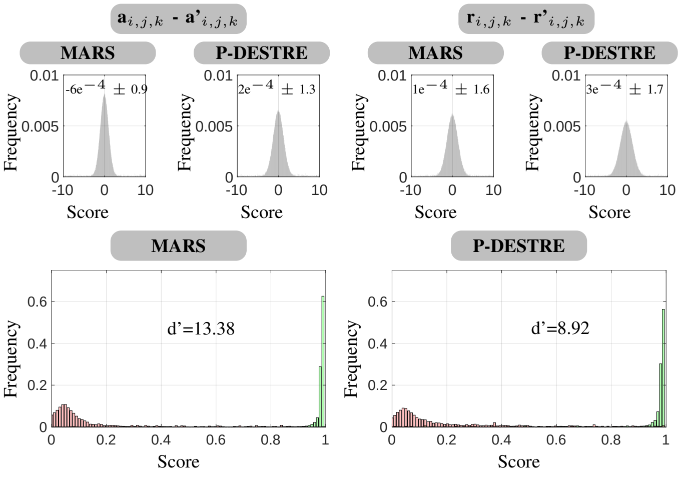

Fig. 11. Top row: histograms of the point-by-point residuals between outputs resulting from the same input data (excluding the noise channel. Results are given for the de-identified ($\mathbf{a}_{i,j,k}$ - $\mathbf{a}'_{i,j,k}$, at left) and reconstructed ($\mathbf{r}_{i,j,k}$ - $\mathbf{r}'_{i,j,k}$, at right) domains. Bottom row: decision environments for MARS (at left) and P-DESTRE (at right) sets, when the genuine pairwise comparisons were exclusively composed of different reconstructed images **r/r'** resulting from the same input data.

left) and reconstructed domains (at right). Also, we report the decision environments for MARS and P-DESTRE sets, when the genuine pairs were exclusively composed of **x** elements and various of their reconstructed versions **r**.

The residuals obtained were small in all cases, and almost undetectable under visual inspection: the mean deviation was close to 0 in both sets, with slightly higher standard deviations observed for **r/r'** elements, probably as a result of additive deviations. Importantly, the IDs inferred for different **r/r'** resulting from the same input were always invariant. This is confirmed by the absolute separation between the *impostors/genuine* distributions, and the extremely peaked distributions for the *genuine* scores. These observations support the extremely low probability of switching identities in multiple $\mathbf{x} \rightarrow \mathbf{a} \rightarrow \mathbf{r}$ mappings. Also, the varying features among different **x** elements and the necessity of their proper reconstruction (imposed by the $\mathcal{L}_{\text{mse}}$ term) justifies why **a** elements are different between identities (i.e., too similar **a** elements for different IDs will not enable different destiny-reconstructions **r**). This is seen both in terms of visual perception and of the IDs inferred by a facial recognizer.

G. Pose, Background and Facial Expressions Consistency

For photorealism purposes, not only the pose of $\mathbf{x}_{.}/\mathbf{a}_{.}$ elements should be consistent, but also the background features in both images should be similar and even the facial expression should agree. According to our loss formulation and experiments, we observed that the distribution loss term $\mathcal{L}_{\text{dis}}$ plays a key role in guaranteeing such consistencies. To quantitatively perceive the head pose consistency, we compared the 3D pose estimation values (*yaw*, *pitch* and *roll*) obtained by the Deep Head Pose [46] method for **x/a** elements. As the model was not specifically trained for each dataset, errors in pose inference were relatively frequent and covered about 20% of the samples of the BIODI set, 7% of the elements in MARS and 28% of the P-DESTRE images. These cases were rejected under human inspection. For the remaining cases, we measured the absolute difference between the 3D angle values, obtaining average *yaw* errors of 0.177 ± 0.091, 0.184 ± 0.087, 0.140 ± 0.075 (BIODI, MARS, P-DESTRE), *pitch* errors of 0.101 ± 0.068, 0.120 ± 0.070, 0.113 ± 0.055 and *roll* errors 0.021 ± 0.006, 0.0.25 ± 0.005, 0.022 ± 0.004 (in radians), which are illustrated in the second row/forth column cell of Fig. 12. Regarding the background consistency, we used the cropped head ROIs as input of a skin-hair-background segmentation method [37]. According to the obtained segmentation masks, we zeroed all pixels deemed to correspond to skin/hair and considered the remaining information as background (denoted by $\mathbf{x}^{(b)}$, $\mathbf{a}^{(b)}$). Next, we measured the point-by-point residuals (Euclidean distance in the CIE 1976 L*a*b* color space) between each pair, obtaining the results shown at the rightmost column of the top row. Finally, regarding the facial expressions consistency, in order to objectively perceive the agreement between the facial expressions of **x** and corresponding **a** elements, we used the [66] method to estimate the facial expressions (considering exclusively the Neutral (N), Disgust (D), Surprise (Su) and Smile (Sm) classes) in the MARS, BIODI and P-DESTRE sets, for **x/a** elements. The normalized confusion matrix is shown at the bottom right corner, with the rows providing the predictions for **x** and the columns providing the corresponding predictions for **a** elements. Inside each cell, the relative frequency of a pair of predictions is given, observing the agreement of **x/a** labels in about 82% of the cases.

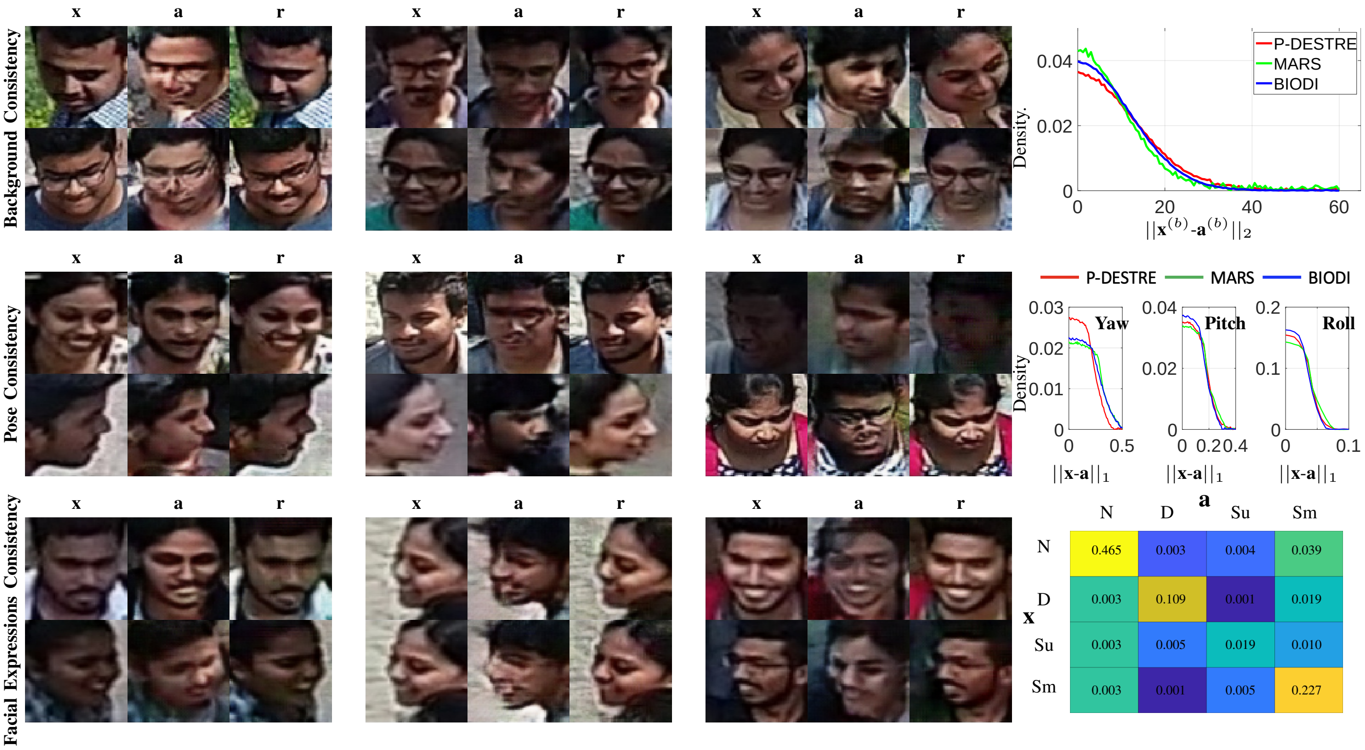

Fig. 12. Examples illustrating the background consistency (upper row), pose consistency (middle row) and facial expressions consistency (bottom row) between the raw samples **x** and their de-identified versions **a**. The forth column provides a density estimate of the point-by-point ℓ_2-norm residuals between the background regions of the original **x** and de-identified **a** images (top row), the density estimates of the absolute residuals between the pose parameters estimated for textbfx/textbfa (middle row) and the confusion matrix for the facial expressions inferred by [66] for corresponding **x** (rows)/**a** (cols) elements.

H. Cross-Domain Adaptability in Surveillance Environments

Being heavily data-driven, the models proposed in this paper are particularly sensitive to changes in domain, i.e., tend to face difficulties when the domains of the data used in learning/inference are different. This section discusses the cross-domain adaptability of our method, keeping in mind that we constrained our analysis to visual surveillance environments. Hence, we considered the MARS and P-DESTRE sets (i.e., with ID information available), and evaluated the decreases in face recognition performance for the cross-domain setting: using the P-DESTRE (P) in learning/MARS (M) for test (and *vice-versa*). We compared the intra- and cross-domain settings from the temporal consistency, diversity and samples reconstruction perspectives. Results are summarized in Fig. 13, in terms of the decidability (14) of the decision environments and of the Euclidean residuals between the original and the reconstructed elements. These values should be compared to the performance reported in Fig. 6 (face recognition), Fig. 7 (temporal consistency) and Fig. 10 (diversity). The images at the bottom provide examples of the typical appearance of the worst cross-domain configuration (M → P). Overall, our perception was that the reconstruction remained effective - particularly in terms of IDs - yet the photorealism of the de-identified samples in the cross-domain setting was substantially lower than the observed for the intra-domain setting. In this context, the background consistency was particularly affected, which was justified by the difficulty of the encoder to generate backgrounds notoriously different of the seen in the training phase.

	Intra-domain		**Cross-domain**	
Experiment	**M → M**	**P → P**	**P → M**	**M → P**
Face recognition (d') **x/a'**	0.264 ± 0.020	0.161 ± 0.023	0.206 ± 0.059	0.150 ± 0.065
Diversity (d') $\mathbf{a}_{i,j,.}/\ \mathbf{a}_{i,j',.}$	0.489 ± 0.038	0.321 ± 0.028	0.403 ± 0.041	0.305 ± 0.044
Tempor. Consist. (d') $\mathbf{a}_{i,j,t}/\mathbf{a}_{i,j,t+k}$	2.378 ± 0.041	1.429 ± 0.052	1.350 ± 0.070	0.947 ± 0.073
Reconst. Error (ℓ_2-norm) $\|\mathbf{x}-\mathbf{r}\|_2$	17.05 ± 8.90	19.99 ± 10.03	28.80 ± 13.48	31.58 ± 16.36

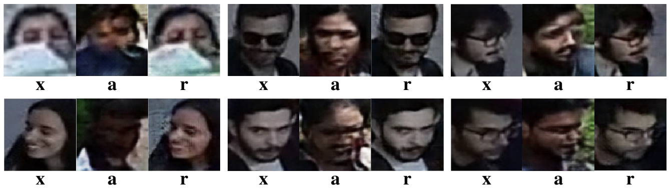

Fig. 13. Cross-domain adaptability between visual surveillance datasets. We compare the decidability d' of the 'Face Recognition', 'Temporal Consistency' and 'Diversity' decision environments in the intra-domain (M → M and P → P) and cross-domain (M → P and P → M) settings. Also, the reconstruction errors in both domains are given, along with examples of the poorest cross-domain configuration observed (M → P).

The temporal consistency of the $\mathbf{a}_{i,j,t}/\mathbf{a}_{i,j,t+k}$ elements also decreased evidently in the cross-domain setting. In opposition, the face recognition and diversity measurements yielded relatively similar values in the intra- and cross-domain experiments. This points that in a cross-domain scenario, the de-identified faces will still be diverse and not likely to be successfully matched with the corresponding IDs in the original domain, even though the de-identified data is not as photorealistic as in the intra-domain setting.

It should be noted that we constrained our analysis to visual surveillance datasets, where - in spite of being acquired in different conditions/environments - datasets evidently share some properties among them (e.g., outdoor lighting and relatively poor resolution). As reported in section IV-I, we observed that in case of substantially different learning/test domains, using learning sets of higher resolution does not contribute for more pleasant de-identified/ elements nor sharpen reconstructed samples, which can be considered a limitation of the proposed solution.

I. Ablation Experiments, Difficult Cases and Limitations

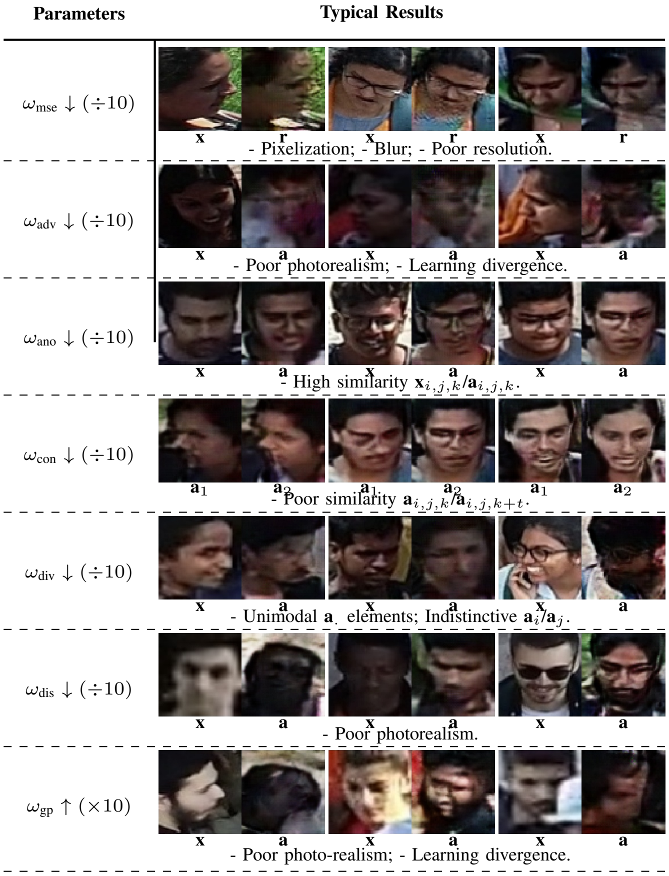
Parameters
Typical Results
$\omega_{\text{mse}} \downarrow (\div 10)$
x r x r x r
- Pixelization; - Blur; - Poor resolution.
$\omega_{\text{adv}} \downarrow (\div 10)$
x a x a x a
- Poor photorealism; - Learning divergence.
$\omega_{\text{ano}} \downarrow (\div 10)$
x a x a x a
- High similarity $\mathbf{x}_{i,j,k}/\mathbf{a}_{i,j,k}$.
$\omega_{\text{con}} \downarrow (\div 10)$
$\mathbf{a}_1$ $\mathbf{a}_2$ $\mathbf{a}_1$ $\mathbf{a}_2$ $\mathbf{a}_1$ $\mathbf{a}_2$
- Poor similarity $\mathbf{a}_{i,j,k}/\mathbf{a}_{i,j,k+t}$.
$\omega_{\text{div}} \downarrow (\div 10)$
x a x a x a
- Unimodal $\mathbf{a}$. elements; Indistinctive $\mathbf{a}_i/\mathbf{a}_j$.
$\omega_{\text{dis}} \downarrow (\div 10)$
x a x a x a
- Poor photorealism.
$\omega_{\text{gp}} \uparrow (\times 10)$
x a x a x a
- Poor photo-realism; - Learning divergence.

Fig. 14. Ablation experiments. Typical variations in the results of the proposed method with respect to changes in each term used in the loss formulation.

The most important variations in the results with respect to changes in the parameterizations of the loss function are illustrated in Fig. 14. The left column gives the change in one parameter with respect to the *optimal* configuration (described in sec. IV-A), and the images in the right column illustrate the typical failure cases. In this experiment, every parameter was orthogonally decreased (↓) or augmented (↑) one order of magnitude. At first, when the ω_{mse} weight was decreased, the reconstructed samples started to appear pixelized and blurred. As an effect of the variation of ω_{ano} weight, **x/a** resemble each other in a much more evident way, and in some cases the $\mathbf{U}_e$ model works practically as an identity operator. Decreasing the weights of the adversarial discriminator ω_{adv} has a catastrophic effect in the **a** results, that completely loose their *face* appearance. When decreasing the ω_{div} parameter, the de-identified images tend to look alike their **x** counterparts, while the ω_{con} weight stresses the temporal consistency requirements. By decreasing the value of the ω_{dis} parameter, the resulting **a** elements have very different color/brightness distributions with respect to **x**, which strongly decreases the photorealism. Also, another catastrophic change typically occurs when the maximum gradient δ_{gp} allowed for adjusting weights of the adversarial discriminator increases, causing the divergence of the training phase.

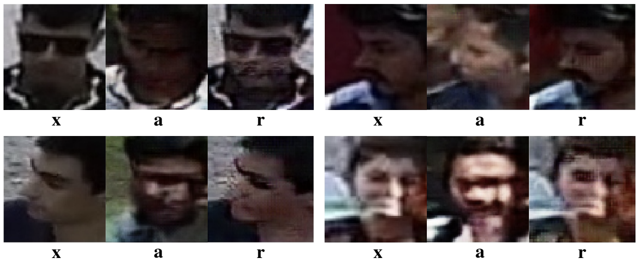

Fig. 15. Results obtained in case of *significantly different* features between the learning/test domains. The YouTube faces dataset was used in the learning phase, with input size extended to 256 × 256 for all models, while inference was done in the P-DESTRE set. In such setting, the amount of additional information used in the learning phase doesn't contribute for visually pleasant de-identified elements and for sharpen reconstructed samples.

Finally, we used the YouTube video face dataset as learning data, with input size changed to 256 × 256 for the $\mathbf{D}_a$, $\mathbf{U}_e$, $\mathbf{U}_d$ and $\mathbf{D}_f$ models. Then, the P-DESTRE set was used in inference. Here, the goal was to perceive if such additional amounts of learning data contributes to obtain de-identified elements that are visually pleasant and sharpen reconstructed samples, when compared to the visual surveillance learning/test sets configuration. As illustrated in Fig. 15, the results were considered (under visual inspection) even poorer than when visual surveillance data were used in the learning/test phases. Hence, we concluded that the relative similarity between the learning/test domains is more important than the amount of information available in each learning element. This feature is regarded as a limitation/constraint of the proposed solution to obtain visually pleasant de-identified results.

V. Conclusions

This paper addressed the security/privacy balance in visual surveillance environments. While data protection regulations forbid the public disclosure of personal sensitive information, there are scenarios, such as crime scene investigation, where the actual identification of subjects is of most importance. Accordingly, we described a solution composed of one *public* module, that detects the faces in each frame and creates their de-identified versions, where the ID information is surrogated in a photorealistic and seamless way. Such elements are overlapped in the data stream, which can be published without compromising subjects' privacy. This process runs *in situ*, such that no privacy-sensitive information is passed through the network. Next, upon a security incident, a *private* module - available to security agencies - is able to reconstruct the original scene and disclose the actual identity of the subjects there.

The proposed solution is landmarks-free and suitable for visual surveillance data. We designed a two-stage learning process, with a conditional generative adversarial network composed of two entities (an *encoder* and a *decoder*) that have the common goal of fooling an adversarial opponent. Their joint optimization enables to intrinsically share knowledge about the features that should be hidden in the encoded data in order to later assure proper reconstruction. The whole process generates realistic faces that preserve pose, lighting, background and facial expressions. Also, we keep full control over the facial attributes that are preserved/changed between the raw and de-identified streams. The experiments were conducted in three visual surveillance datasets, and support the usability of the proposed solution, conditioned by a relative similarity between the domains of the data used in the learning/inference phases.

ACKNOWLEDGEMENTS

This work is funded by FCT/MEC through national funds and co-funded by FEDER - PT2020 partnership agreement under the projects UIDB/50008/2020, POCI-01-0247-FEDER-033395 and C4: Cloud Computing Competence Centre.

REFERENCES

[1] P. Agrawal and P. Narayanan. Person De-Identification in Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no.3, pag. 299–310, 2011. 3

[2] J. Bao, D. Chen, F. Wen, H. Li and G. Hua. Towards open-set identity preserving face synthesis. In proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2018.00702, 2018. 3

[3] V. Blanz, K. Scherbaum, T. Vetter and H.-P. Seidel. Exchanging faces in images. *Computer Graphics Forum*, vol. 23, no. 3, pag. 669–676, 2004. 2

[4] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur and S. Nayar. Face Swapping: Automatically Replacing Faces in Photographs. *ACM Transactions on Graphics*, vol. 27, issue 3, doi: 10.1145/1360612.1360638, 2008. 2

[5] M. Boyle, C. Edwards and S. Greenberg. The Effects of Filtered Video on Awareness and Privacy. In proceedings of the *ACM Conference on Computer Supported Cooperative Work*, pag. 1–10, 2000. 1, 2

[6] K. Brkic, I. Sikiric, T. Hrkac and Z. Kalafatic. I know that person: Generative full body and face de-identification of people in images. In proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, doi: 10.1109/CVPRW.2017.173, 2017. 2

[7] D. Butler, J. Huang, F. Roesner and M. Cakmak. The privacy-utility tradeoff for remotely tele-operated robots. In proceedings of the *Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, doi: 10.1145/2696454.2696484, 2015. 6, 7

[8] J. Cao, Y. Li and Z. Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, pag. 4290–4299, 2018. 3

[9] Q. Cao, L. Shen, W. Xie, O. Parkhi and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In proceedings of the *IEEE Conference on Automatic Face and Gesture Recognition*, doi: 10.1109/FG.2018.00020, 2018. 7, 8

[10] A. Chattopadhyay and T. Boult. PrivacyCam: a Privacy Preserving Camera Using ucLinox on the Blackfin DSP. In proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2007.383413, 2007. 3

[11] K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlasic, W. Matusik and H. Pfister. Video face replacement. *ACM Transactions on Graphics*, vol. 30, no. 6, pag. 1–10, 2011. 3

[12] A. Dehghan,A., S. ssari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2015.7299036, 2015. 6

[13] J. Deng, J. Guo and S. Zafeiriou. Single-Stage Joint Face Detection and Alignment. In proceedings of the *IEEE/CVF International Conference on Computer Vision Workshop*, doi: 10.1109/ICCVW.2019.00228, 2019. 2

[14] T. Denemark, M. Boroumand and J. Fridrich. Steganalysis Features for Content-Adaptive JPEG Steganography. *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 8, pag. 1736–1746, 2016. 2, 4, 5

[15] L. Du, M. Yi, E. Blasch and H. Ling. GARP-face: Balancing privacy protection and utility preservation in face de-identification. In proceedings of the *IEEE International Joint Conference on Biometrics*, doi: 10.1109/BTAS.2014.6996249, 2014. 2

[16] 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, pag. 1168–1174, 2008. 2, 3

[17] P. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pag. 1627–1645, 2010. 6

[18] O. Gafni, L. Wolf and Y. Taigman. Live Face De-Identification in Video. In proceedings of the *IEEE International Conference on Computer Vision*, pag. 9378–9387, https://arxiv.org/abs/1911.08348v1, 2019. 3

[19] R. Gross, L. Sweeney, J. Cohn, F. de la Torre and S. Baker. Face De-identification. In: Senior A. (eds) Protecting Privacy in Video Surveillance. Springer, doi: 10.1007/978-1-84882-301-3_8, 2009. 2

[20] X. Gu, W. Luo, M. Ryoo and Y. Lee. Password-conditioned Anonymization and Deanonymization with Face Identity Transformers. *ArXiv*, https://arxiv.org/abs/1911.11759, 2019. 3

[21] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. Courville. Improved training of Wasserstein GANs. In proceedings of the *Advances in Neural Information Processing Systems* conference, pag. 5769–5779, 2017. 4

[22] K. He, G. Gkioxari, P. Dollar and R. Girshick, Mask r-CNN. In proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, pag. 2961–2969, 2017.

[23] Z. He, W. Zul, M. Kan, S. Shan and X. Chen. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing*, vol. 28, no. 11, pag. 5464–5478, 2019. 3

[24] H. Hukkelas, R. Mester and F. Lindseth. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In proceedings of the *International Symposium on Visual Computing*, Lecture Notes in Computer Science, vol 11844, doi: 10.1007/978-3-030-33720-9_44, 2019. 2

[25] P. Isola, J-Y. Zhu, T. Zhou and A. Efros. Image-to-image translation with conditional adversarial networks. In proceedings of the *IEEE International Conference on Biometrics*, doi: 10.1109/ICB.2015.7139096, 2015. 2, 6

[26] A. Jourabloo, X. Yin and X. Lu. Attribute preserved face de-identification. In proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2017.632, 2017. 2

[27] T. Karras, S. Laine and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2019.00453, 2019. 3

[28] I. Korshunova, W. Shi, J. Dambre and L. Theis. Fast face-swap using convolutional neural networks. In proceedings of the *IEEE International Conference on Computer Vision*, doi: 10.1109/ICCV.2017.397 2017 2

[29] S. Kumar, E. Yaghoubi, A. Das, B. Harish and H. Proença. The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, Re-Identification and Search from Aerial Devices. *ArXiv*, https://arxiv.org/abs/2004.02782v1, 2020. 6

[30] Y. Li and S. Lyu. De-identification Without Losing Faces. In proceedings of the *ACM Information Hiding and Multimedia Security Workshop*, doi: 10.1145/3335203.3335719 2019. 3

[31] M. Maximov, I. Elezi and L. Leal-Taixé. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. In proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: https://arxiv.org/abs/2005.09544v1, 2020. (in press) 2, 3

[32] B. Meden, Z. Emersic, V. Struc and P. Peer. k-Same-Net: k-Anonymity with Generative Deep Neural Networks for Face De-identification.*MDPI Entropy*, vol. 20, no. 60, doi: 10.3390/e20010060, 2018. 2

[33] Mrityunjay and P. Narayanan. The De-Identification Camera. In proceedings of the *Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, pag. 192–195, 2011. 3

[34] E. Newton, L. Sweeney and B. Malin. Preserving privacy by de-identifying facial images.*IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pag. 232–243, 2005. 2

[35] C. Neustaedter, S. Greenberg and M. Boyle. Blur Filtration Fails to Preserve Privacy for Home-Based Video Conferencing. *ACM Transactions on Computer Human Interaction*, vol. 13, issue 1, pag. 1–36 2006. 1, 2

[36] P. Phillips. Privacy operating characteristic for privacy protection in surveillance applications. Audio- and Video-Based Biometric Person Authentication (T. Kanade, A. Jain, and N. Ratha, eds.), Lecture Notes in Computer Science, Springer pag. 869–878, 2005. 2

[37] H. Proença and João C. Neves. Soft Biometrics: Globally Coherent Solutions for Hair Segmentation and Style Recognition based on Hierarchical MRFs. *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pag. 1637–1645, doi: 10.1109/TIFS.2017.2680246, 2017. 10

[38] G-J. Qi, L. Zhang, H. Hu and M. Edraki. Global versus Localized Generative Adversarial Nets. In proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2018.00164, 2018. 1, 3

[39] G-J. Qi. Loss-Sensitive General Adversarial Networks on Lipschitz Densities, *International Journal of Computer Vision*, vol. 128, pag 1118–1140, doi: 10.1007/s11263-019-01265-2, 2020. 1

[40] J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement. https://arxiv.org/abs/1804.02767v1, 2018.

[41] K. Regmi and A. Borji. Cross-view image synthesis using conditional GANs. In proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2018.00369, 2018. 2

[42] S. Ren, K. He, R. Girshick and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pag. 1137–1149, 2017. 2, 4, 5

[43] Z. Ren, Y. Lee and M. Ryoo. Learning to Anonymize faces for Privacy Preserving Action Detection. In proceedings of the *European Conference on Computer Vision*, pag. 639–655, 2018. 3

[44] O. Ronneberger, P. Fischer and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597, 2015 2, 6

[45] M. Ryoo, B. Rothrock, C. Fleming and H. Yang. Privacy-preserving human activity recognition from extreme low resolution. In proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pag. 4255–4262, 2017. 6, 7

[46] N. Ruiz, E. Chong and J. Rehg. Fine-Grained Head Pose Estimation Without Keypoints. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, doi: 10.1109/CVPRW.2018.00281, 2018. 10

[47] B. Samarzija and S. Ribaric. An Approach to the De-Identification of Faces in Different Poses. In proceedings of *Special Session on Biometrics, Forensics, De-identification and Privacy Protection*, pag. 21–26, 2014. 2, 3

[48] 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 proceedings of the *IEEE/RSJ International Conference on Intelligent Robots and Systems*, doi: 10.1109/IROS.2007.4399122, 2009. 3

[49] J. Seo, S. Hwang and Y-H. Suh. A Reversible Face De-Identification Method based on Robust Hashing. In proceedings of the *International Conference on Consumer Electronics*, doi: 10.1109/ICCE.2008.4587904, 2008. 2

[50] A. Senior. Privacy Protection in a Video Surveillance System. *Protecting Privacy in Video Surveillance*, pag. 35–47, doi: 10.1007/978-1-84882-301-3_3, 2009. 1

[51] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems*, vol. 10, no. 5, pag. 557–570, 2002. 2

[52] W. Shen and R. Liu. Learning residual images for face attribute manipulation. In proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2017.135, 2017. 3

[53] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt and B. Schiele. A hybrid model for identity obfuscation by face replacement. In proceedings of the *European Conference on Computer Vision*, pages 570–586, 2018. 3

[54] Q. Sun, L. Ma, S. Joon, L. Van Gool, B. Schiele and M. Fritz. Natural and effective obfuscation by head inpainting. In proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: *10.1109/CVPR.2018.00530*, 2018. 3

[55] D. Tao, Y. Guo, Y. Li and X. Gao. Tensor Rank Preserving Discriminant Analysis for Facial Recognition.*IEEE Transactions on Image Processing*, vol. 27, no. 1, pag. 325–334, 2018. 1

[56] D. Tao, J. Cheng, K. Yue and L. Wang. Domain-Weighted Majority Voting for Crowdsourcing.*IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pag. 163–174, 2019. 1

[57] W. Wu, D. Tao, H. Li, Z. Yang and J. Cheng. Deep features for person re-identification on metric learning.*Pattern Recognition*, vol. 110, doi: 10.1016/j.patcog.2020.107424, 2021. 1

[58] T. Winkler and B. Rinner. TrustCAM: Security and Privacy-Protection for an Embedded Smart Camera based on Trusted Computing. In proceedings of the *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, pag. 593 – 600, 2010. 3

[59] L. Wolf, T. Hassner and I. Maoz. Face Recognition in Unconstrained Videos with Matched Background Similarity. In proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: *10.1109/CVPR.2011.5995566*, 2011. 6

[60] T. Xiao, J. Hong and J. Ma. ELEGANT: Exchanging Latent Encodings with GAN for Transferring Multiple Face Attributes. In proceedings of the *European Conference on Computer Vision*, doi: 10.1007/978-3-030-01249-6_11, 2018. 3

[61] M. Yamac, M. Ahishali, N. Passalis, J. Raitoharju, B. Sankur and M. Gabbouj. Reversible Privacy Preservation using Multi-level Encryption and Compressive Sensing. In proceedings of the *27th European Signal Processing Conference*, doi: 10.23919/EUSIPCO.2019.8903056, 2019. 3

[62] B. Yan, M. Pei and Z. Nie. Attributes Preserving Face De-Identification. In proceedings of the *IEEE International Conference on Computer Vision*, dos: 10.1109/ICCVW.2019.00154, 2019. 2

[63] W. Yang, K. Wang and W. Zuo. Neighbourhood Component Feature Selection for High-Dimensional Data. *Journal of Computers*, vol. 7, no. 1, pag. 161–168, 2012. 9

[64] 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, pag. 1499–1503, 2016. 7

[65] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang and D. Metaxas. Stack-GAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In proceedings of the *IEEE International Conference on Computer Vision*, doi: 10.1109/ICCV.2017.629, 2017. 2

[66] F. Zhang, T. Zhang, Q. Mao and C. Xu. Joint Pose and Expression Modelling for Facial Expression Recognition. In proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: *10.1109/CVPR.2018.00354*, 2018. 10, 11

[67] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang and Q. Tian. MARS: A Video Benchmark for Large-Scale Person Re-Identification. In proceedings of the *European Conference on Computer Vision*, part VI, pag. 868–884, 2016. 6

[68] Y. Zhong, J. Sullivan and H. Li. Face attribute prediction using off-the-shelf deep learning networks. CoRR, abs/1602.03935, 2016. 3

[69] S. Zhu, R. Urtasun, S. Fidler, D. Lin and C. Loy. Be your own prada: Fashion synthesis with structural coherence. In proceedings of the *IEEE International Conference on Computer Vision*, doi: 10.1109/ICCV.2017.186, 2017. 2

[70] J-Y. Zhu, T. Park, P. Isola and A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In proceedings of the *IEEE International Conference on Computer Vision*, doi: 10.1109/ICCV.2017.244, 2017. 3

[71] X. Zhu, H. Liu, Z. Lei, H. Shi, F. Yang, D. Yi, G. Qi and S. Li. Large-Scale Bisample Learning on ID versus Spot Face Recognition. *International Journal of Computer Vision*, vol. 127, no. 6-7, doi: 10.1007/s11263-019-01162-8, 2019. 1