Cross-Domain Face Synthesis using a Controllable GAN

Fania Mokhayeri Kaveh Kamali Eric Granger
Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
Dept. of Systems Engineering, École de technologie supérieure, Montreal, Canada

Abstract

The performance of face recognition (FR) systems for video surveillance has been shown to improve when the design data is augmented through synthetic face generation. This is true, for instance, with pair-wise matchers (e.g., deep Siamese networks) that rely on a reference gallery, typically with one still image per individual. However, generating synthetic images based on stills (from the source domain) may not improve performance during operations due to the domain shift w.r.t. the target domain. Moreover, despite the emergence of Generative Adversarial Networks (GANs) for realistic synthetic generation, it is often difficult to control the conditions under which synthetic faces are generated. In this paper, a cross-domain face synthesis approach is proposed that integrates a new Controllable GAN (C-GAN). It employs an off-the-shelf 3D face model as a simulator to generate facial images under various poses. The simulated images and noise are input to the C-GAN for realism refinement. It relies on an additional adversarial game as a third player to preserve the identity and specific facial attributes of the refined images. This allows generating realistic synthetic face images that reflect capture conditions in the target domain, while controlling the GAN output such that faces may be generated under desired pose conditions. Experiments were performed using videos from the Chokepoint and COX-S2V datasets, and a deep Siamese network for FR with a single reference still per person. Results indicate that the proposed approach can provide a higher level of accuracy compared to state-of-the-art approaches for synthetic data augmentation[1].

1. Introduction

Recent advances in deep learning have significantly increased the performance of still-to-video face recognition (FR) systems applied in video monitoring and surveillance. One of the pioneering techniques in this area is FaceNet [25]. It uses a deep Siamese network architecture, where the same CNN feature extractor can be trained through similarity learning to perform pair-wise matching between query (video) and reference (still) faces. Despite many recent advances, FR with a single sample per person (SSPP) remains a challenging problem in video-based security and surveillance applications. In such cases, the performance of deep learning models for FR can decline significantly due to the limited robustness of matching to a single reference still captured during enrolment [2]. One effective solution to alleviate the aforementioned problem is extending the gallery using synthetic face images.

Some recent research [19, 28, 20, 22] augments reference galleries using synthetic images generated from 3D models. Tran et al. [30] proposed a face synthesis technique where CNN is employed to regress the 3D model parameters to overcome the shortage of training data. Although their results are encouraging, the synthetic face images may not be realistic enough to represent intra-class variations of target domain capture conditions. The synthetic images generated in this way are highly correlated with the original facial stills from enrolment, and there is typically a domain shift between the distribution of synthetic faces and that of faces captured in the target domain. Models trained on these synthetic images, often fail to generalize well when matched to real images captures in the target domain. Mokhayeri et al. [21] proposed an algorithm for domain-specific face synthesis (DSFS) that exploits the intra-class variation information available from the target domain.

Generative adversarial networks (GANs) have recently shown promising results for the synthesis of realistic face images [1, 4, 9]. For instance, DA-GAN [33] has been proposed for automatically generating augmented data for FR in unconstrained conditions. One of the challenging issues in GAN-based face synthesis models is the difficulty of controlling images they generate since a random distribution is used as the input to generators. Modified GAN architectures, like the conditional GAN, have attempted to address this issue by setting conditions on the generative and discriminative networks for image synthesis [15, 17, 29]. However, the mapping of conditional GANs does not con-

[1]Code is available: https://github.com/faniamokhayeri/C-GAN.

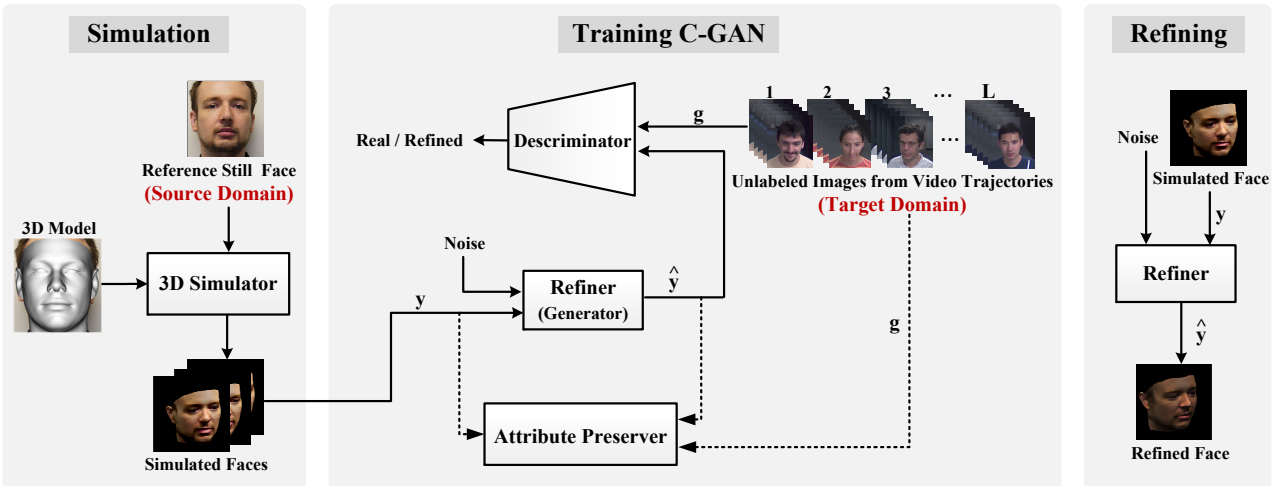


Figure 1. An overview of the proposed cross-domain face synthesis approach based on the C-GAN. The 3D simulator generates simulated faces, $\mathbf{y}$, with the arbitrary pose. The refiner is trained using $\mathbf{y}$, the generic set, $\mathbf{g}$, and random noise to generate refined images, $\hat{\mathbf{y}}$, under the target domain capture conditions, and while specifying the pose of $\mathbf{y}$ using an additional adversarial game.

strain the output to the target manifold, thus the output can be arbitrarily off the target manifold. Generating identity-preserving faces is another unsolved challenge with GAN models for face synthesis.

In this paper, we propose a cross-domain face synthesis approach that relies on a new controllable GAN (C-GAN). It extends the original GAN by using an additional adversarial game as the third player to the GAN, competing with the refiner (generator) to preserve the specific attributes, and accordingly, providing control over the face generation process. As depicted in Figure 1, C-GAN involves three main steps: (1) generating simulated face images via 3D morphable model [3] rendered under a specified pose, (2) refining the realism of the simulated face images using an unlabeled generic set to adapt synthetic face images from the source domain to appear as if drawn from the target domain, and (3) preserving the specific attributes of the simulated face images during the refinement through another adversarial network. Using C-GAN, a set of realistic synthetic facial images are generated that represent gallery stills under the target domain with high consistency, while preserving their identity and allowing to specify the pose conditions of synthetic images. The refined synthetic face images are then used to augment the reference gallery to boost the performance of FR with SSPP. The main contribution of this paper is a novel cross-domain face synthesis approach that integrates C-GAN to leverage an additional adversarial game as third player into the GAN model, producing highly consistent realistic face images in a controllable manner. Additionally, we show that using the images generated by C-GAN as additional design data within a Siamese network allows to improve still-to-video FR performance under unconstrained capture conditions.

For proof-of-concept experiments, the performance of the proposed and baseline face synthesis methods are evaluated using a "recognition via generation" framework [33] on videos from the publicly available Chokepoint and COX-S2V datasets. In a particular implementation, we extend the reference gallery of a deep Siamese network for still-to-video FR.

2. Related Work

GANs for Realistic Face Synthesis. Recently, Generative Adversarial Networks (GANs) [9] have shown promising performance in face synthesis. Existing methods typically formulate GAN as a two-player game, where a discriminator D distinguishes face images from the real and synthesized domains, while a generator G reduces its capacity to discriminate by synthesizing a face of realistic quality. Their competition converges when the discriminator is unable to differentiate these two domains. Benefiting from GAN, the FaceID-GAN [26] treats a face identity classifier as the third player, competing with the generator by distinguishing the identities of the real and synthesized faces. The major shortcoming of GAN models for face synthesis is that they may produce images that are inconsistent due to the weak global constraints. To reduce this gap, Shrivastava *et al.* developed SimGAN that learns a model using synthetic images as inputs instead of random noise vectors [27]. Our work draws inspiration from SimGAN specialized with face synthesis. With vanilla GAN, it is difficult to control the generator output. Recently, conditional GANs have added condition information to the generative network, and the discriminative network for conditional image synthesis [15, 17]. Tran *et al.* proposed DR-GAN that inputs a pose code and a random noise vector to the dis-

criminator in order to generate a face of the same identity with the target pose that can fool the discriminator [29]. In the CAPG-GAN [13], a couple-agent discriminator is introduced which forms a mask image to guide the generator during the learning process, and provides a flexible controllable condition during inference. The bottleneck of conditional GANs is the regression of the generator may lead to arbitrarily large output errors, which makes it unreliable for real-world applications [6]. This paper aims to address the above problems by augmenting the GAN refiner with a domain-invariant feature extractor.

Domain-Invariant Representations. Recently, some methods have been developed to produce domain-invariant feature representations from a single input [34]. One of the most popular approaches in this area is the domain-adversarial neural network which integrates a gradient reversal layer into the standard architecture to ensure a domain-invariant feature representation [7]. They introduced a domain confusion loss term to learn domain-invariant features. Haeusser *et al.* [10] produce statistically domain-invariant embedding by reinforcing associations between source and target data directly in embedding space. A slightly different approach is presented in [8], where common feature assimilation is achieved implicitly by using a decoder to reconstruct the input source and target images. In a similar spirit, [24] uses a generator from the encoded features to generate samples which follow the same distribution as the source dataset.

3. Proposed Approach

In the following, the set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{d \times N}$ denote a gallery set composed of n reference still ROIs belonging to one of k different classes in the source domain, where d is the number of pixels representing a ROI and $N = kn$ is the total number of reference still ROIs; $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_M\} \in \mathbb{R}^{d \times M}$ and $\mathbf{H} = \{\mathbf{h}^1, \ldots, \mathbf{h}^i, \ldots, \mathbf{h}^M\} \in \mathbb{R}^{(kp) \times M}$ denote the simulated set and the corresponding one-hot labels, where M is the number of the simulated ROIs and kp is number of 3D simulated classes (k identity class with p pose). The label associated with $\mathbf{y}_i$ is defined as $\mathbf{h}^i = \{h_d^i, h_p^i\}$, where h_d represents the label for identity and h_p for pose. $\mathbf{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_i, \ldots, \mathbf{g}_L\} \in \mathbb{R}^{d \times L}$ denote a generic set composed of L unlabeled video ROIs in the target domain. The objective of C-GAN model is to generate realistic face images with high consistency while specifying the main attributes, in particular pose h_p and identity, shown in synthetic images and preserving the identity h_d.

Figure 2 depicts the overall C-GAN process within the approach for cross-domain face synthesis. Our approach is divided into three stages: (1) 3D simulation, (2) training the refiner, (3) refiner inference. In the first stage, the 3D model of each reference still image is reconstructed via a 3D simulator and rendered under a specified pose. The rendered images, $\mathbf{Y}$, are then imported to the refiner (generator) to recover the information inherent in the target domain. In contrast to the vanilla GAN formulation [9], in which the generator is conditioned only on a noise vector, our model's generator is constrained on both a noise vector (z) and simulated image.

During the training stage, the refiner is trained to produce realistic images through an adversarial game with a discriminator network, D_R. The discriminator classifies a refined image as real/fake image. The refiner is further encouraged to generate realistic images while preserving the identity and capture conditions of $\mathbf{Y}$ by augmenting the refiner with a domain invariant feature extractor [7]. The feature extracting is applied on both input and output of the refiner and the Euclidean distance of the two features is considered as an additional loss. The feature extractor F must be invariant with respect to $\mathbf{Y}$ and $\mathbf{G}$ while including all identity and pose information. For this purpose, an additional adversarial game between another discriminator and the feature extractor is employed to train the feature extractor to be domain invariant. The second discriminator D_F takes the output of the domain-invariant feature extractor and distinguish between the features extracted from the the refined images and real images. In order to guarantee that the extracted features include all the information of identity and pose, an identity-pose classifier predicts identity and pose of the refined images while being trained simultaneously on the labeled 3D simulated images, $\mathbf{Y}$. In this way, the target domain variations are effectively transferred onto the reference still images while specifying the pose shown in synthetic images, and without losing the consistency. The refiner in the proposed C-GAN shares ideas with methods for unsupervised domain adaptation [7], where labeled still images from the source domain and unlabeled video images from the target domain are used to learn a domain-invariant embedding. We minimize the difference between the refined images and generic set while preserving the joint distribution information (on identity and pose). To stabilize the training process of such dual-agent GAN model, we impose a boundary equilibrium regularization term. Once the synthetic images have been generated, they can be leveraged to augment the reference gallery of any pair-wise matcher like FaceNet for still-to-video FR.

3.1. 3D Simulator:

The simulated image set, $\mathbf{Y} \in \mathbb{R}^{d \times M}$, is formed by reconstructing the 3D face model of reference ROIs, $\mathbf{x}_i$, using a customized version of the 3DMM [3] in which the texture fitting of the original 3DMM is replaced with image mapping for simplicity [21]. The shape model is defined as a

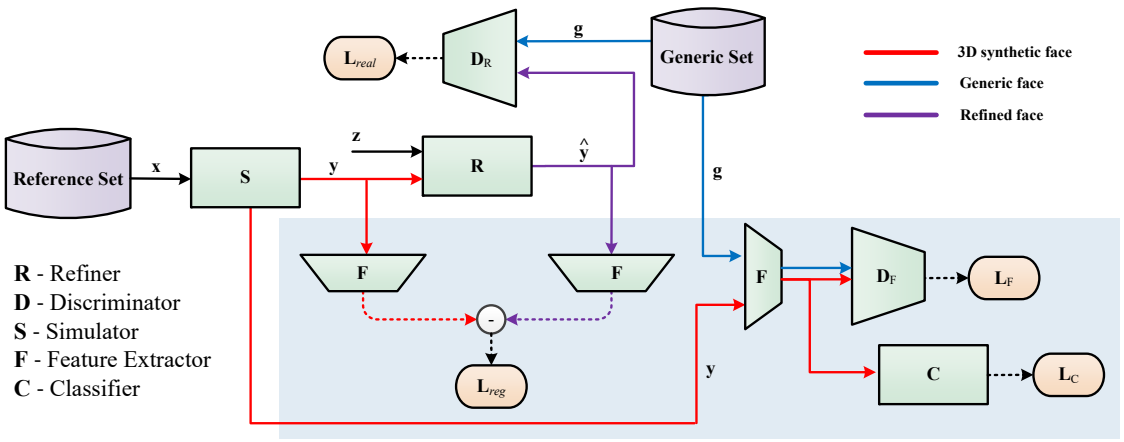


Figure 2. Illustration of our proposed C-GAN architecture. It incorporates a simulator, S, a refiner, R, two discriminators, D_R and D_F, and a classifier, C, constrained by the simulated and noise vector. The blue box indicates the attribute-preserving module.

convex combination of shape vectors of a set of examples:

$$\mathbf{s} = \bar{\mathbf{s}} + \sum_{k=1}^{m-1} \alpha_k . \hat{\mathbf{s}}_k \ , \tag{1}$$

where $\hat{\mathbf{s}}_k$, $1 \leq k \leq m$ is the basis vector, m is the number of the basis vectors, and $\alpha_k \in [0, 1]$ are the shape parameters. The optimization algorithm presented in [3] is employed to find optimal α_k^i, for each $\mathbf{x}_i$. Then, the material layer of $\mathbf{x}_i$ is extracted and projected to the 3D geometry of $\mathbf{x}_i$. Given the 3D facial shape and texture, novel poses can be rendered under various pose by adjusting the parameters of a camera model. During the rendering procedure, the 3D face is projected onto the image plane with weak perspective projection:

$$\mathbf{y}^j = f * \lambda * \mathbf{R}^j * (\bar{\mathbf{s}} + \sum_{k=1}^{m-1} \alpha_k^i . \hat{\mathbf{s}}_k) + \mathbf{t}_{2d}^j \ , \tag{2}$$

where $\mathbf{y}^j$ is the j^{th} reconstructed pose of $\mathbf{x}_i$, f is the scale factor, λ is the orthographic projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, $\mathbf{R}^i$ is the rotation matrix constructed from rotation angles and $\mathbf{t}_{2d}^j$ is the translation vector.

3.2. C-GAN Network Structure:

The main part of C-GAN is the refiner (R) which improves the realism of a 3D simulator's output, $\mathbf{Y}$, using unlabeled images in the target domain, $\mathbf{G}$, while controlling their specific facial appearance (e.g. pose). The refiner, R, consists of an encoder R_{enc} and a decoder R_{dec}. The encoder R_{enc} aims to learn an identity and attribute representation from a face image $\mathbf{y}$: $\mathbf{f}(\mathbf{y}) = R_{enc}(\mathbf{y})$. The decoder R_{dec} aims to synthesize a face image $\hat{\mathbf{y}} = R_{dec}(\mathbf{f}(\mathbf{y}), \mathbf{z})$ where $\mathbf{z} \in \mathbb{R}^{N^z}$ is the noise modeling other variance besides identity or attribute (e.g. pose). The goal of R is to fool D_R to classify $\hat{\mathbf{y}}$ as a generic image.

We adopt CASIA-Net [32] for R_{enc} and D_R where batch normalization and exponential linear unit are applied after each convolutional layer. A fully connected layer with logistic loss is added at the output of D_R. R_{enc} and R_{dec} are bridged by the to-be-learned identity representation $\mathbf{f}(\mathbf{y}) \in \mathbb{R}^{320}$, which is the AvgPool output in R_{enc} network. $\mathbf{f}(\mathbf{y})$ is concatenated with a random noise $\mathbf{z}$ and fed to R_{dec}. A series of fractionally-strided convolutions [23] transforms the $(320 + N_z)$-dim concatenated vector into a realistic image $\hat{\mathbf{y}} = R(\mathbf{y}, \mathbf{z})$, which is the same size as $\mathbf{y}$.

An encoder, F, with the same structure as R_{enc} is used at the input and output of R to compare the domain invariant features of the simulated images to the refined images. As mentioned before, the discriminator, D_F, and classifier, C, are used to train F. Both C and D_F consist two fully-connected 1024 unit layers. A fully-connected softmax loss for k identity and p pose classification is added to C while a 1 unit fully-connected logistic layer is added to D_F. Table 1 shows the neural network structures of R_{enc}, D_R, F and R_{enc}.

3.3. Training the Refiner:

Let a refined image be denoted by $\hat{\mathbf{y}}_i$, then $\hat{\mathbf{y}} = R(\theta_R; \mathbf{y})$ where θ_R is the function parameters. The key requirement is that the refined image, $\hat{\mathbf{y}}$, must look like a real generic image preserving the identity and pose information from the 3D simulator. To this end, θ_R is learned by minimizing a combination of two losses:

$$\mathcal{L}_R(\theta_R, \theta_F) = \sum_i \mathcal{L}_{real}(\theta_R; \mathbf{y}_i, \mathbf{G}) + \lambda \mathcal{L}_{reg}(\theta_F; \mathbf{y}_i) \tag{3}$$

The first part of the cost, $\mathcal{L}_{real}$, adds realism to the simulated images, while the second part, $\mathcal{L}_{reg}$, preserves the identity and pose information.

The adversarial loss used for training the refiner network, R, is responsible for fooling D_R into classifying the refined

Table 1. Network structure of the proposed C-GAN architecture.

R_{enc} and D_R			R_{dec}		
Layer	Filter/Stride	Output Size	Layer	Filter/Stride	Output Size
			FConv52	$3 \times 3/1$	$6 \times 6 \times 320$
Conv11	$3 \times 3/1$	$96 \times 96 \times 32$	FConv52	$3 \times 3/1$	$6 \times 6 \times 160$
Conv12	$3 \times 3/1$	$96 \times 96 \times 64$	FConv51	$3 \times 3/1$	$6 \times 6 \times 256$
Conv21	$3 \times 3/2$	$48 \times 48 \times 64$	FConv43	$3 \times 3/2$	$12 \times 12 \times 256$
Conv22	$3 \times 3/1$	$48 \times 48 \times 64$	FConv42	$3 \times 3/1$	$12 \times 12 \times 128$
Conv23	$3 \times 3/1$	$48 \times 48 \times 128$	FConv41	$3 \times 3/1$	$12 \times 12 \times 192$
Conv31	$3 \times 3/2$	$24 \times 24 \times 128$	FConv33	$3 \times 3/2$	$24 \times 24 \times 192$
Conv32	$3 \times 3/1$	$24 \times 24 \times 96$	FConv32	$3 \times 3/1$	$24 \times 24 \times 96$
Conv33	$3 \times 3/1$	$24 \times 24 \times 192$	FConv31	$3 \times 3/1$	$24 \times 24 \times 128$
Conv41	$3 \times 3/2$	$12 \times 12 \times 192$	FConv23	$3 \times 3/2$	$48 \times 48 \times 128$
Conv42	$3 \times 3/1$	$12 \times 12 \times 128$	FConv22	$3 \times 3/1$	$48 \times 48 \times 64$
Conv43	$3 \times 3/1$	$12 \times 12 \times 256$	FConv21	$3 \times 3/1$	$48 \times 48 \times 64$
Conv51	$3 \times 3/2$	$6 \times 6 \times 256$	FConv13	$3 \times 3/2$	$96 \times 96 \times 64$
Conv52	$3 \times 3/1$	$6 \times 6 \times 160$	FConv12	$3 \times 3/1$	$96 \times 96 \times 32$
Conv53	$3 \times 3/1$	$6 \times 6 \times 320$	FConv11	$3 \times 3/1$	$96 \times 96 \times 1$
AvgPool	$6 \times 6/1$	$1 \times 1 \times 320$			
FC (D_R only)		1			

images as real. This problem is modeled a two-player minimax game, and update the refiner network, R, and the discriminator network. D_R updates its parameters by minimizing the following loss:

$$\mathcal{L}_D(\phi) = -\sum_i \log(D_R(\phi; \hat{\mathbf{y}}_i)) - \sum_j \log(1 - D_R(\phi; \mathbf{g}_j)) \tag{4}$$

where $D_R(\cdot)$ is the probability of the input being a refined image, and $1 - D_R(\cdot)$ that of a real one. For training this network, each mini-batch consists of randomly sampled $\hat{\mathbf{y}}_\mathbf{i}$ and $\mathbf{g}_\mathbf{i}$.

The realism loss function employs the trained discriminator D_R as follows:

$$\mathcal{L}_{real}(\theta_R) = \log(1 - D_R(R(\theta_R; \mathbf{y}_i))) \tag{5}$$

By minimizing this loss function, the refiner forces the discriminator to fail classifying the refined images as synthetic. In order to preserve the annotation information of the 3D simulator, we use a self-regularization loss that minimizes difference between a feature transform of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$,

$$\mathcal{L}_{reg}(\theta_F) = \|F(\hat{\mathbf{y}}_i, \theta_F) - F(\mathbf{y}_i, \theta_F)\| \tag{6}$$

where F is the mapping from image space to a feature space, and $\|\cdot\|$ is the ℓ_2 norm.

Another adversarial game is employed to train the feature extractor network parameters (θ_F). For this purpose, the classifier, $C(.)$, assigns identity and pose information labels ($\mathbf{h}^i$) to a set of features extracted by F. In this way, F learns to extract the features that are domain-invariant and consist information of identity and pose. F and C are updated based on the identity and pose labels of $\mathbf{Y}$ in a traditional supervised manner. F is also updated using the adversarial gradients from D_F so that the feature learning and image generation processes co-occur smoothly.

$$\mathcal{L}_C(\theta_C) = -\sum_i \sum_{j=1}^{c} \mathbf{h}_j^i \log(C(\theta_C; F(\hat{\mathbf{y}}_i))) \tag{7}$$

$$\mathcal{L}_{D_F}(\gamma) = -\sum_i \log(D_F(\gamma; F(\hat{\mathbf{y}}_i))) - \sum_i \log(1 - D_F(\gamma; F(\mathbf{g}_j))) \tag{8}$$

Given a realistic simulated images $\hat{\mathbf{y}}_i$ as input, D_F outputs a binary distribution optimized by minimizing a binary cross entropy loss $\mathcal{L}_F$.The gradients are generated using the following loss functions:

$$\mathcal{L}_F(\theta_F) = \sum_i \log(1 - D_F(F(\theta_F \hat{\mathbf{y}}_i))) \tag{9}$$

where the F and D_F parameters are learned by minimizing $\mathcal{L}_F(\theta_F)$ and $\mathcal{L}_{D_F}(\gamma)$ alternately. We leave γ fixed while updating the parameters of F, and we fix θ_F while updating D_F.

4. Experimental Analysis

4.1. Evaluation Methodology:

The performance of the proposed and baseline methods was evaluated using two datasets for still-to-video FR. Both are comprised of a still image and several videos per person. The **Chokepoint** [31] consists of 25 subjects walking through portal 1 and 29 subjects in portal 2. Videos are recorded over 4 sessions one month apart. An array of 3 cameras are mounted above portal 1 and portal 2 that capture the entry of subjects during 4 sessions. The videos have frame rate of 30 fps and the image resolution is 800x600 pixels. In total, the dataset consists of 54 video sequences and 64,204 face images. The **COX-S2V** dataset [14] contains 1000 individuals, with 1 high-quality still image and 4 lower-resolution video sequences per individual simulating video surveillance scenario. The video frames are captured by 4 cameras mounted at fixed locations. In each video, an individual walks through a designed S-shape route with changes in illumination, scale, and pose.

FR performance under SSPP scenario was assessed via the "recognition via generation" framework to validate our hypothesis that by adding photo-realistic synthetic faces to a references gallery set can address the visual domain shift, and accordingly improve the accuracy. Since photographic results are also an indicator of qualitative performance, the visual quality is also compared in our experiment. Additionally, we compare our results with those obtained by flow-based Frontalization [11]. During the enrollment of an individual to the system, q simulated ROIs for each reference still ROI are generated under different poses using the conventional 3DMM [3]. The images are then refined

using the controlled GAN that projects the capture conditions of the target domain on them while preserving their pose and identity. The gallery is formed using the original reference still ROIs along with the corresponding synthetic ROIs. During the operational phase, FR is performed using a deep Siamese network that is pre-trained using the VGG-Face2 dataset with Inception Resnet V1 architecture. The CNN feature extractors in this model is trained using stochastic gradient descent and AdaGrad with standard back-propagation [25]. Finally, given the CNN feature vectors produced for the query (video) and reference (still) faces, pair-wise matching is performed using the k-NN classifier based on Euclidean distance.

In all experiments with Chokepoint and COX-S2V datasets, 5 and 20 target individuals are selected, respectively, to populate the watch-list, using one high-quality still image. Videos of 10 and 50 of unlabeled individuals of Chokepoint and COX-S2V datasets, respectively, along with videos of the individuals who are already enrolled in the watch-list are used for testing. The rest of videos individuals that are assumed to come from unlabeled persons are used as a generic set. In order to obtain representative results, this process is repeated 5 times with a different random selection of watch-lists and the average performance is reported with standard deviation over all the runs. The average performance of the proposed and baseline system for still-to-video FR is presented by measuring the partial area under ROC curve, pAUC(20%) (using the AUC at $0 < FPR \leq 20\%$), and mean average precision, mAP. We further employed the Frechet Inception Distance (FID) [12] to quantitatively measure the quality of synthetic faces.

4.2. Results and Discussion:

Figure 3 shows examples of the synthetic face images generated using our proposed cross-domain face synthesis approach on 4 reference still ROIs and a generic set of the Chokepoint dataset. In these examples, the ROIs refined using C-GAN are shown to preserve their pose variations. Figure 4 compares the qualitative results obtained with state-of-the-art techniques. Synthetic images are generated for (a) reference still ROIs using: (b) 3DMM [3], (c) 3DMM-CNN [30], (d) DSFS [21], and (e) our proposed approach with C-GAN.

Table 2 shows the average accuracy of a deep Siamese network for still-to-video FR that relies on the proposed and baseline methods for generating synthetic face images to augment the reference gallery. The baseline system is designed with an original reference still ROI alone. For our proposed approach, the synthetic faces are generated with 5° step size within a range of ± 5 to ± 60 degrees in yaw, pitch, and roll. Consequently, we generate a total of $q = 73$ synthetic face images in our experiments. For reference, the still-to-video FR system based on frontalization is also evaluated. Results indicate that by adding extra synthetic ROIs generated with C-GAN allows to outperform baseline systems. The pAUC and mAP accuracy increases by about 3%, typically with $q = 73$ synthetic pose ROIs for Chokepoint and COX-S2V datasets. Results suggest that that leveraging target domain information within the GAN framework while controlling its pose and identity can efficiently mitigate the impact of visual domain shift.

Figure 5 shows the average pAUC(20%) and mAP accuracy obtained with the deep Siamese network of still-to-video FR when increasing the number of synthetic ROIs per each individual. Adding synthetic ROIs generated under various capture conditions allows to significantly outperform the baseline system designed with the original reference still ROI alone. As shown in Figure 5, accuracy trends to stabilize to its maximum value when the size of the synthetic faces is greater than $q = 73$ with our approach.

Frechet Inception Distance (FID) [12][5] has recently been proposed to evaluate the performance of image synthesis tasks quantitatively where lower FID score indicates the smaller Wasserstein distance between two distributions. Inception V3 model is employed to extract feature vectors from images. Table 3 show the FID between the real and the synthesized faces across different yaw which demonstrates the effectiveness of our proposed cross-domain synthetic generation method.

To further evaluate the effectiveness of the refiner in our C-GAN, we use the t-SNE [18] projection to visualize the deep features of simulated, refined and real faces in a 2D space. Figure 6 shows there is significant difference between the distribution of 3D simulated and real face. However, after refining the 3D simulated images, the distributions of refined and real images become more similar.

Figure 7 compares the performance of Siamese networks for FR when adding 73 selected synthetic ROIs generated with the proposed, versus 73 randomly selected images (without condition). For reference, FR based on 3DMM face synthesizing is also evaluated. Results in this figure indicate that the C-GAN with a specified range outperforms other models. FR accuracy is higher when the gallery is designed using the representative views than based gallery comprised of randomly selected synthetic faces per person. The proposed C-GAN can therefore adequately generate representative facial ROIs for the reference gallery.

Ablation Study. To evaluate the components of C-GAN (D_F , D_R , C), the model is trained by removing these modules while fixing the training process and all parameters. Recognition accuracy is evaluated on the synthetic images generated from each variant. We observe (Table 4) that the accuracy will decrease by about 3% if one module is not used.

Complexity: Time complexity is estimated empirically as the amount of time required to match a pair of facial ROIs.

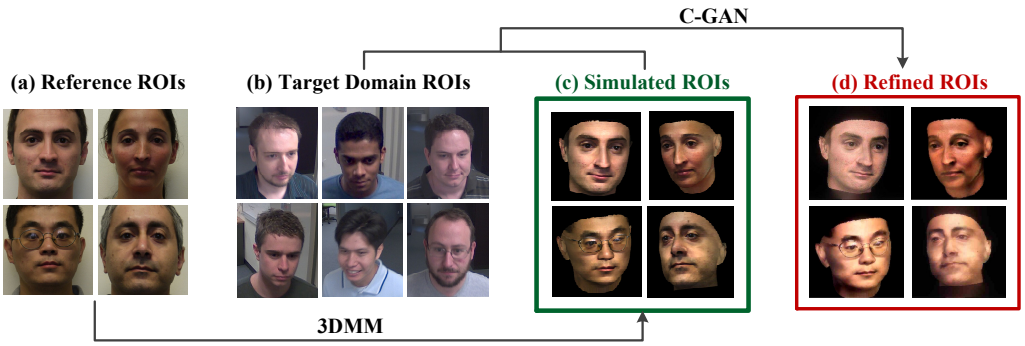


Figure 3. Examples of the synthetic faces obtained with the proposed approach on Chokepoint database (ID#1, ID#5, ID#6, ID#16). The simulated images (c) are refined based on the target domain capture conditions (b) while preserving the identity of reference stills (a) under specific pose.

Table 2. Average pAUC and mAP accuracy of the Siamese network using the proposed and baseline methods for data augmentation. The '# synth' columns show the minimum number of synthetic samples needed to attain the highest level of accuracy.

Techniques	**Chokepoint database**			**COX-S2V database**		
	pAUC(20%)	**mAP**	**# Synth**	**pAUC**(20%)	**mAP**	**# Synth**
Baseline	0.908±0.018	0.861±0.020	N/A	0.912±0.017	865±0.016	N/A
3DMM [3]	0.917±0.023	0.877±0.025	73	0.928±0.026	872±0.027	73
3DMM-CNN [30]	0.915±0.025	0.873±0.028	73	0.922±0.024	871±0.028	73
DSFS [2] [21]	0.923±0.018	0.880±0.019	17	0.934±0.021	896±0.022	14
SimGAN [27]	0.942±0.025	0.901±0.023	73	0.948±0.023	904±0.020	73
DR-GAN [29]	0.931±0.019	0.893±0.017	73	0.939±0.016	903±0.017	73
FaceID-GAN [26]	0.936±0.023	0.905±0.019	73	0.942±0.018	911±0.022	73
Frontalization [11]	0.919±0.020	0.884±0.019	N/A	0.926±0.017	892±0.020	N/A
C-GAN (Ours)	0.951±0.023	0.917±0.022	73	0.957±0.019	925±0.019	73

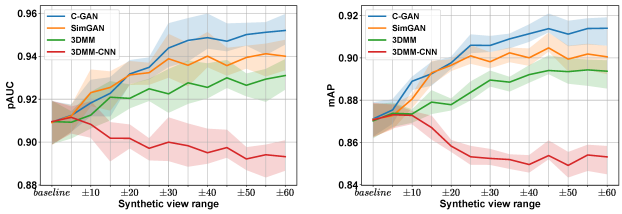
Figure 4. Examples of facial images generated using state-of-the-art face synthesizing methods on Chokepoint dataset (ID#23, ID#25).

Table 3. FID across different views with Chokepoint and COX-S2V datasets.

Technique	**Chokepoint data**			**COX-S2V data**		
	±5°	±15°	±45°	±5°	±15°	±45°
3DMM [3]	22.3	23.4	25.7	20.5	21.4	21.7
3DMM-CNN [30]	49.5	53.2	61.4	42.2	50.7	53.2
DSFS [21]	21.4	22.7	24.5	17.9	21.8	23.1
C-GAN (Ours)	20.9	22.1	23.8	17.3	20.9	21.5

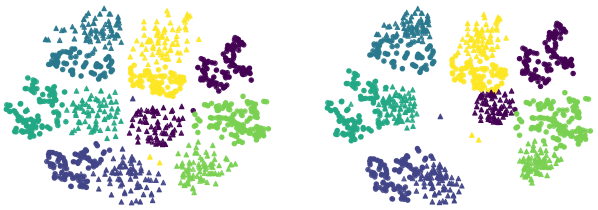
Table 4. The results of ablation study with Chokepoint and COX-S2V datasets.

Accuracy	Removed Module					
	Chokepoint data			COX-S2V data		
	D_F	D_R	C	D_F	D_R	C
pAUC	0.905±0.022	0.901±0.023	0.882±0.020	0.908±0.021	0.902±0.025	0.891±0.027
mAP	0.873±0.024	0.868±0.021	0.854±0.019	0.885±0.018	0.875±0.020	0.859±0.024



(a) (b)

Figure 5. Average pAUC(20%) (a) and mAP (b) accuracy on the Chokepoint database of the proposed and baseline techniques versus the number of synthetic ROIs included in the reference gallery.



(a) (b)

Figure 6. t-SNE visualization. Circles represent the generic set. Triangles in (a) represent 3D simulated faces while triangles in (b) represent refined faces.

Table 5. Average matching time over videos ROIs of the Chokepoint and COX-S2V datasets.

Techniques	Matching Time (sec)	
	Chokepoint	COX-S2V
Siamese Network [16]		
· 1 frontal reference still / person	6.4	13.2
· +73 uniform synthetic / person	129.7	186.1
· +100 random synthetic / person	152.3	211.5
Frontalization [11]	12.7	16.3

The average running time is measured with randomly selected probe ROIs using a PC workstation with an Intel Core i7 CPU (3.41GHz) processor and 16GB RAM. Table 5 shows average matching time of a deep Siamese networks over videos ROIs of the Chokepoint and COX-S2V datasets. The table shows that time complexity grows with the gallery size. Since our approach can provide a compact set or images to improve accuracy, it represents an interesting trade-off between accuracy and complexity.

[2]The DSFS technique employs a clustering on target capture conditions to find the optimal number of samples required for FR.

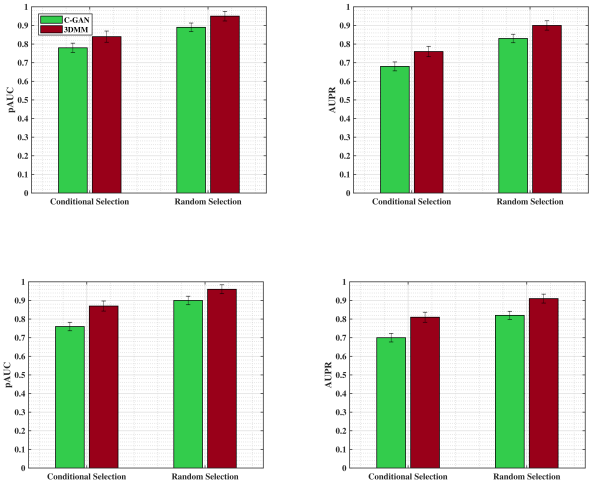


Figure 7. Average pAUC(20%) and AUPR accuracy for Siamese network using the proposed and 3DMM face synthesis with q specified and randomly selected faces on Chokepoint (a,b) and COX-S2V (c,d) datasets. Error bars are standard deviation.

5. Conclusion

In this paper, a cross-domain face synthesis approach with a new C-GAN model is proposed for data augmentation that generates highly consistent, realistic and identity preserving synthetic face images under specific pose conditions. The proposed model allows to mitigate the impact of some common issues with the original GAN model for data augmentation, such as lack of control and inconsistency. C-GAN leverages an additional adversarial game as third player to encourage the refiner during the inference to specify the capture conditions shown in synthetic images in a controllable manner. This allows augmenting to the gallery of a deep Siamese network with a diverse, yet compact set of synthetic views relevant to the target domain. Experimental results obtained using the Chokepoint and COX-S2V datasets suggest that the synthetic face images based on C-GAN allow us address visual domain shift, and thereby improve the accuracy of still-to-video FR system, with no need to generate a large number of synthetic face images. A future direction is to simulate and control other facial appearance (e.g. illumination and expression) during the face synthesis process. This can be further used to augment a dataset with representative images to train a deep neural network for still-to-video FR.

References

[1] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. In *CVPR 2018*.

[2] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau. Dynamic ensembles of exemplar-svms for still-to-video face recognition. *Pattern Recognition*, 69:61–81, 2017.

[3] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[4] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR 2019*.

[5] J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun. 3D aided duet GANs for multi-view face image synthesis. *IEEE Trans. on Information Forensics and Security*, 2019.

[6] G. Chrysos, J. Kossaifi, and S. Zafeiriou. Robust conditional generative adversarial networks. *ICLR 2019*.

[7] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

[8] M. Ghifary, W. Kleijn, M. Zhang, D. Balduzzi, and W. Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV 2016*.

[9] I. Goodfellow and et al. Generative adversarial nets. In *NIPS 2014*.

[10] P. Haeusser, T. Frerix, A. Mordvintsev, and D. Cremers. Associative domain adaptation. In *ICCV 2017*.

[11] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR 2015*.

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS 2017*.

[13] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun. Pose-guided photorealistic face rotation. In *CVPR 2018*.

[14] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen. A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Trans. on Image Processing*, 24(12):5967–5981, 2015.

[15] P. Isola, J. Zhu, T. Zhou, and A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR 2017*.

[16] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML workshop 2015*.

[17] J. Lin, Y. Xia, T. Qin, Z. Chen, and T. Liu. Conditional image-to-image translation. In *CVPR 2018*.

[18] L. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9:2579–2605, 2008.

[19] F. Mokhayeri and E. Granger. Robust video face recognition from a single still using a synthetic plus variational model. In *FG 2019*.

[20] F. Mokhayeri and E. Granger. Video face recognition using siamese networks with block-sparsity matching. *IEEE Trans. on Biometrics, Behavior, and Identity Science*, 2019.

[21] F. Mokhayeri, E. Granger, and G.-A. Bilodeau. Domain-specific face synthesis for video face recognition from a single sample per person. *IEEE Trans. on Information Forensics and Security*, 14(3):757–772, 2019.

[22] M. Parchami, S. Bashbaghi, E. Granger, and S. Sayed. Using deep autoencoders to learn robust domain-invariant representations for still-to-video face recognition. In *AVSS 2017*.

[23] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[24] S. Sankaranarayanan, Y. Balaji, C. Castillo, and R. Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *CVPR 2018*.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR 2015*.

[26] Y. Shen, P. Luo, J. Yan, and X. Wang, X.and Tang. FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis. In *CVPR 2018*.

[27] T. T. O. S. J. W. W. Shrivastava, A. Pfister and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR 2017*.

[28] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Perez, M. Zollhofer, and C.Theobalt. FML: Face model learning from videos. In *CVPR 2018*.

[29] L. Q. Tran, X. Yin, and X. Liu. Representation learning by rotating your faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2018.

[30] A. Tran Tuan, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *CVPR 2017*.

[31] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *CVPR Workshop 2011*.

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

[33] J. Zhao and et al. Dual-agent GANs for photorealistic and identity preserving profile face synthesis. In *NIPS 2017*.

[34] J. Zhu, R. Zhang, D. Pathak, T. Darrell, A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *NIPS 2017*.