Learning Face Recognition Unsupervisedly by Disentanglement and Self-Augmentation

Yi-Lun Lee, Min-Yuan Tseng, Yu-Cheng Luo, Dung-Ru Yu, Wei-Chen Chiu
National Chiao Tung University, Taiwan

***Abstract*— As the growth of smart home, healthcare, and home robot applications, learning a face recognition system which is specific for a particular environment and capable of self-adapting to the temporal changes in appearance (e.g., caused by illumination or camera position) is nowadays an important topic. In this paper, given a video of a group of people, which simulates the surveillance video in a smart home environment, we propose a novel approach which unsupervisedly learns a face recognition model based on two main components: (1) a triplet network that extracts identity-aware feature from face images for performing face recognition by clustering, and (2) an augmentation network that is conditioned on the identity-aware features and aims at synthesizing more face samples. Particularly, the training data for the triplet network is obtained by using the spatiotemporal characteristic of face samples within a video, while the augmentation network learns to disentangle a face image into identity-aware and identity-irrelevant features thus is able to generate new faces of the same identity but with variance in appearance. With taking the richer training data produced by augmentation network, the triplet network is further fine-tuned and achieves better performance in face recognition. Extensive experiments not only show the efficacy of our model in learning an environment-specific face recognition model unsupervisedly, but also verify its adaptability to various appearance changes.**

I. Introduction

Face recognition has been a long-standing and fundamental task for the research area of computer vision and robotics, and it is widely used in our daily life, e.g. surveillance and security control. With the development of internet-of-things (IoT) and cloud technologies, the face recognition is now getting integrated into the smart home and home robot system, and runs for the applications like home security, elderly healthcare, baby monitoring, and family activity recognition. Under this scenario, instead of being able to recognize millions of people as in public surveillance system, the goal of face recognition model turns into well identifying a small group of people (e.g. family members) in a specific environment (e.g. home), which is exactly what we would like to address in this paper, and we take the unconstrained videos that satisfy this problem scenario as our target data.

Even though the rapid development of deep learning advances the frontier of face recognition upon having large-scale supervised training data, it is still a great challenge to recognize faces in unconstrained videos where there could be non-frontal faces, different resolutions for the facial regions across multiple shots, or considerably large variance for the appearance of a person's face. In particular, annotating each unconstrained video to collect the training data is also expensive and not practical due to the high diversity of the videos.

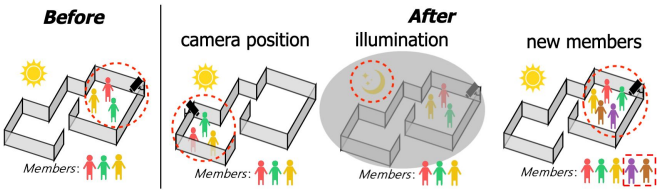


Fig. 1: In this paper we aim to learn a face recognition model that is able to automatically adapt to various environmental and appearance changes, such as camera position, illumination, and even different target groups. This important feature of the face recognition system is in demand for smart home or home robot applications nowadays.

Some up-to-date research works on Multiple Object/Target Tracking (MOT or MTT) address the aforementioned problems of doing face recognition in unconstrained videos. For instance, [1] utilizes the spatiotemporal tracklets and adopts metric learning techniques to learn the discriminative features of facial regions which improve the long-range tracking. [2] use different body parts to assist the face tracking, and apply Gaussian Process to improve the performance of clustering face images. However, these approaches heavily rely on the quality of the initial tracklets and the performance could drop drastically when the target group is extremely different from the one used in training.

In this paper, we propose a face recognition model which unsupervisedly learns the identity-aware feature representation of face images in the unconstrained videos by leveraging the spatiotemporal characteristics in a video as well as the augmented training data provided by our augmentation network. Two key ideas behind motivate our model designs: First, an identity-encoder projects face images into an embedding space where the euclidean distance between samples in this space represents the semantic of identity similarity, which is learned by exploiting the algorithm of triplet-based metric learning. In other words, the faces of the same identity should be close to each other, while those of different identities ought to be as far away from each other as possible. In order to eliminate the expensive effort of manually annotating groundtruth data as supervision, we refer to the physical constraints in a video to produce the training data of negative pairs for the metric learning, i.e. the persons appear at the same frame are definitely with different identities. While for the positive pairs, we extract the feature of face images by a pre-trained VGG-based network, take the nearest neighbors of every face, and assume they are of the same identity. According to

this idea, we thus achieve unsupervised learning. Second, in order to let face recognition model have better adaptability to the changes in the unconstrained videos (examples as illustrated in Figure 1) and avoid getting over-fitted to the face samples drawn for metric learning, we proposed an augmentation network to enrich the data distribution used for training the identity-aware feature representation. This augmentation network is realized by extending the CVAE-GAN [3] architecture into the CVAE-InfoGAN one, which we are going to detail later. With taking identity-aware features as the condition, our CVAE-InfoGAN disentangles out the features that are irrelevant to the identity, hence the augmentation network is capable of synthesizing face images of various appearance but the same identity. With the richer data obtained by disentanglement and self-augmentation, our identity-encoder model can be further refined and in results gets better performance in face recognition.

In brief, our model takes advantage of both the spatiotemporal characteristics from the video and the self-augmentation mechanism to recognize faces without having any supervision of face identities. In experiments, we quantify the performance of face recognition and make comparison with several baselines. We show that our augmentation network not only overcomes the limitation of getting over-fitting to face samples, but also has the capacity of adapting to the variation of environment or appearance changes.

II. Related Works

Learning Face Representations. Face recognition has been widely used in various areas, such as identity authentication [4], [5], human tracking [2], [6], and public security [7], [8]. As the recent renaissance of deep learning, plenty research works [9], [10], [11], [12] in face recognition achieve remarkable performance on aforementioned tasks by learning powerful feature representation of face images. For example, for the task of identity classification, VGGFace [9] and FaceNet [10] employ the metric learning techniques to learn a network for face feature extraction, and discriminate human faces according to the feature distance in the embedding space. While most of the deep-learning-based approaches heavily rely on annotated large-scale training dataset and focus on recognizing millions of faces, our problem scenario instead is to learn a video-specific face recognition model in which the training data is automatically discovered from the video itself and the number of identities in a video is relatively small but their appearances could change temporally.

Disentangled Image Generation. Data augmentation is a well-known technique in deep learning which aims at generating various data samples that are never seen before in order to enrich the distribution of training data and enhance the generalizability of the model learnt. As deep generative models such as Generative Adversarial Networks (GANs [13]) or Variational Autoencoders (VAEs [14]) has made impressive progress in image generation recently, they becomes straightforwardly popular choices for realizing data augmentation [15], [16], [17]. Particularly, in most of the problems of supervised learning scenario, as the labels/annotations are provided with data samples, the conditional generative models have won a wide research interest since they are capable of generating synthetic data with labels. For instance, CVAE [18], CGAN [19], and AC-GAN[20] learn to disentangle the latent space into two individual parts, one is related to the labels and taken as the condition while the another part is modelling the factors which are irrelevant to the labels. Recently, Bao *et al* [3] propose a CVAE-GAN network that learns a controllable disentanglement on images, and it is able to modify the fine-grained attributes or the attribute-invariant appearance of a given image and produce realistic output. Liu *et al* [21] utilizes adversarial tricks to learn the disentanglement of a face image into identity-distilled and identity-dispelled features. Different from these methods, our proposed method learns disentangled features composed of identity-aware and identity-irrelevant features without any supervision of identity labels, and further utilizes these features to generate new face images for data augmentation.

III. Proposed Method

As motivated above, our goal in this paper is to recognize multiple persons' faces of a specific small group within an unconstrained video, based on an unsupervised-learning scenario, and maintain the recognition performance when the environment or appearance changes drastically. To achieve this, we propose a model which is composed of an **Identity Encoder** E^{id} and an **augmentation network**, and the overall training procedure consists of three sequential stages (as shown in Figure 2), where we detail them in the following.

A. Stage-I: Training Identity-Encoder E^{id}

The identity-encoder E^{id} aims at extracting the identity-aware feature of face images and is trained to discover an embedding space where the euclidean distance between embedded features stands for the similarity of identities (i.e. smaller distance leads to a higher chance of belonging to the same person). This is achieved by using the metric learning technique (i.e. triplet network [22] in our approach), where we need to provide it the training data composed of negative and positive pairs (i.e. two face images of different identities or of the same identity respectively). We then exploit the spatiotemporal characteristic of the face samples within a video so as to build up the training data, as described below.

Negative pairs. Given a video that consists of a group of N persons, it is simple to discover that the faces co-exist in the same frame (i.e. appear simultaneously) must belong to different identities. Therefore, let x_k^i be a face of the identity i in frame k, we can generate a set of negative pairs, denoted as $P^- = \{x_k^i, x_k^j\},\ \forall i, j = 1, ..., N,\ i \neq j,\ \forall k = 1, ..., F$, where F denotes the number of frames in the video.

Positive pairs. Unlike [1], [2] which uses both pre-trained features and tracklets to find the faces likely of the same identity in the video, here we simply adopt a pre-trained network to extract features of face images in order to find out the positive pairs. To be detailed, the pre-trained network is of the VGG architecture [23] and trained with

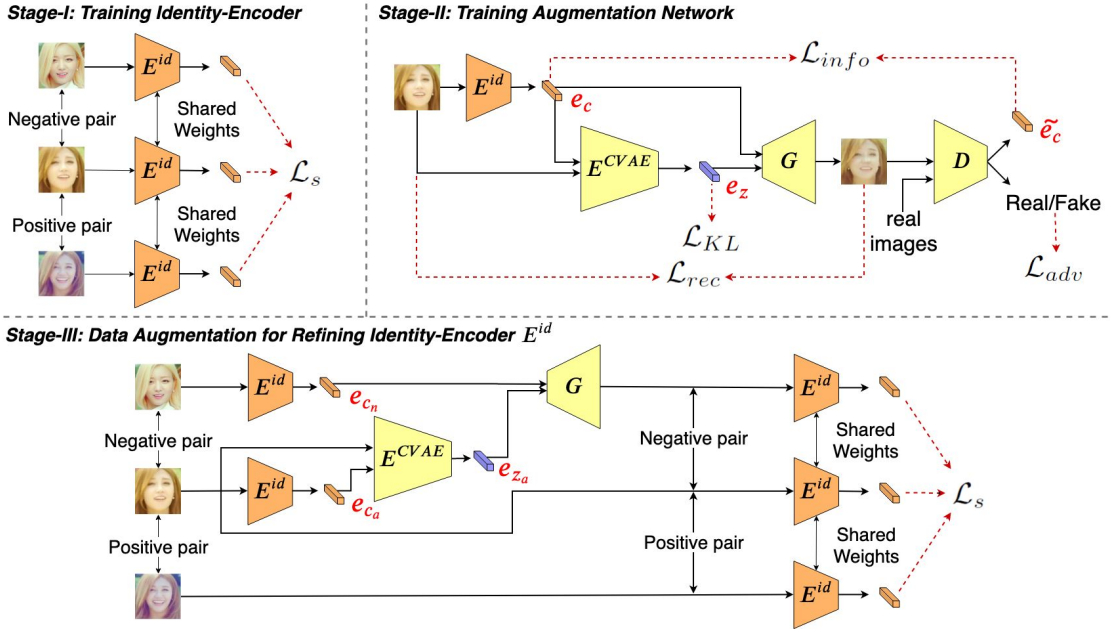


Fig. 2: Overview of our proposed method. The training procedure is composed of three stages. First, we train identity-encoder E^{id} based on the triplet loss, where the training data is automatically discovered from the spatiotemporal characteristics of the target video. Second, upon being conditioned on the identity-aware feature e_c obtained from E^{id}, our augmentation network stemmed from CVAE-InfoGAN learns to disentangle a face image into identity-aware e_c and identity-irrelevant e_z features and is able to synthesize augmented data according to the given condition e_c. Finally, with having more triplet produced by the augmented data, E^{id} is further fine-tuned to achieve better performance in face recognition.

CASIA-WebFace dataset [24]. Once all the face samples in the given video are mapped into the feature presentations by the pre-trained network, for each face x_k^i we apply K-Nearest Neighbor (KNN) algorithm to find its K most similar ones from other frames which are temporally away from the frame k by a threshold τ (where τ is set as 15 in our experiments). The positive pairs are denoted as $P^+ = \{x_k^i, x_l^i\} \in,\ \forall k, l = 1, ..., F,\ \ |k-l| > 15,\ \forall i = 1, ..., N$. Please note that the purpose of having the threshold τ is to avoid that case we always get the nearest neighbors from the consecutive frames in which the built positive pairs would be less informative for the metric learning.

With the positive and negative pairs, we construct the triplets to train our identity-encoder E^{id} for learning the identity-aware feature of face images. Here we adopt the symmetric triplet loss [1] as our objective function. Given a triplet $T = (x_k^i, x_l^i, x_k^j)$, where $(x_k^i, x_l^i) \in P^+,\ (x_k^i, x_k^j) \in P^-$, there are three distances in the embedding space produced by E^{id}: $d(E^{id}(x_k^i), E^{id}(x_l^i))$, $d(E^{id}(x_l^i), E^{id}(x_k^j))$, and $d(E^{id}(x_k^i), E^{id}(x_k^j))$, where the first one is a positive pair and the last two forms negative pairs, and d represents the euclidean distance. The symmetric triplet loss $\mathcal{L}_s$ is defined as

$$\mathcal{L}_s = \ max[0,\ d(E^{id}(x_k^i), E^{id}(x_l^i)) - \frac{1}{2}(d(E^{id}(x_k^i), E^{id}(x_k^j)) + d(E^{id}(x_l^i), E^{id}(x_k^j))) + \alpha] \quad (1)$$

which encourages E^{id} to distinguish different identities by ensuring that the mean of the distances for the negative pairs, i.e. (x_k^i, x_k^j) and (x_l^i, x_k^j), are larger than that of the positive pair (x_k^i, x_l^i) by a margin α. The margin α is set to 0.8 in all our experiments. The architecture of our identity-encoder E^{id} follows the one in VGG network [23] with replacing its fully connected layer by a subnetwork composed of BatchNorm2d, Linear, ReLu, and BatchNorm1d layers.

B. Stage-II: Training Augmentation Network

We notice that in an unconstrained video, it is common to see that most of the video frames only contains subset of persons or even only has a single person. The triplets built under this situation is insufficient to train E^{id} for learning the identity-aware feature e_c which is discriminative enough for good face recognition. In order to resolve this issue, we propose an augmentation network, which is basically the integration between conditional VAE [18] and InfoGAN [25], to perform data augmentation for enriching the training data distribution and improving E^{id}. The architecture of the augmentation network is illustrated in the upper-right part of Figure 2. First, the conditional VAE part takes the identity-aware feature e_c obtained from E^{id} as condition, learns E^{CVAE} to decompose the latent space of face images x into identity-aware features e_c and identity-irrelevant ones e_z (e.g. pose, lighting, expression, etc.), and trains the generator G to synthesize images $\tilde{x}$ given e_c and e_z. Second, the generator G and a discriminator D with an auxiliary classifier together form the InfoGAN part which improves the image quality of synthesized images $\tilde{x}$ and further boost the disentanglement between e_c and e_z. The learning of the augmentation network is based on the following objective functions.

Reconstruction Loss. Given a face image x and its corresponding identity-aware feature $e_c = E^{id}(x)$ as the input for the conditional VAE, the synthesized image $\tilde{x} = G(e_c, e_z)$

should well reconstruct x, where $e_z = E^{CVAE}(x, e_c)$, in order to ensure that e_c and e_z together can encode most of the important information of x. Moreover, as CVAE-GAN [3], we encourage the matching between $f_D(x)$ and $f_D(\tilde{x})$ where $f_D(\cdot)$ denotes the feature extracted from the last convolution layer of discriminator D. The reconstruction loss $\mathcal{L}_{rec}$ is:

$$\mathcal{L}_{rec} = \frac{1}{2}(\|x - \tilde{x}\|_2^2 + \|f_D(x) - f_D(\tilde{x})\|_2^2) \quad (2)$$

KL Divergence Loss. As CVAE [18], we impose a Gaussian prior $p(e_z) = \mathcal{N}(0, 1)$ on the distribution of identity-irrelevant features e_z, by a KL-divergence loss $\mathcal{L}_{KL}$:

$$\mathcal{L}_{KL} = KL(q(e_z \mid x) \mid\mid p(e_z)) \quad (3)$$

Adversarial Loss. The typical adversarial loss function $\mathcal{L}_{adv}$ is used to train both G and D for making the generated face image more realistic and with high-quality:

$$\mathcal{L}_{adv} = \mathbb{E}_{x\sim\mathbb{P}_x} logD(x) + \mathbb{E}_{\tilde{x}\sim\mathbb{P}_{\tilde{x}}} log(1 - D(\tilde{x})) \quad (4)$$

where $\mathbb{P}_x$ and $\mathbb{P}_{\tilde{x}}$ are the distribution of real data x and synthetic samples $\tilde{x}$.

Information Loss. As the input of generator G comes from the disentangled e_c and e_z, in order to ensure that G does take the information from the condition e_c to synthesize $\tilde{x}$ as well as enhance the independence between e_c and e_z, we adopt the InfoGAN [25] idea to attach an auxiliary classifier Q onto the discriminator D, in which Q aims to predict $\tilde{e_c}$ from $\tilde{x} = G(e_c, e_z)$. As $\tilde{e_c}$ should be close to corresponding e_c, the information loss $\mathcal{L}_{info}$ thus is defined as:

$$\mathcal{L}_{info} = \|e_c - \tilde{e_c}\|_2^2 \quad (5)$$

The overall objective $\mathcal{L}_{aug}$ for the augmentation network is the weighted sum of the aforementioned losses mentioned above, i.e. $\mathcal{L}_{aug} = \lambda_1 L_{rec} + \lambda_2 L_{KL} + \lambda_3 L_{adv} + \lambda_4 L_{info}$, where λ_1, λ_2, λ_3, λ_4 are hyper-parameters and set to $\{1, 0.0003, 1, 1\}$ respectively in our experiments. The network architecture of both G and D basically follows the DCGAN [26] paper, where Q simply consists of two convolution layers and is attached to D. E^{CVAE} contains the same architecture as D with having two linear blocks in the end to estimate the mean and the variance of e_z.

C. Stage-III: Data Augmentation for Refining E^{id}

After learning the augmentation network, we are now able to generate face samples of the given identity-aware feature e_c with various appearance driven by randomly-sampled e_z, and construct more triplet samples for improving the efficacy of identity-encoder E^{id}. For instance, let c_k^i and z_k^i be the identity-aware and identity-irrelevant features of x_k^i respectively. Given a triplet $T = (x_k^i, x_l^i, x_k^j)$, we use E^{id} and E^{CVAE} to extract their latent features and replace z_k^j with z_k^i for further generation by G. As demonstrated in the bottom part of Figure 2, we can now obtain a new triplets $(x_k^i, x_l^i, \tilde{x}_k^j)$ for data augmentation. Based on the triplets obtained from both real (as in Stage-I) and synthetic data, the identity-encoder E^{id} is getting refined and the effectiveness of the identity-aware feature e_c is improved accordingly on the faces in the unconstrained video. When applying our model in daily life, we could easily identify a novel face based on the trained model and old face data via nearest-neighbor algorithm. Our source code and experimental settings are released at `https://github.com/YiLunLee/Unsupervised-Face-Recognition`.

IV. Experiments

A. Datasets and Evaluation Metric

Music Video Dataset (MVD) [1] consists of 8 edited music videos. These videos include the challenges for face recognition such as dramatic variations in facial appearance, frequent scene changes, and rapid camera movement. We use the face detection algorithm provided by [1], remove some false alarm by *dlib* face detection toolkit, and resize them into size of 128×128 for our experiments.

Extension Datasets. Here we create three extension datasets in order to simulate the case of having various environmental and appearance changes.

(1) Apink-NoNoNo dataset. We choose from MVD dataset [1] a video (i.e., "Apink-Petal") of an idol group "Apink", and additionally collect another video of this group (i.e., "Apink-NoNoNo") from the web, in order to mimic the scenario of having dramatic change in the environment and facial appearance for the same group of persons. After using the *dlib* toolkit to detect faces in the video "Apink-NoNoNo", we label them manually for the evaluation purpose.

(2) Apink-DayNight dataset. We generate another collection of face images by randomly decreasing the illumination of face images (based on the color-jitter function provided in pytorch) in the "Apink" video from MVD dataset, in order to mimic the illumination change.

(3) Apink-Pussycat dataset. Moreover, we combine two videos, "Apink" and "PussycatDolls" from MVD dataset for simulating the case of adding more persons into the face recognition model, as Apink and PussycatDolls are two different idol groups.

In the left of Table II, III, and IV we provide the t-SNE visualization for the face images from these three extension datasets, based on the VGG16 features (output of first linear layers of VGG16 model pretrained on ImageNet [27]) which is not identity-specific but related to the appearance of faces. The distribution discrepancy between two videos of each extension dataset verify that these datasets are not trivial and requires the adaptation to carry out.

Evaluation Metric based on Clustering Purity. The Weighted Clustering Purity (WCP) [1] is used to quantify the efficacy of the face feature representations learnt from different face recognition models. Basically, if a face feature better encodes the identity information, then the face samples belonging to the same identity are more likely to be grouped together after applying clustering. WCP is defined as $\frac{1}{N}\sum_{c\in C} n_c \cdot p_c$, where N denotes the total number of faces in the video, C is the total number of clusters, n_c is the number of faces belonging to a cluster $c \in C$, and its purity p_c is measured as a fraction of the largest number of faces from the same person to n_c.

Videos	T-ara	Pussycat Dolls	Bruno Mars	Hello Bubble	Darling	Apink	Westlife	Girls Aloud
# of Identities / Faces	6 / 7510	6 / 6921	11 / 11757	4 / 3227	8 / 7201	6 / 6590	4 / 7787	5 / 7355
Siamese [1]	0.69	0.77	0.88	0.54	0.46	0.48	0.54	0.67
Triplet [1]	0.68	0.77	0.83	0.60	0.49	0.60	0.52	0.67
SymTriplet [1]	0.69	0.78	**0.90**	0.64	0.70	0.72	0.56	0.69
VGG-Features	0.55	0.79	0.85	0.51	0.60	0.54	0.73	0.92
SymTriplet [1] Redo	**0.81**	0.79	0.80	0.66	0.77	0.71	0.93	0.94
Stage-I	0.71	0.82	0.85	0.64	0.70	0.75	0.70	0.90
Longer Stage-I	0.80	0.67	0.85	0.68	**0.78**	0.68	**0.94**	0.95
Our Full Model	**0.81**	**0.84**	0.87	**0.69**	0.77	**0.88**	0.88	**0.97**

TABLE I: Evaluation of face recognition performance on the music videos from MVD dataset [1], based on WCP metric.

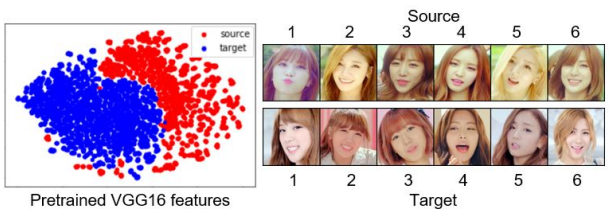
B. Quantitative Evaluation

Face Recogntion by Clustering We evaluate the effectiveness of different identity-aware representations of face images and make comparison among the ones obtained from various approaches, as indicated in Table I. The WCP performances of ***Siamese***, ***Triplet***, and ***SymTriplet*** are obtained from [1], representing different techniques of metric learning; ***VGG Features*** stands for the features extracted by the VGG-based pre-trained network used in our Stage-I to discover positive pairs; ***SymTriplet Redo*** re-trains the model proposed by [1] with our triplets for the fair comparison; ***Stage-I*** uses identity-aware features learnt after Stage-I of our proposed method; ***Longer Stage-I*** utilizes the same group the triplets built in Stage-I (without any augmented triplet generated from the augmentation network) to train E^{id} longer until achieves the number of iterations as used in Stage-III. Please note that we do not use any tracking technique to build the triplets in our Stage-I, while [1] does. We can see that the performance of the features obtained from our ***Stage-I*** is already comparable with the best one from [1] (i.e. ***SymTriplet***). Moreover, ***Our Full Model*** with having E^{id} refined by the synthetic triplets produced by the augmentation network further improves to outperform other baselines on most of the videos. The worse performance of ***Longer Stage-I*** indicates the potential problem of getting over-fitted (e.g. Pussycat Dolls and Apink videos) to the insufficient training triplets built in Stage-I, and simultaneously demonstrates that our augmentation network in the full model is able to enrich the data distribution of the triplets thus help to eliminate the issue of over-fitting.

Adaptation Ability As in each of the extension datasets there are two videos, we sequentially adjust the ratio of data samples from these two videos to experiment the process of adapting from one (as the source video) to another (as the target video). For the experiments on each extension dataset, we directly train our Stage-I and Full Model on the target video from scratch as the oracles, denoted as **Oracle-I** and **Oracle-FM**. In order to verify the adaptation ability of our model (Stage-I and Full Model), we use the E^{id} learnt from the source video as the initialization, and keep updating them upon sequentially increasing the ratio of the samples drawn from the target video, where they are denoted as **Adapt-I** and **Adapt-FM**. The **SymTriplet** baseline [1] (named as SymTriplet) is trained based on the same dataset and used here for comparison. We can see from the quantitative results provided in Table II and III that our full model with having augmentation network to synthesize richer training data is able to adapt to the appearance and illumination changes, and reaches the similar performance as oracle when the data samples are fully from the target video. In particular, for the Apink-DayNight dataset, with only seeing 25% of the target video (ill-lighted "Apink"), our model has achieve quite high performance (up to 0.87 in WCP) on the target. This result could be due to the less discrepancy between the source and target videos, as demonstrated in the correspoding t-SNE visualization. When being fully adapted to the target video, our full model (i.e., Adapt-FM) is able to provide superior performance in comparison to other baselines, thus demonstrates the efficacy and adaptation ability of our proposed method. In the experiments on the Apink-Pussycat dataset, as the face recognition model should now learn to not only recognize the old members in the source video (i.e., "Apink") but also the new ones in the target video (i.e., "PussycatDolls"), we directly perform evaluation on three sets: the source video, target video, and the whole Apink-Pussycat dataset. As shown in the quantitative results of Table IV, our Full model has demonstrated its ability of adaptation for not only preserving the performance on "Apink" but also learning the representation that is discriminative on "PussycatDolls".

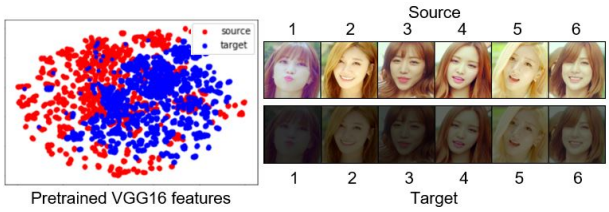
C. Qualitative Evaluation

Take the face images in the "Apink" video from MVD dataset [1] as example, we visualize the distribution of identity-aware features extracted under different settings of E^{id}, as shown in Figure 3, where different colors stands for the groundtruth identity labels. We can see that our full model gives highest purity and clearest separation between identities. We further provide examples to demonstrate the efficacy of our augmentation network in learning disentanglement between identity-aware and identity-irrelevant features as well as generating realistic images. In Figure 4, each row represents the faces with various appearance (i.e. having different e_z) of the same identity (i.e. having the same e_c); while each column (except the left-most one which is the real images as input) in contrast demonstrates images of different identities but with the same identity-irrelevant appearance. We can observe that, within each column, every face is facing to the same direction, illustrating that our augmentation



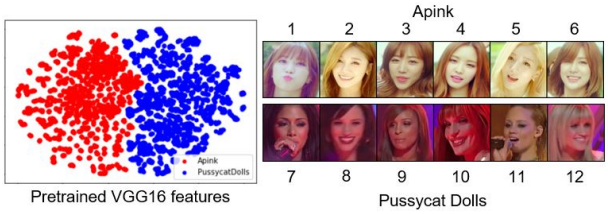
WCP on target (source: "Petal", target: "NoNoNo")					
source/target ratio	100/0	75/25	50/50	25/75	0/100
Oracle-I	-	-	-	-	0.69
Oracle-FM	-	-	-	-	**0.74**
SymTriplet [1]	0.41	0.48	0.60	0.65	0.63
Adapt-I	0.39	0.53	0.66	0.68	0.69
Adapt-FM	0.41	0.62	0.71	0.70	**0.74**

TABLE II: Experiment results on the Apink-NoNoNo extension dataset



WCP on target (source: "Apink", target: ill-lighted "Apink")					
source/target ratio	100/0	75/25	50/50	25/75	0/100
Oracle-I	-	-	-	-	0.71
Oracle-FM	-	-	-	-	0.83
SymTriplet [1]	0.26	0.67	0.61	0.76	0.77
Adapt-I	0.30	0.62	0.75	0.81	0.85
Adapt-FM	0.29	0.87	0.91	0.76	**0.90**

TABLE III: Experiment results on the Apink-DayNight extension dataset



WCP on Apink-Pussycat			
Videos	Apink	PussycatDolls	Apink-Pussycat
Oracle-I	0.66	0.68	0.69
Oracle-FM	0.72	0.81	**0.75**
SymTriplet [1]	0.76	0.82	0.66
Adapt-I	0.73	0.75	0.67
Adpat-FM	**0.86**	**0.86**	0.74

TABLE IV: Experiment results on the Apink-Pussycat extension dataset.

network successfully extracts the identity-irrelevant features (like pose here). Also, with the same identity-aware feature, the faces in each row represent the same person as the input face on the left-most column.

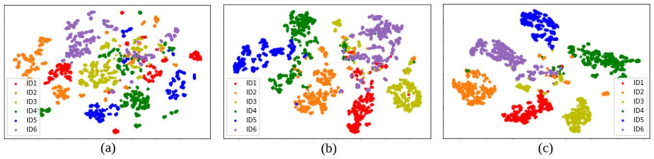


Fig. 3: t-SNE visualization of identity-aware features extracted under different settings of E^{id}. (a) VGG features. (b) Stage-I (c) Our full model.

V. Conclusions

A face recognition model is proposed to learn identity-aware face representation in the unconstrained videos without any supervision of identities by using the physical properties of the video, and make a self-improvement based on the synthesized data provided by our augmentation network. Experiments on several datasets and the ablation study show not only the good performance of our model but also its ability of adapting to the changing environment or appearance. Our proposed method would be beneficial to various applications, such as smart home, elderly healthcare, and home robot.

Fig. 4: Examples produced by augmentation network. The left-most column is the input face images. From the second to the last column, each column represents face images of different identities e_c but with the same identity-irrelevant appearance e_z. Each row shows the images with different appearance e_z but of the same given input identity e_c.

VI. Acknowledgement

This project is supported by MOST-108-2636-E-009-001, MOST-108-2634-F-009-007, and MOST-108-2634-F-009-013. We are grateful to the National Center for High performance Computing for computer time and facilities.

REFERENCES

[1] S. Zhang, Y. Gong, J.-B. Huang, J. Lim, J. Wang, N. Ahuja, and M.-H. Yang, "Tracking persons-of-interest via adaptive discriminative features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[2] C.-C. Lin and Y. Hung, "A prior-less method for multi-face tracking in unconstrained videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Cvae-gan: Fine-grained image generation through asymmetric training," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] M. E. Fathy, V. M. Patel, and R. Chellappa, "Face-based active authentication on mobile devices," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[5] D. Crouse, H. Han, D. Chandra, B. Barbello, and A. K. Jain, "Continuous authentication of mobile user: Fusion of face image and inertial measurement unit data," in *2015 International Conference on Biometrics (ICB)*, 2015.

[6] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, "Arttrack: Articulated multi-person tracking in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] P. Hu, H. Ning, T. Qiu, H. Song, Y. Wang, and X. Yao, "Security and privacy preservation scheme of face identification and resolution framework using fog computing in internet of things," *IEEE Internet of Things Journal*, 2017.

[9] O. M. Parkhi, A. Vedaldi, A. Zisserman, *et al.*, "Deep face recognition." in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *ArXiv:1801.07698*, 2018.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[15] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2018.

[16] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," in *ArXiv:1711.04340*, 2017.

[18] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[19] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *ArXiv:1411.1784*, 2014.

[20] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[21] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang, "Exploring disentangled feature representation beyond face identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[22] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*, 2015.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

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

[25] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

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

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, 2015.