IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 1, 2020, accepted August 16, 2020, date of publication August 19, 2020, date of current version August 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3017779

PA-GAN: A Patch-Attention Based Aggregation Network for Face Recognition in Surveillance

MING LIU[1], JINJIN LIU[1,2], PING ZHANG[1], AND QINGBAO LI[1]
[1]State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China
[2]School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China

Corresponding author: Ming Liu (lm_puree@outlook.com)

This work was supported in part by the National Natural Science Foundation of China under Project 61802432; and in part by the National Social Science Fund of China under Project 15AJG012.

ABSTRACT Face recognition in unconstraint surveillance is a complicated problem on account of motion blur, expression variations and low resolution. Recent works have demonstrated that patch-attention is strictly more powerful than convolution in recognition models. In this study, we investigate the task of unconstraint surveillance face recognition. First, a Patch-Attention Generative Adversarial Network (PA-GAN) model is devised to aggregate some robust features on behalf of a set of raw surveillance frames, which not only increases the recognition accuracy but also reduces the computational costs of face matching. Second, an improved center loss function combined with abundant unlabeled surveillance faces is utilized to accurately classify the known identities. With the proposed method, the discriminativeness of the face representations is largely enhanced. Finally, the proposed method is verified in two widely used datasets, IJB-A dataset and QMUL-SurvFace dataset to demonstrate the effectiveness. Evaluation of the algorithm performances in comparison with other state-of-the-art methods indicates that the proposed design can achieve competitive accuracy on both the verification and identification protocols.

INDEX TERMS Face recognition, video surveillance, attention model, generative adversarial network.

I. INTRODUCTION

During recent decades, video-based face recognition (FR) has received considerable attention in both academia and industry due to its wide range of various security systems and law enforcement applications. One most significant thing is the successful use of the face recognition technology by public security systems to arrest escaping criminals and search for missing person. How to quickly and accurately identify the unique information of enormous faces in videos is of great significance to the development of security field. Although the compelling progress in deep learning and computer vision, it is still a great challenge to match surveillance face images in different modalities, especially in open-set scenario [1]. There have been varieties of efforts about video-based face recognition [2]–[4]. However, most of them focus on learning an image-level face representation or aggregating face representations through simple pooling from favorable viewing angles. Due to the considerable discrepancy between source and the target domains, one challenge is that the face recognition model trained on unconstrained high-quality data often degrades significantly for surveillance face recognition. Furthermore, public surveillance cameras are installed far away from the recognition subjects, resulting in a lower resolution of the human face. It is well known that deep learning model is data-driven. Only when the training sets and the test set have similar distribution, the model can achieve satisfied results. The performance of the face recognition system would be degraded if a same weight is given to both the low-quality images and other high-quality images. Therefore, a qualified network should be able to reduce the impact of such distracting images and focus on the informative ones. Although the unconstrained still-based recognition models struggle to extract valuable information from images. This type of methods may be limited in practical usage. The experimental results [1] show that the results of applying the still-based recognition method to video-based recognition are very bad. Because the quality of the human face image captured in the actual monitoring environment is very different from the high-resolution human face image, which means the data distribution is inconsistent.

It is not only difficult but also labor intensive to directly label the data samples collected in the monitoring environment. As the number of surveillance cameras increasing,

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

the captured videos will need to be processed automatically. We all know that sufficient training data is critical to applying deep learning methods to new target areas. Many deep face recognition models depend on millions-scale training sets. There is not yet a huge database of surveillance video faces. And tagging such a large data set is also a tedious task. Therefore, part of the works [6], [7] fine-tuned their pre-trained deep convolutional neural networks on a small amount of target domain data through the method of transfer learning. Although these models can obtain a high accuracy on popular benchmark, they achieve unsatisfied results in practical in video surveillance. That experiment, for example, was a total failure when testing the VGGFace model [8] on an unconstrained video dataset. If we compare every picture in probe with the gallery, we may get a better result. However, calculating the eigenvector similarity of vectors in all templates costs tremendous memory space and computation.

In this article, we will deal with unconstrained low-quality face recognition based on surveillance video. That is more compatible with the real-world venues, which is perhaps the most interesting for forensic and surveillance systems applications. As opposed to previous works, we inspired by the attentional patterns of human visual mechanisms. Therefore, we devise a flexible patch-attention modeling which learns more discriminative representation yet keeps greater efficiency. Specifically, we consider a practical protocol in surveillance environment: surveillance-to-still, where the query is a surveillance video and each subject have a single frontal still image in the gallery. The main contributions of this article can be summarized as follows:

♦ We propose an efficient Patch-Attention Generative Adversarial Network (PA-DAN) which aggregates each frame adaptively to a few representations for surveillance face recognition. It significantly reduces computational cost and wisely leverages the useful surveillance information.

♦ The proposed PA-GAN utilizes the unlabeled faces to augment training sets and elaborately designs face loss functions. This is beneficial for open-set recognition.

♦ Experimental results on two challenging surveillance benchmarks IJB-A [5] and QMUL-SurvFace [1], demonstrate that our framework improves the recognition accuracy and accelerate the recognition speed simultaneously.

II. RELATED WORK

A. VIDEO FACE RECOGNITION

As the numerous surveillance data and video media producing, video face recognition system has many practical applications. Video actually consists of many frames, so video-based face recognition can be treated as set-based recognition. This work was pioneered by Phillips [9] in 1996. However, due to few benchmarks were available, the development of video-based face recognition was slowed down. Nowadays, there are numerous potential uses of the systems with surveillance FR capability in real-world environments and bringing it back into focus. Existing methods about video FR are simply split into two computation stages:

1) IMAGE-AGGREGATION APPROACHES

In particular, image-aggregation methods consist of three levels:(a) Image-level; (b) Representation-level; (c) Component-wise. The common idea of these methods is to fuse the feature vectors of multiple images into a single substitute at different levels. In ref. [2], authors use two attention blocks through supervised learning adaptively aggregate the face features to one fixed-dimension convex hull feature. It results that high-quality face made more contribution to the final feature and favors the face images more discriminative. The study [3] firstly provided a component-wise aggregation, which controls the normalized quality of corresponding feature pooling multiple frames together. Some of the works like ref. [10] and ref. [11] unlike the simple pooling strategies [12], [13], such as max pooling and average pooling, they presented a similar module to predict a quality score for each feature vector and aggregates the vectors weighted the assigned scores. Zhao *et al.* [4] employed a dense subgraph in place of handcrafting the face medias. Each dense subgraph discovered a sub-set of face media that are with small intra-set variance but discriminative from other subject faces. This provided a comprehensive and concise face representations, reducing the impact of media inconsistencies and greatly improving face recognition performance based on unconstrained sets. In these methods, when a high percentage of low-quality images are present, it will cause their performances falling off.

2) ROBUST FEATURE EXTRACTION

Indeed, ref. [14] provided a ScatterNet coding deep features from much fewer labelled examples rapidly to tackle the intricate work of age-invariant face recognition in real-world videos. They also built a self-dataset, Celebrities Video Aging (CVA), it promotes the development of innovative age-invariant methods. Gong *et al.* [15] learned multiple attentions from video context to solve low quality video face recognition by embedding context-awareness combined with recurrent neural network. In ref. [16], authors considered finding the focus of videos as a Markov decision process and leveraged a deep reinforcement network to make better use of temporal information. Reference [17] proposed a generic graphical algorithm, in which a contextual connecting formulate between high-quality and low-quality faces is designed. In ref. [18], to relieve the deficiency of raw surveillance faces, training data were processed by adding artificially motion blur by tow kernel filters. It used an end-to-end ensemble trunk-branch CNN to learn pose-invariant and occlusion-robust representations for efficiently video face recognition.

B. ATTENTION MACHINING

While the convolution has undoubtedly been effective as the basic operator in modern image recognition, it is not without drawbacks. Recent works have shown that self-attention may constitute a viable alternative. The developments of effective

self-attention architectures in computer vision hold the exciting prospect of discovering models with different and perhaps complementary properties to convolutional networks. The advantages of attentions over convolution are much elastic mapping, which is an effectual way to make the connection between any part of the input series stronger. There are a number of methods [2], [16], [19], [20] for boosting the accuracy of CNN classification models by employing the attention mechanism. In ref. [20], the author used cascade attention mechanism to guide the different layers of CNN and concatenate them to gain discriminative representation as the input of final linear classifier. In contrast to the aforementioned methods, we combine aggregation method and generative adversarial network together. We apply the attention mechanism on each branch of the generative network for the sake of the discriminative regions for classification.

FIGURE 1. Some samples from the IJB-A [5] dataset. This shows that many factors affect image quality, such as pose, illumination, and expression variation in images.

III. APPROACH

A. MOTIVATION

Surveillance videos with multiple faces in a video clip can be both beneficial and challenging. On the one hand, recognizing each frame results in too much redundant of the same face in video. It leads to wasting of computing resources and excessive false positives. In a certain period of time, dozens of consecutive frames of images have little change in face attitude, which will generate a lot of redundant information; on the other hand, in most frames of video, face pose is not standard and motion blur makes it difficult to be recognized. It is difficult to get accurate results for face detection and recognition of these invalid frames. Therefore, a fundamental issue in surveillance face recognition is to build an excellent pose-invariant eigenvector instead of the original video clips, such that the information across different frames can effectively use to maintain beneficial features while dropping the remaining multiple degenerating video frames.

Attention mechanism plays a critical role in human visual experience. The human visual system can not only detect and recognize objects, but also infer the deep structure of the scene. Some recent works have demonstrated that attention mechanism can also play an important role in computer vision and natural video prediction. The attention model allows the algorithm to model parts of an image or feature that have a greater impact on the final result. These remarkable results inspire us to employ one type of self-attention, patch-attention, to devise a generative adversarial network (PA-GAN) for efficient face representation extraction in surveillance video.

As evident from Figure 1, we yield a novel module, named PA-GAN, which is composed of two pivotal components. Above all, we exploit the residual patch-attention block and shortcut connections to build a generator, which outputs a more discriminative face instead of primal face templates. Secondly, an auto-encoder functions as the discriminator, which is precise to estimate whether the image is generated or selected from the original video. Then, we transfer-learn a similar feature extract network presented in ref. [22], which have a high discrimination power.

B. PATCH -ATTENTION GENERATOR

The advantage of patch attention over convolution operation is much flexible in allocating weights. Based on these analyses, we introduce the patch-attention block to a modern backbone networks ResNet [23] by shortcut connections as the generator G. Patch-attention block is the first strategy to increase the feature extraction ability of the face recognition network. Compared with the standard convolution that each filter operates on all input channels, the attention block [24] is very sparse, and thus it is powerful to replace convolutions entirely. Figure 3 illustrates the processing of the patch-attention block. The self-attention mechanism allows inputs to interact with each other and figures out what they should be paid more attentions. Compared with the standard convolution that mainly concentrates on feature aggregation and feature transformation, patch-attention uses a mapping mechanism to perform feature aggregation. Then feature transformation can be performed by perceptron layers that process each feature vector separately. The input feature tensor is passed through two processing streams. The left evaluates the attention weights by computing function and subsequent mapping. The right applies a linear transformation reducing the dimensionality for efficient processing. The outputs of two streams are aggregated by Hadamard product and expansion.

Skip-connections is the second strategy to increase robustness of model. Our network can better fit the complex correlations between channels and greatly reduce the number of parameters. Contextual information from global and local parts compensates each other and spontaneously benefits face recognition. The hierarchical features within a skip-net are multi-scale in nature owing to the increasing receptive field sizes, which are combined together via skip connections. Such a combined representation comprehensively preserves the contextual information, which is useful for extracting information about the structure of an individual face.

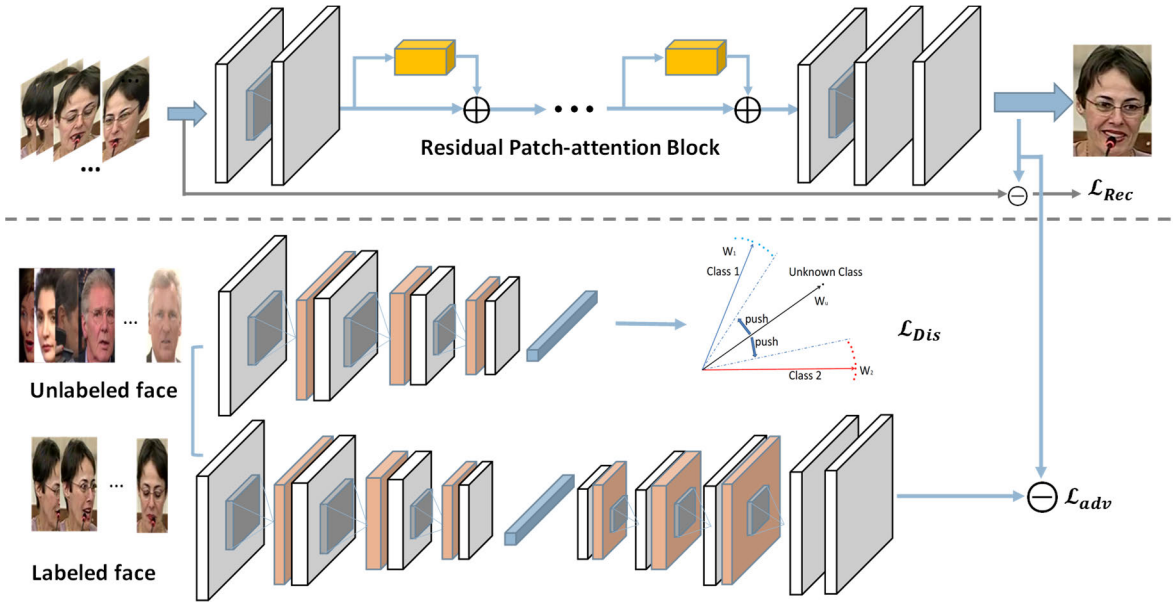


FIGURE 2. Detailed architecture of the proposed pipeline. Top panel: the generator with patch-attention blocks. Bottom panel: a standard discriminator with discriminative loss and adversarial loss.

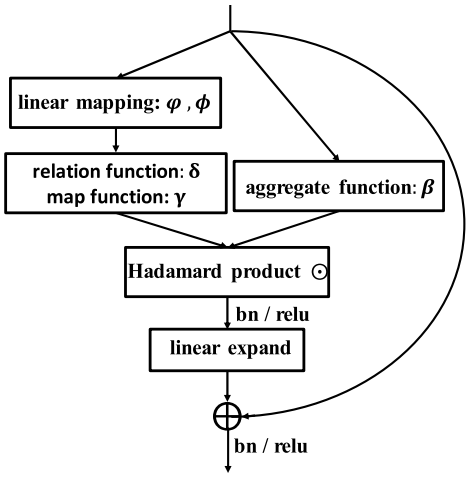


FIGURE 3. The illustration of the patch-attention block. The left branch is used for evaluating the output attention weight; The right branch performs linear transformation on the input to reduce dimension for more efficient processing. φ and ϕ are linear mapping trainable transformations. δ is a relation function. γ is a map function. β is an aggregate function.

For the surveillance face recognition, the original video templates are denoted as $V_i = \{f_i^j, y_i$ and the aggregated face image is denoted as:

$$v_i = \mathbb{G}(V_i) \tag{1}$$

where V_i is the i^{th} individuals ($i = 1, 2, \ldots n$),f_i^j denotes the j^{th} frame of $V_i (j = 1, 2, \ldots m)$, and y_i means the ground truth of V_i.

C. DISCRIMINATOR

The main function of the discriminative model is to constantly improve its discriminant ability by learning from the generative model. We also introduce a discriminator network D, an auto-encoder, consisting of several convolution blocks. The vital demand for discriminator is that the refining face image v_i like a real face image in appearance while reducing the number of the images to be processed. We hope our framework PA-GAN can aggregate video clips into single image while obtain more discriminative ability. To this end, we design a comprehensive loss function to ensure the discriminator supervises generator to produce photorealistic and identity-preserving face image:

$$\mathrm{L} = \lambda_1 \mathrm{L}_{Dis} + \lambda_2 \mathrm{L}_{Rec} + \mathrm{L}_{Adv} \tag{2}$$

where L_{Dis} is the discriminative loss for enhancing the discriminative capacity to distinguish the identity of subjects, L_{Rec} is the reconstruction loss for preserving the identity information, and L_{Adv} is the adversarial loss for adding realism to the synthetic images and alleviating artifacts. λ_1 and λ_2 are trade-off parameters.

D. LOSS FUNCTION

1) DISCRIMINATIVE LOSS

In practice, real-world surveillance FR is an open-set problem. Tens of thousands of labeled people in dataset are only a tiny fraction of the billions of people on the earth. And the trained model generalization ability may be insufficient. There is not enough labeled surveillance person, which hinders the improvement of the model performance to some extent.

To solve the above problems, we first crawled a certain number of unlabeled surveillance video faces from Internet, and then modified Center Loss [25] to optimize the face recognition model with these data. Our approach just needs to ensure that these unlabeled people do not appear in the

labeled data. The addition of unlabeled data can easily expand the number of training samples whist substantially improving the generalization ability. In the training process, each sample feature needs to be close to the center of the corresponding category. For unlabeled classes, they do not belong to any class of the testing sets, so the model needs to reject them. That is, the unlabeled features are far enough away from the center of each classification layer. Each different unlabeled class will get a confidence coefficient $\rho_i (i = 1, 2, \ldots, n)$ denotes the classification result. Ideally, $\rho_1, \rho_2 \ldots \rho_n$ should be lower than the threshold. Consequently, our goal is to:

$$\begin{aligned} &minimize(\rho_1, \rho_2 \ldots \rho_n) \\ &\text{s.t.} \sum_{i=1}^{n} \rho_i = 1 \end{aligned} \tag{3}$$

And then, combining the Center Loss [25] with the discriminative loss together can force the distribution of known classes more sparsely. Detail of discriminative loss is formulated as follows:

$$\mathcal{L}_{Dis} = \mathcal{L}_{Soft} + \frac{\lambda_1}{2} \left(\sum_{i=1}^{n} \left\| x_i - c_{y_i} \right\|_2^2 - \sum_{i=1}^{n} \log(p_i) \right) \tag{4}$$

where L_{Soft} is normal SoftMax function. λ_1 is the hyper-parameters for adjusting the impact of discriminator, c_{y_i} is the center of y_i class of high-dimension feature, and p_i is the probability of x_i belongs to i^{th} class.

2) RECONSTRUCTION LOSS

To ensure that generating the target domain image and retaining the semantic content of the input image, reconstruction loss is included in the training generator. Each subject in the video set contains multiple images. Compared with all the images or any one face with the generated face by G, the training results are obviously not perfect. So, we first calculated the confidence scores for each real face.

$$\theta_i = \frac{\exp[\tau_t * (a_i^T f)]}{\sum_j \exp[\tau_t * (a_i^T f)]} \tag{5}$$

where θ_i denotes the predicted confidence score, τ_t means the selective attenuation item, and a and f are l_2 normalized to achieve boundary equilibrium.

Selective attenuation on the confidence scores of genuine samples in turn increases the corresponding classification losses, which narrows the decision boundary and controls the intra-class affinity and inter-class distance. The reconstruction loss is defined as below:

$$\mathcal{L}_{Rec} = \left\| F(v_i) - \sum\nolimits_{j=1}^{m} \theta_i \cdot F(f_i^j) \right\|^2 \tag{6}$$

where F is the function of face feature extraction. In this way, we would like to make the composite face feature is much closer to the original video center.

3) ADVERSARIAL LOSS

In order to produce more realistic images, as same with the previous GAN network, the generator makes completion with discriminator through adversarial loss. The generator wants to confuse the discriminator by generating outputs similar to the real samples, and the discriminator wants to accurately determine whether the image is true or false. At the same time, we use Wasserstein distance loss as a counter loss.

$$\mathcal{L}_{adv} = \beta \mathbb{E}\left[\log \mathbb{D}(A)\right] + (1 - \beta)\, \mathbb{E}(\log(1 - \mathbb{D}(v_i))) \tag{7}$$

where β denotes the identity ground truth, $\mathbb{D}(v_i)$ means the probability that synthetic face v_i is directly chosen from the raw video V_i. The primary superiority of this method is that it is able to offer a robust aggregated keyframe representation that can accurately mitigate the original surveillance clips with noisy information. In this way, we can efficiently minimize the distance between aggregated face and the original surveillance frames. Algorithm 1 describes the overall training process.

Algorithm 1 Learning Algorithm in PA-GAN With $\mathcal{L}_{Dis} + \mathcal{L}_{Rec} + \mathcal{L}_{Adv}$

Input: Training video samples $\{V_i | i = 1, 2, \ldots, n$, initialized n class centers $\{c_{yi} | i = 1, 2 \ldots, n\}$, learning rate l_r, hyperparameters λ_1 and λ_2, iterative number I_t.
Output: Generator Network $\mathbb{G}$

1: Initialize $\mathbb{G}$, $\mathbb{D}$ with pretrained model
2: Initialize feature extraction model F
3: **for** t < I_t **do**
  1: Generate a same-identity face from a video $v_i = \mathbb{G}(V_i)$
  2: Extract the simulated face feature $\tilde{f}_i = F(v_i)$
  3: Calculate the joint loss by $\mathcal{L}^t = \lambda_1 \mathcal{L}^t_{Dis} + \lambda_2 \mathcal{L}_{Rec} \mathcal{L}^t_{Rec} + \mathcal{L}^t_{Adv}$
  4: Calculate the backpropagation error $\frac{\partial \mathcal{L}^t}{\partial v_i^t}$ for each sample i by $\frac{\partial \mathcal{L}^t}{\partial v_i^t} = = \lambda_1 \frac{\partial \mathcal{L}^t_{Dis}}{\partial v_i^t} + \lambda_2 \frac{\partial \mathcal{L}^t_{Rec}}{\partial v_i^t} + \frac{\partial \mathcal{L}^t_{Adv}}{\partial v_i^t}$
  5: Update c_{yi} for each center and layer weights
  6: t = t +1
end for
return Network $\mathbb{G}$

E. FEATURE EXTRACTION NETWORK

With the continuous improvement of the performance of static unrestricted face recognition methods, more discriminative features can be extracted by the deep convolution face networks trained on enormous data samples. In this work, we build an analogical face feature extraction network presented in [22]. The specific architecture is depicted in the table 1. Each convolutional layer is followed by a ReLU unit [43], except the last one. Additionally, we add two batch normalization layers in the first convolution to mitigate the effects of illumination variations. For handling few generated

TABLE 1. The architecture of feature extractor.

Type/Stride	Filter Size	Type/Stride	Filter Size
Conv11 / S1	3 × 3	Conv41 / S1	3 × 3
Conv12 / S1	3 × 3	Conv42 / S1	3 × 3
Max Pool / S2	2 × 2	Max Pool / S2	2 × 2
Conv 21 / S1	3 × 3	Conv 51 / S1	3 × 3
Conv22 / S1	3 × 3	Conv52 / S1	3 × 3
Max Pool / S2	2 × 2	Avg Pool / S2	2 × 2
Con31 / S1	3 × 3	Dropout	40%
Conv32 / S1	3 × 3	FC	Fully connected
Max Pool / S2	2 × 2	loss	softmax

face images from generator, we append average pooling to the last convolution operation to aggregate the multiple images.

IV. EXPERIMENTAL RESULTS

A. BENCHMARK DATASETS AND PROTOCOLS

In watch-list identification task, most appeared faces are not included of the interest list. Thus, it leads to the open-set protocol. In order to better simulate the real surveillance video recognition, we employ two widely used benchmark datasets IJB-A [5] and QMUL-SurvFace [1] to evaluate the robustness of our models.

1) IJB-A [5]

The primary purpose of IJB-A dataset is to accelerate the frontiers of unconstrained face recognition. It includes not only the still image of the person being photo-graphed, but also video fragments of the person being photographed. It contains 500 subjects approximately 11.4 pictures and 4.2 videos each person. The subjects sample deliberately include a broader geographical distribution. This can effectively increase the recognition rate of the model for different races. Most of the subjects have huge changes in facial expression, illumination and different resolutions. The subjects are also from different countries, regions and races of the world, with a wide range of regions. It is because the IJB-A data set has realistic application features that the data set is very suitable for practical application scenarios. Of course, it also offers great challenges.

2) QMUL-SurvFace[1]

Compared to the previous face recognition benchmarks, QMUL-SurvFace directly sampled from 17 person re-identification datasets, that were collected in various real-world surveillance venues across different sites and multiple countries. It just has 0.46M low-quality images from 15,573 unique subjects with severe blur. This dataset presents the challenges of different training and testing environments, uncontrolled illumination, low resolution, less gallery and test data, head pose orientation and a large number of classes. QMUL-SurvFace is exceptionally characterized by very low-resolution faces typical in video surveillance. The average resolution is 24 × 20 pixel-wise.

In the verification process, we use two type evaluation indexes: (1) the True Accept Rate (TAR) representing the proportion of correct acceptance; and (2) the False Accept Rate (FAR) meaning the proportion of false acceptance. We use the paired TAR@FAR measure. We choose the standard measure as the open-set face identification performance metrics: (1) the False Positive Identification Rate (FPIR), which is the fraction of comparisons between probe templates and non-mate gallery templates which corresponds to a match score exceeding the threshold; and (2) the Ture Positive Identification Rate (FNIR), which is the fraction of probe searches that fail to match a mated gallery template above a score of the threshold.

B. IMPLEMENTATION DETAILS

We detect face area and mark 5 points landmarks by a recent method MTCNN [48], and then use the similarity trans-formation to normalization. Considering the limitations of the training data set, we initialize the input faces of generator less than 20 frames. During training and testing stage, we resize all face images in the methods of bicubic interpolation to the required size 128×128 pixel. Such rescaled images are still of "low resolution" as the underlying resolution is mostly unchanged. We use the stochastic gradient descent with minibatch size 128. We set hyper-parameters, $\lambda_1 = 0.1$, $\lambda_2 = 0.2$. Momentum of 0.9, and weight decay of $1e^{-4}$. In our setting, the learning rate is initialized to $1e^{-2}$, and during fine-tuning, the learning rate is initialized to $1e^{-3}$. We utilize both ResNet26 and ResNet50 [23] as the baselines.

The experiment is implemented by Pytorch framework [26] on a machine with four GeForce RTX2080Ti GPUs and 11GB memory for neural network training.

C. RESULT AND ANALYSIS

1) ABLATION STUDY ON PA-GAN

In this section, to evaluate the efficiency of the PA-GAN, we investigate different architectures and loss functions on IJB-A dataset to verify the improvements of the aforementioned

TABLE 2. The explanation of ResNet34 and patch-attention network.

Layers	ResNet-34	Layers	PaNet-15
Conv1	$7 \times 7, 64, stride\ 2$	Transition	2×2, max pool
		PA1	$\begin{bmatrix} 3 \times 3, & 16 - \text{d pa} \\ 64 - \text{d linear} \end{bmatrix} \times 3$
		Transition	2×2, max pool
Conv2	$\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64 \end{bmatrix} \times 3$	PA2	$\begin{bmatrix} 7 \times 7, & 64 - \text{d pa} \\ 256 - \text{d linear} \end{bmatrix} \times 2$
		Transition	2×2, stride 2 max pool
Conv3	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix} \times 4$	PA3	$\begin{bmatrix} 7 \times 7, & 128 - \text{d pa} \\ 512 - \text{d linear} \end{bmatrix} \times 3$
		Transition	2×2, stride 2 max pool
Conv4	$\begin{bmatrix} 3 \times 3, & 256 \\ 3 \times 3, & 256 \end{bmatrix} \times 6$	PA4	$\begin{bmatrix} 7 \times 7, & 256 - \text{d pa} \\ 1024 - \text{d linear} \end{bmatrix} \times 5$
		Transition	2×2, stride 2 max pool
Conv5	$\begin{bmatrix} 3 \times 3, & 512 \\ 3 \times 3, & 512 \end{bmatrix} \times 3$	PA5	$\begin{bmatrix} 7 \times 7, & 512 - \text{d pa} \\ 2048 - \text{d linear} \end{bmatrix} \times 2$
		Transition	2×2, stride 2 max pool
average pool, fc, softmax		average pool, fc, softmax	

TABLE 3. Improvement for each component of PA-GAN.

Backbone	Adversarial loss	Discriminative loss	Reconstruction loss	Accuracy	Params
ResNet-34	✓			91.22±0.90	36.4M
	✓	✓		92.45±0.76	
	✓		✓	92.35±0.96	
	✓	✓	✓	93.21±0.81	
PANet-15	✓			93.16±0.88	33.1M
	✓	✓		94.44±0.70	
	✓		✓	94.24±0.82	
	✓	✓	✓	95.21±0.65	

constituents. We begin from the baseline ResNet34 with basic adversarial loss, then gradually add other losses of the model. For fair comparison the effects of each network module, we use semblable network architecture, PANet15, following the same strategy. From Table 3 we observe that: 1) using ResNet-34 as the baseline, a well discriminative ability is achieved. 2) by adding discriminative loss and reconstruction loss, the performance obviously improved, which indicates that both losses can encourage model to have a high discrimination power. As an illustration from the results, we find that $\mathcal{L}_{Dis}$ is more powerful than $\mathcal{L}_{Rec}$, increased by 1.25% and 1.10% respectively. 3) setting the PANet-15 as the backbone, the recognition accuracy has been significantly improved by 2%. Notably, due to the consecutive transition layers, it reduces the parameter befittingly.

2) RESULTS ON IJB-A

To demonstrate the advantage of PA-GAN, we tested the proposed methods on the IJB-A dataset [5]. The performance comparison in terms of TAR@FAR, TPIR@FPIR and Rank-N on IJB-A are reported in Table 2 and III. In general, the CNN+MaxPool performs worst among the baseline methods. Although most images of IJB-A collected by unconstrained environments, the image quality keeps a high standard. CNN+AvgPool method performs slightly better @FAR $= 0.1$, but it drops a lot in the more rigorous @FAR $=$ imple 0.001.

Intuitively, our PA-GAN19 always achieves compelling search results in TAR@FAR $= 0.01$ and Rank1, which well proves it is robust to extract unconstrained face feature. In the light of these results, model B achieves a consistently superior accuracy (TAR and TPIR) than model A on both 1:1 face verification and 1: N face identification. PA-GAN outperforms all its baselines by appreciable margins, especially on the low FAR cases. For example, in the verification task, the TARs of our PA-GAN at FARs of 0.01 and 0.001 are respectively 0.968 and 0.923, improves the accuracy by 1.10% and 2.00% over the second best in verification

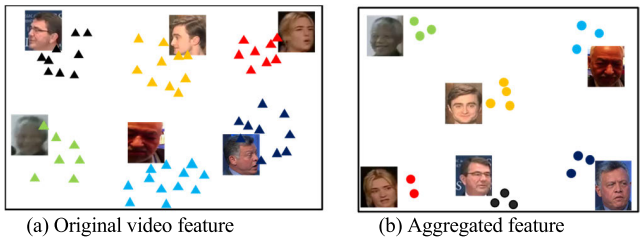
TABLE 4. Comparative performance analysis on IJB-A benchmark for verification. Symbol "–" represents the result not available.

Method	1: 1 Verification		
	TAR@FAR = 0.1	TAR@FAR = 0.01	TAR@FAR = 0.001
LSFS [30]	0.895±0.0013	0.733±0.034	0.514±0.060
DR-GAN [31]	-	0.774±0.027	0.539±0.043
DCNN [32]	0.967±0.009	0.838±0.042	-
Triplet Embedding [34]	0.964±0.005	0.900±0.010	0.813±0.020
VGGFace [8]	-	0.805±0.030	-
VGGFace2[35]	0.985±0.002	0.958±0.004	0.904±0.020
NAN [2]	0.978±0.003	0.941±0.008	0.881±0.011
Template Adaptation [27]	0.979±0.004	0.939±0.013	0.836±0.027
CNN+Maxpool	0.601±0.024	0.345±0.025	0.202±0.029
CNN+Avepool	0.977±0.004	0.913±0.014	0.771±0.064
Model A (PA-GAN 15)	0.968±0.005	0.960±0.008	0.911±0.014
Model B (PA-GAN 19)	0.989±0.002	0.969±0.003	0.923±0.013

TABLE 5. Comparisons performance of PA-GAN with Baselines on IJB-A. The TPIR vs. FPIR and The Rank-N Accuracies Are Presented.

Method	1: N Identification			
	FPIR = 0.1	FPIR = 0.01	Rank1	Rank5
LSFS [30]	0.613±0.032	0.383±0.063	0.820±0.024	0.929±0.013
DR-GAN [31]	-	0.774±0.027	0.855±0.015	0.947±0.011
DCNN [32]	0.790±0.033	0.577±0.094	0.903±0.012	0.965±0.008
Triplet Embedding [34]	0.863±0.014	0.753±0.030	0.932±0.010	-
VGGFace [8]	0.670±0.031	0.461±0.077	0.913±0.011	-
VGGFace2[35]	0.946±0.004	0.883±0.038	0.982±0.004	0.993±0.002
NAN [2]	0.917±0.009	0.817±0.041	0.958±0.025	0.980±0.013
Template Adaptation [27]	0.882±0.016	0.774±0.049	0.928±0.010	0.977±0.004
CNN+Maxpool	0.924±0.005	0.842±0.035	0.975±0.004	0.991±0.005
CNN+Avepool	0.942±0.003	0.882±0.032	0.982±0.004	0.992±0.002
Model A (PA-GAN 15)	0.949±0.008	0.891±0.036	0.972±0.006	0.985±0.004
Model B (PA-GAN 19)	0.972±0.004	0.930±0.040	0.987±0.003	0.990±0.002

task (@FAR = 0.001) respectively. This demonstrates that the synthesized faces by PA-GAN are photorealistic with well-preserved identity information. These video-like faces can be represented by the original video faces.



(a) Original video feature (b) Aggregated feature

FIGURE 4. Visualization of original video feature and aggregated feature. It is easy to see the aggregated features have more discriminative characteristic in a compact space.

On the other hand, NAN [2] and TP [27] trained their models on sufficient datasets over 2M face images getting impressive performance. Nevertheless, tour models were just trained on the original CASIA-WebFace [28] which comprises about 500K images. To illustrate the validity of our PA-GAN, we further visualize the aggregated face and the original face in two-dimensional space in Figure 4. This shows that PA-GAN is able to store identity information well while reducing computational costs. Generally, the PA-GAN is better than most of the other methods.

3) RESULTS ON QMUL-SURVFACE

In addition, we apply our patch-attention block and elaborate face losses to a more intricate surveillance venue. Notably, the images between CASIA-WebFace [28] and QMUL-SurvFace datasets have large domain gap. We use the domain transfer method of [29] to reconstruct the CASIA-WebFace [28]. The transferred images not only well preserve the distinctive information but also well fit the type of the low-quality surveillance video. This method can compensate the deficiency of training data. We firstly pre-train the still face recognition model on transferred CASIA-WebFace [28], then fine-tune on QMUL-SurvFace. For verification, PA-GAN gains a uniformly higher-performance (TAR) by 3.00-5.6% for TAR@FAR = 0.001-0.1 than other deep face models. Next, Model B shows higher accuracy than model A with improvement of 6.9 - 11.4% TAR@FAR = 0.001-0.1.

Table 6 and Table 7 show the testing results of open-set identification performance on QMUL-SurvFace. In verification task, even though existent best method CenterFace [25] failed to fully meet expectations at TAR@FAR = 0.01, 0.001. Despite the low-quality testing images, our models strive to enhance by 2.4% - 13.3%. In identification task, the performance of the model trained only with the QMUL-SurvFace dataset is worst. This again suggests that using limited number of samples training deep face

TABLE 6. Face verification accuracy on QMUL-SurvFace. The TAR vs. FAR are Reported.

Model	TAR@FAR				AUC
	30%	10%	1%	0.1%	
DeepID2[36]	0.806	0.600	0.282	0.134	0.841
CenterFace[25]	0.952	0.860	0.533	0.268	0.948
FaceNet[37]	0.946	0.799	0.403	0.127	0.935
VGGFace [8]	0.832	0.630	0.201	0.040	0.850
SphereFace[21]	0.800	0.636	0.341	0.156	0.839
Model A(PA-GAN15)	0.945	0.796	0.525	0.240	0.942
Model B(PA-GAN19)	0.963	0.890	0.584	0.334	0.972

TABLE 7. Comparisons performance of PA-GAN with baselines on IJB-A. The TPIR vs. FPIR and the Rank-N accuracies are presented.

Training data	QMUL-SurvFace				CASIA (refined)+QMUL-SurveFace			
Metrics	TPIR20(%) FPIR			AUC (%)	TPIR20(%) FPIR			AUC (%)
	30%	20%	10%		30%	20%	10%	
DeepID2[36]	12.5	8.1	3.3	20.8	13.8	9.5	4.6	21.5
CentreFace[25]	26.2	20.0	12.2	34.6	29.1	22.7	14.5	39.2
FaceNet[37]	10.6	7.9	3.6	18.9	13.9	9.6	4.9	21.8
SphereFace[21]	18.8	13.5	7.0	26.6	23.0	17.2	9.3	29.6
Ours (PA-GAN15)	23.5	17.8	11.4	31.7	26.2	20.7	14.7	34.7
Ours (PA-GAN19)	30.2	22.1	14.6	37.6	33.8	26.2	16.8	42.1

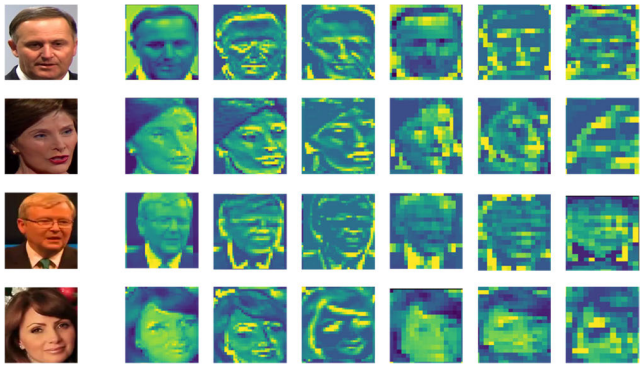


FIGURE 5. Some feature mapping visualization examples of the patch-attention blocks. We found that more attention was focused on the facial organs.

recognition model cannot deal with the challenging problem of unconstrained low-quality face recognition. We observe that the PA-GAN acquire very competitive performance compared with recent proposed methods, by the improvement of 2.3%, 2.5% and 4.7% at TPIR@FPIR = 0.1-0.3, respectively.

In such case, our experimental results confirmed PA-GAN is more practical to extract robust feature, and its aggregated feature representation is more favorable for the video face recognition task. It illustrates that patch-attention can enhance the discriminative ability by adaptively focusing on the feature mapping. Moreover, we gain better results by further augmenting training data. Indeed, the models trained on the transferred CASIA-WebFace [28] show better performance than the original results provided by [1], with the improvement of 4.5% and 2.0%, respectively. But they are still far from the practical demands of the intelligent surveillance system.

V. CONCLUSION

In order to recognize face in surveillance efficiently, a novel Patch-Attention based Generative Adversarial Network (PA-GAN) is proposed in this article. PA-GAN combines patch-attention learning model and unlabeled face training to exactly discard the misleading frames and aggregates the useful information of an input video. One promising potential function of the PA-GAN is for shrinking intra-class distance and enlarging inter-class distance in the feature space. Furthermore, runtime is reduced as we only need to pass a few output images through feature extraction network for recognition. Experimental results on two widely used datasets demonstrate the effectiveness of our framework.

REFERENCES

[1] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," 2018, *arXiv:1804.09691*. [Online]. Available: https://arxiv.org/abs/1804.09691

[2] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5216–5225.

[3] S. Gong, Y. Shi, N. D. Kalka, and A. K. Jain, "Video face recognition: Component-wise feature aggregation network (C-FAN)," in *Proc. Int. Conf. Biometrics (ICB)*, Crete, Greece, Jun. 2019, pp. 1–8.

[4] J. Zhao, J. Li, X. Tu, F. Zhao, Y. Xin, J. Xing, H. Liu, S. Yan, and J. Feng, "Multi-prototype networks for unconstrained set-based face recognition," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Macao, China, Aug. 2019, pp. 4397–4403.

[5] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark a," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1931–1939.

[6] G. S. Ekladious, H. Lemoine, E. Granger, k. Kamali, and S. Moudache, "Dual-triplet metric learning for unsupervised domain adaptation in video-based face recognition," presented at the 19th Int. Joint Conf. Neural Netw., Jul. 2020.

[7] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Conf. Learn. Represent.*, Toulon, France, Apr. 2017, pp. 1–13.

[8] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, Swansea, U.K., 2015, pp. 1–12.

[9] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[10] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2144–2157, Dec. 2014.

[11] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4694–4703.

[12] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa, "An end-to-end system for unconstrained face verification with deep convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Boston, MA, USA, Dec. 2015, pp. 118–126.

[13] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller, "One-to-many face recognition with bilinear CNNs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, New York, NY, USA, Mar. 2016, pp. 1–9.

[14] S. Bodhe, P. Kapse, and A. Singh, "Real-time age-invariant face recognition in videos using the scatternet inception hybrid network (SIHN)," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Seoul, South Korea, Oct. 2019, pp. 1112–1120.

[15] S. Gong, Y. Shi, and A. Jain, "Low quality video face recognition: Multi-mode aggregation recurrent network (MARN)," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 1027–1035.

[16] Y. Rao, J. Lu, and J. Zhou, "Attention-aware deep reinforcement learning for video face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3951–3960.

[17] J. Zheng, R. Yu, J.-C. Chen, B. Lu, C. Castillo, and R. Chellappa, "Uncertainty modeling of contextual-connections between tracklets for unconstrained video-based face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 703–712.

[18] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 1002–1014, Apr. 2018.

[19] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.

[20] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," in *Proc. Conf. Learn. Represent.*, Vancouver, BC, Canada, Apr./May 2018, pp. 1–14.

[21] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6738–6746.

[22] J.-C. Chen, R. Ranjan, S. Sankaranarayanan, A. Kumar, C.-H. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa, "Unconstrained still/video-based face verification with deep convolutional neural networks," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 272–291, Apr. 2018.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[24] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," presented at the 30th IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2020.

[25] Y. Wen, K. Zhang, and Z. Li, "A discriminative feature learning approach for deep face recognition," in *Proc. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 499–515.

[26] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. D. Vito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 1–4.

[27] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 1–8.

[28] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923

[29] J. Liu, Q. Li, P. Zhang, G. Zhang, and M. Liu, "Unpaired domain transfer for data augment in face recognition," *IEEE Access*, vol. 8, pp. 39349–39360, 2020.

[30] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1122–1136, Apr. 2017.

[31] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1283–1292.

[32] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep CNN features," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Placid, NY, USA, Mar. 2016, pp. 1–9.

[33] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4838–4846.

[34] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Niagara Falls, NY, USA, Sep. 2016, pp. 1–8.

[35] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, Xi'an, China, May 2018, pp. 67–74.

[36] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 1988–1996.

[37] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 815–823.