Noname manuscript No.
(will be inserted by the editor)

Learning Discriminative Aggregation Network for Video-based Face Recognition and Person Re-identification

Yongming Rao, Jiwen Lu, Jie Zhou

Received: date / Accepted: date

Abstract In this paper, we propose a discriminative aggregation network (DAN) method for video-based face recognition and person re-identification, which aims to integrate information from video frames for feature representation effectively and efficiently. Unlike existing video aggregation methods, our method aggregates raw video frames directly instead of the features obtained by complex processing. By combining the idea of metric learning and adversarial learning, we learn an aggregation network to generate more discriminative images compared to the raw input frames. Our framework reduces the number of image frames per video to be processed and significantly speeds up the recognition procedure. Furthermore, low-quality frames containing misleading information can be well filtered and denoised during the aggregation procedure, which makes our method more robust and discriminative. Experimental results on several widely used datasets show that our method can generate discriminative images from video clips and improve the overall recognition performance in both the speed and the accuracy for video-based face recognition and person re-identification.

Keywords Face recognition · person re-identification · metric learning · adversarial learning · video-based recognition

Yongming Rao[1]
E-mail: raoyongming95@gmail.com

Jiwen Lu[1]
E-mail: lujiwen@tsinghua.edu.cn

Jie Zhou[1]
E-mail: jzhou@tsinghua.edu.cn

[1]Department of Automation, Tsinghua University, State Key Lab of Intelligent Technologies and Systems, and Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, 100084, China. Partial of this work was presented in (Rao et al, 2017)

1 Introduction

Video-based face recognition and person re-identification have been attracting increasing efforts and interests over the past few years (Wolf et al, 2011; Hu et al, 2014a; Beveridge et al, 2013; Chen et al, 2015; Yang et al, 2016a; Li et al, 2014a; Wen et al, 2016; Schroff et al, 2015; Parkhi et al, 2015; Chen et al, 2016a; Zheng et al, 2016; Wang et al, 2016), which has many potential practical applications such as visual surveillance and scalable video search. Compared to image-based face recognition and person re-identification, video-based recognition is more challenging because there are many noisy frames in face and person videos which contain unfavorable poses and viewing angles. Furthermore, as the video usually consists of plenty of frames (e.g. more than 100 frames), it brings considerable computational burdens for the state-of-the-art video-based recognition methods such as the deep neural networks. Therefore, it is desirable to present a framework that can denoise the original videos by extracting useful information from noisy data and reducing the overall runtime. In other words, a new framework which can aggregate the information from raw videos and keep the same or even higher discriminative ability for efficient video-based face recognition and person re-identification is desirable and required. Since faces in video usually have similar appearance but vary in pose, image quality, occlusion and etc, it is possible to integrate the information across video frames and denoise the input video at the same time.

There have been varieties of efforts on integrating information from different image frames to represent the whole video (Huang and Van Gool, 2016; Huang et al, 2016; Yang et al, 2016a; Lu et al, 2016; Cevikalp and Triggs, 2010). However, most of them focus on extracting features from raw video frames, which means that feature extraction is performed at first before the matching procedure. This kind of strategy will harm the recognition performance because

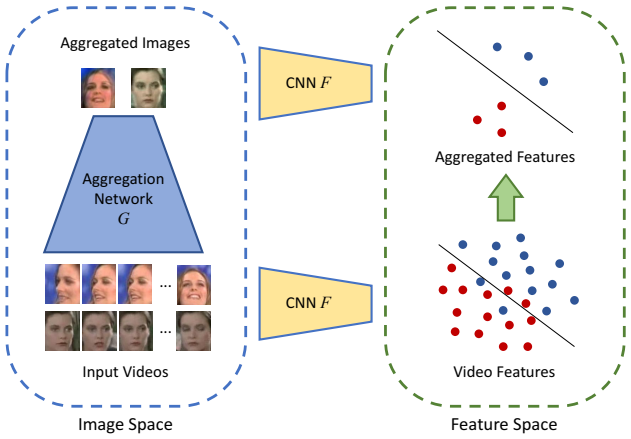

Fig. 1 The basic idea of our proposed video frame aggregation method, where we take video-based face recognition as an example. For each face video clip, we integrate the information of each video to produce few synthesized face images with the discriminative aggregation network before feature extraction. The supervision signal of our proposed framework makes the synthesized images more discriminative than original frames in the feature space. Having completed the video frame aggregation procedure, we only need to pass the few aggregated face images into the feature extraction network and thus greatly speed up the overall face recognition system.

some image frames of low quality may mislead the system into wrong decisions, which cannot be easily distinguished in the feature space because such information is usually lost during the feature extraction stage. Therefore, it is important to conduct video frame aggregation before feature extraction for video-based face recognition and person re-identification.

Generative adversarial networks (GAN) have achieved great success in many fields of computer vision (Goodfellow et al, 2014; Chen et al, 2016b; Radford et al, 2015; Isola et al, 2016; Zhang et al, 2016a; Larsen et al, 2015; Reed et al, 2016). Inspired by the basic idea of adversarial learning, we propose a GAN-like aggregation network which takes an video clip as the input and reconstruct a single image as the output. However, the output image produced by the generative adversarial network is only visually similar to the original data, but does not guarantee any discriminative power. On the other hand, metric learning (Hu et al, 2014a; Schroff et al, 2015; Guillaumin et al, 2009; Wen et al, 2016) has been one of the most discriminative techniques in face recognition and person re-identification, which maps samples into a semantic feature space where they can be well distinguished. By combining metric learning with adversarial learning, we are able to train a generative model that can produce photo-realist images and provide even stronger discriminative ability simultaneously.

In this paper, we propose a discriminative aggregation network (DAN) method for video-based face recognition and person re-identification, where the overall framework is shown in Fig. 1. By combining metric learning and adversarial learning, DAN can aggregate the useful information of an input video into one or several images that are discriminative in the feature space for recognition. Since the number of images to be processed is greatly reduced, our framework significantly speeds up the whole recognition system. Unlike existing methods which extract features from raw video frames before other information fusion strategies, our framework directly fuses the information from one raw video into several images, and can thus distinguish low quality frames and denoise the input video simultaneously. For video clips with large pose variations (e.g. video tracklets in the person re-identification task), we propose a video-level spatial transformer network (V-STN) to align video frames before aggregation, which is a video extension of spatial transformer networks (Jaderberg et al, 2015) and aligns video clips by utilizing cross-frame information. As a part of DAN, V-STN can also be learned in an end-to-end manner without extra labels and stabilize the training of DAN.

The contributions of this work are summarized as follows:

1. We propose a framework to aggregate video clips before feature extraction for video-based face recognition and person re-identification. By combining the idea of metric learning and adversarial learning, we develop a discriminative aggregation network (DAN) to boost the recognition performance and reduce the computational complexity simultaneously.
2. We propose a video-level spatial transformer network (V-STN) as an important part of DAN to align video clips before aggregation and utilize cross-frame information, and apply it for video-based person re-identification to handle large pose and viewpoint variations.
3. We conduct extensive video-based face recognition and person re-identification experiments with detailed ablation studies to demonstrate the effectiveness of our proposed approach. Experimental results on four widely used datasets including the YouTube Face (YTF) (Wolf et al, 2011), Point-and-Shoot Challenge (PaSC) (Beveridge et al, 2013), YouTube Celebrities (YTC) (Kim et al, 2008), IARPA Janus Benchmark-A (IJB-A) (Klare et al, 2015), IARPA Janus Benchmark-B (IJB-B) (Whitelam et al, 2017) and Motion Analysis and Re-identification Set (MARS) datasets (Zheng et al, 2016) are presented to show that DAN can accelerate the recognition speed and improve the recognition performance simultaneously.

2 Related Work

In this section, we briefly review four related topics: 1) video-based face recognition, 2) video-based person re-identification,

3) deep metric learning, and 4) conditional image generation.

2.1 Video-based Face Recognition

Existing video-based face recognition methods (Yang et al, 2016a; Lu et al, 2016; Wang et al, 2012; Lu et al, 2013; Hu et al, 2011; Huang et al, 2014, 2015; Lu et al, 2015; Taigman et al, 2014; Schroff et al, 2015; Parkhi et al, 2015; Sun et al, 2015; Wen et al, 2016; Chen et al, 2012; Hu et al, 2011) can be mainly categorized into two classes: still-based and video-based. For the first category, each video is considered as a set of images and the information from different frames is integrated for recognition. These methods are designed to solve the general still-based face recognition, where they can be easily applied into video-based face recognition (Taigman et al, 2014; Schroff et al, 2015; Parkhi et al, 2015; Sun et al, 2015; Wen et al, 2016; Ding and Tao, 2017). We consider this type of methods as the basis of video-based face recognition, and our model is built upon this type of methods. For the second category, each video is modeled as an image set, and the similarity between videos is computed by the properties of image sets. In previous works, image-set-based models have a variety of forms. For example, Cevikalp *et al.* and Hu *et al.* modeled image sets as affine hulls (Cevikalp and Triggs, 2010; Hu et al, 2011). Huang *et al.* calculated the distances between image sets using the distances between SPD manifolds (Huang et al, 2015; Huang and Van Gool, 2016). Lu *et al.* represented image sets as a set of n-order statistics (Wang et al, 2012; Lu et al, 2013). Yang *et al.* proposed an attention-based model to aggregated features of image sets (Yang et al, 2016a). Chen *et al.* built a dictionary-based method to model temporal information in video (Chen et al, 2012). For methods from both classes, the key challenge is how to represent a video as a single feature. In their works, they first represent frames in videos using handcrafted feature vectors or deep neural networks, and then aggregate these features. To obtain robust and discriminative representation for face recognition, there have been many efforts to train a more powerful and robust still-based face recognition model such as (Cao et al, 2018; Wright et al, 2009; Ding and Tao, 2017; Wang et al, 2014). Different from these works that try to improve recognition model, we focus on improving the quality of input information. In this work, we represent face videos in a different way, where we aggregate image frames at the beginning and speed up the recognition process by using an adversarial learning approach.

2.2 Video-based Person Re-identification

In recent years, video-based person re-identification methods have attracted great attention in computer vision (Zheng et al, 2016; Wang et al, 2016; Xiao et al, 2016; Zhong et al, 2017; Tesfaye et al, 2017; Hermans et al, 2017; Zhou et al, 2017; Lin et al, 2017). Similar to video-based face recognition, existing video-based person re-identification methods can be also divided into two major groups: generic methods for solving both still-based and video-based person re-identification, and methods exploit the characteristics from image sequences or image sets for person re-identification. Feature matters in the person re-identification problem. For example, (Xiao et al, 2016) proposed a domain guided dropout method to combine training data from different datasets, which is the first step to solve the insufficient data in person re-identification. In (Hermans et al, 2017), an modified triplet loss was used to achieve better optimization for feature extractor. Sequential information has been proven to be useful to boost the video-based person re-identification performance. For example, (Zheng et al, 2016; Wang et al, 2016; Zhou et al, 2017) combined the spatial and temporal information in person videos and observed the importance of sequential information in video-based person re-identification. Furthermore, structural information of the camera network was exploited in (Lin et al, 2017), which is also helpful for both image-based and video-based person re-identification. However, efficient models for video-based re-identification have not been visited yet, which is desirable to improve the recognition speed of a practical person re-identification system.

2.3 Deep Metric Learning

A variety of metric learning algorithms have been proposed in recent years, and many of them have been successfully applied to improve the performance of face recognition and person re-identification systems (Hu et al, 2014a; Schroff et al, 2015; Guillaumin et al, 2009; Wen et al, 2016). However, most previous metric learning methods learn a linear mapping to project samples into a new feature space, which suffer from the nonlinear relationship of face and person samples, which usually lie on or nearby a nonlinear manifold. Deep metric learning (Hu et al, 2014a) aims to produces discriminative features through the combination of deep learning and metric learning, where its key idea is to explicitly learn a set of hierarchical nonlinear transformations to map samples into other feature space for matching, which unify feature learning and metric learning as a joint learning framework. These deep metric learning methods have shown state-of-the-art performance for various visual understanding applications such as face recognition, person

re-identification, cross-modal matching, and image set classification. For example, Hu *et al*. (Hu et al, 2014a) employed a fully connected network to achieve parametric metric learning. Schroff *et al*. presented a triplet loss function for feature embedding. Wen *et al*. proposed the center loss function to improve the faces distribution in feature space. Different from deep metric learning methods, we propose a new *image embedding* method to guide an aggregation network to synthesize discriminative images for recognition tasks.

2.4 Conditional Image Generation

Conditional image generation aims to generate image based on input condition for a specific purpose. Image super-resolution (Dong et al, 2014, 2016) is one of the most important application of image generation techniques. Similar with the idea of super-resolution, in this work, we try to improve the quality of input video via a generative model. Goodfellow *et al.* (Goodfellow et al, 2014) proposed the idea of generative adversarial networks (GAN), which has attracted great attention in computer vision in recent years (Chen et al, 2016b; Radford et al, 2015; Isola et al, 2016; Zhang et al, 2016a; Larsen et al, 2015; Reed et al, 2016). Compared to conventional generative models, GAN has shown promising performance for generating sharper images, which demonstrated strong abilities of photo-realistic image synthesis and has been applied in many vision tasks. For example, Larsen *et al.* combined a variational autoencoder (VAE) (Kingma and Welling, 2013) with GAN to take the advantages from both models and learned a high-level abstract visual features embedding (Larsen et al, 2015). Zhang *et al.* developed the idea of deep convolutional GAN (Radford et al, 2015) and text-to-image synthesis (Reed et al, 2016) and achieved impressive results on image synthesis (Zhang et al, 2016a), Isola *et al.* studied on a variety of image-to-image translation applications in computer vision by combining the traditional n-norm distance loss and adversarial loss (Isola et al, 2016). Ledig *et al.* employed a GAN-like network by using a loss function defined by high-level features to improve the perceptual quality of image super-resolution. However, little progress has been made in exploiting the idea of adversarial learning for recognition tasks. In our work, we combine the idea of adversarial learning with metric learning to aggregate photo-realistic images discriminatively for boosting the video-based face recognition and person re-identification performance.

3 Approach

In this section, we first formulate the problem of video-based face recognition and person re-identification with the proposed discriminative aggregation network. Then we show the overall framework and the training methodology of our proposed method. Lastly, we present the proposed V-STN method for video clip alignment before aggregation, which aims to handle the large pose and viewpoint variations of persons in videos across different cameras.

3.1 Problem Definition

We first take video-based face recognition as an example to illustrate the basic idea of our proposed method. Video-based face recognition aims to recognize whether a face video belongs to a certain subject. Such videos usually contain more than 100 frames (like videos in the YouTube Face dataset) and brings considerable computational burdens for existing methods. The goal of our discriminative aggregation network (DAN) framework is to aggregate a long video into one or a few frames while still remains or increases the discriminative power, which can be used for effective and efficient face recognition.

Specifically, we denote our goal as the following objectives:

$$V^m \rightarrow X^n \tag{1}$$

subject to

$$m > n \tag{2}$$

and

$$Dis(F(X_p), F(X_n)) > Dis(F(V_p), F(V_n)), \tag{3}$$

where V^m is the input video with m frames and X^n is the aggregated n images, with m is much greater than n. The subscripts p and n refer to positive and negative samples and F is the feature extraction network. We used a function Dis to evaluate the discriminative ability between positive and negative samples. This means that we can greatly reduced the number of images to be processed with DAN, while the aggregated images still have more discriminative ability in the feature space of certain CNN F. By maximizing the discriminative ability of synthesized images and limiting the number of input frames, only the most informative frames are remained after aggregation, which enables our model to identify low quality frames and denoise input videos.

3.2 Overall Framework

Fig. 2 shows the overall framework of our proposed discriminative aggregation network (DAN) framework. DAN consists of 3 sub-networks, which are defined as the aggregation (generator) network G, the discriminator network D and the feature extraction network F. We denote the whole video as V. For the ease of implementations, at each time we aggregate a subset S of V into a single image, so that the input of G

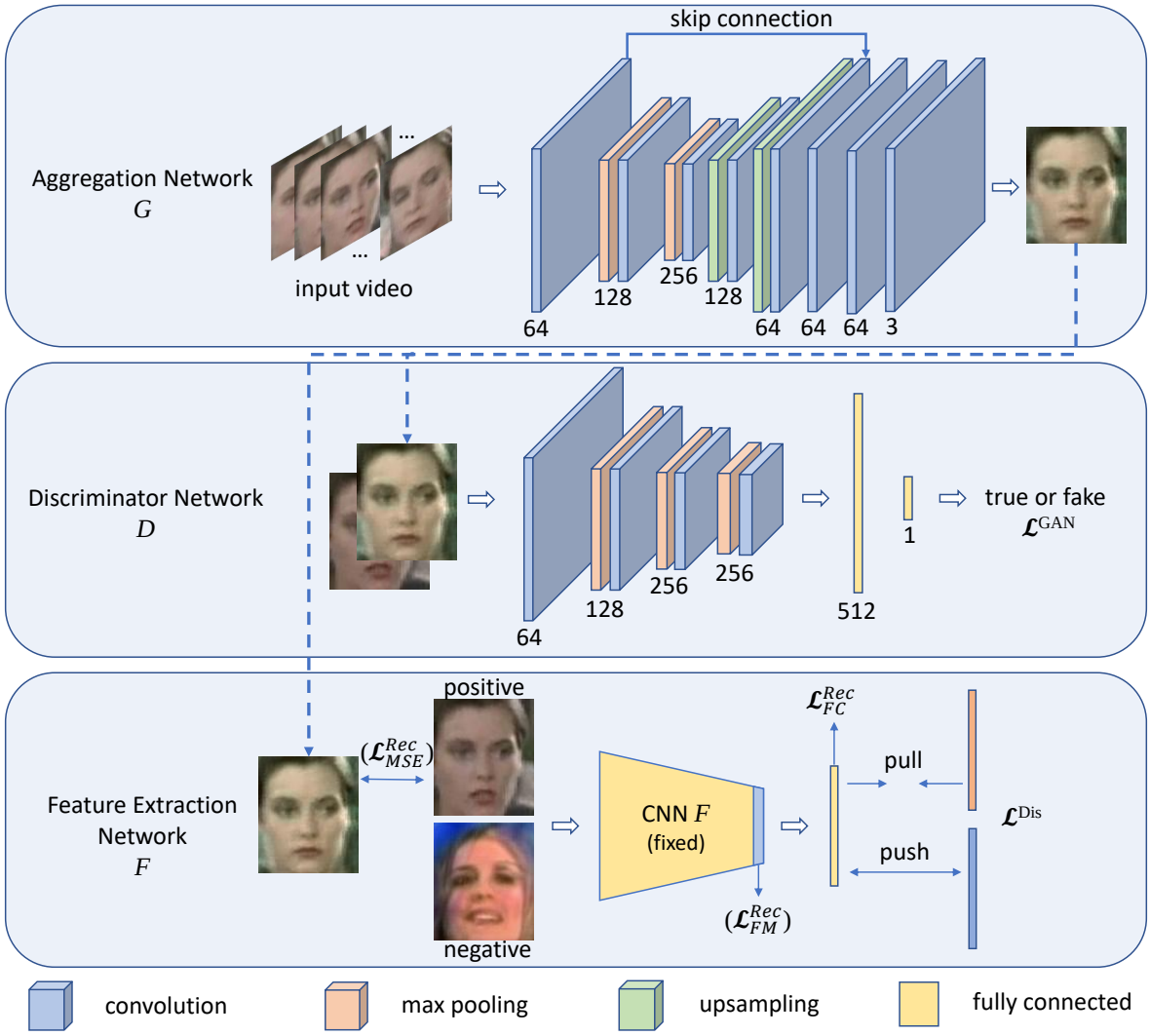

Fig. 2 Detailed architecture of our proposed framework. The numbers are either the feature map channel for convolutional blocks or feature dimension for fully connected layers. The synthesized image has the same height and width as the input video frame. The output of aggregation network is then fed into the discriminative network for adversarial learning, and the feature extraction network to increase the discrimination. Different losses are applied at different places as illustrated in this figure.

is a subset S and the output is a single discriminative image X. The discriminator D aims to judge whether the image is generated by G or selected from the original video, forming adversarial learning with G. The feature generator network F extracts features from the aggregated images, and tries to make the feature as discriminative as possible in the feature space.

The aggregation network G starts with several convolution blocks into smaller feature maps, and then reconstructs the aggregated output image with several deconvolution blocks, where the kernel size of all convolution layer is 3×3. We also add a skip connection between the first and high-level feature maps by following another GAN based framework (Ledig et al, 2016). For ease of implementation, input frames are directly concatenated along their channel dimension. The discriminator network D consists of several convolution blocks and finally produces 1 output denoting whether the image is generated or selected from the original video. For the aggregation network, each convolutional block consists of a standard convolution layer and a batch normalization layer (Ioffe and Szegedy, 2015). For the discriminator network, each convolutional block is a standard convolution layer. The kernel size and stride of convolutional layers is 3×3 and 1×1 respectively. All max pooling layers have the kernel size of 2×2 and a stride of 2. For the upsampling layer, we use a bilinear filter with stride 2 to upscale feature maps. All hidden layers in both G and D use PReLU (He et al, 2015) as activations, and the output layer of G uses the `tanh` nonlinearity to produce normalized pixel values. The output layer of D use the `sigmoid` nonlinearity to produce the possibil-

ity of whether the input image is true or synthesized. G and D are trained iteratively so that they can provide loss signal to each other to reach an optimal balance, where the generated output of G cannot be distinguished from ground-truth images. For the feature extraction network F, we use the network provided by the author of (Wen et al, 2016), which is a residual convolutional network (He et al, 2016). We keep F unchanged during the training process.

3.3 Loss Function

We expect that our framework DAN can aggregate a video clip into a single image while at the same time gain more discriminative power. To achieve this, we design the following loss function:

$$\mathscr{L} = \lambda \mathscr{L}^{Dis} + \eta \mathscr{L}^{Rec} + \gamma \mathscr{L}^{GAN} \tag{4}$$

where $\mathscr{L}^{Dis}$ is the discriminative loss, $\mathscr{L}^{Rec}$ is the reconstruction loss, and $\mathscr{L}^{GAN}$ is the adversarial loss. We set the weight of adversarial loss γ to 0.01 following (Isola et al, 2016) and set λ and η to 1 in all of our experiments.

3.3.1 Discriminative Loss

Samples from face and person datasets consists of positive video pairs and negative video pairs. We use the term (X,P) for positive pairs and (X,N) for negative pairs, where X is the aggregated image, P and N are positive and negative samples randomly chosen from the other video clips respectively.

To make the generated image discriminative, we propose the discriminative loss as follows:

$$\mathscr{L}^{Dis} = \begin{cases} (||F(X) - F(P)||^2 - \alpha)_+ & y = 1 \\ (\beta - ||F(X) - F(N)||^2)_+ & y = 0 \end{cases} \tag{5}$$

and

$$\alpha = \min_{A \in S} ||F(A) - F(P)||^2 \tag{6}$$

where y is the label either 1 or 0 denoting positive or negative pairs, F is the feature extraction network, S is the subset clip to be aggregated, and A is one of a frame in it. We use the Euclidean distance to compute the distance between two feature representations. α is the smallest distances between all frames in S and P. β is a manually set constant margin. The subscript $+$ means $\max(0,\cdot)$.

The basic idea of this loss is that if we sample a positive video pair from training data, and take a subset S of one video for aggregation and randomly sample a frame P from the other video, we expect the aggregated image X is closer to P than any other frame from the original video subset S in the feature space of F. Contrarily, if negative sample is considered, we expect the distance between generated X and N is greater than a certain margin. With such a loss function, we guarantee the feature of aggregated image extracted by F is more discriminative than original frames.

3.3.2 Reconstruction Loss

Since we reconstruct a face image from a compressed representation, we need to exert reconstruction loss on the output image. Here we compared three forms of reconstruction losses as shown in Fig. 2.

Pixel-wise MSE loss is the most widely used objective function for existing frameworks like (Dong et al, 2016; Shi et al, 2016), which is calculated as:

$$\mathscr{L}^{Rec}_{MSE} = \frac{1}{N_I} ||I - X||^2_{\mathscr{F}} \tag{7}$$

where I is the original image and X is the reconstructed one. N_I is the number of total pixels in an image.

Another reconstruction loss is the one proposed in (Ledig et al, 2016), which focuses on the feature map difference between reconstructed or original image, as shown in the bottom part in Fig. 2. The loss function is defined as follows:

$$\mathscr{L}^{Rec}_{FM} = \frac{1}{N} \sum_{i=1}^{n} ||\phi_i(I) - \phi_i(X)||^2_{\mathscr{F}} \tag{8}$$

where ϕ maps image to its high-level feature maps, and in our case. We use the convolutional part of feature extraction network F as ϕ. The subscript i denotes the index of channel, with totally n feature maps. N_{FM} is the number of total entries of feature maps.

We cannot naively define the above two forms of reconstruction loss, as there are multiple images in the input S. For ease of implementation, we choose I according to the following rule:

$$I = \begin{cases} \text{argmin}_{A \in S} ||F(A) - F(P)||^2 & y = 1 \\ \text{argmax}_{A \in S} ||F(A) - F(N)||^2 & y = 0 \end{cases} \tag{9}$$

However, the two forms of reconstruction loss both focus on visually similarity, from a shallow to upper level. They can guarantee visual characteristics but not semantic information or discriminative power. DAN focuses on the feature representation extracted from the aggregated image, so it is naturally to apply reconstruction loss to the feature embedding. Hence, we propose the following reconstruction loss:

$$\mathscr{L}^{Rec}_{FC} = ||F(X) - \text{mean}(F(V^m))||^2 \tag{10}$$

where F is the feature extraction network, and V^m is the original video consisting of m frames. We expect that the feature of reconstructed image is close to the mean of features extracted from V per frame to reduce the intra-class distance.

We present detailed analysis on these three forms of reconstruction losses in Section 4.3.

Algorithm 1 Minibatch Stochastic Gradient Descent.

Input: Training video pairs, learning rate lr, iterative number I_t, and parameter λ, η.
Output: Aggregation network G

1: Initialize G with MSE pretrained model.
2: Initialize D with pretrained model.
3: Load model of F.
4: **for** $iter < I_t$ **do**
5: **for** k steps **do**
– Sample a video V from the training set, and aggregate a subset S into image $X = G(S)$.
– Sample a frame A from the subset S
– Update the discriminator by ascending its stochastic gradient:

$$\nabla \mathscr{L}^{GAN}$$

6: **end for**
– Sample a video sampling V from the training set, and aggregate a subset S into image $X = G(S)$.
– Calculate the reconstruction target of $\mathscr{L}^{Rec}$ from selected V
– Update the aggregation by descending its stochastic gradient:

$$\nabla \mathscr{L} = \nabla(\mathscr{L}^{Dis} + \mathscr{L}^{Rec} + \mathbb{E}_{\text{batch}}[\log(1 - D(G(S))])$$

7: **end for**
8: **return** Neural network G

3.3.3 Adversarial Loss

In addition to the reconstruction loss, we also add the adversarial loss to our framework as widely adopted in GAN-based frameworks (Chen et al, 2016b; Radford et al, 2015; Isola et al, 2016; Zhang et al, 2016a; Larsen et al, 2015; Reed et al, 2016). This encourages G to generate aggregated outputs that are close to the natural distribution, by forming adversarial learning with G. The loss is defined based on the possibility whether an image comes from the original video, denoted as:

$$\begin{aligned}\mathscr{L}^{GAN} &= \mathbb{E}_{A \sim p_{\text{train}}(A)}[\log D(A)] \\ &+ \mathbb{E}_{V^m \sim p_{\text{train}}(V^m)}[\log(1 - D(G(V^m))]\end{aligned} \tag{11}$$

Here $D(G(V^m))$ is the probability that the aggregated image $G(V^m)$ is a natural image taken from the original video V^m. The goal of D is to maximize L^{GAN} while G tends to minimize it. D and G play the minimax game until reaching a balanced state. D and G are trained iteratively following commonly used settings.

The overall training procedure of our DAN method is summarized in **Algorithm 1**.

3.4 V-STN for Video-based Person Re-identification

For video-based person re-identification, frames in video clips usually have large pose and location variations, because of changing human pose and imperfect human detections (for example, human bounding boxes in (Zheng et al, 2016) were generated by the inaccurate DPM detector (Felzenszwalb et al, 2010)). However, convolutional neural networks are lack of the ability to be spatially invariant to the input date. Since the above-proposed DAN is a fully convolutional model, large variations in video clips will significantly harm the performance of our model and lead to unstable training process. To tackle this problem, we propose an extension of spatial transformer networks (Jaderberg et al, 2015), called video-level STN as a pre-processing module for input video clips, which takes the video clips as input and produces *aligned* video frames to provide spatial transformation capabilities for DAN.

The spatial transformer (Jaderberg et al, 2015) is an operator to perform spatial transformation on the input data, which consists of three components: localisation network f_{loc}, grid generator and sampler. The operator produces an affine transformation parameters θ for the input feature maps $U \in \mathbb{R}^{W \times H \times C}$ with width W, height H and C channels and applies the corresponding transformation $\mathscr{T}_\theta$ on U. Our proposed video-level spatial transformer network (V-STN) is built upon the spatial transformer operator. Different from the original spatial transformer operator, the input data of our framework is a video clip instead of a single image. To utilize the cross-frame information for video clip alignment, we design a video-level spatial transformer operator, which is detailed as follows.

Given a video clip $X \in \mathbb{R}^{N \times W \times H \times C}$ with N frames, the video-level localisation network $\hat{f}_{loc}$ aims to estimate the affine transformation parameters $\hat{\theta} \in \mathbb{R}^{N \times 2 \times 3}$ for video clip alignment. Then, a video-level grid generator and a video-level sampler are used to apply transformation parameters $\hat{\theta}$ to corresponding frames in X and produce an aligned video clip $\hat{X}$.

The video-level localisation network $\hat{f}_{loc}$ begins with several convolutional blocks into smaller feature maps, and then produce affine transformation parameters $\hat{\theta}$ with two attached fully connected layers. Each convolutional block consists of a 3×3 convolutional layer with stride 2, a batch normalization layer and a PReLU activation. The detailed architecture of V-STN is illustrated in Fig. 3.

Following (Jaderberg et al, 2015), we train the V-STN model in an end-to-end manner as a part of DAN, so that the video-level localisation network can be directly optimized for the above objectives via back propagation. More technical details about the gradient expression of STN can be found in (Jaderberg et al, 2015).

4 Experiments

We conducted experiments to evaluate our proposed DAN and V-STN methods on four widely used datasets for video-based face recognition and person re-identification, including YouTube Face (YTF) (Wolf et al, 2011), Point-and-Shoot

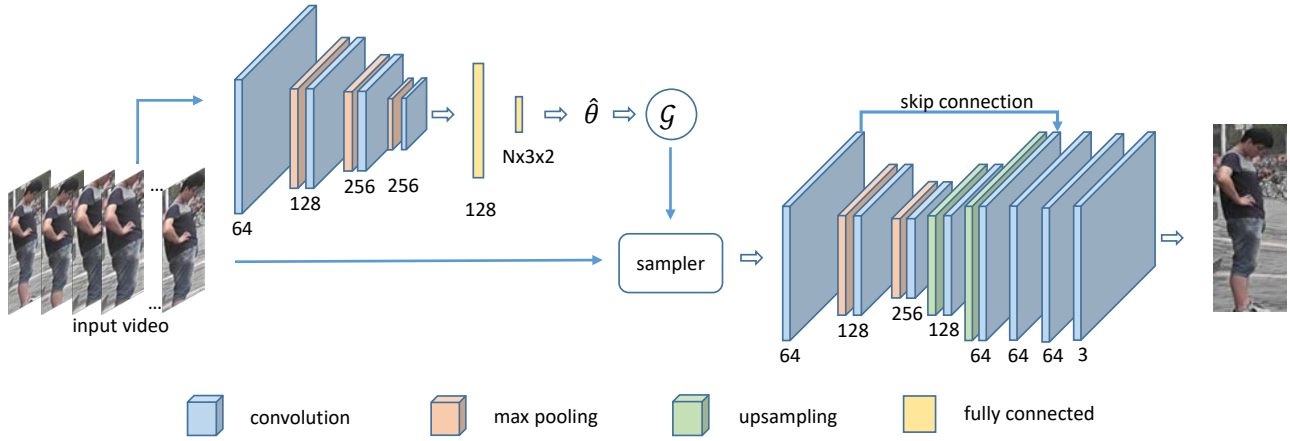

Fig. 3 Detailed architecture of V-STN. The numbers are either the feature map channel for convolutional blocks or feature dimension for fully connected layers. N is the number of frames in input video clip.

Table 1 Detailed information of the feature extraction network for video-based face recognition on YTF, PaSC and YTC. We present the number of parameters, FLOPs and face verification accuracy (%) on the LFW and YTF datasets.

# Parameters	FLOPs	LFW accuracy	YTF accuracy
2.75×10^7	2.9×10^9	97.96	93.16

Challenge (PaSC) (Beveridge et al, 2013), YouTube Celebrities (YTC) (Kim et al, 2008), IARPA Janus Benchmark-A (IJB-A) (Klare et al, 2015), IARPA Janus Benchmark-B (IJB-B) (Whitelam et al, 2017) and Motion Analysis and Re-identification Set (MARS) (Zheng et al, 2016). Specifically, we evaluated our DAN on YTF, PaSC and YTC datasets for face recognition, and DAN with V-STN on MARS for person re-identification.

4.1 Datasets and Protocols

YTF: The YouTube Face (YTF) dataset is a widely used video face dataset, which contains 3,425 videos of 1,595 different subjects. In this dataset, there are many challenging videos, including amateur photography, occlusions, problematic lighting, pose and motion blur. The length of face videos in this dataset vary from 48 to 6,070 frames, and the average length of videos is 181.3 frames. In experiments, we followed the standard verification protocol and tested our method for unconstrained face $1:1$ verification with the given 5,000 video pairs. These pairs are equally divided into 10 splits, and each split has around 250 intra-personal pairs and around 250 inter-personal pairs.

PaSC: The Point-and-Shoot Challenge (PaSC) dataset contains 2,802 videos of 265 subjects. In this dataset, videos have different distances to the camera, viewpoints, the sensor types and etc. The dataset is composed of two parts, in which videos are taken by control and handheld cameras respectively. Compared to the YTF dataset, PaSC is more challenging because faces in this dataset have full pose variations. We followed the standard $1:N$ verification protocol and tested our method on both control and handheld parts of the dataset.

YTC: The YouTube Celebrities (YTC) dataset contains 1,910 videos of 47 subjects and the number of frames varies from 8 to 400. We followed the protocol of standard ten-fold cross validation and randomly selected 3 videos for training and 6 videos for testing for each subject in each fold. We used the dataset to evaluate the performance of our method on the video-based face identification task.

IJB-A* and *IJB-B: The IARPA Janus Benchmark-A (IJB-A) consists of 5,712 images and 2,085 videos from 500 subjects, with an average of 11.4 images and 4.2 videos per subject. The IJB-A dataset is challenging since all images and videos in this dataset are captured from unconstrained environment and vary in expression and image qualities. The IARPA Janus Benchmark-B (IJB-B) datasets is an extension of IJB-A, where more than 21,800 images from 1,845 subjects and 55,000 frames from 7, 011 videos are contained in this dataset. In the IJB-A and IJB-B datasets, images and video frames that belong to the same subject are grouped into several ”templates”, where recognition algorithm needs to perform face verification or identification on these templates instead of face images or frames. For the IJB-A dataset, our model was trained on training set of each split and evaluated on the corresponding test set for both face verification and identification tasks. In the IJB-B dataset, two gallery sets that are disjoint from each other are provided. Since training set is not provided in the IJB-B dataset, we divided all templates into two subsets for training and evaluation respectively, where the training set is disjoint with the gallery set. Therefore, two different models are trained for each gallery

Table 2 Detailed information of the datasets used for training the feature extraction network for person re-identification.

Dataset	#identities	#training images	#testing images	# camera	detection method
CUHK01 (Li and Wang, 2013)	971	1552	388	2	hand
CUHK03 (Li et al, 2014b)	1467	21012	5252	5	DPM
PRID (Hirzer et al, 2011)	385	2997	749	2	-
Shinpuhkan (Kawanishi et al, 2014)	24	18004	4500	16	-
VIPeR (Gray et al, 2007)	632	506	126	2	hand
3DPeS (Baltieri et al, 2011)	193	420	104	6	hand
iLIDS (Zheng et al, 2009)	119	194	48	-	hand
Market-1501 (Zheng et al, 2015)	1501	12,936	19,732	6	DPM
Our combination	3380	63536	5252	-	-

Table 3 Detailed information of feature extraction network for video-based person re-identification. We present the number of parameters, FLOPs and rank-1 accuracy (%) on the CUHK03 and MARS datasets.

#Parameters	FLOPs	rank-1 CUHK03	rank-1 MARS
2.56×10^7	6.1×10^8	87.87	84.29

using the corresponding training set, and the 1:N identification performance is measured across two gallery sets for both mixed media and video face recognition tasks.

MARS: The Motion Analysis and Re-identification Set (MARS) dataset is an extended version of the Market1501 dataset for video-based person re-identification, which contains 20,478 tracklets of 1,268 identities captured by 6 different camera views. Bounding boxes of this dataset is automatically generated by the DPM person detection detector (Felzenszwalb et al, 2010)). Due to the inaccuracy of the detector, this dataset contains 3,248 distractors. We used the dataset to evaluate our DAN with V-STN on the video-based person re-identification task by following the evaluation codes and splits provided by the authors of MARS.

4.2 Implementation Details

In this section, we present the implementation details of our approach from five aspects: 1) data pre-precessing, 2) feature extraction networks, 3) training details, 4) testing details, and 5) experimental environments.

Data Pre-processing: For video-based face recognition, we employed the MTCNN method (Zhang et al, 2016b) to detect 5 points landmarks for each face frame per video by following (Wen et al, 2016). If the detection fails, we used the landmarks provided by datasets. Moreover, we used similarity transformation to align faces according to the landmarks, and cropped and resized faces to remove the background information. In our experiments, we use 224×224 face images for IJB-A and IJB-B and use 112×96 images for other datasets according to the input sizes of feature extraction networks. For video face identification task (Task 5) of IJB-B, since only the first frame in each video is annotated, we employ the face detection algorithm described in (Yang et al, 2016b) to find faces in the following frames. Having detected all faces in videos, we use a pre-trained face recognition provided by the author of (Cao et al, 2018) to remove faces that belong to different subject from the first frame by applying a threshold. After filtering , we obtained 7,110 templates that are composed of video frames, where the length of templates vary from 1 to 1241 frames and the average length is 13.2 frames. For video-based person re-identification, the bounding boxes provided by dataset were directly used for training and testing. As mentioned above, we resized each bounding box into 144×56 pixels. In order to reduce the influences of different lengths of videos, each video was re-sampled to 200 frame. In practice, our aggregation network can be applied to images of arbitrary size since it is fully convolutional. In our experiments, we only computed the cosine similarity by using feature vectors of frames or images directly to measure the recognition performance for both the video-based face recognition and person re-identification tasks. Here we did not use the horizontal flip, multi-cropping, PCA and re-ranking tricks for all experiments. The reason is that most state-of-the-art methods also have not utilized these tricks and we expect to compare our method with these methods fairly.

Feature Extraction Networks: For the video-based face recognition task in YTF, PaSC and YTC, we used the still face recognition network trained by the supervision signal of the joint softmax loss and center loss (Wen et al, 2016) provided by authors of (Wen et al, 2016). Architecture of the network is a 29-layer residual network that is not applied bottleneck structure, where the kernel size and stride of all convolutional layers are set to 3×3 and 1, and the kernel size and stride of all max pooling layers are set as 2×2 and 2, respectively. The network used the PReLU function (He et al, 2015) as activations. We present the detailed information of the network in Table 1. For the face verification task, we followed the standard protocol of the LFW (Huang et al, 2007) and YTF (Wolf et al, 2011) datasets. Since the

Table 4 Comparisons of the baseline person re-identification network with state-of-the-art results.

	CUHK03			MARS	
Method	rank-1	rank-5	rank-10	mAP	rank-1
MARS (Zheng et al, 2016)	-	-	-	49.3	68.3
Guided Dropout (Xiao et al, 2016)	75.3	-	-	-	-
Joint Spatial and Temporal (Zhou et al, 2017)	-	-	-	50.7	70.6
Triplet Loss (Hermans et al, 2017)	75.5	95.2	99.2	67.7	79.8
Our	87.87	96.78	98.18	69.58	84.29

IJB-A and IJB-B datasets contain more challenging face images and frames, we employ more powerful face recognition network (Cao et al, 2018) as the feature extraction network, which is a SENet-50 (Hu et al, 2018) model trained on an unconstrained face dataset called VGGFace2 (Cao et al, 2018). For the video-based person re-identification task, we used a feature extractor CNN by following (Xiao et al, 2016). To build a strong baseline network to demonstrate the effectiveness of proposed method, we adopted an ImageNet pre-trained ResNet-50 (He et al, 2016) network as the backbone CNN. Following (Xiao et al, 2016), we only trained a single model by using a combination of serval person re-identification datasets (Xiao et al, 2016): CUHK01 (Li and Wang, 2013), CUHK03 (Li et al, 2014b), PRID (Hirzer et al, 2011), Shinpuhkan (Kawanishi et al, 2014), VIPeR (Gray et al, 2007), 3DPeS (Baltieri et al, 2011), iLIDS (Zheng et al, 2009) and additional Market-1501 (Zheng et al, 2015) dataset. CUHK01 was captured by using two camera views and contains 1552 images totally. CUHK03 consists of five different pairs of camera views, which has more than 14,000 images of 1467 subjects. PRID extracts pedestrian images from recorded trajectory video frames with two camera views, which contains 1134 identities in total. Shinpuhkan is a large-scale dataset with more than 22,000 images of only 24 subjects captured by 16 cameras. The other 3 datasets are relatively small, and we keep them in our training data to maintain the diversity. In our implementations, an extra dataset Market-1501 was used, which is another large-scale dataset with more than 32,000 images of 1501 subjects. We find that adding this dataset can greatly improve the ability of feature extraction network, since the Market-1501 dataset can provide rich information on intra-personal variations. At the training stage, we used both the training and the testing data of all of above-mentioned dataset for training except CUHK03 and Market-1501, and the test set of CUHK03 was used for validation. The detailed information of these datasets and training data of our model are summarized in Table 2. In our implementations, each image was resized into 144×56 pixels as (Xiao et al, 2016), which has been proven to be a good trade-off between computational complexity and accuracy. At the training stage, horizontal flipping was used for data augmentation. We employed a joint softmax loss and triplet loss as supervision signal to train the network, where the margin of triplet loss was set to 0.3 and the size of mini-batch was set to 128. In each mini-batch, we sampled 4 images for each identity and thus each mini-batch contains exactly 32 identities. A standard SGD optimizer with momentum 0.9 and weight decay of 0.0001 was used. The network was trained for 80 epoches with initial learning 0.01, and we decreased the learning rate by 10 at 40 and 60 epoch. We further evaluated our feature extractor on the test sets of CUHK03 and MARS by following the standard protocol, which is presented in Table 3. We compared our baseline model with other works on the test sets of CUHK03 and MARS in Table 4.

Training Details: We set the input of aggregation network as 20 video frames in our implementations, which is a good trade-off of efficiency and complexity. When training the adversarial networks, we followed the standard approach (Goodfellow et al, 2014) and set k as 1. We alternately updated one step for the discriminator network and one step for the aggregation network. To optimize our proposed networks, we employed the mini-batch stochastic gradient descent (SGD) with the batch size of 16 and applied the Adam (Kingma and Ba, 2014) optimizer. We set the learning rate, β_1 and β_2 as 0.0001, 0.5 and 0.999, respectively. We used the aggregation and discriminator networks pre-trained by the supervision signal of the MSE loss as initialization before using the other reconstruction losses and discriminative losses to avoid the local optima. Each network was trained for $10,000$ iterations. We turned off the update of batch normalization parameters during the test time to ensure that the output depends only on the input (Ioffe and Szegedy, 2015). To improve the ability and robustness of DAN model, input video frames are chosen in random order, thus the order of input frames will not affect the quality of synthesized image. The parameter λ and η in equation 4 is set as 1.0, and the weight of GAN loss γ is set as 0.01, which is the same as (Isola et al, 2016). For PaSC and YTC which have relatively small training sets, we fine-tuned the model trained on all videos of YTF to report the recognition result. For IJB-A and IJB-B, we choose to aggregate video frames in template instead of the whole template, because: 1) compared to images, video frames are the most redundant

Table 5 Comparisons of the average verification accuracy (%) of our method with the state-of-the-art face verification results on the YTF dataset.

Method	Accuracy
LM3L (Hu et al, 2014b)	81.3 ± 1.2
DDML (Hu et al, 2014a)	82.3 ± 1.2
EigenPEP (Li et al, 2014a)	84.8 ± 1.4
DeepFace-single (Taigman et al, 2014)	91.4 ± 1.1
DeepID2+ (Sun et al, 2015)	93.2 ± 0.2
FaceNet (Schroff et al, 2015)	95.12 ± 0.39
Deep FR (Parkhi et al, 2015)	97.3
NAN (Yang et al, 2016a)	95.72 ± 0.64
Wen et al. (Wen et al, 2016)	94.9
CNN	93.16 ± 0.97
DAN	94.28 ± 0.69
CNN-finetuned	94.12 ± 0.76
DAN-finetuned	95.01 ± 0.60

Table 6 Comparisons of the verification rate (%) of our method with the other state-of-the-art results on the PaSC dataset at a false accept rate (FAR) of 0.01.

Method	Control	Handheld
PittPatt	48.00	38.00
DeepO2P (Ionescu et al, 2015)	68.76	60.14
VGGFace	78.82	68.24
SPDNet (Huang and Van Gool, 2016)	80.12	72.83
GrNet (Huang et al, 2016)	80.52	72.76
CNN	90.78	78.67
DAN	92.06	80.33
CNN-finetuned	93.76	91.34
DAN-finetuned	94.88	92.12

Table 7 Comparisons of the classification accuracy (%) of our method with the other state-of-the-art results on the YTC dataset.

Method	Accuracy
MDA (Wang and Chen, 2009)	67.2 ± 4.0
LMKML (Lu et al, 2016)	70.31 ± 2.52
MMDML (Lu et al, 2015)	78.5 ± 2.8
GJRNP (Yang et al, 2016b)	81.3 ± 2.0
DRM-WV (Hayat et al, 2015)	88.32 ± 2.14
CNN	96.79 ± 1.27
DAN	97.32 ± 0.71
CNN-finetuned	96.88 ± 1.12
DAN-finetuned	97.70 ± 0.72

and computationally dense part in face recognition system; 2) our method focuses on reduce redundant computational cost on video frames that have similar face appearance but vary in pose, image quality, occlusion and etc. Moreover, aggregating the whole template that include faces from different ages and backgrounds is difficult to learn and will significantly harm recognition perform while the diversity of faces in input template is reduced. For mixed media tasks in IJB-A and IJB-B, each video was re-sampled to 20 frames, then a DAN model was trained to aggregate these frames to a single image. Following the practice in (Cao et al, 2018), we generate descriptor of each template by averaging descriptor of each media, where the descriptor of each media is compute by averaging the CNN feature vectors of images in that media. For video identification task in IJB-B, we did not re-sample input video and aggregated every 20 frames to a image, thus feature vector of a input video can be obtained by averaging the CNN features of aggregated images. A DAN model with V-STN was trained for the MARS dataset, and we report the results on the test set of MARS by following the standard protocol.

Testing Details: For all of these datasets, we firstly used our proposed method to aggregate the whole video into 10 images by aggregating every 20 frames to a single image following their order. Then we use the feature extraction network and the mean-pooling operation to represent each video as a single feature vector. For the face verification task, we used the cosine similarity and threshold comparison, where thresholds were computed from the training set. For the classification task, we computed the cosine similarity between examples in the training set and the testing set and decided the categories according to the nearest neighbor rule.

Experimental Environments: For the video-based face recognition task, our models were trained and tested by using the Python interface of Caffe (Jia et al, 2014) on a Tesla K80 GPU. For video-based person re-identification, our method was implemented using the PyTorch interface (Paszke et al, 2017) on a GTX 1080Ti GPU.

4.3 Results and Analysis

4.3.1 Comparisons with State-of-the-Arts

Tables 5-10 show the recognition results of different methods on the YTF, PaSC, YTC, IJB-A, IJB-B and MARS datasets, respectively. For all these six datasets, we report the average accuracy obtained by our framework, denoted as **DAN** in these tables. We also report the results by directly passing all the video frames through feature extraction CNN F with mean pooling for comparison, which is denoted as **CNN** in these tables.

The results show that DAN outperforms the original CNN for dense feature extraction on all datasets. This is a strong baseline with high computation complexity, showing that the aggregated images produced by DAN are more discriminative than original video frames. These results also show

Table 8 Comparisons of mixed media verification and identification performance of our method with recent state-of-the-art results on IJB-A dataset.

	1:1 verification TAR (%)			1:N identification accuracy (%)		
Method	FAR=0.1	FAR=0.01	FAR=0.001	rank-1	rank-5	rank-10
Pooling Faces (Hassner et al, 2016)	-	81.9	63.1	82.8	92.1	94.3
Disentangled (Tran et al, 2017)	-	77.4 ± 2.7	53.9 ± 4.3	85.5 ± 1.5	94.7 ± 1.1	-
Domain Adaptation (Sohn et al, 2017)	97.0 ± 0.1	86.4 ± 0.7	64.9 ± 2.2	89.5 ± 0.3	95.7 ± 0.2	96.8 ± 0.2
Template Adaptation (Hermans et al, 2017)	97.9 ± 0.4	93.9 ± 1.3	83.6 ± 2.7	92.8 ± 1.0	97.7 ± 0.4	98.6 ± 0.3
NAN (Yang et al, 2016a)	97.8 ± 0.3	94.1 ± 0.8	88.1 ± 1.1	95.8 ± 0.5	98.0 ± 0.5	98.6 ± 0.3
VGGFace2 (Cao et al, 2018)	99.0 ± 0.2	96.8 ± 0.6	92.1 ± 1.4	98.2 ± 0.4	99.3 ± 0.2	99.4 ± 0.1
CNN	98.1 ± 0.4	94.0 ± 1.2	90.3 ± 2.1	97.7 ± 0.5	98.9 ± 0.2	99.3 ± 0.1
DAN	98.3 ± 0.4	94.1 ± 0.9	91.0 ± 1.4	98.0 ± 0.4	99.0 ± 0.2	99.3 ± 0.1

Table 9 Comparisons of mixed media and video identification performance of our method with recent state-of-the-art results on IJB-B dataset.

	1:N mixed media accuracy (%)			1:N video accuracy (%)		
Method	rank-1	rank-5	rank-10	rank-1	rank-5	rank-10
IJB-B (Whitelam et al, 2017)	79.0	85.0	90.0	-	-	-
VGGFace2 (Cao et al, 2018)	90.2 ± 3.6	94.6 ± 2.2	95.9 ± 1.5	-	-	-
CNN	89.1 ± 3.5	93.5 ± 2.0	95.0 ± 1.3	70.3 ± 3.8	78.6 ± 3.2	80.2 ± 2.5
DAN	89.9 ± 3.0	93.7 ± 1.8	95.2 ± 1.2	73.2 ± 3.0	80.4 ± 2.7	82.2 ± 2.0

Table 10 Comparisons of the mAP and rank-1 accuracy (%) of our method with the other state-of-the-art results on the MARS dataset.

Method	mAP	rank-1
Compact Appearance (Zhang et al, 2017)	-	55.5
Multi-target (Tesfaye et al, 2017)	-	68.2
MARS (Zheng et al, 2016)	49.3	68.3
Joint Spatial and Temporal (Zhou et al, 2017)	50.7	70.6
Quality Aware (Liu et al, 2017)	51.7	73.7
Re-ranking (Zhong et al, 2017)	68.5	73.9
Triplet Loss (Hermans et al, 2017)	67.7	79.8
CNN	69.58	84.29
DAN with V-STN	70.23	84.65

the robustness and denoising ability of the proposed DAN method.

Compared to previous state-of-the-art methods, DAN outperforms all of them on PaSC, YTC and MARS. On YTF, DAN achieves competitive but not the best result. This is largely due to the baseline CNN which is comparatively weaker than those of (Schroff et al, 2015) and (Yang et al, 2016a). But the gained improvement over baseline CNN result has already proven the effectiveness. To further show the effectiveness of the proposed method, we fine-tuned the baseline CNN model on the training set of the corresponding video face datasets following the practice in (Ding and Tao, 2017) and supervised by the triplet loss with the learning rate of 0.001 and keep all other experiment settings unchanged. The results are shown in Table 5-7, where the results measured on fine-tuned models are referred as *CNN-finetuned* and *DAN-finetuned*. We can see that our method can still improve final results based on the fine-tuned CNN model. These results prove that our method generalizes well with different feature embedding models.

Experimental results IJB-A and IJB-B are shown in Table 8 and Table 9 respectively, where performance is reported using the true accept rates (TAR) at different false positive rates (FAR=0.1, 0.01, 0.001) for verification task and rank-n accuracy (n=1, 5, 10) for identification task. It can be observed that our method can consistently outperform the baseline method and achieve competitive results compared to recently proposed state-of-the-art methods. These results show that our method generalizes well on different datasets and can handle unconstrained faces. For the mixed media recognition task, since the final descriptor of template is a combination of image and video features, our method can only slightly improve the baseline method. However, DAN can reduce around 70% redundant computation cost on feature extraction (67.7% on IJB-A and 71.2% on IJB-B). It can be observed that CNN baseline of video-based identification is much lower than other tasks. We believe it is because that the gallery sets of IJB-B only consisted of images, and the domain gap between image and video can severely harm recognition performance. Since our method can help video frames to have better embedding in feature space and close the gap between frames and images, our method can significantly improve recognition performance in this task.

Table 11 Runtime analysis on YTF with the 29-layers Residual network.

Method	Runtime(ms)	Processed frames
CNN	819.7	181
Random + CNN	42.0	10
DAN	126.1	200

Table 12 Runtime analysis on MARS with ResNet-50. Note that we report the average inference time on all videos on the test set of MARS.

Method	Runtime(ms)
CNN	124.32
Random + CNN	7.77
DAN with V-STN	10.53

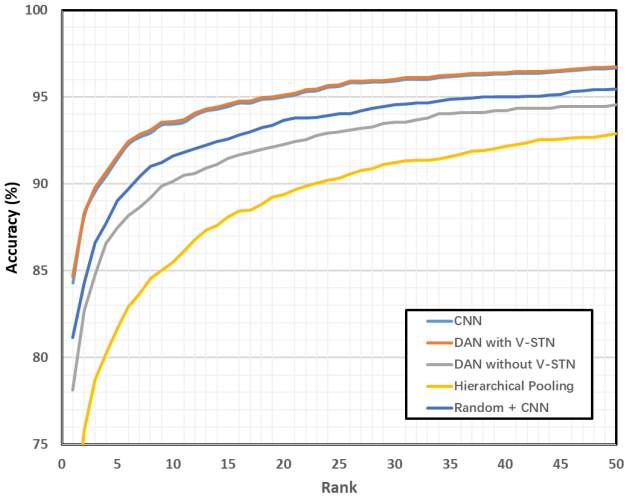

Fig. 4 CMC curve on MARS. DAN with V-STN surpasses all other baseline models and V-STN can greatly improve the performance of our previous DAN model without V-STN.

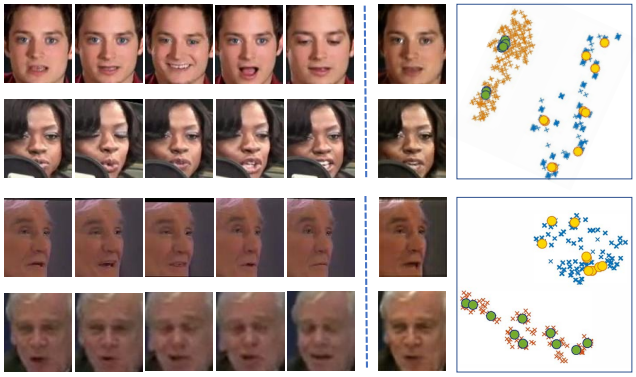

Fig. 5 The examples of original video frames and the aggregated images (on the left), and the distribution of their features after t-sne (Maaten and Hinton, 2008) (on the right). The crossings refer to original video frames and the dots refer to synthesized images. From the distribution we can see that DAN can decrease the intra-class distance while increase the inter-class distance.

4.3.2 Runtime Analysis

Efficiency is one advantage of our framework. To show the efficiency of our method, we give a short analysis of the runtime of our method. For dense feature extraction baselines, we calculated that the average frame number of the YTF dataset is 181.3, which is used to measure the runtime. For the random CNN protocol, we randomly selected 10 frames from the video and measure the forward time. For our DAN, we measured the overall time including aggregating 200 frames into 10 images with DAN and the forward time of CNN. Similar results can be also observed on the MARS dataset with the V-STN model. The results are shown in Table 11 and Table 12, respectively.

From these tables, we see that our DAN is much faster than the baseline method with the dense feature extraction, and has only little overhead compared to the random sampling baseline. Moreover, DAN achieves the best recognition performance among all these compared methods, which further shows the effectiveness and efficiency of our proposed DAN framework.

4.3.3 Ablation Study

To further understand and analyze the performance of our proposed method, we conducted several ablation experiments as follows.

Investigation of the influence of frame numbers: To investigate into the influence of frame numbers, we formed two subsets of original video frames besides dense CNN feature pooling: (1) randomly sampling the same number of frames as generated by DAN from the original video, which is 10 in our case; (2) performing mean pooling on similar faces and summarizing the whole video as 10 frames. We measured the performance by the mean pooling on corresponding CNN features and denoted results as **Random + CNN** and **Hierarchical Pooling** respectively. Results with different inference methods are presented in Table 13. We see from this table that our proposed DAN method consistently outperforms all other inference methods on all four datasets, including dense feature extraction, random sampling and hierarchical pooling methods. Note that for video clips with large pose variances such as videos from the MARS dataset, directly performing the mean pooling operator on original frames will significantly harm the recognition performance, which shows that our method is more robust and effective than other methods.

Investigation of Loss Functions: To investigate the effectiveness of different terms of the loss function of our DAN, we conducted experiments with different variations of DAN with different combinations of these loss terms. Here

Table 13 Investigation of the influence of frame numbers used in our method and the corresponding accuracy (%).

	YTF	PaSC		YTC	MARS	
Method	Accuracy	Control	Handheld	Accuracy	mAP	rank-1
CNN	93.16 ± 0.97	90.78	78.67	96.79 ± 1.27	69.58	84.29
Random + CNN	92.80 ± 1.17	89.12	78.03	96.63 ± 1.31	67.23	81.22
Hierarchical Pooling	93.15 ± 1.12	89.83	78.23	96.78 ± 1.25	52.12	69.22
DAN	$\mathbf{94.28 \pm 0.69}$	**92.06**	**80.33**	$\mathbf{97.32 \pm 0.71}$	**70.23**	**84.65**

Table 14 Investigation of different loss functions and the corresponding accuracy (%).

Adversarial loss $\mathscr{L}^{GAN}$	Discriminative loss $\mathscr{L}^{Dis}$	Reconstruction loss $\mathscr{L}^{Rec}_{MSE}$	Reconstruction loss $\mathscr{L}^{Rec}_{FM}$	Reconstruction loss $\mathscr{L}^{Rec}_{FC}$	Accuracy
		✓			91.38 ± 0.74
✓					92.50 ± 0.96
✓		✓			92.36 ± 0.90
✓			✓		92.46 ± 0.97
✓				✓	92.92 ± 0.81
✓	✓	✓			93.02 ± 0.88
✓	✓		✓		93.16 ± 0.93
✓	✓			✓	$\mathbf{94.28 \pm 0.69}$

we analyze the effects of each loss function with detailed experiments on YTF. The results are shown in Table 14. As shown in this table, we see that training with the MSE reconstruction loss provides the basic baseline, where the discriminative ability is reduced and significant performance drop is obtained. Moreover, introducing the adversarial loss contributes to more realistic and therefore more discriminative images can be obtained than the MSE loss for recognition. However, it is worse that the alternative which used the dense CNN feature extraction baseline. Therefore, combining the adversarial loss and the reconstruction loss can further improve the performance slightly.

For the reconstruction loss, we provided comparisons between three forms of loss $\mathscr{L}^{Rec}$: the MSE loss, the feature map loss and the feature embedding loss. The MSE loss focuses on low level visual characteristics, and thus can make little contribution to the discriminative power of the extracted feature. The feature map loss exerts supervision on high level activation map and is closer to perceptual similarity. Such characteristics can help to distinguish person in some degree, but still cannot guarantee the distribution in the final feature embedding. On the contrary, our proposed $\mathscr{L}^{Rec}_{FC}$ directly supervises the feature embedding itself, and introduces metric learning into the training, thus can make the aggregated images even more dividable in the feature space.

The most important observation is that bringing the discriminative loss $\mathscr{L}^{Dis}$ to the system can greatly boost the recognition performance, which is also the main contribution of our work. By combining the discriminative loss $\mathscr{L}^{Dis}$ and the feature embedding based reconstruction loss $\mathscr{L}^{Rec}_{FC}$, we can obtain the best result beyond the CNN baseline.

Table 15 Ablation experimental results of our methods with and without the V-STN model, where the mAP and rank-1 accuracy (%) on MARS are reported.

Method	mAP	rank-1
DAN without V-STN	61.27	78.12
DAN with V-STN	70.23	84.65

Effectiveness of V-STN: To demonstrate the effectiveness of our proposed V-STN method for video-based person re-identification, we trained a DAN model without V-STN. We present the comparison of these two models in Table 15. It can be observed that results of DAN significantly decrease by 9% and 6% of mAP and accuracy respectively, when we removed V-STN from the DAN model. We also report the CMC curve on MARS by following the standard protocol of (Zheng et al, 2016) and Fig. 4 shows the results of our method with and without the V-STN. We see that our proposed DAN model with V-STN surpasses all other baseline models and V-STN greatly improves the recognition performance of our previous DAN without V-STN model.

4.3.4 Visualization

To investigate the effectiveness of the proposed DAN, we visualize some results in Fig. 5. The visualization results consist of 2 parts: raw video frames and aggregated images, as well as their distributions in the feature space after reduc-

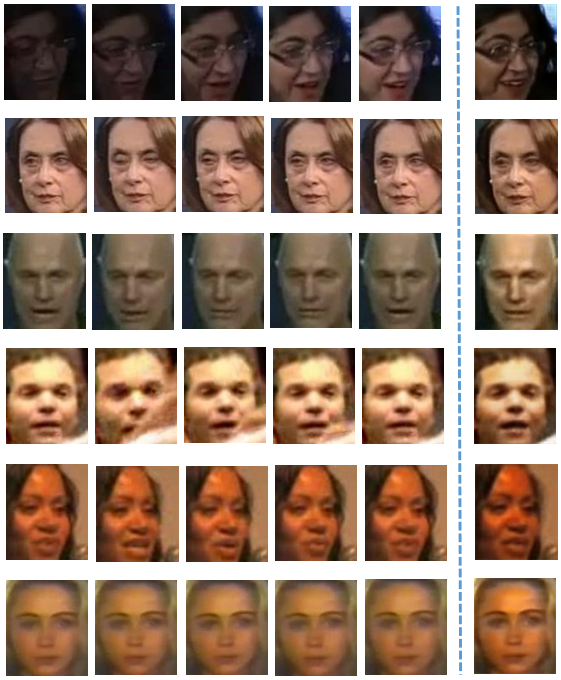

Fig. 6 Visual results on YTF. We presented the original video frames (on the left) and the aggregated images (on the right). Input 20 video frames are sampled every 4 frames.

ing the feature dimension into 2 with t-SNE (Maaten and Hinton, 2008). We see that the aggregated images are visually similar to the original images, with very good quality including good positions, viewing angles, illuminations, etc., which are very important to recognition. Bad quality frames with blurring or unfavorable viewing angles are denoised during the aggregation procedure.

We plot the distributions of original videos with 200 frames and the aggregated 10 images. We see that DAN enlarges the margin between negative video pairs, especially the first example, and reduce the intra-class distance. This clearly demonstrates that aggregated images by DAN have better discriminative power and robustness than the original videos. Fig. 6 and Fig. 7 show more visual results of our generative model on YTF and MARS, respectively. It can be observed that the synthesized images are visually better than input frames and our proposed DAN can denoise the low-quality frames.

4.4 Limitations

Some failure cases on YTF and MARS are presented in Fig. 8. In the first case, the DAN model fails to remove unseen occlusion such as subtitle in video. In the second case, DAN model fails to capture useful information from very noisy input video. Since the aggregation model is trained without additional labels and "good" examples are not ex-

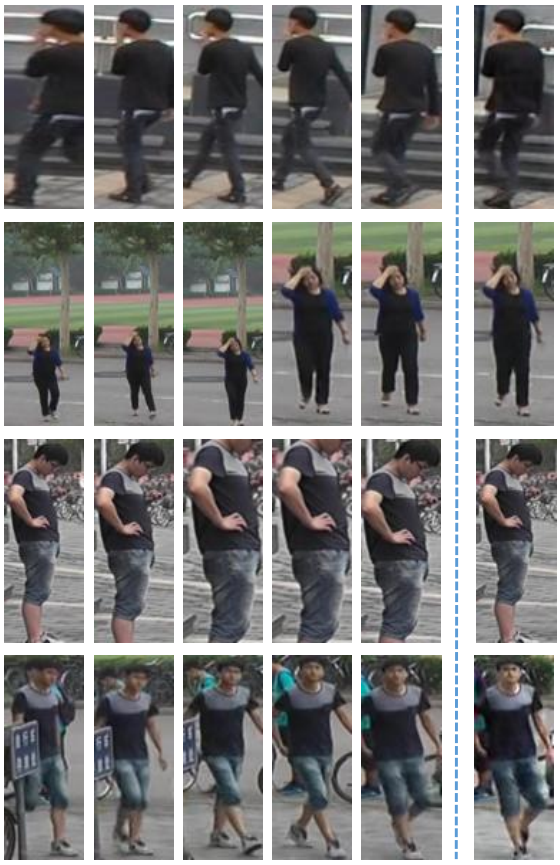

Fig. 7 Visual results on MARS. We presented the original video frames (on the left) and the aggregated images (on the right). Input 20 video frames are sampled every 4 frames.

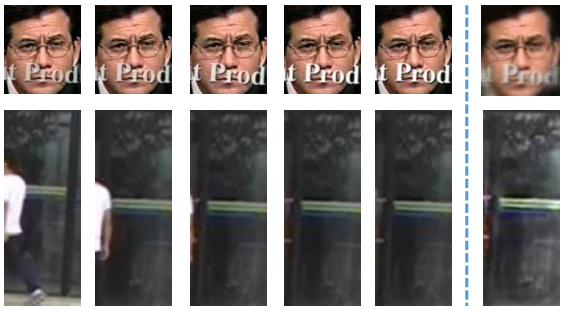

Fig. 8 Example Failure cases on the YTF and MARS datasets. We presented the original video frames (on the left) and the aggregated images (on the right). Input 20 video frames are sampled every 4 frames. These examples are selected as some of the worst generated images on these datasets. Failure cases are usually caused by unusual or very noisy inputs.

plicitly presented to model, DAN model cannot remove unusual noise in videos.

Since our method focuses on reducing redundant computational cost of video-based face recognition system and is designed for integrate information across frames in the same video, where faces have similar appearance and background, our method can hardly aggregate generic image sets such as "templates" in the IJB-A and IJB-B datasets, which may have large variations in ages and environments and contain few redundant images.

5 Conclusion

In this paper, we have proposed a discriminative aggregation network (DAN) method for effective and efficient video-based face recognition and person re-identification. By combining metric learning and adversarial learning, our DAN can aggregate useful information of an input video into one or few more discriminative images in the feature space, which can be used for both face recognition and person re-identification. To our best knowledge, DAN is one of the first aggregation frameworks that takes raw video frames as input instead of feature embedding. With our aggregation framework, the generated images have smaller intra-class distances and greater inter-class distances in the feature space, contributing to the discriminative power and robustness of the recognition system. Furthermore, runtime is greatly reduced as we only need to pass few output images through the feature extraction network for face recognition and person re-identification. Experimental results on four widely used datasets have been proposed to demonstrate the effectiveness of our framework.

Acknowledgements This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, in part by the National Natural Science Foundation of China under Grants 61672306, U1713214, 61527808, in part by the National 1000 Young Talents Plan Program, in part by the National Basic Research Program of China under Grant 2014CB349304, and in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564.

Compliance with Ethical Standards

Conflict of interest: The author confirms that this article content has no conflicts of interest.
Ethical approval: This article does not contain any studies with human participants or animals.

References

Baltieri D, Vezzani R, Cucchiara R (2011) 3dpes: 3d people dataset for surveillance and forensics. In: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding, ACM, pp 59–64

Beveridge JR, Phillips PJ, Bolme DS, Draper BA, Givens GH, Lui YM, Teli MN, Zhang H, Scruggs WT, Bowyer KW, et al (2013) The challenge of face recognition from digital point-and-shoot cameras. In: BTAS, pp 1–8

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

Cevikalp H, Triggs B (2010) Face recognition based on image sets. In: CVPR, pp 2567–2573

Chen JC, Ranjan R, Kumar A, Chen CH, Patel VM, Chellappa R (2015) An end-to-end system for unconstrained face verification with deep convolutional neural networks. In: ICCVW, pp 118–126

Chen JC, Patel VM, Chellappa R (2016a) Unconstrained face verification using deep cnn features. In: WACV, pp 1–9

Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P (2016b) Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: NIPS, pp 2172–2180

Chen YC, Patel VM, Phillips PJ, Chellappa R (2012) Dictionary-based face recognition from video. In: European Conference on Computer Vision, Springer, pp 766–779

Ding C, Tao D (2017) Trunk-branch ensemble convolutional neural networks for video-based face recognition. PAMI

Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: ECCV, Springer, pp 184–199

Dong C, Loy CC, He K, Tang X (2016) Image super-resolution using deep convolutional networks. T-PAMI 38(2):295–307

Felzenszwalb PF, Girshick RB, McAllester D (2010) Cascade object detection with deformable part models. In: CVPR, IEEE, pp 2241–2248

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: NIPS, pp 2672–2680

Gray D, Brennan S, Tao H (2007) Evaluating appearance models for recognition, reacquisition, and tracking. In: PETS, Citeseer, vol 3, pp 1–7

Guillaumin M, Verbeek J, Schmid C (2009) Is that you? metric learning approaches for face identification. In: ICCV, pp 498–505

Hassner T, Masi I, Kim J, Choi J, Harel S, Natarajan P, Medioni G (2016) Pooling faces: template based face recognition with pooled face images. In: CVPRW, pp 59–67

Hayat M, Bennamoun M, An S (2015) Deep reconstruction models for image set classification. PAMI 37(4):713–727

He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV, pp 1026–1034

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR, pp 770–778

Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv preprint arXiv:170307737

Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on Image analysis, Springer, pp 91–102

Hu J, Lu J, Tan YP (2014a) Discriminative deep metric learning for face verification in the wild. In: CVPR, pp 1875–1882

Hu J, Lu J, Yuan J, Tan YP (2014b) Large margin multi-metric learning for face and kinship verification in the wild. In: ACCV, pp 252–267

Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks

Hu Y, Mian AS, Owens R (2011) Sparse approximated nearest points for image set classification. In: Computer vision and pattern recognition (CVPR), 2011 IEEE conference on, IEEE, pp 121–128

Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, Univer-

sity of Massachusetts, Amherst
Huang Z, Van Gool L (2016) A riemannian network for spd matrix learning. arXiv preprint arXiv:160804233
Huang Z, Wang R, Shan S, Chen X (2014) Learning euclidean-to-riemannian metric for point-to-set classification. In: CVPR, pp 1677–1684
Huang Z, Wang R, Shan S, Li X, Chen X (2015) Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In: ICML, pp 720–729
Huang Z, Wu J, Van Gool L (2016) Building deep networks on grassmann manifolds. arXiv preprint arXiv:161105742
Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167
Ionescu C, Vantzos O, Sminchisescu C (2015) Matrix backpropagation for deep networks with structured layers. In: ICCV, pp 2965–2973
Isola P, Zhu JY, Zhou T, Efros AA (2016) Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:161107004
Jaderberg M, Simonyan K, Zisserman A, et al (2015) Spatial transformer networks. In: NIPS, pp 2017–2025
Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. In: ACM-MM, pp 675–678
Kawanishi Y, Wu Y, Mukunoki M, Minoh M (2014) Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In: 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision, vol 5, p 6
Kim M, Kumar S, Pavlovic V, Rowley H (2008) Face tracking and recognition with visual constraints in real-world videos. In: CVPR, pp 1–8
Kingma D, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:13126114
Klare BF, Klein B, Taborsky E, Blanton A, Cheney J, Allen K, Grother P, Mah A, Jain AK (2015) Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: CVPR, pp 1931–1939
Larsen ABL, Sønderby SK, Larochelle H, Winther O (2015) Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:151209300
Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al (2016) Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:160904802
Li H, Hua G, Shen X, Lin Z, Brandt J (2014a) Eigen-pep for video face recognition. In: ACCV, pp 17–33
Li W, Wang X (2013) Locally aligned feature transforms across views. In: CVPR, pp 3594–3601
Li W, Zhao R, Xiao T, Wang X (2014b) Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR, pp 152–159
Lin J, Ren L, Lu J, Feng J, Zhou J (2017) Consistent-aware deep learning for person re-identification in a camera network. In: CVPR, pp 5771–5780
Liu Y, Yan J, Ouyang W (2017) Quality aware network for set to set recognition. In: CVPR, vol 2, p 8
Lu J, Wang G, Moulin P (2013) Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In: ICCV, pp 329–336
Lu J, Wang G, Deng W, Moulin P, Zhou J (2015) Multi-manifold deep metric learning for image set classification. In: CVPR, pp 1137–1145
Lu J, Wang G, Moulin P (2016) Localized multifeature metric learning for image-set-based face recognition. TCSVT 26(3):529–540
Maaten Lvd, Hinton G (2008) Visualizing data using t-sne. JMLR 9(Nov):2579–2605
Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: BMVC, vol 1, p 6
Paszke A, Gross S, Chintala S, Chanan G (2017) Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration
Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434
Rao Y, Lin J, Lu J, Zhou J (2017) Learning discriminative aggregation network for video-based face recognition. In: ICCV, pp 3781–3790
Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H (2016) Generative adversarial text to image synthesis. In: ICML, vol 3
Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: CVPR, pp 815–823
Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR, pp 1874–1883
Sohn K, Liu S, Zhong G, Yu X, Yang MH, Chandraker M (2017) Unsupervised domain adaptation for face recognition in unlabeled videos. In: CVPR, pp 3210–3218
Sun Y, Wang X, Tang X (2015) Deeply learned face representations are sparse, selective, and robust. In: CVPR, pp 2892–2900
Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: Closing the gap to human-level performance in face verification. In: CVPR, pp 1701–1708
Tesfaye YT, Zemene E, Prati A, Pelillo M, Shah M (2017) Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. arXiv preprint arXiv:170606196
Tran L, Yin X, Liu X (2017) Disentangled representation learning gan for pose-invariant face recognition. In: CVPR, vol 3, p 7
Wang J, Lu C, Wang M, Li P, Yan S, Hu X (2014) Robust face recognition via adaptive sparse representation. IEEE transactions on cybernetics 44(12):2368–2378
Wang R, Chen X (2009) Manifold discriminant analysis. In: CVPR, pp 429–436
Wang R, Guo H, Davis LS, Dai Q (2012) Covariance discriminative learning: A natural and efficient approach to image set classification. In: CVPR, pp 2496–2503
Wang T, Gong S, Zhu X, Wang S (2016) Person re-identification by discriminative selection in video ranking. IEEE transactions on pattern analysis and machine intelligence 38(12):2501–2514
Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: ECCV, pp 499–515
Whitelam C, Taborsky E, Blanton A, Maze B, Adams JC, Miller T, Kalka ND, Jain AK, Duncan JA, Allen K, et al (2017) Iarpa janus benchmark-b face dataset. In: CVPR Workshops, pp 592–600
Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity. In: CVPR, pp 529–534
Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. IEEE transactions on pattern analysis and machine intelligence 31(2):210–227
Xiao T, Li H, Ouyang W, Wang X (2016) Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR, pp 1249–1258
Yang J, Ren P, Chen D, Wen F, Li H, Hua G (2016a) Neural aggregation network for video face recognition. arXiv preprint arXiv:160305474
Yang M, Wang X, Liu W, Shen L (2016b) Joint regularized nearest points for image set based face recognition. IVC
Zhang H, Xu T, Li H, Zhang S, Huang X, Wang X, Metaxas D (2016a) Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. arXiv preprint arXiv:161203242
Zhang K, Zhang Z, Li Z, Qiao Y (2016b) Joint face detection and alignment using multitask cascaded convolutional networks. SPL 23(10):1499–1503

Zhang W, Hu S, Liu K (2017) Learning compact appearance representation for video-based person re-identification. arXiv preprint arXiv:170206294

Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. In: ICCV, pp 1116–1124

Zheng L, Bie Z, Sun Y, Wang J, Su C, Wang S, Tian Q (2016) Mars: A video benchmark for large-scale person re-identification. In: ECCV, Springer, pp 868–884

Zheng WS, Gong S, Xiang T (2009) Associating groups of people. In: BMVC, vol 2

Zhong Z, Zheng L, Cao D, Li S (2017) Re-ranking person re-identification with k-reciprocal encoding. arXiv preprint arXiv:170108398

Zhou Z, Huang Y, Wang W, Wang L, Tan T (2017) See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In: CVPR, IEEE, pp 6776–6785