Unsupervised Domain Adaptation for Face Recognition in Unlabeled Videos

Kihyuk Sohn[1] Sifei Liu[2] Guangyu Zhong[3] Xiang Yu[1] Ming-Hsuan Yang[2] Manmohan Chandraker[1,4]
[1]NEC Labs America [2]UC Merced [3]Dalian University of Technology [4]UC San Diego

Abstract

Despite rapid advances in face recognition, there remains a clear gap between the performance of still image-based face recognition and video-based face recognition, due to the vast difference in visual quality between the domains and the difficulty of curating diverse large-scale video datasets. This paper addresses both of those challenges, through an image to video feature-level domain adaptation approach, to learn discriminative video frame representations. The framework utilizes large-scale ***unlabeled*** *video data to reduce the gap between different domains while transferring discriminative knowledge from large-scale labeled still images. Given a face recognition network that is pretrained in the image domain, the adaptation is achieved by (i) distilling knowledge from the network to a video adaptation network through feature matching, (ii) performing feature restoration through synthetic data augmentation and (iii) learning a domain-invariant feature through a domain adversarial discriminator. We further improve performance through a discriminator-guided feature fusion that boosts high-quality frames while eliminating those degraded by video domain-specific factors. Experiments on the YouTube Faces and IJB-A datasets demonstrate that each module contributes to our feature-level domain adaptation framework and substantially improves video face recognition performance to achieve state-of-the-art accuracy. We demonstrate qualitatively that the network learns to suppress diverse artifacts in videos such as pose, illumination or occlusion without being explicitly trained for them.*

1. Introduction

Motion of objects or the observer in a video sequence is a powerful cue for perceptual tasks such as shape determination or identity recognition [5, 10]. For face recognition in computer vision, recent years have seen the success of emerging approaches in video face analysis [37, 19, 4, 38, 20] and several image-based face recognition engines [32, 24] on video face recognition benchmarks [37, 17]. But it is arguably true that these efforts have been outpaced by image-based face recognition engines that perform comparably or even better than human perception in certain settings [32, 29, 26, 15]. For example, a verification accuracy of

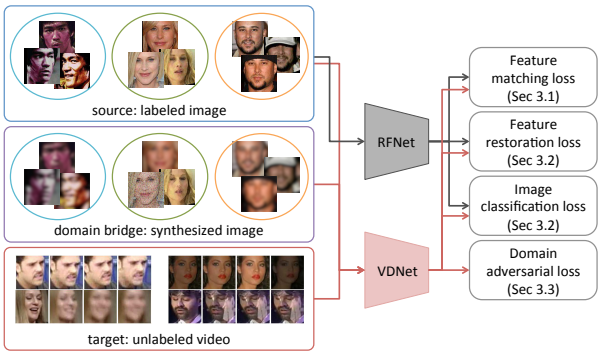

Figure 1. We propose an unsupervised domain adaptation method for video face recognition using large-scale unlabeled videos and labeled still images. To help bridge the gap between two domains, we introduce a new domain of synthesized images by applying a set of image transformations specific to videos such as motion blur to labeled images that simulates a video frame from still image. We utilize images, synthesized images, and unlabeled videos for domain adversarial training. Finally, we train a video domain-adapted network (VDNet) with domain adversarial loss (Section 3.3) as well as by distilling knowledge from pretrained reference network (RFNet) through feature matching (Section 3.1), feature restoration and image classification (Section 3.2) losses.

95.12% is reported by [26] on the YouTube Faces dataset, much lower than 99.63% on the LFW dataset.

Besides better understanding of training convolutional neural networks (CNNs), a key ingredient to the success of image-based face recognition is the availability of large-scale datasets of labeled face images collected from the web [40]. Thus, one source of difficulty for video face recognition is attributable to the lack of similar large-scale labeled datasets. The number of images used to train the state-of-the-art face recognition engines varies from 200K [29] to 200M [26] collected with at least 10K different identities or as many as 8M. In contrast, large-scale labeled video database is publicly available to date, such as YouTube face dataset (YTF) [37], only contains 3.4K videos in total from 1.5K different subjects. Although more frames may be labeled, it is difficult to collect a dataset with as many variations without a surge in dataset size and labeling effort.

An avenue for overcoming the lack of labeled training data in the video domain is to transform labeled still face images so that they look like images captured from videos.

Video frames are likely to be degraded for multiple reasons such as motion or out-of-focus blur, compression noise or scale variations. Approaches such as [6] augment image-based training data with synthetic blur kernels and noise, to demonstrate moderate improvement in video face recognition. However, attempting to bridge the domain gap between images and videos with such an approach faces fundamental challenges – first, it is non-trivial to sufficiently enumerate all types of blur kernels that degrade visual quality in videos, and second, it is not possible to model the transformation from images to videos with sufficient accuracy.

In this work, we propose a data-driven method for image to video domain adaptation for video face recognition. Instead of collecting a labeled video face dataset, we utilize large-scale unlabeled video data to reduce the gap between video and image domains, while retaining the discriminative power of large-scale labeled still images. To take advantage of labeled image data, Section 3.1 proposes to transfer discriminative knowledge by distilling the distance metric through feature matching, from a reference network (RFNet) trained on a web-face dataset [40] to our video face network (VDNet). A further avenue to leverage image domain labels is through the domain-specific data augmentation of Section 3.2, whereby we degrade still images using synthetic motion blur, resolution variation, or video compression noise. Then, we train VDNet to be able to *restore* the original representation of an image extracted from RFNet.

While the above augmentation is useful, its effectiveness is limited by the fact that types of artifacts in videos are too diverse to be enumerated. In Section 3.3, we further regularize VDNet to reduce the domain gap by introducing a discriminator that learns to distinguish different domains, without any supervision such as identity labels or instance-level correspondence. Once trained, the score output by this discriminator is a measure of the confidence in the similarity of the feature representation of a video frame to that of a still image. This is a useful ability, since poor performance in video face recognition can often be attributed to some frames in a sequence that are of substantially poor quality. Consequently, Section 3.4 proposes a discriminator-guided weighted feature fusion to aggregate frames in each video, by assigning higher weights to "image-like" frames, that potentially have better quality among the others. Figure 1 illustrates our proposed framework.

In Section 5, we extensively evaluate the proposed framework on the YouTube Faces (YTF) dataset to demonstrate performance that surpasses prior state-of-the-art. We present ablation studies that demonstrate the importance of each of the above components. Interestingly, degradation factors such as blur, illumination or occlusions, automatically emerge in qualitative visualizations of frames within a sequence ranked by domain discriminator scores.

The main contributions of this work are:

- We present a novel unsupervised domain adaptation algorithm from images to videos for face recognition in unlabeled videos.
- We develop a feature-level domain adaptation to learn VDNet by distilling discriminative knowledge from pre-trained RFNet through feature matching.
- We propose a domain adversarial learning method that modulates the VDNet to learn a domain-invariant feature without needing to enumerate all causes of domain gap.
- We design a method to train with synthetic data augmentation for feature-level restoration and to help the discriminator to discover domain differences.
- We use the confidence score of the discriminator to develop an unsupervised feature fusion method that suppresses low quality frames.
- We demonstrate the superiority of VDNet over existing methods with extensive experiments on YTF dataset, achieving state-of-the-art verification accuracy. We also demonstrate performance gains over baseline methods on the IJB-A dataset without supervised fine-tuning.

2. Related Work

Our work falls into the class of problems on unsupervised domain adaptation [22, 8, 33, 9] that concerns adapting a classifier trained on a source domain (e.g., web images) to a target domain (e.g., video) where there is no labeled training data for target domain to fine-tune the classifier. Among those, feature space alignment and domain adversarial learning methods are closely related to our approach.

The basic idea of feature space alignment is to minimize the distance between domains in the feature space through learning a transformation of source to target features [8, 25, 7, 35, 43], or a joint adaptation layer that embeds features into a new domain-invariant space [22, 33]. Specifically, Tzeng et al. [33] use two CNNs for source and target domain with shared weights and the network is optimized for classification loss in the source domain as well as domain difference measured by the maximum mean discrepancy (MMD) metric. Gupta et al. [13] consider a similar network architecture for cross-modality supervision transfer.

There also exists a body of work on unsupervised domain adaptation and transfer with adversarial learning [11], where the domain difference is measured by a discriminator network $\mathcal{D}$ [41, 21, 30]. For example, [41, 30] consider cross-domain transfer of images from one style to another without instance-level correspondence between domains using adversarial loss. Coupled GAN [21] constructs individual networks for each domain with partially shared higher-layer parameters for generator and discriminator to generate coherent images of two domains. Unlike the above works that generate images in the target domain, we consider feature-level domain adaptation.

For feature-level domain adaptation using adversarial

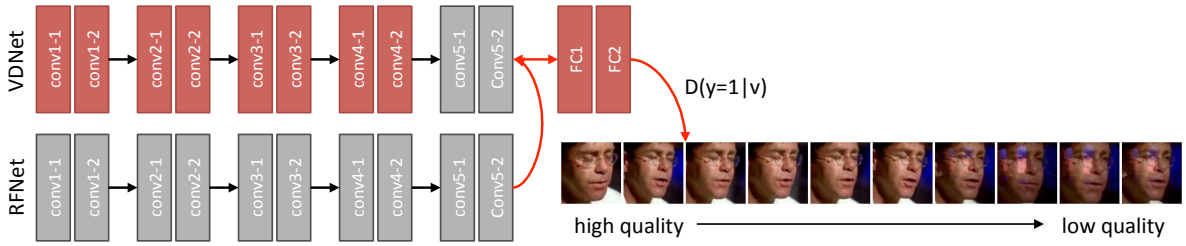

Figure 2. An illustration of network architecture for RFNet, VDNet and discriminator ($\mathcal{D}$). The red and gray blocks denote the trainable and fixed modules, respectively. VDNet not only shares the network architecture with RFNet, but also is initialized with the same network parameters. Once trained, $\mathcal{D}$ can sort the frames in a video sequence by indicating whether a frame is similar to images compatible to a face recognition engine and rejects those frames that are extremely ill-suited for face recognition.

learning, domain adversarial neural network (DANN) [9] appends domain classifier to high-level features and introduces a gradient reversal layer for end-to-end learning via backpropagation while avoiding cumbersome minimax optimization of adversarial training. The goal of DANN is to transfer discriminative classifier from source to target domain, which implicitly assumes the label spaces of two domains are equivalent (or at least the label space of target domain is the subset of that of source domain). Our work is to transfer discriminative *distance metric* and hence there is no such restriction in label space definition. In addition, we propose domain-specific synthetic data augmentation to further enhance the performance of domain adaptation and use discriminator outputs for feature fusion.

3. Domain Adaptation from Image to Video

As previewed in Section 1, curating large-scale video datasets with identity labels is an onerous task, but there do exist such datasets for still images. This makes it natural to consider image to video domain adaptation. However, the representation gap is a challenging one to bridge due to blur, compression, motions and other artifacts in videos. This section tackles the challenge by introducing a set of domain adaptation objectives that allow our video face recognition network (VDNet) to be trained on large-scale unlabeled videos in $\mathcal{V}$, while taking advantage of supervised information from labeled web-face images in $\mathcal{I}$.

3.1. Distilling Knowledge by Feature Matching

To take advantage of labeled web-face images, we train VDNet by distilling discriminative knowledge from a face recognition engine pretrained on a labeled web-face dataset, which we call a reference network (RFNet). Unlike previous works that distill knowledge through class probabilities [14], we do so by matching feature representations between two networks, since we do not have access to labeled videos. Let $\phi(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^K$ be a feature generation operator of VDNet and $\psi(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^K$ be that of RFNet. The feature matching (FM) loss is defined on an image $x \in \mathcal{I}$ as:

$$\mathcal{L}_{\text{FM}} = \frac{1}{|\mathcal{I}|} \sum_{x \in \mathcal{I}} \|\phi(x) - \psi(x)\|_2^2 \qquad (1)$$

The FM loss allows VDNet to maintain a certain degree of discriminative information for face identity recognition. With regards to network structure, VDNet can be very flexible as long as the matching feature has the same dimensionality with that of RFNet. In practice, we use the same network architecture between VDNet and RFNet. Moreover, we initialize the network parameters of VDNet with RFNet and freeze network parameters for a few higher layers to further maintain discriminative information learned from labeled web-face images, as illustrated in Figure 2. Note that while more complex distillation methods and architectures are certainly possible, our intent is simply a strong initialization for VDNet, for which these choices suffice.

3.2. Adaptation via Synthetic Data Augmentation

Data augmentation has been widely used for training very deep CNNs with limited amount of training data as it prevents overfitting and enhances generalization ability. In addition to generic data transformations such as random cropping or horizontal flips, applying data transformation that is specific to the target domain has been shown to be effective [6]. To generalize to video frames, we consider data augmentation by applying transformations such as linear motion blur, image resolution (scale) variation or video compression noise, which are the most typical causes of quality degradation in video. We train VDNet to "restore" the original RFNet representation of an image without data augmentation through the feature restoration (FR) loss:

$$\mathcal{L}_{\text{FR}} = \frac{1}{|\mathcal{I}|} \sum_{x \in \mathcal{I}} \mathbb{E}_{B(\cdot)} \left[\|\phi(B(x)) - \psi(x)\|_2^2 \right] \qquad (2)$$

where $B(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^D$ is an image transformation kernel and $\mathbb{E}_{B(\cdot)}$ is the expectation over the distribution of $B(\cdot)$. In this work, we consider three types of image transformations with the following parameters:

- Linear motion blur: kernel length is randomly selected in $(5, 15)$ and kernel angle is selected in $(10, 30)$.
- Scale variation: we rescale an image as small as $\frac{1}{6}$ of the original image size.
- JPEG compression: the quality parameter is set randomly in $(30, 75)$.

These augmentations are applied in sequence to an image with probability of 0.5 for each noise process.

Taking advantage of labeled training examples from image domain, one can also use standard metric learning objectives to learn discriminative metric that generalizes to low-quality images defined by aforementioned blur kernels. We adopt N-pair loss [27], which is shown to be effective at learning deep distance metric from large number of classes. Given N pairs of examples from N different classes $\{(x_i, x_i^+)\}_{i=1}^N$ with individual synthetic data augmentation $B_i(\cdot)$, the N-pair loss is defined as follows:

$$\mathcal{L}_{\text{IC}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\phi(B_i(x_i^+))^\top\psi(x_i))}{\sum_{n=1}^{N}\exp(\phi(B_i(x_i^+))^\top\psi(x_n))} \quad (3)$$

We note that N-pair loss could be one example of an objective function for metric learning with synthetic augmentation, but can be replaced with other standard metric learning objectives such as contrastive loss [2] or triplet loss [26].

3.3. Adaptation via Domain Adversarial Learning

Although data augmentation has been successful in many computer vision applications, the types of transformation between source and target domains are not always known, that is, there might be many unknown factors of variation between two domains. Moreover, modeling such transformations is challenging even if they are known, so we may need to resort to an approximation in many cases. Therefore, it is difficult to close the gap between two domains. Rather than attempting to exhaustively enumerate or approximate different types of transformations between two domains, we learn them from large-scale unlabeled data and facilitate the recognition engine to be robust to those transformations.

Adversarial learning [11] provides a good framework to approach the above problem, whereby the generator, that is, VDNet, is regularized to close the gap between two domains, where the domain difference is captured by the discriminator. The adversarial loss with two domains $\mathcal{I}$ and $\mathcal{V}$ is defined over the expectation of all training samples:

$$\begin{aligned}\mathcal{L}_{\text{D}} = &-\mathbb{E}_{x\in\mathcal{I}}\big[\log\mathcal{D}(y=1|\phi(x))\big] \\ &-\mathbb{E}_{x\in\mathcal{V}}\big[\log\mathcal{D}(y=2|\phi(x))\big] \quad (4)\\ \mathcal{L}_{\text{Adv}} = &-\mathbb{E}_{x\in\mathcal{V}}\big[\log\mathcal{D}(y=1|\phi(x))\big] \quad (5)\end{aligned}$$

The discriminator ($\mathcal{D}$) is defined on top of VDNet that already induces highly abstract features from a deep CNN. Thus, the architecture of $\mathcal{D}$ can be very simple, such as two or three fully-connected layer networks. Unlike several recent applications of adversarial frameworks on image translation [18, 31], $\mathcal{D}$ is not distinguishing between generated and real images in pixel space, rather between feature representations. We argue that this is desirable due to the relative maturity of feature learning for face recognition as opposed to high-quality image generation.

Note that adversarial loss allows utilizing a large volume of unlabeled video data to train VDNet without additional labeling effort. However, the loss can only match representations between two domains in a global manner and the effect would be marginal if the contrast between two domains is small or the discriminator cannot distinguish them well. As a result, we can still take advantage of synthetic data augmentation to guide the discriminator, either to realize the difference between domains or to discriminate additional domain differences from known synthetic transformations. This naturally leads us to two different discriminator types, one with two-way classifier between image ($\mathcal{I}$) and synthesized image and video ($B(\mathcal{I})\cup\mathcal{V}$) or the other with a three-way classifier among image, synthesized image, and video.

Two-way $\mathcal{D}$. We use a two-way softmax classifier as $\mathcal{D}$ to discriminate between the image domain ($y=1$) and the domain of synthesized images and videos ($y=2$). While the original images are from the image domain, both synthetically degraded images as well as random video frames are trained to belong to the same domain as follows:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}} = &-\mathbb{E}_{x\in\mathcal{I}}\big[\log\mathcal{D}(y=1|\phi(x))\big] \\ &-\mathbb{E}_{x\in B(\mathcal{I})\cup\mathcal{V}}\big[\log\mathcal{D}(y=2|\phi(x))\big] \quad (6)\\ \mathcal{L}_{\text{Adv}} = &-\mathbb{E}_{x\in B(\mathcal{I})\cup\mathcal{V}}\big[\log\mathcal{D}(y=1|\phi(x))\big] \quad (7)\end{aligned}$$

Since the contrast between two classes becomes apparent by including synthetic images for the second class, the transformations in the video domain that are similar to synthetic image transformations can be easily restored.

Three-way $\mathcal{D}$. We use a three-way softmax classifier as $\mathcal{D}$ to discriminate images ($y=1$), synthesized images ($y=2$) and video frames ($y=3$) into three different categories.

$$\begin{aligned}\mathcal{L}_{\mathcal{D}} = &-\mathbb{E}_{x\in\mathcal{I}}\big[\log\mathcal{D}(y=1|\phi(x))\big] \\ &-\mathbb{E}_{x\in\mathcal{B}(\mathcal{I})}\big[\log\mathcal{D}(y=2|\phi(x))\big] \quad (8)\\ &-\mathbb{E}_{x\in\mathcal{V}}\big[\log\mathcal{D}(y=3|\phi(x))\big] \\ \mathcal{L}_{\text{Adv}} = &-\mathbb{E}_{x\in B(\mathcal{I})\cup\mathcal{V}}\big[\log\mathcal{D}(y=1|\phi(x))\big] \quad (9)\end{aligned}$$

Unlike the two-way network, the three-way network aims to distinguish video frames from not only the image domain but also synthetically degraded images. Therefore, it may not learn a VDNet with as strong restoration capability to synthetic transformations as with two-way discriminator, but aims to find additional factors of variation between image or synthetic image and video domains.

Overall, the objective function is written as follows:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \alpha\mathcal{L}_{\text{FR}} + \beta\mathcal{L}_{\text{IC}} + \gamma\mathcal{L}_{\text{Adv}} \quad (10)$$

3.4. Discriminator-Guided Feature Fusion

As noted by Yang et al. [39], the quality evaluation of each frame is important for video face recognition since not all frames contribute equally. Moreover, it is important to

discount frames that are extremely noisy due to motion blur or other noise factors, in favor of those that are better for recognition. Our discriminator is already trained to distinguish still images from blurred ones or video frames, so its output may already be used as a confidence score at for each frame being a high quality image. Training with the domain contrast between image, blurred image and video, the discriminator is ready to provide a confidence score at test time, for each frame being a "high-quality web image" ($\mathcal{D}(y = 1|\phi(v))$). Specifically, with the confidence score from the discriminator, the aggregated feature vector for a video V with frames v is represented as a weighted average of feature vectors as follows:

$$\phi_V = \frac{\sum_{v \in V} \mathcal{D}\left(y = 1|\phi\left(v\right)\right) \cdot \phi\left(v\right)}{\sum_{v \in V} \mathcal{D}\left(y = 1|\phi\left(v\right)\right)}. \tag{11}$$

Note that this target domain of web images comes with large-scaled labeled training examples to train a discriminative face recognition engine. Thus, the discriminator serves a dual role of guiding both the feature-level domain adaptation and a fusion weighted by confidence in the fitness of a frame for a face recognition engine.

4. Implementation Details

We provide detailed information of network architectures for the RFNet, the VDNet, and the discriminator.

4.1. Face Recognition Engine

Our face recognition engine is also on a deep CNN trained on CASIA-webface dataset [40]. The network architecture is similar to the ones used in [40, 27], which contains 10 layers of 3×3 convolution followed by ReLU nonlinearities with 4 max pooling layers with stride 2 and one average pooling layer with stride 7, except for that our network uses strided convolution to replace max pooling and maxout units [12] for every other convolution layer instead of ReLU layers. Please see supplementary material for more details. The model is trained with the deep metric learning objective called N-pair loss [27] as described in Section 3.2. Our implementation is based on Torch [3] and $N = 1080$ (N-pair loss pushes (N-1) negative examples at the same time while pulling a single positive example) is used on 8 GPUs for training. Faces are detected and aligned using keypoints [42] and 100×100 gray-scale image patches randomly cropped from 110×110 resized face images are fed to network for training. The model achieves 98.85% verification accuracy on the Labeled Faces in the Wild dataset [15].

The RFNet is the same as face recognition engine and the parameters are fixed. The VDNet is initialized the same as RFNet but the parameters are updated for all layers except for the last two convolution layers, as illustrated in Figure 2.

4.2. Discriminator

We apply a similar network architecture of $\mathcal{D}$ for two and three-way discriminators. For the discriminator $\mathcal{D}$, we use a simple neural network with two fully-connected layers ($320 - 160 - \text{ReLU} - 3$) as shown in Figure 2. For two-way networks, we replace the output channel of last fully-connected layer from three to two. We train the network using Adam optimizer [16] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate of 0.0003, while setting $\alpha = \beta = \gamma = 1$. Details of our network architecture and hyper parameters can be found in supplementary material.

5. Experimental Results

We evaluate the performance of our proposed unsupervised domain adaptation framework, by first providing the baseline methods in Section 5.1, and then performing an ablation study for each component of the proposed approach on YouTube Faces (YTF) dataset [37] in Section 5.2. We further evaluate the model trained on the YTF dataset to IJB-A [17], which demonstrates the generalization capabilities of the proposed approach, as presented in Section 5.3.

5.1. Evalutation Protocol

The standard application of image-based face recognition engine for video face recognition is to first apply the face recognition engine to each frame and then aggregate extracted features from individual frames to obtain a single representation of videos. For baselines, we follow the standard protocol of extracting L2 normalized features from each frame and its horizontally flipped image followed by temporal average pooling over frames per video. For discriminator-guided feature fusion, we follow Equation (11) to obtain a video representation from individual L2 normalized features. We compute inner product between two video representations for similarity metric.

5.2. YouTube Faces Dataset

The YTF dataset contains 3425 videos of unconstrained face images from 1595 different people with the average length of 181.3 frames per video. Ten folds of video pairs are available for verification experiments, where each fold is composed of 250 positive and negative pairs with no overlapping identity between different folds. We use videos in 8 training folds in addition to CASIA-webface dataset to train a VDNet, but no identity label from video is used.

Six networks (A-F) with different combinations of feature matching (FM), feature restoration (FR) with various data augmentation methods, adversarial training (Adv) and discriminator-guided feature fusion are presented in Table 1.

Feature matching loss. The FM loss enforces VDNet to learn similar representations as those produced by RFNet on labeled still images, which is one of the key contributors to a good initialization for our method. Combination of image classification (IC) and FM reduces to training the baseline model. The effectiveness of FM loss can be seen by comparing accuracies of model A and B in Table 1. A significant

Table 1. Video face recognition accuracy and standard error on the YTF dataset. Image-classification loss (IC), feature matching loss (FM), feature restoration loss (FR) and adversarial loss (Adv) are applied for training. For feature restoration loss, we consider three types of data augmentation, namely, linear motion blur (M), scale variation (S), or JPEG compression noise (C). The best performer and those with overlapping standard error are boldfaced.

Model	IC	FM	FR	Adv	fusion	1 (fr/vid)	5 (fr/vid)	20 (fr/vid)	50 (fr/vid)	all
baseline			–		–	91.12±0.318	93.17±0.371	93.62±0.430	93.74±0.443	93.78±0.498
					✓	–	93.30±0.362	93.72±0.428	93.80±0.444	93.94±0.493
A	✓	–	M/S	–	–	91.37±0.334	92.97±0.381	93.42±0.399	93.43±0.384	93.32±0.443
B	✓	✓	M/S	–	–	91.44±0.348	93.46±0.392	93.84±0.433	93.95±0.443	93.94±0.507
C	✓	✓	M/S/C	–	–	91.68±0.320	93.52±0.323	93.94±0.337	93.90±0.361	93.82±0.383
D	✓	✓	–	two-way	–	91.38±0.350	93.74±0.354	94.04±0.375	94.23±0.379	94.36±0.346
E	✓	✓	M/S/C	two-way	–	92.39±0.315	94.72±0.306	**95.13**±0.263	**95.13**±0.286	**95.22**±0.319
					✓	–	94.73±0.270	**95.14**±0.229	**95.13**±0.261	**95.16**±0.284
F	✓	✓	M/S/C	three-way	–	92.17±0.353	94.44±0.343	**94.90**±0.345	**94.98**±0.354	**95.00**±0.415
					✓	–	94.52±0.356	**95.01**±0.352	**95.15**±0.370	**95.38**±0.310

performance drop is observed for the model trained without FM loss but only with the FR loss, when it is evaluated with more number of frames per video. We hypothesize that while feature restoration loss drives the VDNet to match the representation of low-quality images to their high-quality counterpart, the representation of the original high-quality images are severely damaged, causing the model to lose its superior performance on high-quality images. By requiring the network to work well on both high-quality as well as low-quality images with FM loss, we observe significant improvement when evaluated with larger number of frames per video (for example, from 93.32% of model A to 93.94% of model B with all frames per video).

Feature restoration loss. We consider three types of data augmentation described in Section 3.2, with different combinations presented with model B and C in Table 1. Overall, FR loss moderately improves accuracy compared to the baseline models. Specifically, we observe that compression noise is quite effective at feature-level restoration when used along with linear motion blur and scale variations.

When combined with adversarial loss, we observe more significant improvement with feature restoration loss. For example, model E and F reduce the relative error by 11.7% and 9.2% compared to model D, respectively, on single frame per video evaluation regime and 15.6% and 13.0% when using randomly selected 50 frames per video for evaluation.

Domain adversarial loss. In addition to feature restoration loss through synthetic data augmentation, domain adversarial training between high-quality web images and the videos contributes to reducing the gap between two domains. To demonstrate its effectiveness, we train the model only with Adv loss with random video frames as the additional input source with the "fake" labels (while random face images associate to the "real" ones), and compared to the baseline model. As shown in Table 1, model D consistently outperforms the baseline one with different number of random frames per video. Note that the "two-way" in model D denotes a binary classification between random face images and video frames, without any artificially degraded sample.

When feature restoration loss is used for training, we consider two types of discriminators since it introduces an additional data domain, namely a synthetic image domain, besides the existing two domains of image and video. First, we merge synthetically degraded images into video domain and the discriminator is still a two-way classifier (model E). Next, synthetic images are considered as their own domain, which leads to a three-way discriminator among image, synthetic image, and video (model F). In comparison to model C which is trained without adversarial loss, both models E and F significantly improve the recognition performance (e.g, from 93.82% to 95.22% or 95.00% for model E and F, respectively). When frame-level features are aggregated by discriminator-guided fusion, the three-way model F improves its performance to 95.38%, which is highly competitive to the performance of previous state-of-the-art face recognition engines such as FaceNet [26] (95.12%), CenterFace [36] (94.9%), or CNN with different feature aggregation methods [39] (e.g., 95.20% with average pooling) as shown in Table 3. Note that the evaluation protocol of prior works is the same as our baseline model, but their networks are either much deeper or trained on significantly larger number of training images.

Discriminator-guided feature fusion. The proposed fusion strategy selectively adopts high-quality frames while discarding poorer ones in order to further improve the recognition accuracy. To reflect the quality of each frame, the feature fusion module aggregates all frames of a video using a weighted average of feature vectors based on the normalized likelihood as in (11). We quantitatively and qualitatively evaluate the discriminator-guided fusion in Table 1 and Figure 3. By applying $\mathcal{D}$ from the three-way network, the model F and the baseline model present consistent improvements. In contrast, the fusion strategy for the two-way network in model E (second sub-row) has marginal effect. This indicates that the three-way network learns a multi-class discriminator that can better distinguish among input sources.

Qualitative visualization of guided fusion. In Figure 3, we demonstrate qualitative results for the feature fusion us-

Table 2. 1:1 verification and rank-K identification accuracy and standard error on IJB-A dataset. The model F is compared to the baseline method described in Section 4.1. Ten image crops (4 corners + 1 center + horizontal flip) are evaluated and fused together with uniform weights or discriminator confidence score. We also evaluate after removing images with significant localization errors (*).

Model	fusion	1:1 Verification TAR			1:N Identification Rank-K Accuracy		
		FAR=0.001	FAR=0.01	FAR=0.1	Rank-1	Rank-5	Rank-10
baseline	–	0.539±0.013	0.773±0.008	0.954±0.002	0.864±0.004	0.951±0.003	0.970±0.002
baseline*	–	0.646±0.012	0.846±0.005	0.968±0.001	0.902±0.003	0.959±0.003	0.971±0.001
F	–	0.531±0.016	0.800±0.008	0.963±0.002	0.869±0.003	0.954±0.003	0.970±0.002
F	✓	0.584±0.018	0.828±0.008	0.962±0.001	0.879±0.004	0.955±0.003	0.970±0.002
F*	✓	0.649±0.022	0.864±0.007	0.970±0.001	0.895±0.003	0.957±0.002	0.968±0.002
Wang et al. [34]	–	0.510±0.019	0.729±0.011	0.893±0.004	0.822±0.007	0.931±0.004	–
$DCNN_{all}$ [1]	–	–	0.787±0.014	0.947±0.003	0.860±0.007	0.943±0.005	–
Yang et al. [39]	Mean L2	0.688±0.025	0.895±0.005	0.978±0.001	0.916±0.004	0.973±0.002	0.980±0.001
Yang et al. [39]	NAN	0.860±0.004	0.933±0.003	0.979±0.001	0.954±0.002	0.978±0.001	0.984±0.001

Table 3. Comparison on the YTF dataset with other unsupervised domain adaptation methods and state-of-the-art methods.

Unsupervised DA		SOTA (image-based)	
baseline	93.78	DeepFace [32]	91.4
PCA	93.56	FaceNet [26]	95.12
CORAL [28]	94.50	CenterFace [36]	94.9
Ours (F)	**95.38**	CNN+AvePool [39]	95.20

ing the three-way discriminator scores. Each row shows the top and bottom scored frames. It is evident that high-quality frames score higher than low-quality ones. More importantly, we observe that the notions of quality are diverse and encompass factors of variation such as pose, blur, lighting and occlusions. This supports our hypothesis that there are several causes of domain gap between images and videos, so an adversarially trained discriminator is better than one that relies on enumerating all possible factors. Our analysis is reminiscent of that in [39], but we learn the quality of video frames in an unsupervised manner without identity labels, whereas [39] utilizes identity labels to assign a higher score to a frame that contributes more to classification.

Comparison with other unsupervised DA methods. We study the effectiveness of our proposed method in comparison to other works on unsupervised domain adaptation such as PCA feature transform or Correlation Alignment (CORAL) [28] methods. We extract features from both still images and video frames using RFNet and apply PCA feature transform while retaining 90% of the total variation. For CORAL, we calculate the mean $(\mu_{\mathcal{I}}, \mu_{\mathcal{V}})$ and covariance $(C_{\mathcal{I}}, C_{\mathcal{V}})$ of the features from two domains on the training set and transform features, for individual frames, as follows:

$$\phi(v) \leftarrow (\phi(v) - \mu_{\mathcal{V}}) C_{\mathcal{V}}^{-\frac{1}{2}} C_{\mathcal{I}}^{\frac{1}{2}} + \mu_{\mathcal{I}} \qquad (12)$$

The results in Table 3 show that simple feature transform method like PCA does not work well since it does not distinguish between two domains when computing the transformation matrix. On the other hand, CORAL demonstrates moderate improvement upon baseline by matching the first and second-order statistics of representations between two domains. Our method is also based on feature distribution matching between two domains through a discriminator, but allows learning a more complete transformation through end-to-end training of deep networks with a combination of losses, which results in more substantial improvements.

5.3. IJB-A Dataset

IJB-A [17] is a benchmark dataset for face recognition in the wild. It contains a mix of 5397 still images and 20412 video frames sampled from 2042 videos of 500 different subjects. The existence of video frames allows set-to-set comparison for verification, which opens up a new challenge for the face recognition problem. It is more challenging than LFW or YTF due to many factors of variation such as pose or facial expression as well as various image qualities.

There are 10 splits with about 30k labeled images in each. Due to the small-sized training set, prior works pretrain the network using large-scale labeled datasets and use training set of each split for supervised fine-tuning. While we also use a pretrained face recognition engine for video face recognition, our network is fine-tuned on external video data without identity information and IJB-A is used for evaluation only. This is because the emphasis of our method is to use large-scale unlabeled video data to improve video face recognition rather than fine-tuning with manually labeled data. We first remove 379 videos from the YTF training set as their identities overlap with those of IJB-A test set. We then utilize the YTF dataset without label information for network training and the unlabeled video fine-tuned network is evaluated on 10 splits of IJB-A test set.

We perform experiments in two settings. The 1:1 verification task compares genuine and impostor samples for one-to-one verification, while the 1:N task is to search samples against an enrolled gallery. We report the true acceptance rate (TAR) at different false acceptance rates (FAR) of 0.1, 0.01, and 0.001 for verification and the rank-1, 5, 10 accuracy for identification in Table 2.

Compared to the baseline model, we observe slightly worse performance at FAR=0.001, but significant improvement on both FAR=0.01 and 0.1 (e.g., 77.3% to 82.8% at FAR=0.01). Especially, when discriminator-guided feature fusion is used, the improvement becomes more significant. We also evaluate by following the protocol in [34] as we observe non-negligible amount of localization errors while

(a) top-5 (b) bottom-5

Figure 3. We sort the frames within a sequence in a descending order with respect to the discriminator-guided weights ($\mathcal{D}(y = 1|v)$), and display them by showing the top-5 and bottom-5 instances, respectively. The weights are shown in the upper-left corner of each frame. We visualize eight video sequences that illustrate the following video quality degradation: **blurring**, **compression noise**, **lighting variation**, **video shot-cut**, **occlusion**, **detection failure**, **pose variation** and **mis-alignment** from the first to the last row.

preprocessing. By removing poor quality images based on localization errors, we obtain much higher verification and identification results. However, the gap between our proposed model and the baseline is reduced, which we believe is due to low-quality images being mostly filtered out, so the baseline method can still work fine or even better. Compared to previous works, the performance of our proposed method is competitive under similar training and testing protocols [34, 1]. Further improvement is expected by training a stronger base network on larger labeled image datasets [39] or with other types of synthetic data augmentations such as pose variation with 3D synthesization [23].

6. Conclusions

Face recognition in videos presents unique challenges due to the paucity of large-scale datasets and several factors of variation that degrade frame quality. In this work, we address those challenges by proposing a novel feature-level domain adaptation approach that uses large-scale labeled still images and unlabeled video data. By distilling discriminative knowledge from a pretrained face recognition engine on labeled still images while adapting to video domain through synthetic data augmentation and domain adversarial training, we learn domain-invariant discriminative representations for video face recognition. Furthermore, we propose a discriminator-guided feature fusion method to effectively aggregate features from multiple frames and effectively rank them in accordance to their suitability for face recognition. We demonstrate the effectiveness of the proposed method for video face verification on the YTF and IJB-A benchmarks. Our future work will further exploit unsupervised domain adaptation to achieve continuous improvements through a growing collection of unlabeled videos.

Acknowledgments

This work is supported in part by the NSF CAREER Grant #1149783 and a gift from NEC Labs America.

References

[1] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. IEEE, 2016. 7, 8

[2] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005. 4

[3] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A Matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011. 5

[4] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *CVPR*, 2013. 1

[5] J. E. Cutting. *Perception with an Eye for Motion*. MIT Press, 1986. 1

[6] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *arXiv:1607.05427*, 2016. 2, 3

[7] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013. 2

[8] B. Fernando, T. Tommasi, and T. Tuytelaars. Joint cross-domain classification and subspace learning for unsupervised adaptation. *Pattern Recoginition*, 2015. 2

[9] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2, 3

[10] J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979. 1

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 2, 4

[12] I. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *ICML*, 2013. 5

[13] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *CVPR*, 2016. 2

[14] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531, 2015. 3

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 1, 5

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[17] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, 2015. 1, 5, 7

[18] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 4

[19] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *CVPR*, 2013. 1

[20] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-pep for video face recognition. In *ACCV*, 2014. 1

[21] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016. 2

[22] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, 2013. 2

[23] I. Masi, A. T. Trn, T. Hassner, J. T. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, 2016. 8

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015. 1

[25] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2

[26] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 4, 6, 7

[27] K. Sohn. Improved deep metric learning with multi-class N-pair loss objective. In *NIPS*. 2016. 4, 5

[28] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016. 7

[29] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014. 1

[30] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *CoRR*, abs/1611.02200, 2016. 2

[31] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. 2017. 4

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1, 7

[33] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014. 2

[34] D. Wang, C. Otto, and A. K. Jain. Face search at scale: 80 million gallery. *arXiv preprint arXiv:1507.07242*, 2015. 7, 8

[35] X. Wang, A. Farhadi, and A. Gupta. Actions ˜ transformations. In *CVPR*, 2016. 2

[36] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016. 6, 7

[37] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011. 1, 5

[38] L. Wolf and N. Levy. The svm-minus similarity score for video face recognition. In *CVPR*, 2013. 1

[39] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *CVPR*, 2017. 4, 6, 7, 8

[40] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. 1, 2, 5

[41] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon. Pixel-level domain transfer. In *ECCV*, 2016. 2

[42] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016. 5

[43] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *ECCV*, 2016. 2