Red Carpet to Fight Club: Partially-supervised Domain Transfer for Face Recognition in Violent Videos

Yunus Can Bilge*[1], Mehmet Kerim Yucel*[1], Ramazan Gokberk Cinbis[2], Nazli Ikizler-Cinbis[1], and Pinar Duygulu[1]

[1]*Department of Computer Engineering, Hacettepe University*
[2]*Department of Computer Engineering, Middle East Technical University*
{*yunuscanbilge,nazli,pinar*}*@cs.hacettepe.edu.tr, mkerimyucel@hacettepe.edu.tr, gcinbis@ceng.metu.edu.tr*

Abstract

In many real-world problems, there is typically a large discrepancy between the characteristics of data used in training versus deployment. A prime example is the analysis of aggression videos: in a criminal incidence, typically suspects need to be identified based on their clean portrait-like photos, instead of their prior video recordings. This results in three major challenges; large domain discrepancy between violence videos and ID-photos, the lack of video examples for most individuals and limited training data availability. To mimic such scenarios, we formulate a realistic domain-transfer problem, where the goal is to transfer the recognition model trained on clean posed images to the target domain of violent videos, where training videos are available only for a subset of subjects. To this end, we introduce the "WildestFaces" dataset, tailored to study cross-domain recognition under a variety of adverse conditions. We divide the task of transferring a recognition model from the domain of clean images to the violent videos into two sub-problems and tackle them using (i) stacked affine-transforms for classifier-transfer, (ii) attention-driven pooling for temporal-adaptation. We additionally formulate a self-attention based model for domain-transfer. We establish a rigorous evaluation protocol for this "clean-to-violent" recognition task, and present a detailed analysis of the proposed dataset and the methods. Our experiments highlight the unique challenges introduced by the WildestFaces dataset and the advantages of the proposed approach.

1. Introduction

People engaging in criminal activities are likely to expose a diverse set of facial expressions/poses. The people in these activities are also likely to move fast, causing the recorded video footage to have significant amount of blur and occlusion. What is even more challenging is that, these people may not necessarily have prior criminal records, therefore may not have recorded *"fight scene footage's"*. They may only have *"clean"* images, such as passport or Facebook-type of photos, that can be used for identification.

In this paper, we formulate this task as transferring the face recognition model from the domain of clean (so-called *Red Carpet*) images to the domain of violent (*Fight Club*) videos [1]. Since the training videos are labeled but scarce and made available only for a subset of people, we refer to the learning problem as *partially-supervised domain-transfer*.

A plethora of studies have focused on face recognition in computer vision literature. Compared to the pioneering works [60, 1, 74, 14, 69, 67], face recognition models that benefit from deep learning-based techniques and concentrate on better formulation of distance metric optimization raised the bar [52, 58, 45, 66, 57, 55, 56, 12, 83, 64]. There has been interest in using additional data (in the form of unlabeled [81, 9] or synthetic data [82]), class-balancing [80] and noisy-data handling [27] to improve face recognition accuracy. In addition to face recognition in still images, video-based face recognition studies have also emerged (see [13] for a recent survey). Ranging from local feature-based methods [37, 44, 38] to manifolds [30] and metric learning [7, 31, 24], recent studies have focused on finding informative frames in image sets [23] and finding efficient ways of feature aggregation [8, 78, 48, 49]. Most of these studies concentrate on relatively easier cases of recognition, where the faces are seen under good lighting conditions and are

*equal contribution

[1]Our dataset is available at https://ycbilge.github.io/wildestFaces.

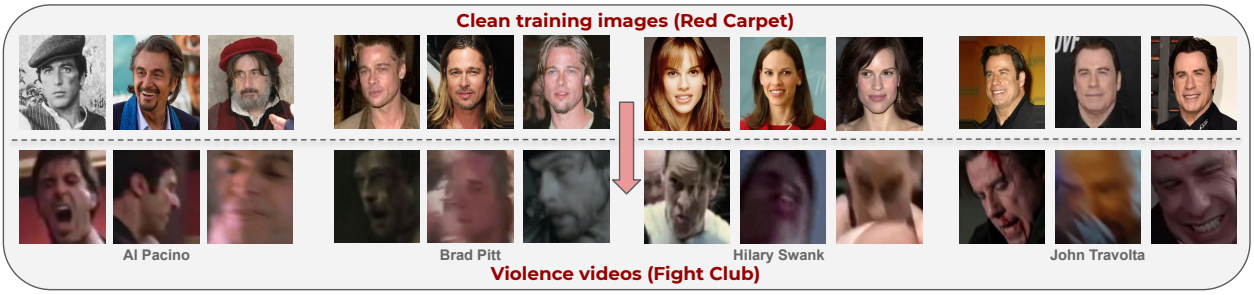


Figure 1. Our focus is the problem of transferring recognition models trained on clean, portrait-like images for person identification in the *wildest* (*e.g.* fight) videos. We introduce the WildestFaces dataset that contains videos of violent scenes from movies, together with the clean *red-carpet* images of the corresponding actors. Importantly, training video examples are available only for a subset of the actors, which leads to a challenging partially-supervised domain transfer problem.

mostly stable with orthogonal viewpoints. In contrast, face recognition in the "wildest" demands more than that.

In order to facilitate research in this direction, we propose a new dataset, referred to as *WildestFaces*. This dataset consists of clips with adverse effects at their extreme, and auxiliary clean facial images of the corresponding people. The videos are collected by manually finding violent scenes in movies of predetermined actors[2], and the clean images are collected from IMDB[3] and similar websites. The task is depicted in Figure 1 with example images from the dataset. The training set provides clean still images of all people and violent videos of only a subset of people. The test set, however, contains novel videos of all people. This setup resembles to that of *generalized zero-shot learning (GZSL)* [73, 5], where there are both seen and unseen classes in testing. We define an evaluation protocol that explicitly measures the success at recognizing people with and without training videos, and, penalizes methods that are poor at any of the two tasks.

To tackle this *partially-supervised domain-transfer* problem, we divide it into two sub-problems and propose two major techniques for each one. First, we leverage stacked affine-transform layers for classifier transfer, which aims to adapt the image-based classification model to the target-domain representations, in a supervised yet data-efficient manner. Second, we propose an attention-driven temporal pooling layer, which aims to enable data-driven adaptation to the face tracks in the video domain. We additionally propose a third self-attention [63] based formulation, with components targeted to both sub-problems. We rigorously evaluate all proposed techniques and show their advantages over a number of state-of-the-art alternatives.

To sum up, the contributions of this paper include: **(1)** A new dataset called WildestFaces that includes a wide range of examples from violent movies; **(2)** A new partially-supervised face recognition task, where classifiers trained on clean image data are evaluated for their ability of recognizing faces in violent videos; **(3)** Rigorous evaluation protocols inspired from the recent developments in related problems (primarily zero-shot learning and few-shot learning); and **(4)** Effective techniques for the proposed partially-supervised, *clean-to-violent* domain-transfer problem.

[2]We use *actor* as a gender-neutral term, following the modern practice.
[3]www.imdb.com

2. Related Work

Face Recognition Datasets: Due to the data-hungry nature of face recognition, there have been many attempts in building large scale datasets. FDDB [32], AFW [85], PASCAL Faces [77], Labeled Faces in the Wild (LFW) [28], Celeb Faces [57], Youtube Faces (YTF) [68], FaceScrub [43], IJB-A [35], MS-Celeb-1M [25], VGG-Face [45], VGG2-Face [4], MegaFace [34] and WIDER Face [79] datasets have been made publicly available for research purposes. Datasets with extreme scales, such as [52] and [58] have also been used but have not been disclosed to the public.

For video face recognition, YouTube Faces [68] is the most widely used dataset. While it contains motion-blurred and low-quality frames, overall the quality of the frames is typically much better than the wildest conditions that we target in our work. Plus, unlike our domain-transfer based image-to-video recognition setup, in YouTube Faces, the primary focus on video-to-video recognition. Other two prominent video face recognition datasets are COX [29] and PasC [2]. Despite their relatively large size, PasC [2] suffers from video location constraints and COX [29] suffers from demographics as well as video location constraints. FaceScrub [43] is a dataset which has resemblance to our case as it also includes actors as individuals. However the dataset only contains actor images rather than videos. The most relevant benchmarks to ours are [33, 15, 53, 36]. However, none of them specifically focus on domain shifts, recognizing *unseen* classes (from a ZSL perspective) or violent settings.

Video Face Recognition: [78] employ attention modules to adaptively aggregate image-based features from frames into a single representation. Instead of aggregating feature representations, [48] opt to aggregate raw frames di-

rectly to produce a synthesized image via generative models that is tailored to be more discriminative. [59] utilizes an encoder-decoder structure in their GAN's generator to achieve a pose-independent identity representation, which is later used to synthesize an image of desired pose. A similar work exploiting attention-like mechanism is presented in [75], where content and visual quality of each image is learned to perform set-wise classification. [49] exploits reinforcement learning to attend to informative frames in videos which are aggregated by a mean-pooling to represent sets as a single feature vector. [71], on the other hand, presents a light-weight network to achieve fast face recognition. [82] employ video-domain only face recognition technique in a fully supervised manner. The method is tailored for single-domain face recognition and for seen classes. In template-based face recognition, a similar effort to produce a single representation is given in [84]. These papers, however, do not address the explicit domain-transfer problem in *clean-to-violent* face recognition problem and operate in the domains of the datasets they are trained on.

A recent work that is most closely related to ours is [42], where the main goal is to identify characters in TV series videos using classifiers trained on clean actor images. Their main focus is different in many ways from ours since [42] (i) presumes that a large number of weakly supervised video examples are available for all characters, (ii) uses voice classifiers and shows that identification is largely influenced by them, and, (iii) leverages similarities across different scenes within a single dataset during recognition.

Domain Adaptation: Due to its dual-domain nature, clean-to-violent face recognition poses a domain-transfer problem. Having found application areas in primarily computer vision tasks [72, 70, 65], supervised [6, 18, 61] and unsupervised variants [17, 19, 76, 41] of domain adaptation techniques have surfaced in recent years. There are several approaches pertinent to the task, such as feature space alignment [54], supervised feature transformation [11, 40], adversarial approaches [22], encoder-decoder structures [3, 19] and many others. For a detailed review in domain adaptation, readers are referred to [10].

The main difference between the mainstream domain-adaptation tasks and our problem definition is that in domain adaptation, it is typically presumed that (labeled or unlabeled) target-domain training examples are available for all classes of interest, which is not realistic for our clean-to-violent recognition problem. To this end, our work aims to (i) address a partially supervised domain-transfer problem, (ii) handle unseen class recognition, (ii) introduce evaluation protocols for partially-supervised transfer and (iii) learn to handle noisy sequences. A similar work to ours in domain adaptation literature is [41], however ours differ from this work by the factors listed above and ours is also geared towards dual-domain (image-to-video and clean-to-Wildest) face recognition.

Zero-shot learning: In zero-shot learning (ZSL), the classifiers are learned over seen classes and then extended to unseen classes of which labeled data is not accessible, by means of auxiliary data such as attributes or textual descriptions. Generalized zero-shot learning (GZSL) [73, 5] extends the test protocol of ZSL to include seen and unseen classes together, as it is more natural to assume cooccurence of these classes in general. Different from regular GZSL, the auxiliary information is in the form of a set of labelled images, as opposed to attributes or textual descriptions as mostly used in the mainstream zero-shot learning research.

Overall, the proposed problem setup is at the intersection of GZSL and supervised domain adaptation, where training video data is available only for a subset of classes.

3. WildestFaces Dataset

To the best of our knowledge, there is no publicly available dataset which is composed of fight and dispute videos, with annotated human faces. We introduce the *WildestFaces* dataset collected by focusing on violent movie scenes of celebrities. Below, we give the details of the dataset and the collection procedure.

Videos. We first created a list of actors appearing in movies with violence. We then collected videos of them from YouTube using a variety of scene settings; *e.g*. car chase, fist fights, gun fights, heated arguments, *etc*. This abundance in scene settings provide an inherent variety of occlusions, poses, background clutter and motion blur. Videos, with an average 25 FPS are then divided into shots with a maximum duration of 10 seconds.

In total, for 64 selected actors, 2,186 shots (64,242 frames) from 410 videos are collected. We annotated the face regions by applying a face detector and manually correcting its mistakes. In this process, no frames were filtered out due to adverse conditions; and we labeled even extremely tiny, occluded, frontal/profile and blurred faces.

Clean images. In order to employ classifier transfer from clean images to violence videos, we also collected images of actors taken under normal conditions such as red carpet images. We primarily use IMDB-WIKI [50], from which we acquire the images of 62 celebrities that overlap with the video subjects. For the remaining two subjects, we collect images from the Internet. In total, we obtain 8069 images of 64 subjects. A detailed analysis of the dataset is presented in Section 5.

4. Partially-supervised domain-transfer

We assume that during training, there are two sets: (i) the source domain training set $\mathcal{D}_x = (x_i, y_i)_{i=1}^{n_x}$ with n_x still image examples, and, (ii) the target video domain training set $\mathcal{D}_v = (v_j, y_j)_{j=1}^{n_v}$ with n_v video examples. Each exam-

ple (x_i, y_i) in the image training set $\mathcal{D}_x$ contains the facial image $x_i \in \mathcal{X}$ and the corresponding person label y_i. Each example (v_j, y_j) in the video training set $\mathcal{D}_v$ contains the facial image sequence $v_j \in \mathcal{V}$ of length $|v_j|$, with frames denoted as $v_j = (v_j[1], ..., v_j[|v_j|])$, and the label y_j.

The crucial detail is the difference between set of classes spanned by these two training resources: while $\mathcal{D}_x$ provides examples for the set $\mathcal{Y}$ of all classes, $\mathcal{D}_v$ provides examples only for a subset of them. However, at test time, an input video may belong to any class in $\mathcal{Y}$. Inspired from the similarity to the generalized zero-shot learning problem (as discussed in Section 2), we refer to the set of classes having both image and video training examples as the *seen* classes and denote them by $\mathcal{Y}_{\text{seen}}$, and, the remaining set of classes having only image domain examples as the *unseen* classes and denote them by $\mathcal{Y}_{\text{unseen}}$. We denote the number of seen and unseen classes by c_s and c_u, respectively [4]. The final goal is to learn a classifier scoring function $f_v : \mathcal{V} \rightarrow \mathbb{R}^{|\mathcal{Y}|}$ that maps a video-domain input to the vector of per-class confidence scores for all classes. We divide this task in two sub-problems and propose methods towards each one in the following sections: (i) classifier transfer, (ii) temporal adaptation. We additionally propose a third self-attention [63] based approach that aims to tackle both sub-problems.

4.1. Classifier Transfer

Let $\phi : \mathcal{X} \rightarrow \mathbb{R}^{d_x}$ be an image-domain feature extractor that maps each input face image to a d_x-dimensional vector, and, let $\Psi : \mathcal{V} \rightarrow \mathbb{R}^{d_v}$ be a video-domain feature extractor that maps each input face video to a d_v-dimensional vector. Throughout our experiments, we use a pre-trained VGG-face network [45] as ϕ (Section 5). We propose a number of Ψ alternatives in Section 4.2. In classifier transfer, the goal is to adapt a classifier pre-trained in the source domain using ϕ representation, to the target domain with representation Ψ via the restricted set of examples for the c_s classes.

We start by defining the source domain classifier. For simplicity, we use a linear model for source-domain classification, parameterized by the matrix $W = [w_1, ..., w_{|\mathcal{Y}|}] \in \mathbb{R}^{d_x \times |\mathcal{Y}|}$. The model is trained on the source domain dataset $\mathcal{D}_x$ via regularized loss minimization:

$$\min_W R(W) + \sum_{i=1}^{n_x} \ell(\phi(x_i)^\intercal W, y_i) \tag{1}$$

where $\ell(\cdot, y)$ is the soft-max cross-entropy loss function, and, $R(W)$ is ℓ_2-regularization in our experiments.[5]

We formalize the classifier-transfer problem as the task of learning a transformation $\tau : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_v}$ by minimizing the regularized loss on the target-domain dataset $\mathcal{D}_v$:

$$\min_\tau R(\tau) + \sum_{j=1}^{n_v} \ell(\tau(\Psi(v_j))^\intercal W, y_j) \tag{2}$$

where $R(\tau)$ represents the regularization applied to the transfer model. In this framework, τ has the responsibility of transforming the input $\Psi(v)$ video-domain representation to a $\phi(x)$-like image-domain representation and make it compatible with the classification layer W. When training on the $\mathcal{D}_v$ dataset, we deliberately keep the classification layer W fixed, in order to keep the class models w_j intact and compatible with each other, and, minimize the risk of learning a bias towards the subset of classes seen in $\mathcal{D}_v$.

The first classifier-transfer technique that we consider is the **fully-connected classifier-transfer** layer:

$$\tau_{\text{fc}}(\Psi(v)) = Q\Psi(v), \tag{3}$$

where τ_{fc} is instantiated by a fully-connected layer (fc) $Q \in \mathbb{R}^{d_x \times d_v}$, and, linearly transforms the d_x dimensional image representation to the d_v dimensional face-track vector.

While the aforementioned approach looks simple and promising, it performs poorly in practice: the number of parameters in Q is typically too high to be trained properly unless the target-domain dataset $\mathcal{D}_v$ is large-scale, which is infeasible in most practical scenarios, including ours. For instance, when VGG-face ϕ descriptors are being used and Ψ is defined as the average of per-frame VGG-face descriptors, Q contains 4096^2 ($\sim$ 16M) parameters. As a result, it quickly leads to over-fitting, and yields a poor trade-off between the seen and unseen class performance in the target-domain (Section 5).

In our preliminary experiments, we have investigated a number of common regularization techniques, including ℓ_2, drop-out, batch-normalization and explicit rank regularization, and in all cases, we have observed very similar poor generalization behavior for the fc based classifier transfer.

To avoid these difficulties, we propose the **affine classifier-transfer** layer, which is built upon a much lower-complexity affine model:

$$\tau_{\text{affine}}(\Psi(v)) = \alpha \odot \Psi(v) + \beta, \tag{4}$$

which implements feature scaling via applying Hadamard Product with the vector α, and, shifting by the vector β, which are trained according to Eq. 2. The underlying assumption here is that the source-domain and target-domain representations are of the same dimensionality and are sufficiently correlated so that an affine transform can provide the necessary correction. Fortunately, this assumption is met in most practical Ψ definitions, including temporal average pooling and attentive temporal pooling, which are explained in the following section.

[4] We adapted the GZSL nomenclature, as clean images and training videos are akin to class descriptions and seen class examples, respectively.

[5] While we exclude the regularization weight from the equations for brevity, we tune it on the validation set and utilize in our experiments.

We propose two extensions to the affine classifier-transfer layer. First, we propose the **stacked affine classifier-transfer** where the affine transform is followed by the ReLU activation and then another affine transform. Second, we propose the **residual stacked affine classifier-transfer** (rsa) layer, which includes a residual connection:

$$\tau_{\text{rsa}}(\Psi(v)) = \alpha_2 \odot \max(\alpha_1 \odot \Psi(v) + \beta_1, 0) + \beta_2 + \Psi(v) \quad (5)$$

where (α_1, β_1) and (α_2, β_2) are the parameters of the first and the second affine transforms, respectively. In Section 5, we thoroughly evaluate all major cases of classifier-transfer and the combinations with temporal adaptation techniques.

4.2. Temporal Adaptation

We now continue within the framework defined in the previous section, and propose techniques for obtaining a video representation $\Psi(v)$ suitable for the clean-to-violent domain-transfer task.

For clarity, we start with the simple **temporal average pooling** scheme. In this case, the representation of a face track is obtained by taking average of the per-frame features extracted using the image-domain feature extractor ϕ:

$$\Psi_{\text{AvgPool}}(v) = \frac{1}{|v|} \sum_{t=1}^{|v|} \phi(v[t]) \quad (6)$$

While temporal average pooling is a versatile technique, the resulting representation is likely to be dominated by the heavily motion-blurred faces in a track, plus, it is likely to handle multiple poses poorly. Similarly, temporal max pooling, an obvious alternative, is likely to be negatively affected by these factors.[6]

To handle the multi-modality and noise in face tracks, we aim to learn a data-driven temporal representation optimized for the clean-to-violent domain-transfer task. For this purpose, we propose **attentive temporal pooling** (ATP), inspired from [78]. The intuition behind this model is to exploit the hidden pose information in a trainable fashion (unlike other pooling strategies which require additional input or manual invertention [26, 16]) to extract useful information in the noisy sequences of video frames. The proposed approach consists of two main components: (i) an attention layer, and, (ii) a attention-weighted pooling layer. Attention module learns to promote the informative parts of given image sequences. Through the pooling layer, the overall sequence information is aggregated.

More formally, assuming that per-frame descriptors are extracted using ϕ, we define the attention weight matrix $A = [a_1, ..., a_K] \in \mathbb{R}^{d_x \times K}$, where K can be interpreted as the hyper-parameter defining the number of canonical appearance modes used in the attention model. The attention function $\Gamma(v)$ computes a $|v| \times K$ attention matrix, whose t-th frame, k-th mode value is given by applying a temporal softmax over the raw attention scores:

$$[\Gamma(v)]_{t,k} = \frac{\exp\left[\phi(v[t])^\intercal a_k\right]}{\sum_{t'=1}^{|v|} \exp\left[\phi(v[t'])^\intercal a_k\right]}. \quad (7)$$

The k-th column of the resulting matrix can be considered as a weight distribution over the frames. We use these weights in temporal pooling to obtain K different representations, *i.e.* a $d_x \times K$ dimensional matrix given by $\Phi(v)\Gamma(v)$, where $\Phi(v) = [\phi(v[t])]_{t=1}^{|v|} \in \mathbb{R}^{d_x \times |v|}$ is the matrix of all per-frame ϕ representations of the facial video v. These per-mode descriptors are then aggregated into a single d_x-dimensional vector using average-pooling (Figure 2). The overall operation can equivalently be expressed as:

$$\Psi_{\text{ATP}}(v) = \frac{1}{K} \Phi(v)\Gamma(v)\mathbf{1}_K \quad (8)$$

where $\mathbf{1}_K$ is the K-dimensional vector of all ones. This expression also reveals that the ATP scheme effectively assigns an attention weight to each frame, where all unnormalized per-frame weights are given by $\Gamma(v)\mathbf{1}_K$.

Using the previously defined domain-transfer framework, we learn the ATP model jointly with the classifier-transfer model τ on the dataset $\mathcal{D}_v$:

$$\min_{\tau, \Psi_{\text{ATP}}} R(\tau) + R(\Psi_{\text{ATP}}) + \sum_{j=1}^{n_v} \ell(\tau(\Psi_{\text{ATP}}(v_j))^\intercal W, y_j) \quad (9)$$

which corresponds to learning a data-driven temporal representation for the domain-transfer task.

4.3. Self-attention based domain-transfer

In addition to the techniques that we propose for classifier-transfer and temporal-adaptation, we define another baseline method, a self-attention [63] based formulation that aims to jointly tackle both sub-problems. Self-attention mechanism aims to capture the internal structure of a sequence by learning the inter-element relations. Below we briefly explain the way we adapt it to the domain-transfer problem, and, refer to [63] for a full specification of the original approach.

A self-attention layer consists of three transforms for computing the *key*, *query* and *value* tensors for each element. The attention weight of each element (*i.e.* face) in a sequence w.r.t. each other element is computed based on the per-element query and key embeddings, and the attention-driven representation of each element is obtained by computing the attention-weighted sum of all per-element value embeddings. In this sense, self-attention has certain similarities to ATP (Eq. 8), with two major differences: (i) while

[6] In fact, we empirically observe that max-pooling performs similar to or worse than average-pooling except when self-attention is being used. We do not report simple max-pooling results in Section 5 for brevity.

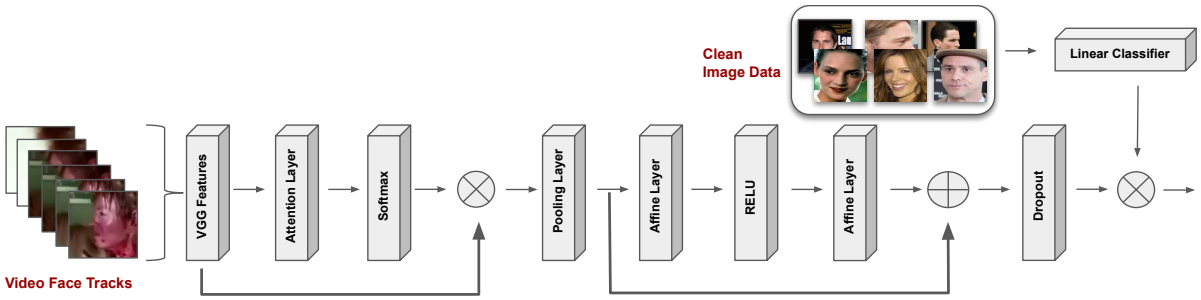


Figure 2. The architecture of the proposed domain transfer approach based on Attentive Temporal Pooling (ATP) layer for temporal-adaptation, combined with residual stacked affine classifier-transfer layer.

ATP learns K canonical attention-references, self-attention uses each element within a sequence as an attention-reference on its own, (ii) while ATP yields a weighted summation of original per-frame descriptors, self-attention additionally learns a descriptor transformation, which can be considered as the classifier-transfer layer.

In the context of our domain-transfer problem, we have observed that it is beneficial to (i) carefully tune the output dimensionality of key and query transforms on the validation set, (ii) utilize *position-wise* feed-forward network component [63], (iii) use max-pooling (instead of average-pooling) to aggregate the final per-element embeddings to a single vector. We set the value dimensionality to d_x, so that the classifier layer can be applied to the resulting video representation Ψ. We learn the parameters of the network on $\mathcal{D}_v$ pretty much the same way as in ATP training (Eq. 9).

5. Experimental Results

In this section, we present a detailed analysis of the WildestFaces dataset, the experimental setup, evaluation protocols, and, the evaluation of the proposed approaches.

5.1. Dataset analysis

In Figure 3, k-means centers of all Al Pacino images are shown for FaceScrub [43] and WildestFaces datasets. It can be seen that WildestFaces has a wide spectrum of adverse effects as its cluster centers are not recognizable. WildestFaces offers a good distribution of blur levels, pose variance, and a noticeable age variance, where approximately half of all shots are occluded. Dataset splits are summarized in Table 1. Below we present the detailed statistics.

Scale. Faces below 100 pixels are accounted as *small*, in between 100 to 300 pixels as *medium*, and larger than 300 pixels as *large*. Scale statistics given in Figure 4(a) shows that *medium* size is more common.

Blur. Inspired from [47], we perform contrast normalization and grayscale conversion. These images are convolved with a 3x3 Laplacian Kernel, and variance of the result is used to produce a blurness value. Blur values are used to empirically find a threshold to categorize images in blur levels. Blur statistics are shown in Figure 4(b).

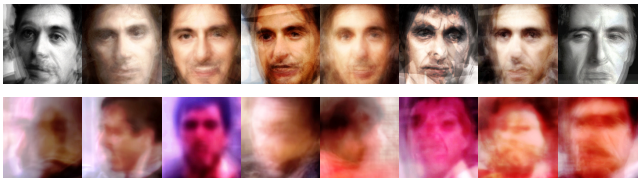


Figure 3. K-Means cluster centers (with $k = 8$ for Al Pacino images in FaceScrub [43] and *WildestFaces* datasets are shown in first and second row. Average faces from *WildestFaces* are hardly recognizable, indicating a large degree of variance in adverse effects. Images are histogram equalized for convenience. Better viewed when zoomed in.

Age. For each individual, we measure the differences between the dates of their earliest and latest movies. We observe age variations up to 40 years (Figure 4(c)), where the average variation is 13 years.

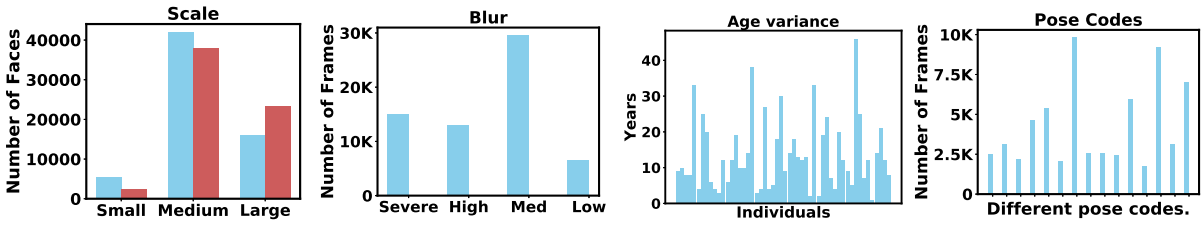
Occlusion. We randomly select 250 shots and label them according to the amount of occlusion present. We observe that 20% have *no occlusion*, 28% have *medium* and 52% have *significant* occlusion.

Pose. We use [51] to find orientations of the faces and then quantize them using k-means to find the pose codes. Figure 4(d) shows the distribution of various pose codes. We observe that pose variance is a major challenge.

5.2. Experimental setup

Supervised evaluation is not fully realistic in our case; not every individual may have a criminal record history and corresponding fight or dispute video footage(s). The only available means of identification can be clean images. Our evaluation protocol mimics this scenario, where the test set includes videos of individuals that are not seen before.

Evaluation protocol and metrics. Training, test and validation sets for WildestFaces are split person-wise, where classes with fewer per-class sample counts are selected as *unseen* classes. For the partially-supervised domain-transfer evaluation, inspired from the recently developed evaluation protocols for generalized zero-shot learning [73, 5] and generalized few-shot learning [20], we use the following protocol and metrics: the recognition model has access to the still image training examples of all 64 classes, the training videos of 40 classes, and the validation videos of additional 10 persons. At test time, an input image may



(a) Scale statistics. (b) Blur statistics. (c) Age variance. (d) Pose statistics.

Figure 4. *WildestFaces* Statistics. (a) Blue and red correspond to width and height, respectively. (b) Blur statistics indicate an emphasis on medium blur. (c) and (d) show that the pose and age variances are high.

Table 1. Dataset splits. The upper two rows and the bottom one correspond to WildestFaces and IMDB datasets, respectively.

	Train	Validation	Test
Shots	1156	495	535
Images	35051	14520	14671
Images	6428	n/a	1641

belong to any one of the 64 classes. The normalized accuracy values over the test images of seen and unseen classes are averaged separately. The final performance score is defined as the harmonic mean (test-h) of the seen and unseen class accuracies. The same procedure is also utilized to obtain validation set harmonic mean score (val-h).

Implementation details. We use VGGFace [45] for image representation. We tune all the hyper-parameters based on the val-h metric and decide whether to use dropout or not for each method separately. We use early stopping to introduce further regularization. While training, we take samples from each class with probability inversely proportional to its number of training examples to deal with class imbalance. The models are implemented in PyTorch [46], parameters are initialized using Xavier [21], and the results are averaged over 10 runs to mitigate the occasional fluctuations.

5.3. Results and discussion

5.3.1 Classifier Transfer

We evaluate the effectiveness of fc and affine transform based classifier transfer methods. We consider two main domain-adaptation baselines, based on MMD [39] and the adversarial training method of ADDA [62]. For both, we consider the fixed VGG feature extractor as the source domain mapping and aim to learn a target-to-source mapping that can transform the video representation to the source domain. We define their fc and affine transform based versions, which yields four domain adaptation baselines.

Results are shown in Table 2. Fully-connected classifier layer (fc) performs poorly, due to heavy overfitting as a result of having a large number of trainable parameters, despite our tuning efforts. While MMD-affine improves over MMD-fc, neither method improves over results w/o any transfer layer. ADDA-fc fails to converge even with one fc layer (not shown for brevity). ADDA-affine, on the other hand, proves effective and improves from 27.3 to 31.7. The proposed affine transfer further improves to 35.4. We also train MMD-affine and ADDA-affine baselines with train and validation videos, but observe only neglibible improvements despite adding examples from 10 new classes.

Table 2. Comparison of classifier-transfer methods (with one affine layer and temporal average pooling).

	seen	unseen	harmonic
Random	2.5	4.1	3.1
No transfer	29.3	25.5	27.3
Fully-connected	20.2	7.2	10.3
MMD-fc [39]	25.8	22.1	23.6
MMD-affine	28.0	23.2	25.2
ADDA-affine [62]	35.2	29.3	31.7
Affine (Ours)	**39.6**	**32.2**	**35.4**

In Table 3, we experiment with the stacked affine classifier-transfer, and residual stacked classifier-transfer layers together with AvgPool and ATP temporal adaptation techniques. As can be seen, amongst different variations, 2-layer residual stacked classifier-transfer (rsa) layer works the best. In the rest of the experiments, we continue with 2-layer rsa as the classifier-transfer method.

5.3.2 Temporal Adaptation

In Table 3, average pooling, majority voting and ATP are evaluated. Compared to vanilla AvgPool, vanilla ATP increases accuracy more than 5 points. ATP with 2-layer residual affine layer increases val-h and test-h even further, from 30.6 and 27.3 to 47.4 and 39.3, respectively. Fine-tuning the IMDB classifier with WildestFaces training set increases the accuracy by another 5 points to 45.8.

We compare our proposed method with another aggregation method DAN [48], which is amongst the state-of-the-art methods for video face recognition. DAN [48] aggregates the information of an input video into one or few discriminative image(s) by using a GAN-based approach. For each face sequence, we generate an image using DAN model pre-trained on Youtube Faces(YTF) [68] dataset. Example images generated by DAN [48] is shown in Figure 5. Inevitably, the images generated by this GAN-based model

Table 3. Comparison of temporal-adaptation techniques for three different affine classifier-transfer models. Majority-voting is an image-to-image baseline (image-level classifier transfer). † are cases where the IMDB classifier is fine-tuned with WildestFaces training set.

	None		1-layer affine		2-layer affine		2-layer res affine (rsa)	
	val-h	test-h	val-h	test-h	val-h	test-h	val-h	test-h
Maj. Voting	30.5	24.6	37.7	28.8	31.4	26.3	43.11	31.9
AvgPool	30.6	27.3	43.6	**35.4**	45.7	**35.4**	**46.6**	34.9
ATP (ours)	36.5	32.6	44.8	39.3	42.9	35.3	**47.4**	**39.3**
Maj. Voting†	38.1	30.0	39.3	30.0	37.1	30.3	46.2	34.6
AvgPool†	40.6	32.7	45.2	35.3	43.6	36.5	**48.4**	**40.5**
ATP (ours)†	43.0	35.3	45.1	36.0	49.8	42.2	**51.8**	**45.8**

Table 4. Comparison of temporal-adaptation techniques for video representation. We report the separate accuracies of seen and unseen classes, together with their harmonic mean. † represents the case that the IMDB classifier is fine-tuned with WildestFaces training set.

	w/o classifier-transfer			w/ classifier-transfer		
	seen	unseen	harmonic	seen	unseen	harmonic
AvgPool	29.3	25.5	27.3	36.7	33.2	34.9
DAN [48]	5.0	2.9	3.7	5.2	6.5	5.6
Self-attention [63]	**37.2**	**34.5**	**35.8**	37.1	34.7	35.9
ATP (ours)	35.3	30.3	32.6	41.6	37.3	39.3
ATP (ours) †	34.5	31.2	32.7	**47.1**	**44.6**	**45.8**

are not precise, due to noisy input sequences (Figure 5).

DAN [48] fails to extend to different domains and unseen classes (see Table 4). Self-attention [63] adaptation performs well yet enjoys slight improvements with classifier transfer. We argue this is due to its implicit classifier transfer mechanism (multi-head attention) as its high complexity can be harmful in our data-sparse setting. Ultimately, ATP outperforms other baselines with a clear margin.

6. Conclusion

In common surveillance scenarios, one may only have access to a clean photo of a person but may need to recognize the person in an unconstrained setting. In line with such scenarios, we study the partially-supervised domain-transfer problem within the context of face recognition, where algorithms are evaluated for their ability to recognize people in videos with violence, based on clean train images.

We introduce the WildestFaces dataset that contains adverse effects at their extreme, such as blur, pose diversity, occlusions and resolution, and a principled evaluation protocol. Towards tackling the partially-supervised domain transfer, we propose (i) affine layers for classifier transfer, and, (ii) attention-based pooling for temporal adaptation. Compared to a number of strong baselines, including a self-attention based model, we show the proposed techniques outperform the baselines. We also highlight the challenges of this newly introduced dataset and the problem definition.

Acknowledgments. This work was supported in part by the TUBITAK Grant 116E445. We thank Oguzhan Oguz for his help in collecting and annotating the dataset.

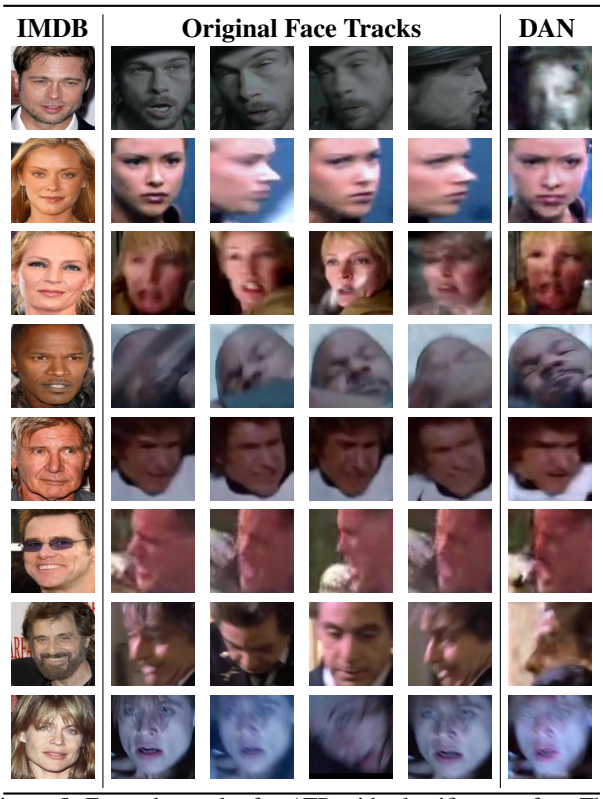


Figure 5. Example results for ATP with classifier-transfer. The first and last 4 rows depict examples for correct and incorrect classifications, respectively. The corresponding DAN [48] generated images (the rightmost column) are mostly noisy and imprecise.

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2006. 1

[2] J Ross Beveridge, P Jonathon Phillips, David S Bolme, Bruce A Draper, Geof H Givens, Yui Man Lui, Mohammad Nayeem Teli, Hao Zhang, W Todd Scruggs, Kevin W Bowyer, et al. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE International Conference on Biometrics*, pages 1–8, 2013. 2

[3] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 343–351, 2016. 3

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2

[5] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European Conference on Computer Vision*, pages 52–68. Springer, 2016. 2, 3, 6

[6] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 521–530, 2017. 3

[7] Gong Cheng, Peicheng Zhou, and Junwei Han. Duplex metric learning for image set classification. *IEEE Trans. on Image Processing*, 27(1):281–292, 2018. 1

[8] Aruni Roy Chowdhury, Tsung-Yu Lin, Subhransu Maji, and Erik Learned-Miller. One-to-many face recognition with bilinear cnns. In *IEEE Winter Conf. on Appl. of Comput. Vis.*, pages 1–9, 2016. 1

[9] Daniel Coelho de Castro and Sebastian Nowozin. From face recognition to models of identity: A bayesian approach to learning about unknown identities from unsupervised data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 745–761, 2018. 1

[10] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *CoRR*, abs/1702.05374, 2017. 3

[11] Gabriela Csurka, Boris Chidlowskii, Stéphane Clinchant, and Sophia Michel. Unsupervised domain adaptation with regularized domain instance denoising. In *European Conference on Computer Vision*, pages 458–466, 2016. 3

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1

[13] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology (TIST)*, 7(3):37, 2016. 1

[14] Gareth J Edwards, Timothy F Cootes, and Christopher J Taylor. Face recognition using active appearance models. In *European conference on computer vision*, pages 581–595, 1998. 1

[15] Claudio Ferrari, Stefano Berretti, and Alberrto Del Bimbo. Extended youtube faces: a dataset for heterogeneous open-set face identification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3408–3413. IEEE, 2018. 2

[16] Claudio Ferrari, Stefano Berretti, and Alberto Del Bimbo. Discovering identity specific activation patterns in deep descriptors for template based face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019. 5

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 3

[18] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 1358–1367, 2017. 3

[19] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613, 2016. 3

[20] Spyros Gidaris and Nikos Komodakis. Dynamic Few-Shot Visual Learning Without Forgetting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 6

[21] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 7

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2672–2680, 2014. 3

[23] Gaurav Goswami, Romil Bhardwaj, Richa Singh, and Mayank Vatsa. Mdlface: Memorability augmented deep learning for video face recognition. In *IEEE International Joint Conference on Biometrics*, pages 1–7, 2014. 1

[24] Gaurav Goswami, Mayank Vatsa, and Richa Singh. Face verification via learned representation on feature-rich video frames. *IEEE Transactions on Information Forensics and Security*, 12(7):1686–1698, 2017. 1

[25] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. *Electronic Imaging*, 2016(11):1–6, 2016. 2

[26] Tal Hassner, Iacopo Masi, Jungyeon Kim, Jongmoo Choi, Shai Harel, Prem Natarajan, and Gerard Medioni. Pooling faces: Template based face recognition with pooled face images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 59–67, 2016. 5

[27] Wei Hu, Yangyu Huang, Fan Zhang, and Ruirui Li. Noise-tolerant paradigm for training face recognition cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11887–11896, 2019. 1

[28] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 2

[29] Zhiwu Huang, Shiguang Shan, Ruiping Wang, Haihong Zhang, Shihong Lao, Alifu Kuerban, and Xilin Chen. A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Trans. on Image Processing*, 24(12):5967–5981, 2015. 2

[30] Zhiwu Huang, Ruiping Wang, Shiguang Shan, Xianqiu Li, and Xilin Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *Proc. Int. Conf. Mach. Learn.*, pages 720–729, 2015. 1

[31] Zhiwu Huang, Ruiping Wang, Luc Van Gool, Xilin Chen, et al. Cross euclidean-to-riemannian metric learning with application to face recognition from video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. 1

[32] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009*, 2(7):8, 2010. 2

[33] Nathan D Kalka, Brianna Maze, James A Duncan, Kevin OConnor, Stephen Elliott, Kaleb Hebert, Julia Bryan, and Anil K Jain. Ijb–s: Iarpa janus surveillance video benchmark. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9. IEEE, 2018. 2

[34] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4873–4882, 2016. 2

[35] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1931–1939, 2015. 2

[36] Vineet Kushwaha, Maneet Singh, Richa Singh, Mayank Vatsa, Nalini Ratha, and Rama Chellappa. Disguised faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2018. 2

[37] Haoxiang Li, Gang Hua, Zhe Lin, Jonathan Brandt, and Jianchao Yang. Probabilistic elastic matching for pose variant face verification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3499–3506, 2013. 1

[38] Haoxiang Li, Gang Hua, Xiaohui Shen, Zhe Lin, and Jonathan Brandt. Eigen-pep for video face recognition. In *Asian Conference on Computer Vision*, pages 17–33, 2014. 1

[39] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 97–105. JMLR.org, 2015. 7

[40] Mingsheng Long, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and S Yu Philip. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1076–1089, 2014. 3

[41] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016. 3

[42] Arsha Nagrani and Andrew Zisserman. From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script. In *British Machine Vision Conference*, 2017. 3

[43] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *IEEE International Conference on Image Processing*, pages 343–347, 2014. 2, 6

[44] Omkar M Parkhi, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. A compact and discriminative face track descriptor. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1693–1700, 2014. 1

[45] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *British Machine Vision Conference*, 2015. 1, 2, 4, 7

[46] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch, 2017. 7

[47] José Luis Pech-Pacheco, Gabriel Cristóbal, Jesús Chamorro-Martinez, and Joaquín Fernández-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *IAPR International Conference on Pattern Recognition*, volume 3, pages 314–317, 2000. 6

[48] Yongming Rao, Ji Lin, Jiwen Lu, and Jie Zhou. Learning discriminative aggregation network for video-based face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3781–3790, 2017. 1, 2, 7, 8

[49] Yongming Rao, Jiwen Lu, and Jie Zhou. Attention-aware deep reinforcement learning for video face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3931–3940, 2017. 1, 3

[50] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. on Computer Vision*, July 2016. 3

[51] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018. 6

[52] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 815–823, 2015. 1, 2

[53] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016. 2

[54] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016. 3

[55] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1988–1996, 2014. 1

[56] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 1

[57] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Hybrid deep learning for face verification. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 1489–1496, 2013. 1, 2

[58] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1701–1708, 2014. 1, 2

[59] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3

[60] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–591, 1991. 1

[61] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 4068–4076, 2015. 3

[62] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 7

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 5998–6008, 2017. 2, 4, 5, 6, 8

[64] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 1

[65] Yifei Wang, Wen Li, Dengxin Dai, and Luc Van Gool. Deep domain adaptation by geodesic distance minimization. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017. 3

[66] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016. 1

[67] Laurenz Wiskott, Norbert Krüger, N Kuiger, and Christoph Von Der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):775–779, 1997. 1

[68] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 529–534, 2011. 2, 7

[69] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009. 1

[70] Chunpeng Wu, Wei Wen, Tariq Afzal, Yongmei Zhang, Yiran Chen, and Hai Li. A compact DNN: approaching googlenet-level accuracy of classification and domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3

[71] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 3

[72] Markus Wulfmeier, Alex Bewley, and Ingmar Posner. Addressing appearance change in outdoor robotics with adversarial domain adaptation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017. 3

[73] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4582–4591, 2017. 2, 3, 6

[74] Shufu Xie, Shiguang Shan, Xilin Chen, and Jie Chen. Fusing local patterns of gabor magnitude and phase for face recognition. *IEEE Trans. on Image Processing*, 19(5):1349–1361, 2010. 1

[75] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. In *British Machine Vision Conference*, 2018. 3

[76] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, volume 3, 2017. 3

[77] Junjie Yan, Xuzong Zhang, Zhen Lei, and Stan Z Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014. 2

[78] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4362–4371, 2017. 1, 2, 5

[79] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5525–5533, 2016. 2

[80] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5713, 2019. 1

[81] Xiaohang Zhan, Ziwei Liu, Junjie Yan, Dahua Lin, and Chen Change Loy. Consensus-driven propagation in massive unlabeled data for face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 568–583, 2018. 1

[82] Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving

profile face synthesis. In *Advances in Neural Information Processing Systems*, pages 66–76, 2017. 1, 3

[83] Yutong Zheng, Dipan K Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5089–5097, 2018. 1

[84] Yujie Zhong, Relja Arandjelović, and Andrew Zisserman. Ghostvlad for set-based face recognition. In *Asian Conf. on Computer Vision*, 2018. 3

[85] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012. 2