Boosting Unconstrained Face Recognition with Auxiliary Unlabeled Data

Yichun Shi Anil K. Jain

Michigan State University

shiyichu@msu.edu, jain@cse.msu.edu

Abstract

In recent years, significant progress has been made in face recognition, which can be partially attributed to the availability of large-scale labeled face datasets. However, since the faces in these datasets usually contain limited degree and types of variation, the resulting trained models generalize poorly to more realistic unconstrained face datasets. While collecting labeled faces with larger variations could be helpful, it is practically infeasible due to privacy and labor cost. In comparison, it is easier to acquire a large number of unlabeled faces from different domains, which could be used to regularize the learning of face representations. We present an approach to use such unlabeled faces to learn generalizable face representations, where we assume neither the access to identity labels nor domain labels for unlabeled images. Experimental results on unconstrained datasets show that a small amount of unlabeled data with sufficient diversity can (i) lead to an appreciable gain in recognition performance and (ii) outperform the supervised baseline when combined with less than half of the labeled data. Compared with the state-of-the-art face recognition methods, our method further improves their performance on challenging benchmarks, such as IJB-B, IJB-C and IJB-S.

1. Introduction

Machine learning algorithms typically assumes that training and testing data come from the same underlying distribution. However, in practice, we would often encounter testing domains that are different from the population where the training data is drawn. Since it is non-trivial to collect data for all possible testing domains, learning representations that are generalizable to heterogeneous testing data is desired [30, 8, 29, 24, 3]. Particularly for face recognition, this problem is reflected by the domain gap between the semi-constrained training datasets and unconstrained testing datasets. Nearly all of the state-of-the-art deep face networks are trained on large-scale web-crawled face images, most of which are high-quality celebrity photos [52, 11]. But in practice, we wish to deploy the trained FR systems for many other scenarios, e.g. unconstrained photos [21, 28] and surveillance [17]. The large degree of face variation in the testing scenarios, compared to the training set, could result in significant performance drop of the trained face models [28, 17].

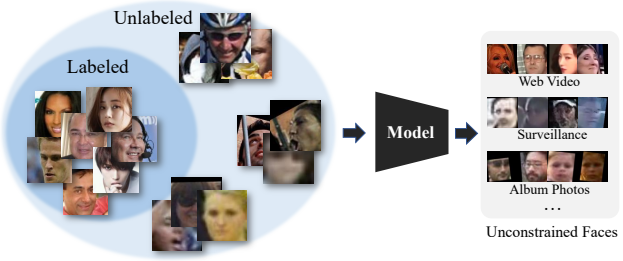


Figure 1: Illustration of the problem settings in our work. Blue circles imply the domains that training face images belong to. By utilizing diverse unlabeled images, we want to regularize the learning of the face embedding for more unconstrained face recognition scenarios.

The simplest solution to such a domain gap problem is to collect a large number of unconstrained labeled face images from different sources. However, due to privacy issue and human-labeling cost, it is extremely hard to collect such a database. Other popular solutions to this problem include transfer learning and domain adaptation, which require domain-specific data to train a model for each of the target domains [31, 7, 26, 40, 35, 19]. However, in unconstrained face recognition, a face representation that is robust to all different kinds of variations is needed, so these domain-specific solutions are not appropriate. *Instead, it would be useful if we could utilize the commonly available, unlabeled data to achieve a domain-agnostic face representation that generalizes to unconstrained testing scenarios* (See Fig. 1). To achieve this goal, we would like to ask the following questions in this paper:

- Is it possible to improve model generalizability to unconstrained faces by introducing more diversity from auxiliary unlabeled data?

- What kind of and how much unlabeled data do we need?
- How much performance boost could we achieve with the unlabeled data?

In this paper, we propose such an semi-supervised framework for learning robust face representations. The unlabeled images are collected from a public face detection dataset, i.e. WiderFace [51], which contains more diverse types (sub-domains) of face images compared to typical labeled face datasets used for training. To utilize the unlabeled data, the proposed method jointly regularizes the embedding model from feature space and image space. We show that adversarial regularization can help to reduce domain gaps caused by facial variations, even in the absence of sub-domain labels. On the other hand, an image augmentation module is trained to discover the hidden sub-domain styles in the unlabeled data and apply them to the labeled training samples, thus increasing the discrimination power on difficult face examples. To our knowledge, this is the first study to use a heterogeneous unlabeled dataset to boost the model performance for general unconstrained face recognition. The contributions of this paper are summarized as below:

- A semi-supervised learning framework for generalizing face representations with auxiliary unlabeled data.
- An multi-mode image translation module is proposed to perform data-driven augmentation and increase the diversity of the labeled training samples.
- Empirical results show that the regularization of unlabeled data helps to improve the recognition performance on challenging testing datasets, e.g. IJB-B, IJB-C, and IJB-S.

2. Related Work

2.1. Deep Face Recognition

Deep neural networks are widely adopted in the ongoing research in face recognition [43, 41, 36, 27, 25, 12, 33, 46, 5]. Taigman et al. [43] were the first to propose using deep convolutional neural network for learning face representations. The subsequent studies have explored different loss functions to improve the discrimination power of the learned feature representation. A number of studies proposed to use metric learning methods for face recognition [36, 38]. Recent work has been trying to achieve discriminative embeddings with a single identification loss function where proxy/prototype vectors are used to represent each class in the embedding space [25, 46, 47, 33, 5, 54, 42].

2.2. Semi-supervised Learning

Classic semi-supervised learning involves a small number of labeled images and a large number of unlabeled images [23, 34, 22, 44, 49, 53, 1, 39]. The goal is to improve the recognition performance when we don't have sufficient data that are labeled. State-of-the-art semi-supervised learning methods can mainly be classified into four categories. (1) Pseudo-labeling methods generate labels for unlabeled data with the trained model and then use them for training [23]. In spite of its simplicity, it has been shown to be effective primarily for classification tasks where labeled data and unlabeled data share the same label space. (2) Temporal ensemble models maintain different versions of model parameters to serve as teacher models for the current model [22, 44]. (3) Consistency-regularization methods apply certain types of augmentation to the unlabeled data while making sure the output prediction remains consistent after augmentation [34, 1, 39]. (4) Self-supervised learning, originally proposed for unsupervised learning, has recently been shown to be effective for semi-supervised learning as well [53]. Compared with classic semi-supervised learning addressed in the literature, our problem is different in two sense of heterogeneity: different domains and different identities between the labeled and unlabeled data. These differences make many classic semi-supervised learning methods unsuitable for our task.

2.3. Domain Adaptation and Generalization

In domain adaptation, the user has a dataset for a source domain and another for a fixed target domain [31, 7, 26, 35, 19]. If the target domain is unlabeled, this leads to an *unsupervised domain adaption* setting [7, 45, 35, 19]. The goal is to improve the performance on the target domain so that it could match the performance on the source domain. This is achieved by reducing the domain gap between the two datasets in feature space. The problem about domain adaption is that one needs to acquire a new dataset and train a new model whenever there is a new target domain. In *domain generalization*, the user is given a set of labeled datasets from different domains. The model is jointly trained on these datasets so that it could better generalize to unseen domains [30, 8, 29, 24, 3, 10]. Our problem lies in the middle between domain generalization and unsupervised domain adaptation : we want to generalize the model to broader domains, yet instead of multi-domain labeled data, we use unlabeled data from other sources to achieve this goal.

3. Methodology

Generally, in face representation learning, we are given a large labeled dataset $\mathcal{X}$=$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x_i and y_i are the face images and identity labels, respectively. The goal is to learn an embedding model f such that $f(x)$ would be discriminative enough to distinguish between different identities. However, since f is only trained on the domain defined by $\mathcal{X}$, which is usually semi-constrained celebrity photo, it might not gener-

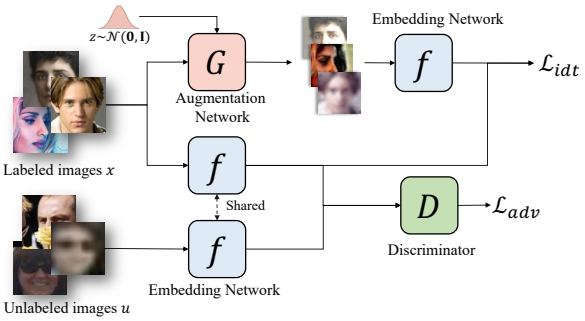
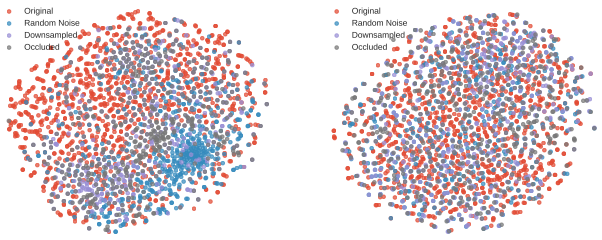


Figure 2: The training framework of the embedding network. In each mini-batch, a random subset of labeled data would be augmented by the augmentation network to introduce additional diversity. The non-augmented labeled data are used to train the feature discriminator. The adversarial loss forces the distribution of the unlabeled features to align with the labeled one.



(a) w/o Domain Adversarial Loss (b) w/ Domain Adversarial Loss

Figure 3: t-SNE visualization of the face embeddings using synthesized unlabeled images. Using part of the MS-Celeb-1M as unlabeled dataset, we create three sub domains by processing the images with either random Gaussian noise, random occlusion or downsampling. (a) different sub-domains show different domain shift in the embedding space of the supervised baseline. (b) with the holistic binary domain adversarial loss, each of the sub-domains is aligned with the distribution of the labeled data.

alize to unconstrained settings. In our framework, we assume the availability of another unlabeled dataset $\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2 \ldots \mathcal{U}_k = \{u_1, u_2, \ldots, u_n\}$, collected from different sources (sub-domains). However, these sub-domain labels may not be available in real applications, thus we do not assume the access to them but instead seek solutions that could automatically leverage these hidden sub-domains. Then, we wish to simultaneously minimize three types of errors:

- Error due to discrimination power within the labeled domain $\mathcal{X}$.
- Error due to feature domain gap between the labeled domain $\mathcal{X}$ and the hidden sub-domains $\mathcal{U}_i$.
- Error due to discrimination power within the unlabeled domain $\mathcal{U}$.

An overview of the framework is shown in Fig. 2.

3.1. Minimizing Error in the Labeled Domain

The deep representation of a face image is usually a point in a hyper-spherical embedding space, where $\|f(x_i)\|^2 = 1$. State-of-the-art supervised face recognition methods all try to find an objective function to maximize the inter-class margin such that the representation could still be discriminative when tested on unseen identities. In this work, we choose to use CosFace loss function [47][46] for training the labeled images:

$$\mathcal{L}_{idt} = -\mathbb{E}_{x_i, y_i \sim \mathcal{X}}[\log \frac{e^{s(W_{y_i}^T f_i - m)}}{e^{s(W_{y_i}^T f_i - m)} + \sum_{j \neq y_i} e^{sW_{y_j}^T f_i}}]. \quad (1)$$

Here s is the hyper-parameter controlling temperature, m is a margin hyper-parameter and W_j is the proxy vector of the j^{th} identity in the embedding space, which is also ℓ_2 normalized. We choose to use CosFace loss function because of its stability and high-performance. It could potentially be replaced by any other supervised identification loss function.

3.2. Minimizing Domain Gaps

The unlabeled dataset $\mathcal{U}$ is assumed to be a diverse dataset collected from different sources, i.e. covering different sub-domains (types) of face images. If we have the access to such sub-domain labels, a natural solution to a domain-agnostic model would be aligning each of the sub-domains with the feature distribution of the labeled images. However, the sub-domain labels might not be available in many cases. In our experiment, we find there is no necessity for pairwise domain alignment. Instead, a binary domain alignment loss is sufficient to align the sub-domains. Formally, given a feature discriminator network D, we could reduce the domain gap via an adversarial loss:

$$\begin{aligned} \mathcal{L}_D = &-\mathbb{E}_{x \sim \mathcal{X}}[\log D(y=0|f(x)] \\ &-\mathbb{E}_{u \sim \mathcal{U}}[\log D(y=1|f(u)], \end{aligned} \quad (2)$$

$$\begin{aligned} \mathcal{L}_{adv} = &-\mathbb{E}_{x \sim \mathcal{X}}[\log D(y=1|f(x)] \\ &-\mathbb{E}_{u \sim \mathcal{U}}[\log D(y=0|f(u)]. \end{aligned} \quad (3)$$

The discriminator D is a multi-layer binary classifier optimized by $\mathcal{L}_D$. It tries to learn a non-linear classification boundary between the two datasets while the embedding network needs to fool the discriminator by reducing the divergence between the distributions of $f(x)$ and $f(u)$. To see the effect of domain alignment loss, we conduct a controlled experiments with a toy dataset. We split the MS-Celeb-1M [11] dataset into labeled images and unlabeled images (no identity overlap). The unlabeled images are then processed with one of the three degradations: random Gaussian noise, random occlusion and downsampling. Thus, we

create three sub-domains in the unlabeled dataset. The corresponding domain shift can be observed in the t-SNE plot in Fig. 3 (a), where the model is trained only on the labeled split. Then, we incorporate the augmented unlabeled images into training with the binary domain adversarial loss. In Fig. 3 (b), we observe that with the binary domain alignment loss, the distribution of each of sub-domains is aligned with the original domain, indicating reduced domain gaps.

3.3. Minimizing Error in the Unlabeled Domains

The domain alignment loss in Section 3.2 helps to eliminate the error caused by domain gaps between unconstrained faces. Thus, the remaining task is to improve the discrimination power of the face representation among the unlabeled faces. Many semi-supervised classification methods address this problem by using pseudo-labeling of unlabeled data [23, 1, 39], but this is not applicable to our problem since our unlabeled dataset does not share the same label space with the labeled one. Furthermore, because of data collection protocols, there is very little chance that one identity would have multiple unlabeled images. Thus, clustering-based methods are also infeasible for our task. Here, we consider to address this issue with a multi-mode augmentation method. Prior studies have shown that an image translation network, such as CycleGAN [55], can be effectively used as a data augmentation module for domain adaptation [13]. The main idea of the augmentation network is to learn the difference between two domains in the image space and then augment the samples from source domain data to create training data with pseudo-labels in the target domain. Since our goal is to generalize the deep face representation to unconstrained faces, which involves a large variety, deterministic method such as CycleGAN would be unsuitable. Therefore, we propose to use a multi-mode image translation network that could discover the hidden domains in the unlabeled data and then augment the labeled training data with different styles. In particular, we need a function G which maps labeled samples x into the image space defined by the unlabeled faces, i.e. $p(x) \rightarrow p(u)$. Then, training the embedding f on $G(x)$ could make it more discriminative in the image space defined by U. There are two requirement of the function G: (1) it should not change the identity of the input image and (2) it should be able to capture different styles that are present in the unlabeled images. Inspired by recent progress in image translation frameworks [55, 16], we propose to train G as a style-transfer network that learns the visual styles during transfer in an unsupervised manner. The network G can then be used as a data-driven augmentation module that generates diverse samples given an input from the labeled dataset. During the training, we randomly replace a subset of the labeled images to be augmented and put them into our identification learning framework. The details of training the augmentation network G is given in Section 3.3.1.

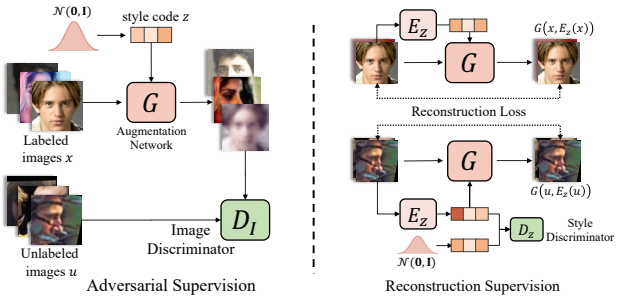


Figure 4: Training framework of the augmentation network G. The two pipelines are optimized jointly during training.

The overall loss function for the embedding network is given by:

$$\mathcal{L} = \lambda_{idt}\mathcal{L}_{idt} + \lambda_{adv}\mathcal{L}_{adv} \tag{4}$$

where L_{idt} also includes the augmented labeled samples.

3.3.1 Multi-mode Augmentation Network

The augmentation network G is a fully convolutional network that maps an image to another. To preserve the geometric structure, our architecture does not involve any downsampling or upsampling. In order to generate styles similar to the unlabeled images, an image discriminator D_I is trained to distinguish between the texture styles of unlabeled images and generated images:

$$\begin{aligned}\mathcal{L}_{D_I} = &- \mathbb{E}_{x\sim\mathcal{X}}[\log D_I(y=0|G(x,z))] \\ &- \mathbb{E}_{u\sim\mathcal{U}}[\log D_I(y=1|u)],\end{aligned} \tag{5}$$

$$\mathcal{L}^G_{adv} = - \mathbb{E}_{x\sim\mathcal{X}}[\log D_I(y=1|G(x,z))]. \tag{6}$$

Here $z \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ is a random style vector to control the styles of the output image, which is injected into the generation process via Adaptive Instance Normalization (AdaIN) [15]. Although the adversarial learning could make sure the output are in the unlabeled space, but it cannot ensure that (1) the content of the input is maintained in the output image and (2) the random style z is being used to generate diverse visual styles, corresponding to differnt sub-domains in the unlabeled images. We propose to utilize an additional reconstruction pipeline to simultaneously satisfy these two requirements. First, we introduce an additional style encoder E_z to capture the corresponding style in the input image, as in [16]. A reconstruction loss is then enforced to keep the consistency of the image content:

$$\mathcal{L}^G_{rec} = \mathbb{E}_{x\sim\mathcal{X}}[\|x - G(x, E_z(x))\|^2] \tag{7}$$

$$+ \mathbb{E}_{u\sim\mathcal{U}}[\|u - G(u, E_z(u))\|^2], \tag{8}$$

Then, during the reconstruction, we add another latent style discriminator D_z to guarantee the distribution of $E_z(u)$

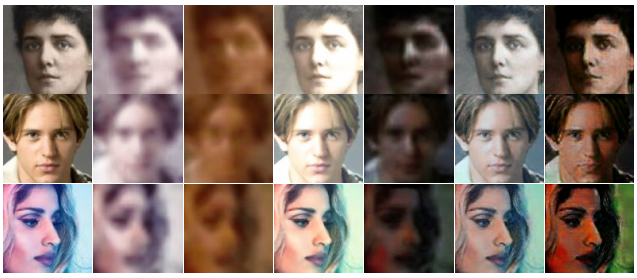


Figure 5: Example generated images of the augmentation network. Each row shows augmented images with different styles for the input in the first column.

align with prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\begin{aligned}\mathcal{L}_{D_z} = &- \mathbb{E}_{u\sim\mathcal{U}}[\log D_z(y=0|E_z(u))] \\ &- \mathbb{E}_{z\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\log D_z(y=1|z)],\end{aligned} \tag{9}$$

$$\mathcal{L}^z_{adv} = - \mathbb{E}_{u\sim\mathcal{U}}[\log D_z(y=1|E_z(u))], \tag{10}$$

The overall loss function of the generator is given by:

$$\mathcal{L}^G = \lambda^G_{adv}\mathcal{L}^G_{adv} + \lambda^G_{rec}\mathcal{L}^G_{rec} + \lambda^z_{adv}\mathcal{L}^z_{adv} \tag{11}$$

A overview of the training framework of G is given in Fig. 4 and example generated images are shown in Fig. 5. The architecture details of different modules are given in the supplementary file.

4. Experiments

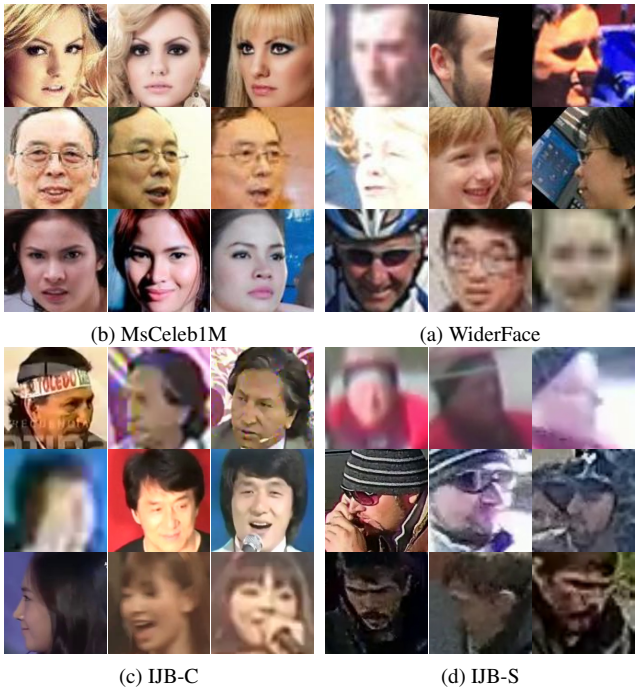
4.1. Implementation Details

Training Details of the Recognition Models All the models are implemented with Pytorch v1.1. We use the RetinaFace [6] for face detection and alignment. All images are transformed into 112×112 pixels. A modified 50-layer ResNet in [5] is used as our architecture. The embedding size is 512 for all models. By default, all the models are trained with $150,000$ steps with a batch size of 256. For semi-supervised models, we use 64 unlabeled images and 192 labeled images in each mini-batch. For models which use the augmentation module, 20% of the labeled images are augmented by the generator network. The scale parameter s and margin parameter m are set to 30 and 0.5, respectively. We empirically set λ_{idt}, λ_{adv} as 1.0 and 0.01.

Training Details of the Generator Models The generator is trained for $160,000$ steps with a batch size of 8 images (4 from each dataset). Adam optimizer is used with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. The learning rate starts with $1e-4$ and drops to $1e-5$ after $80,000$ steps. The detailed architectures are provided in the supplementary material. λ^G_{adv}, λ^G_{rec} and λ^z_{adv} are set to as 1.0, 10.0 and 1.0, respectively.

4.2. Datasets

We use **MS-Celeb-1M** [11] as our labeled training dataset. MS-Celeb-1M is a large-scale public face dataset of celebrity photos. The original dataset is known to contain a large number of noisy labels [2], so we use a cleaned version from ArcFace [5] as training data. After removing the overlapped subjects with the testing sets and duplicate images, we are left with 3.9M images of 85.7K classes. As for unlabeled images, we choose **WiderFace** [51] as our auxiliary training data. WiderFace is a dataset collected by retrieving images from search engines with different event keywords. As a face detection dataset, WiderFace includes a more diverse set of faces (See Fig. 6). Many faces in this dataset still cannot be detected by state-of-the-art detection methods [6]. Thus, we only keep the detectable faces in the WiderFace training set as our training data. Our goal is to close the gap between face detection and recognition engine and improve the general recognition performance for any detectable faces. At the end, we were able to detect about 70K faces from WiderFace, less than 2% of our labeled training data.



(b) MsCeleb1M (a) WiderFace

(c) IJB-C (d) IJB-S

Figure 6: Examples face images of different datasets used in this work. The unconstrained testing datasets (IJB-C, IJB-S) are significantly different from the labeled training data (MsCeleb) with more types and degree of variation. Thus, auxiliary data from WiderFace are utilized to generalize the model.

To evaluate the proposed method, we test on three unconstrained face recognition benchmarks, namely IJB-B, IJB-C and IJB-S. These datasets represent real-world testing scenarios where faces are significantly different from celebrity photos in the training set. The details of these datasets are as follows:

- **IJB-B** [48] includes both high quality celebrity photos taken from the wild and low quality photos or video

frames with large variations of illumination, occlusion, head pose, etc. There are 68,195 images of 1,845 identities in all. We test on both verification and identification protocols of the IJB-B benchmark.

- **IJB-C** [28] is a newer version of IJB-B dataset. It has a similar protocol but with 140,732 images of 3,531 identities.
- **IJB-S** [17] is an extremely challenging benchmark where the images were collected from surveillance cameras. There are in all 202 identities with an average of 12 videos per person. Each person also has 7 high-quality enrollment photos (with different poses) which constitute the gallery. We test on two protocols of the IJB-S dataset, Surveillance-to-Still (**V2S**) and Surveillance-to-Booking (**V2B**), both of which are identification protocols. The difference between them is that in Surveillance-to-Still (V2S) the gallery of each person is a single frontal photo while Surveillance-to-Booking (V2B) uses all 7 registration photos as gallery. To Reduce the evaluation time, the expriments in Sec. 4.3 and Sec. 4.4 are conducted with subsampled frames from each video, whose performance is close to using the whole video (Sec. 5.1).

Although our goal is to improve the recognition performance on unconstrained faces, we would not like to lose the discrimination power in the original domain (high-quality photos). Therefore, during ablation we also evaluate our models on the standard **LFW** [14] protocol, which is a celebrity photo dataset, similar to the labeled training data (MS-Celeb-1M). Note that the accuracy on the LFW protocol is highly saturated, so the main goal is just to check whether there is a significant performance drop on the constrained faces while increasing the generalizability to unconstrained ones. Example images of different datasets are shown in Figure 6.

4.3. Ablation Study

In this section, we conduct an ablation study to quantitatively evaluate the effect of different modules proposed in this paper. In particular, we have two modules to study: Domain Alignment (DA) and Augmentation Network (AN). The performance is shown in Table 1. As we already showed in Fig. 3, domain adversarial loss is able to force smaller domain gaps between the sub-domains in WiderFace and the celebrity faces, even though we do not have access to those domain labels. Consequently, we observe the performance improvement on most of the protocols on IJB-C and IJB-S. Introducing the augmentation network (AN) further helps improving the performance on unconstrained benchmarks, where a multi-mode (MM) augmentation network outperforms a single-model (SM) augmentation network. More details of ablating over the augmentation network can be found in the supplementatry material.

Method	IJB-C (Vrf)			IJB-C (Idt)		IJB-S (V2S)		LFW
	1e-7	1e-6	1e-5	Rank1	Rank5	Rank1	Rank5	Accuracy
Baseline	62.90	82.94	90.73	94.90	96.77	53.23	62.91	99.80
+ DA	72.74	85.33	90.52	94.99	96.75	56.35	**66.77**	**99.82**
+ DA + AN (SM)	74.80	87.58	**91.94**	95.51	97.09	56.98	65.66	99.80
+ DA + AN (MM)	**77.39**	**87.92**	91.86	**95.61**	**97.13**	**57.33**	65.37	99.75

Table 1: Ablation study over different training methods of the embedding network. All models has identification loss by default. "DA", "AN", "SM" and "MM" refer to "Domain Alignment", "Augmentation Network", "Single-mode" and "Multi-mode", respectively.

4.4. Quantity vs. Diversity

Although we have shown in Sec. 4.3 that utilizing unlabeled data leads to better performance on challenging testing benchmarks, generally it shall be expected that simply increasing the number of labeled training data can also have a similar effect. Therefore, in this section, we conduct a more detailed study to answer such a question: *which is more important for feature generalizability: quantity or diversity of the training data?* In particular, we train several supervised models by adjusting the number of labeled training data. For each such model, we also train a corresponding model with additional unlabeled data. The evaluation results are shown in Figure 7.

On the IJB-S dataset, which is significantly different from the labeled training data, we see that the models trained with unlabeled data consistently outperform the supervised baselines with a large margin. In particular, the proposed method achieves better performance than the supervised baseline even when there is only one-fourth of the overall labeled training data (1M vs 4M), indicating the value of data diversity during training. Note that there is a significant performance boost when increasing the number of labeled samples from 0.5M to 1M. However, after that, the benefit of acquiring more labeled data plateaus and in fact it is more helpful to introduce 70K unlabeled data than 3M additional labeled data.

On the IJB-C dataset, for both verification and identification protocols, we observe a similar trend as the IJB-S dataset. In particular, a larger improvement is achieved at lower FARs. This is because the verification threshold at lower FARs is affected by the low quality test data (difficult impostor pairs), which is more similar to our unlabeled data. Another interesting observation is that the improvement margin increases when there is more labeled data. Note that in general semi-supervised learning, we would expect less improvement by using unlabeled data when there is more labeled data. But it is the opposite in our case because the unlabeled data has different characteristics than the labeled data. So when the performance of supervised model saturates with sufficient labeled data, transferring the knowledge from diverse unlabeled data becomes more helpful.

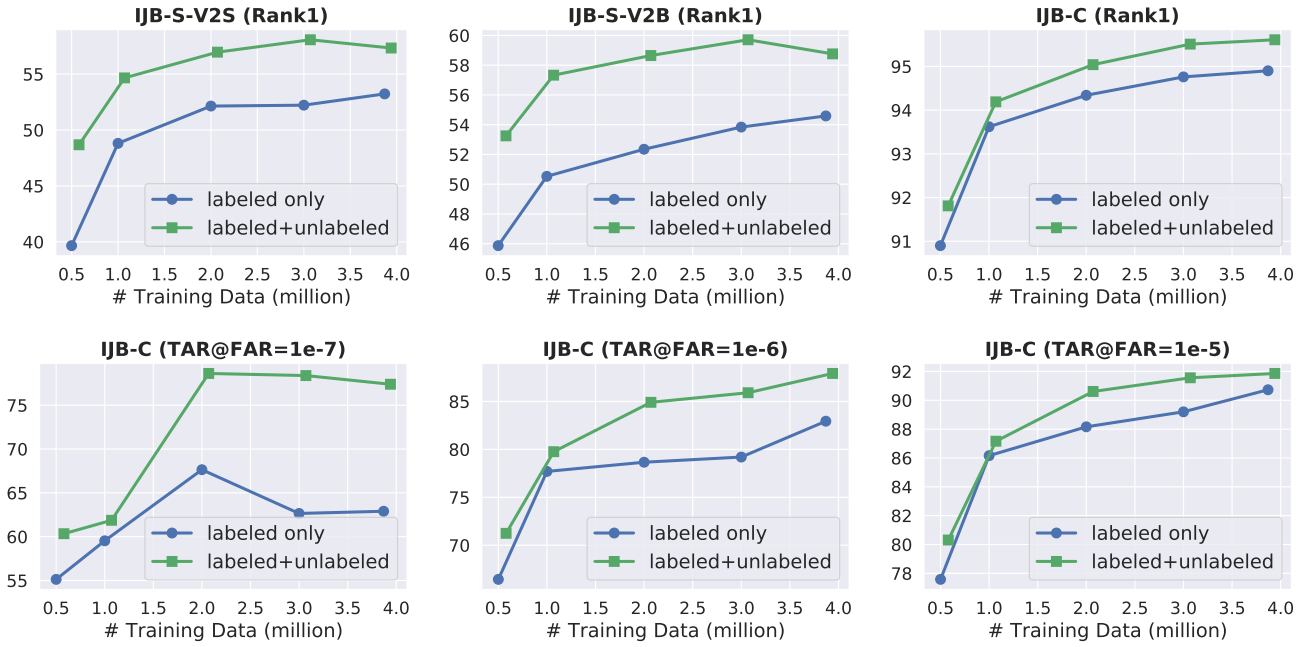


Figure 7: Evaluation Results on IJB-C and IJB-S with different protocols and different number of labeled training data.

For both IJB-S and IJB-C (TAR@FAR=1e-7), we observe that after a certain point, adding more labeled data does not boost performance any more and the performance starts to fluctuate. This happens because the new labeled data does not necessarily help with those hard cases. Based on these results, we conclude that *when the number of labeled training data is small, it is more important to increase the quantity of the labeled dataset. Once there is sufficient labeled training data, the generalizablity of the representation tends to saturate while the diversity of the training data becomes more important*. Additional experimental results on the choice of the unlabeled dataset can be found in Section 5.

5. Choice of the Unlabeled Dataset

In Section 4.4, we discussed on the impact of the quantity/diversity of training data on feature generalizability. A remaining question is how the choice and amount of unlabeled data would affect the performance. Here, we show additional experiments on different choices of unlabeled dataset. In addition to the WiderFace, we consider to utilize two other datasets: MegaFace [20] and CASIA-WebFace [52]. For MegaFace, we only use the distractor images in their identification protocol, which are crawled from album photos on Flicker and present a larger degree of variation compared with the faces in MS-Celeb-1M. CASIA-WebFace, similar to MS-Celeb-1M, is mainly composed of celebrity photos, and therefore it should not introduce much additional diversity. Note that CASIA-

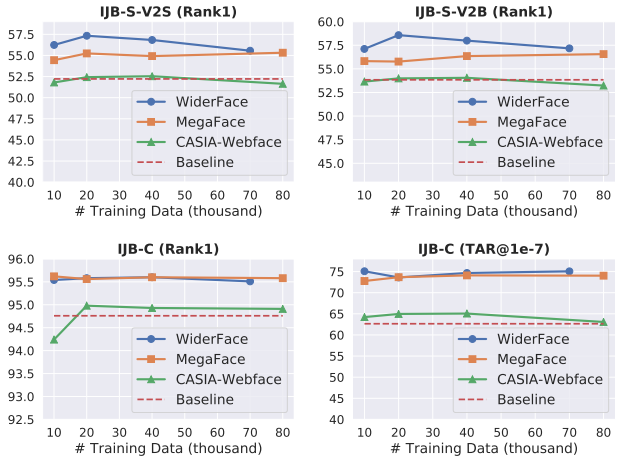


Figure 8: Evaluation Results on IJB-S and IJB-C with different protocols and different number and choice of unlabeled training data. The red line here refers the performance of the supervised baseline which does not use any unlabeled data.

WebFace is a labeled dataset but we ignore its labels for this experiment. The diversity (facial variation) of the three datasets can be ranked as: *WiderFace* > *MegaFace* > *CASIA-WebFace*. Example images of the three datasets are shown in Figure 6. For both MegaFace and CASIA-Webface, we choose a random subset to match the number of the WiderFace. Furthermore, to see the impact of the quantity of unlabeled dataset, we also train the models with different numbers of unlabeled data. Then, we evaluate all the models on IJB-S, IJB-C and LFW. The reason to evalu-

Method	Data	Model	Verification				Identification	
			1e-7	1e-6	1e-5	1e-4	Rank1	Rank5
Cao et al. [2]	13.3M	SE-ResNet-50	-	-	76.8	86.2	91.4	95.1
PFE [37]	4.4M	ResNet-64	-	-	89.64	93.25	95.49	97.17
ArcFace [5]	5.8M	ResNet-50	67.40	80.52	88.36	92.52	93.26	95.33
Ranjan et al. [32]	5.6M	ResNet-101	67.4	76.4	86.2	91.9	94.6	97.5
AFRN [18]	3.1M	ResNet-101	-	-	88.3	93.0	**95.7**	**97.6**
DUL [4]	3.6M	ResNet-64	-	-	90.23	94.2	**95.7**	**97.6**
Baseline	3.9M	ResNet-50	62.90	82.94	90.73	94.57	94.90	96.77
Proposed	4.0M	ResNet-50	**77.39**	**87.92**	**91.86**	**94.66**	95.61	97.13

Table 2: Performance comparison with state-of-the-art methods on the IJB-C dataset.

Method	Data	Model	Verification				Identification	
			1e-6	1e-5	1e-4	1e-3	Rank1	Rank5
Cao et al. [2]	13.3M	SE-ResNet-50	-	70.5	83.1	90.8	90.2	94.6
Comparator [50]	3.3M	ResNet-50	-	-	84.9	93.7	-	-
ArcFace [5]	5.8M	ResNet-50	40.77	84.28	91.66	94.81	92.95	95.60
Ranjan et al. [32]	5.6M	ResNet-101	**48.4**	80.4	89.8	94.4	93.3	96.6
AFRN [18]	3.1M	ResNet-101	-	77.1	88.5	94.9	**97.3**	**97.6**
Baseline	3.9M	ResNet-50	40.12	84.38	**92.79**	**95.90**	93.85	96.55
Proposed	4.0M	ResNet-50	43.38	**88.19**	92.78	95.86	94.62	96.72

Table 3: Performance comparison with state-of-the-art methods on the IJB-B dataset.

ate on LFW here is to see the impact of different unlabeled datasets on the performance in the original domain. The results are shown in Fig. 8. Note that due to the large number of experiments, we do not use augmentation network here. But empirically we found the trends are similar with the data augmentation network.

From Fig. 8, it can be seen that in general, the more diverse the unlabeled dataset is, the more performance boost it leads to. In particular, using CASIA-WebFace as the unlabeled dataset hardly improves performance on any protocol. This is expected because CASIA-WebFace is very similar to MS-Celeb-1M and hence it cannot introduce additional diversity to regularize the training of face representations. Using MegaFace distractors as the unlabeled dataset improves the performance on both IJB-C and IJB-S, both of which have more variations than the MS-Celeb-1M. Using WiderFace as the unlabeled dataset further improves the performance on the IJB-S dataset. Note that all the models in this experiment maintain the high performance on the LFW dataset. In other words, *using a more diverse unlabeled dataset would not impair the performance on the original domain and safely improves the performance on the challenging new domains*. An additional result that we can observe is that the size of the unlabeled dataset does not have a clear effect compared to its diversity.

5.1. Comparison with State-of-the-Art FR Methods

In Table 2 we show more complete results on IJB-C dataset and compare our method with other state-of-the-art methods. In generally, we observe that with fewer labeled training samples and number of parameters, we are able to achieve state-of-the-art performance on most of the protocols. Particularly at low FARs, the proposed method outperforms the baseline methods with a good margin. This is because at a low FAR, the verification threshold is mainly determined by low quality impostor pairs, which are instances of the difficult face samples that we are targeting with additional unlabeled data. Similar trend is observed for IJB-B dataset (Table 3). Note that because of fewer number of face pairs, we are only able to test at higher FARs for IJB-B dataset.

Method	Surveillance-to-Still					Surveillance-to-Booking				
	Rank1	Rank5	Rank10	1%	10%	Rank1	Rank5	Rank10	1%	10%
MARN [9]	58.14	64.11	-	21.47	-	59.26	65.93	-	32.07	-
PFE [37]	50.16	58.33	62.28	31.88	35.33	53.60	61.75	62.97	35.99	39.82
ArcFace[5]	50.39	60.42	64.74	32.39	42.99	52.25	61.19	65.63	34.87	43.50
Baseline	53.23	62.91	67.83	31.88	43.32	54.26	64.18	69.26	32.39	44.32
Proposed	**59.29**	**66.91**	**69.63**	**39.92**	**50.49**	**60.58**	**67.70**	**70.63**	**40.80**	**50.31**

Table 4: Performance on the IJB-S benchmark.

In Table 4 we show the results on two different protocols of IJB-S. Both the Surveillance-to-Still (V2S) and Surveillance-to-Booking (V2B) protocols use surveillance videos as probes and mugshots as gallery. Therefore, IJB-S results represent a cross domain comparison problem. Overall, the proposed system achieve new state-of-the-art performance on both protocols.

6. Conclusions

We have proposed a semi-supervised framework of learning robust face representation that could generalize to unconstrained faces beyond the labeled training data. Without collecting domain specific data, we utilized a relatively small unlabeled dataset containing diverse styles of face images. In order to fully utilize the unlabeled dataset, two methods are proposed. First, we showed that the domain adversarial learning, which is common in adaptation methods, can be applied in our setting to reduce domain gaps between labeled faces and hidden sub-domains. Second, we propose an augmentation network that can capture different visual styles in the unlabeled dataset and apply them to the labeled images during training, making the face representation more discriminative for unconstrained faces. Our experimental results show that as the number of labeled images increases, the performance of the supervised baseline tends to saturate on the challenging testing scenarios. Instead, introducing more diverse training data becomes more important and helpful. In a few challenging protocols, we showed that the proposed method can outperform the supervised baseline with less than half of the labeled data. By training on the labeled MS-Celeb-1M dataset and unlabeled WiderFace dataset, our final model achieves state-of-the-art performance on challenging benchmarks such as IJB-B, IJB-C and IJB-S.

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 2, 4

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE FG*, 2018. 5, 8

[3] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019. 1, 2

[4] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *CVPR*, 2020. 8

[5] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CVPR*, 2019. 2, 5, 8

[6] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 5

[7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. 2015. 1, 2

[8] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2015. 1, 2

[9] Sixue Gong, Yichun Shi, and Anil Jain. Low quality video face recognition: Multi-mode aggregation recurrent network (marn). In *ICCV Workshops*, 2019. 8

[10] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *CVPR*, 2020. 2

[11] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large scale face recognition. In *ECCV*, 2016. 1, 3, 5

[12] Abul Hasnat, Julien Bohné, Jonathan Milgram, Stéphane Gentric, and Liming Chen. Deepvisage: Making face recognition simple yet with powerful generalization skills. *ICCV*, 2017. 2

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 4

[14] 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 07-49, University of Massachusetts, Amherst, October 2007. 6

[15] Xun Huang and Serge J Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 4

[16] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 4

[17] Nathan D. Kalka, Brianna Maze, James A. Duncan, Kevin J. O'Connor, Stephen Elliott, Kaleb Hebert, Julia Bryan, and Anil K. Jain. IJB-S : IARPA Janus Surveillance Video Benchmark . In *BTAS*, 2018. 1, 6

[18] Bong-Nam Kang, Yonghyun Kim, Bongjin Jun, and Daijin Kim. Attentional feature-pair relation networks for accurate face recognition. In *ICCV*, 2019. 8

[19] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, 2019. 1, 2

[20] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016. 7

[21] 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 *CVPR*, 2015. 1

[22] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. 2017. 2

[23] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013. 2, 4

[24] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 1, 2

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 2

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017. 1, 2

[27] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, 2016. 2

[28] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *ICB*, 2018. 1, 6

[29] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017. 1, 2

[30] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. 1, 2

[31] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Trans. on Neural Networks*, 2010. 1, 2

[32] Rajeev Ranjan, Ankan Bansal, Jingxiao Zheng, Hongyu Xu, Joshua Gleason, Boyu Lu, Anirudh Nanduri, Jun-Cheng Chen, Carlos D Castillo, and Rama Chellappa. A fast and accurate system for face detection, identification, and verification. *IEEE Trans. on Biometrics, Behavior, and Identity Science*, 2019. 8

[33] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv:1703.09507*, 2017. 2

[34] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NeurIPS*, 2015. 2

[35] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 1, 2

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[37] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, 2019. 8

[38] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016. 2

[39] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv:2001.07685*, 2020. 2, 4

[40] Kihyuk Sohn, Wendy Shang, Xiang Yu, and Manmohan Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. In *CVPR*, 2017. 1

[41] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014. 2

[42] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. *arXiv:2002.10857*, 2020. 2

[43] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 2

[44] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2

[45] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2

[46] Feng Wang, Weiyang Liu, Haijun Liu, and Jian Cheng. Additive margin softmax for face verification. *arXiv:1801.05599*, 2018. 2, 3

[47] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *CVPR*, 2018. 2, 3

[48] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 90–98, 2017. 5

[49] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv:1904.12848*, 2019. 2

[50] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *ECCv*, 2018. 8

[51] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, 2016. 2, 5

[52] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 1, 7

[53] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *ICCV*, 2019. 2

[54] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *CVPR*, 2019. 2

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 4