Towards Universal Representation Learning for Deep Face Recognition

Yichun Shi[1,2*] Xiang Yu[2] Kihyuk Sohn[2] Manmohan Chandraker[2,3] Anil K. Jain[1]
[1]Michigan State University [2]NEC Labs America [3]University of California, San Diego

Abstract

Recognizing faces in the wild is extremely hard as they appear with diverse variations. Traditional methods either train with specifically annotated target domain data which contains the variations, or introduce unlabeled target domain data to adapt from the training domain. Instead, we propose a universal representation learning face recognition framework, URFace, that can deal with larger variations unseen in the given training data, without leveraging knowledge of the target domain. We firstly synthesize the training data that corresponds to several semantically meaningful variations, such as low resolution, occlusion and head pose. However, directly using the augmented data hinders training convergence, since the augmented samples are usually hard examples. We propose to split the feature embedding into multiple sub-embeddings and associate different confidence values for each sub-embedding to smooth the training procedure. The sub-embeddings are further decorrelated by regularizing classification loss on variations and adversarial loss on different partitions of them. Experiments show that our method achieves state-of-the-art performance on general face recognition datasets such as LFW and MegaFace, while being significantly better on extreme benchmarks such as TinyFace and IJB-S.

1. Introduction

Deep face recognition seeks to map input images to a feature space with small intra-identity distance and large inter-identity distance, which has been achieved by prior works through loss design and datasets with rich within-class variations [29, 41, 17, 39, 4]. However, even very large public datasets manifest strong biases, such as ethnicity [33, 34] or head poses [20, 24, 45]. This lack of variation leads to significant performance drops on challenging test datasets, for example, accuracy reported by prior state-of-the-art [31] on IJB-S or TinyFace [11, 3] are about 30% lower than IJB-A [14] or LFW [10].

Recent works seek to close the domain gap caused by such data bias through domain adaptation, i.e., identifying specific factors of variation and augmenting the training datasets [24], or further leveraging unlabeled data along such nameable factors [33]. While nameable variations are hard to identify exhaustively, prior works have sought to align the feature space between source and target domains [28, 34]. Alternatively, individual models might be trained on various datasets and ensembles to obtain good performance on each [19]. All these approaches either only handle specific variations, or require access to test data distributions, or accrue additional run-time complexity to handle wider variations. In contrast, we propose learning a single "universal" deep feature representation that handles the variations in face recognition without requiring access to test data distribution and retains run-time efficiency, while achieving strong performance across diverse situations especially on low-quality images (see Figure 1).

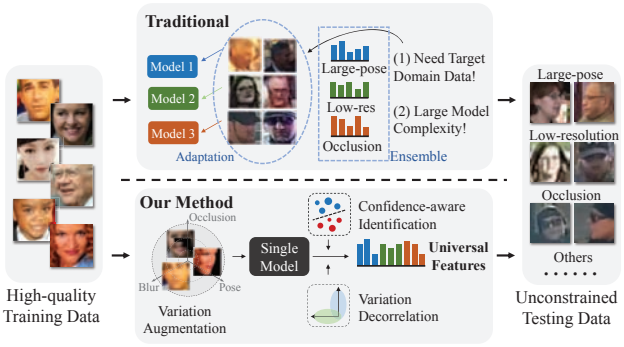


Figure 1: Traditional recognition models require target domain data to adapt from the high-quality training data to conduct unconstrained/low-quality face recognition. Model ensemble is further needed for a universal representation purpose which significantly increases model complexity. In contrast, our method works only on original training data without any target domain data information, and can deal with unconstrained testing scenarios.

This paper introduces several novel contributions in Section 3 to learn such a universal representation. First, we note that inputs with non-frontal poses, low resolutions and heavy occlusions are key nameable factors that present challenges for "in-the-wild" applications, for which training data may be synthetically augmented. But directly adding hard augmented examples into training leads to a more difficult

*This work was conducted as part of a summer internship at NEC Labs America.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00685

optimization problem. We mitigate this by proposing an identification loss that accounts for per-sample confidence to learn a probabilistic feature embedding. Second, we seek to maximize representation power of the embedding by decomposing it into sub-embeddings, each of which has an independent confidence value during training. Third, all the sub-embeddings are encouraged to be further decorrelated through two complementary regularization over different partitions of the sub-embeddings, i.e., classification loss on variations and adversarial loss on different partitions. Fourth, we achieve further decorrelation by mining for additional variations for which synthetic augmentation is non-trivial. Finally, we account for the varying discrimination power of sub-embeddings for various factors through a probabilistic aggregation that accounts for their uncertainties.

In Section 5, we extensively evaluate the proposed methods on public datasets. Compared to our baseline model, the proposed method maintains the high accuracy on general face recognition benchmarks, such as LFW and YTF, while significantly boosting the performance on challenging datasets such as IJB-C, IJB-S, where new state-of-the-art performance is achieved. Detailed ablation studies show the impact of each of the above contributions in achieving these strong performance.

In summary, the main contributions of this paper are:

- A method for learning a universal face representation by associating features with different variations, leading to improved generalization on diverse testing datasets.
- A confidence-aware identification loss that utilizes sample confidence during training to leverage hard samples.
- A feature decorrelation regularization that applies both a classification loss on variations and an adversarial loss on different partitions of the feature sub-embeddings, leading to improved performance.
- A training strategy to effectively combine synthesized data to train a face representation applicable to images outside the original training distribution.
- State-of-the-art results on several challenging benchmarks, such as IJB-A, IJB-C, TinyFace and IJB-S.

2. Related Work

Deep Face Recognition: Deep neural networks are widely adopted in current research on face recognition [36, 35, 29, 20, 17, 8, 25, 38, 4, 45]. Taigman *et al*. [36] propose an early deep convolutional neural network for face recognition. Subsequent works have explored different loss functions to improve the discrimination power of the feature representation. Wen *et al*. [41] propose center loss to reduce intra-class variation. A series of works have also proposed to use metric learning for face recognition [29, 32]. Recent works have attempted to achieve discriminative embeddings with a single identification loss function where proxy or prototype vectors are used to represent each class in the embedding space [17, 38, 39, 25, 4].

(a) Blur (b) Occlusion (c) Pose (d) Randomly Combined

Figure 2: Samples with augmentation alongside different variations.

Universal Representation: Universal representation refers to a single model that can be applied to various visual domains (usually different tasks), e.g. object, character, road signs, while maintaining the performance of using a set of domain-specific models [1, 26, 27, 40, 34]. The features learned by such a single model are believed to be more universal than domain-specific models. Different from domain generalization [13, 22, 15, 16, 37], which targets adaptability on unseen domains by learning from various seen domains, universal representation learning does not involve re-training on unseen domains. Several methods focus on increasing the parameter efficiency by reducing the domain-shift with techniques such as conditioned BatchNorm [1] and residual adapters [26, 27]. Based on SE modules [9], Wang *et al*. [40] propose a domain-attentive module for intermediate (hidden) features of a universal object detection network. Our work is different from those methods in two ways: (1) it is a method for similarity metric learning rather than detection or classification tasks and (2) it is model-agnostic. The features learned by our model can then be directly applied to different domains by computing the pairwise similarity between samples of unseen classes.

3. Proposed Approach

In this section, we first introduce three augmentable variations, namely blur, occlusion and head pose, to augment the training data. Visual examples of augmented data are shown in Figure 2 and the details can be found in Section 4. Then in Section 3.1, we introduce a confidence-aware identification loss to learn from hard examples, which is further extended in Section 3.2 by splitting the feature vectors into sub-embeddings with independent confidence. In Section 3.3, we apply the introduced augmentable variations to further decorrelate the feature embeddings. A method for discovering further non-augmentable variations is proposed to achieve better decorrelation. Finally, an uncertainty-guided pairwise metric is proposed for inference.

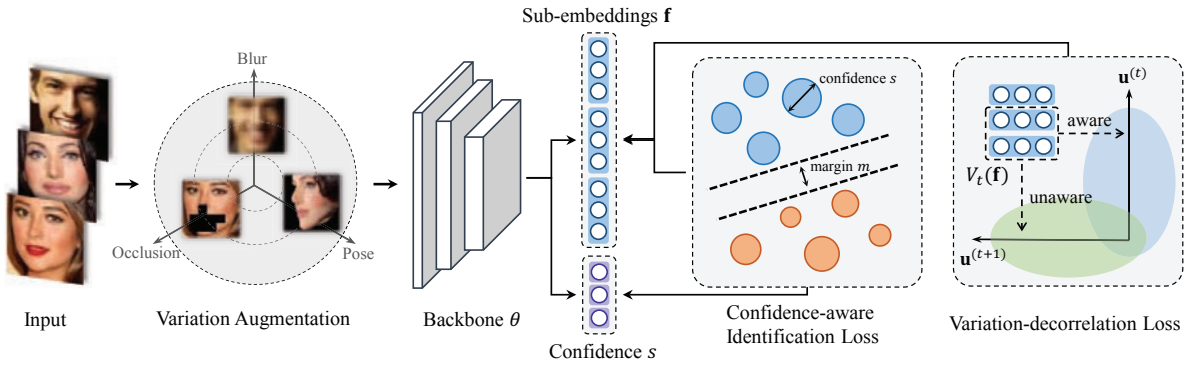


Figure 3: Overview of the proposed method. High-quality input images are first augmented according to pre-defined variations, i.e., blur, occlusion and pose. The feature representation is then split into sub-embeddings associated with sample-specific confidences. Confidence-aware identification loss and variation decorrelation loss are developed to learn the sub-embeddings.

3.1. Confidence-Aware Identification Loss

We investigate the posterior probability of being classified to identity $j \in \{1, 2, \ldots, N\}$, given the input sample $\mathbf{x}_i$. Denote the feature embedding of sample i as $\mathbf{f}_i$ and the j^{th} identity prototype vector as $\mathbf{w}_j$, which is the identity template feature. A probabilistic embedding network θ represents each sample $\mathbf{x}_i$ as a Gaussian distribution $\mathcal{N}(\mathbf{f}_i, \sigma_i^2\mathbf{I})$ in the feature space. The likelihood of $\mathbf{x}_i$ being a sample of class j is given by:

$$p(\mathbf{x}_i|y=j) \propto p_\theta(\mathbf{w}_j|\mathbf{x}_i) = \frac{1}{(2\pi\sigma_i^2)^{\frac{D}{2}}} \exp(-\frac{\|\mathbf{f}_i - \mathbf{w}_j\|^2}{2\sigma_i^2}), \quad (1)$$

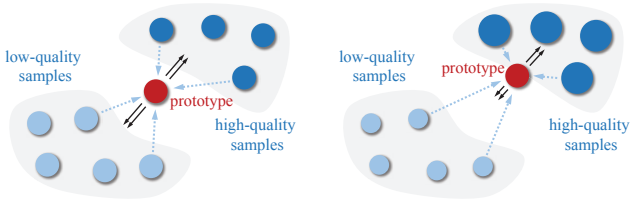
where D is feature dimension. Further assuming the prior of assigning a sample to any identity as equal, the posterior of $\mathbf{x}_i$ belonging to the j^{th} class is derived as:

$$p(y=j|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|y=j)p(y=j)}{\sum_{c=1}^{N} p(\mathbf{x}_i|y=c)p(y=c)} = \frac{\exp(-\frac{\|\mathbf{f}_i-\mathbf{w}_j\|^2}{2\sigma_i^2})}{\sum_{c=1}^{N}\exp(-\frac{\|\mathbf{f}_i-\mathbf{w}_c\|^2}{2\sigma_i^2})}, \quad (2)$$

For simplicity, let us define a confidence value $s_i = \frac{1}{\sigma_i^2}$. Constraining both $\mathbf{f}_i$ and $\mathbf{w}_j$ on the ℓ_2-normalized unit sphere, we have $\frac{\|\mathbf{f}_i-\mathbf{w}_j\|^2}{2\sigma_i^2} = s_i(1 - \mathbf{w}_j^T\mathbf{f}_i)$ and

$$p_(y=j|\mathbf{x}_i) = \frac{\exp(s_i\mathbf{w}_j^T\mathbf{f}_i)}{\sum_{c=1}^{N}\exp(s_i\mathbf{w}_c^T\mathbf{f}_i)}. \quad (3)$$

The effect of confidence-aware posterior in Equation 3 is illustrated in Figure 4. When training is conducted among samples of various qualities, if we assume the same confidence across all samples, the learned prototype will be in the center of all samples. This is not ideal, as low-quality samples convey more ambiguous identity information. In



(a) w/o confidence (b) w/ confidence

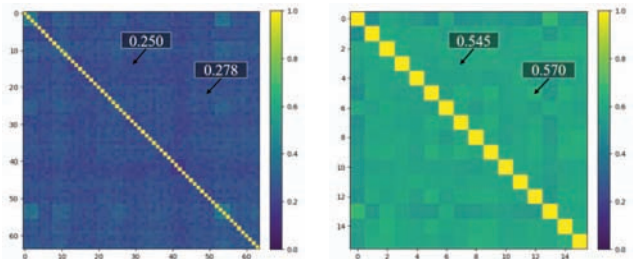
Figure 4: Illustration of confidence-aware embedding learning on quality-various data. With confidence guiding, the learned prototype is closer to high-quality samples which represents the identity better.

contrast, if we set up sample-specific confidence s_i, where high-quality samples show higher confidence, it will push the prototype $\mathbf{w}_j$ to be more similar to high-quality samples in order to maximize the posterior. Meanwhile, during update of the embedding $\mathbf{f}_i$, it provides a stronger push for low-quality $\mathbf{f}_i$ to be closer to the prototype.

Adding loss margin [39] over the exponential logit has been shown to be effective in narrowing the within-identity distribution. We also incorporate it into our loss:

$$\mathcal{L}'_{idt} = -\log \frac{\exp(s_i\mathbf{w}_{y_i}^T\mathbf{f}_i - m)}{\exp(s_i\mathbf{w}_{y_i}^T\mathbf{f}_i - m) + \sum_{j\neq y_i}\exp(s_i\mathbf{w}_j^T\mathbf{f}_i)}, \quad (4)$$

where y_i is the ground-truth label of $\mathbf{x}_i$. Our confidence-aware identification loss (C-Softmax) is different from cosine loss[39] as follows: (1) each image has an independent and dynamic s_i rather than a constant shared scalar and (2) the margin parameter m is not multiplied by s_i. The independence of s_i allows it to gate the gradient signals of $\mathbf{w}_j$ and $\mathbf{f}_i$ during network training in a sample-specific way, as the confidence (degree of variation) of training samples can have large differences. Though samples are specific, we aim to learn a homogeneous feature space such that the metric across different identities is consistent. Thus, allowing s_i to compensate for the confidence difference of the samples, we



(a) sub-embedding of size 8 (b) sub-embedding of size 32

Figure 5: The correlation matrices of sub-embeddings by splitting the feature vector into different sizes. The correlation is computed in terms of distance to class center.

expect m to be consistently shared across all the identities.

3.2. Confidence-Aware Sub-Embeddings

Though the embedding $\mathbf{f}_i$ learned through a sample-specific gating s_i can deal with sample-level variations, we argue that the correlation among the entries of $\mathbf{f}_i$ itself is still high. To maximize the representation power and achieve a compact feature size, decorrelating the entries of the embedding is necessary. This encourages us to further break the entire embedding $\mathbf{f}_i$ into partitioned sub-embeddings, each of which is further assigned a scalar confidence value.

Illustrated in Figure 3, we partition the entire feature embedding $\mathbf{f}_i$ into K equal-length sub-embeddings as in Equation 5. Accordingly, the prototype vector $\mathbf{w}_j$ and the confidence scalar s_i are also partitioned into the same size K groups.

$$\begin{aligned} \mathbf{w}_j &= [\mathbf{w}_j^{(1)T}, \mathbf{w}_j^{(2)T}, \ldots, \mathbf{w}_j^{(K)T}], \\ \mathbf{f}_i &= [\mathbf{f}_i^{(1)T}, \mathbf{f}_i^{(2)T}, \ldots, \mathbf{f}_i^{(K)T}], \\ \mathbf{s}_i &= [s_i^{(1)}, s_i^{(2)}, \ldots, s_i^{(K)}], \end{aligned} \tag{5}$$

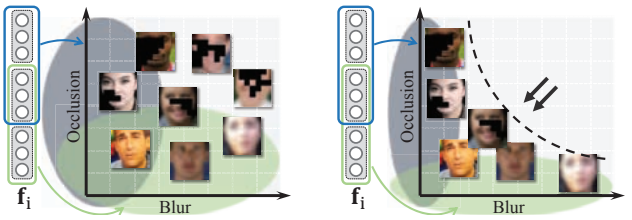
Each group of sub-embeddings $\mathbf{f}_i^{(k)}$ is ℓ_2 normalized onto unit sphere separately. The final identification loss thus is:

$$\mathcal{L}_{idt} = -\log \frac{\exp(\mathbf{a}_{i,y_i} - m)}{\exp(\mathbf{a}_{i,y_i} - m) + \sum_{j \neq y_i} \exp(\mathbf{a}_{i,j})}, \tag{6}$$

$$\mathbf{a}_{i,j} = \frac{1}{K} \sum_{k=1}^{K} s_i^{(k)} \mathbf{w}_j^{(k)T} \mathbf{f}_i^{(k)}. \tag{7}$$

A common issue for neural networks is that they tend to be "over-confident" on predictions [6]. We add an additional l_2 regularization to constrain the confidence from growing arbitrarily large:

$$\mathcal{L}_{reg} = \frac{1}{K} \sum_{k=1}^{K} s_i^{(k)2}. \tag{8}$$



(a) variation-correlated features (b) variation-decorrelated features

Figure 6: The variation decorrelation loss disentangles different sub-embeddings by associating them with different variations. In this example, the first two sub-embeddings are forced to be invariant to occlusion while the second two sub-embeddings are forced to be invariant to blur. By pushing stronger invariance for each variation, the correlation/overlap between two variations is reduced.

3.3. Sub-Embeddings Decorrelation

Setting up multiple sub-embeddings alone does not guarantee the features in different groups are learning complementary information. Empirically shown in Figure 5, we find the sub-embeddings are still highly correlated, i.e., dividing $\mathbf{f}_i$ into equal 16 groups, the average correlation among all the sub-embeddings is 0.57. If we penalize the sub-embeddings with different regularization, the correlation among them can be reduced. By associating different sub-embeddings with different variations, we conduct variation classification loss on a subset of all the sub-embeddings while conducting variation adversarial loss in terms of other variation types. Given multiple variations, such two regularization terms are forced on different subsets, leading to better sub-embedding decorrelation.

For each augmentable variation $t \in \{1, 2, \ldots, M\}$, we generate a binary mask V_t, which selects a random $\frac{K}{2}$ subset of all sub-embeddings while setting the other half to be zeros. The masks are generated at the beginning of the training and will remain fixed during training. We guarantee that for different variations, the masks are different. We expect $V_t(\mathbf{f}_i)$ to reflect t^{th} variation while invariant to the others. Accordingly, we build a multi-label binary discriminator C by learning to predict all variations from each masked subset:

$$\begin{aligned} \min_C \mathcal{L}_C &= -\sum_{t=1}^{M} \log p_C(\mathbf{u}_i = \hat{\mathbf{u}}_i | V_t(\mathbf{f}_i)) \\ &= -\sum_{t=1}^{M} \sum_{t'=1}^{M} \log p_C(u_i^{(t')} = \hat{u}_i^{(t')} | V_t(\mathbf{f}_i)) \end{aligned} \tag{9}$$

where $\mathbf{u}_i = [u_i^{(1)}, u_i^{(2)}, \ldots, u_i^{(M)}]$ are the binary labels (0/1) of the known variations and $\hat{\mathbf{u}}_i$ is the ground-truth label. For example, if $t = 1$ corresponds to resolution, $\hat{u}_i^{(1)}$ would be 1 and 0 for high/low-resolution images, respectively. Note that Equation 9 is only used for training the discriminator C. The corresponding classification and adversarial loss of the

embedding network is then given by:

$$\mathcal{L}_{cls} = -\sum_{t=1}^{M} \log p_C(u^{(t)} = \hat{u}_i^{(t)} | V_t(\mathbf{f}_i)) \tag{10}$$

$$\mathcal{L}_{adv} = -\sum_{t=1}^{M}\sum_{t' \neq t}(\frac{1}{2}\log p_C(u^{(t')} = 0|V_t(\mathbf{f}_i)) + \frac{1}{2}\log p_C(u^{(t')} = 1|V_t(\mathbf{f}_i))) \tag{11}$$

The classification loss $\mathcal{L}_{cls}$ to encourage V_t to be variation-specific while $\mathcal{L}_{adv}$ is an adversarial loss to encourage invariance to the other variations. As long as no two masks are the same, it guarantees that the selected subsets V_t is functionally different from other $V_{t'}$. We thus achieve decorrelation between V_t and $V_{t'}$. The overall loss function for each sample is:

$$\min_{\theta} \mathcal{L} = \mathcal{L}_{idt} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{adv}\mathcal{L}_{adv}. \tag{12}$$

During the optimization, Equation (12) is averaged across the samples in the mini-batch.

3.4. Mining for Further Variations

The limited number (three in our method) of augmentable variations leads to limited effect of decorrelation as the number of V_t are too small. To further enhance the decorrelation, as well to introduce more variations for better generalization ability, we aim to explore more variations with semantic meaning. Notice that not all the variations are easy to conduct data augmentation, e.g. smiling or not is hard to augment. For such variations, we attempt to mine out the variation labels from the original training data. In particular, we leverage an off-the-shelf attribute dataset CelebA [18] to train a attribute classification model θ_A with identity adversarial loss:

$$\begin{aligned} \min_{\theta_A} \mathcal{L}_{\theta_A} &= -\log p(l_A|\mathbf{x}_A) - \frac{1}{N_A}\sum_{c}^{N_A}\log p(y_A = c|\mathbf{x}_A) \\ \min_{D_A} \mathcal{L}_{D_A} &= -\log p(y_A = y_{\mathbf{x}_A}|\mathbf{x}_A), \end{aligned} \tag{13}$$

where l_A is the attribute label and y_A is the identity label. $\mathbf{x}_A$ is the input face image and N_A is the number of identities in the CelebA dataset. The first term penalizes the feature to classify facial attributes and the second term penalizes the feature to be invariant to identities.

The attribute classifier is then applied to the recognition training set to generate T new soft variation labels, e.g. smiling or not, young or old. These additional variation binary labels are merged with the original augmentable variation labels as: $\mathbf{u}_i = [u_i^{(1)}, \ldots, u_i^{(M)}, u_i^{(M+1)}, \ldots, u_i^{(M+T)}]$ and are then incorporated into the decorrelation learning framework in Section 3.3.

3.5. Uncertainty-Guided Probabilistic Aggregation

Considering the metric for inference, simply taking the average of the learned sub-embeddings is sub-optimal. This is because different sub-embeddings show different discriminative power for different variations. Their importance should vary according to the given image pairs. Inspired by [31], we consider applying the uncertainty associated with each embedding for a pairwise similarity score:

$$\begin{aligned} score(\mathbf{x}_i, \mathbf{x}_j) = &-\frac{1}{2}\sum_{k=1}^{K}\frac{\left\|\mathbf{f}_i^{(k)} - \mathbf{f}_j^{(k)}\right\|^2}{\sigma_i^{(k)2} + \sigma_j^{(k)2}} \\ &- \frac{D}{2K}\sum_{k=1}^{K}\log(\sigma_i^{(k)2} + \sigma_j^{(k)2}) \end{aligned} \tag{14}$$

Though with Equation 8 for regularization, we empirically find that the confidence learned with the identification loss still tend to be overconfident and hence cannot be directly used for Equation 14, so we fine-tune the original confidence branch to predict σ while fixing the other parts. We refer the readers to [31] for the training details of fine-tuning.

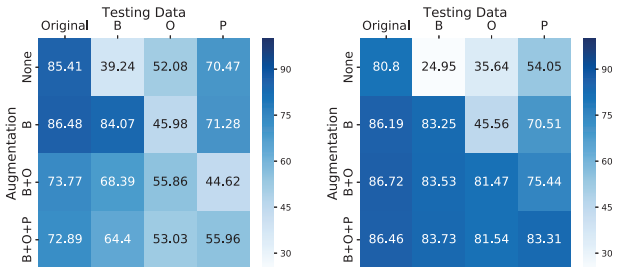
4. Implementation Details

Training Details and Baseline All the models are implemented with Pytorch v1.1. We use the clean list from ArcFace [4] for MS-Celeb-1M [7] as training data. After cleaning the overlapped subjects with the testing sets, we have 4.8M images of 76.5K classes. We use the method in [46] for face alignment and crop all images into a size of 110×110. Random and center cropping are applied during training and testing, respectively, to transform the images into 100×100. We use the modified 100-layer ResNet in [4] as our architecture. The embedding size is 512 for all models, and the features are split into 16 groups for multi-embedding methods. The model C is a linear classifier. The baseline models in the experiments are trained with CosFace loss function [39, 38], which achieves state-of-the-art performance on general face recognition tasks. The models without domain augmentation are trained for 18 epochs and models with domain augmentation are trained for 27 epochs to ensure convergence. We empirically set λ_{reg}, λ_{cls} and λ_{adv} as 0.001, 2.0 and 2.0, respectively. The margin m is empirically set to 30. For non-augmentable variations, we choose $T = 3$ attributes, namely smiling, young and gender.

Variation Augmentation For the low-resolution, we use Gaussian blur with a kernel size between 3 and 11. For the occlusion, we split the images into 7×7 blocks and randomly replace some blocks with black masks. (3) For pose augmentation, we use PRNet [5] to fit the 3D model of near-frontal faces in the dataset and rotate them into a yaw degree between $40°$ and $60°$. All the augmentations are randomly combined with a probability of 30% for each.

(a) Type I (b) Type II (c) Type III

Figure 7: Examples of the three types of datasets. The images are sampled from LFW [10], IJB-A [14], IJB-S [11], respectively.



(a) Baseline (b) Proposed

Figure 8: Testing results on synthetic data of different variations from IJB-A benchmark (TAR@FAR=0.01%). Different rows correspond to different augmentation strategies during training. Columns are different synthetic testing data. "B", "O", "P" represents "Blur", "Occlusion" and "Pose", respectively. The performance of the proposed method is improved in a monotonous way with more augmentations being added.

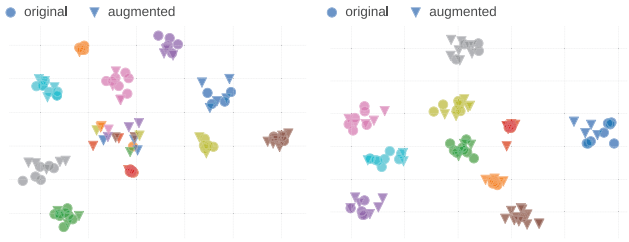
5. Experiments

In this section, we firstly introduce different types of datasets reflecting different levels of variation. Different levels of variation indicate different image quality and thus lead to different performance. Then we conduct detailed ablation study over the proposed confidence-aware loss and all the proposed modules. Further, we show evaluation on those different types of testing datasets and compare to state-of-the-art methods.

5.1. Datasets

We evaluate our models on eight face recognition benchmarks, covering different real-world testing scenarios. The datasets are roughly categorized into three types based on the level of variations:

Type I: Limited Variation LFW [10], CFP [30], YTF [42] and MegaFace [12] are four widely applied benchmarks for general face recognition. We believe the variations in those datasets are limited, as only one or few of the variations being presented. In particular, YTF are video samples with relatively lower resolution; CFP [30] are face images with large pose variation but of high resolution; MegaFace includes 1 million distractors crawled from internet while its labeled images are all high-quality frontal faces from FaceScrub dataset [23]. For both LFW and YTF, we use the unrestricted verification protocol. For CFP, we focus on the frontal-profile (FP) protocol. We test on both verification and identification protocols of MegaFace.



(a) Baseline (b) Proposed

Figure 9: t-SNE visualization of the features in a 2D space. Colors indicate the identities. Original training samples and augmented training samples are shown in circle and triangle, respectively.

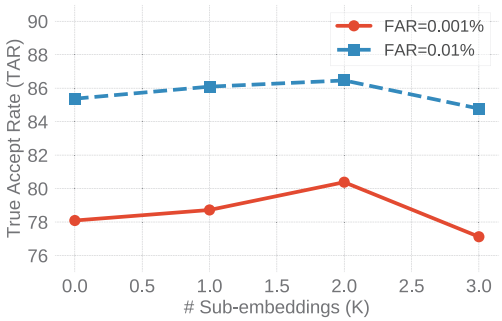


Figure 10: Performance change with respect to difference choice of K.

Type II: Mixed Quality IJB-A [14] and IJB-C [21] include both high quality celebrity photos taken from the wild and low quality video frames with large variations of illumination, occlusion, head pose, etc. We test on both verification and identification protocols of the two benchmarks.

Type III: Low Quality We test on TinyFace [3] and IJB-S [11], two extremely challenging benchmarks that are mainly composed of low-quality face images. In particular, TinyFace only consists of low-resolution face images captured in the wild, which also includes other variations such as occlusion and pose. IJB-S is a video face recognition dataset, where all images are video frames captured by surveillance cameras except a few high-quality registration photos for each person. Example images of the three types of datasets are shown in Figure 7.

5.2. Ablation Study

5.2.1 Effect of Confidence-aware Learning

We train a set of models by gradually adding the nameable variations. The "Baseline" model is an 18-layer ResNet trained on a randomly selected subset of MS-Celeb-1M (0.6M images). The "Proposed" model is trained with the confidence-aware identification loss and $K = 16$ embedding groups. As a controlled experiment, we apply the same type of augmentation on IJB-A dataset to synthesize testing data of the corresponding variations. In Figure 8, "Baseline" model shows decreasing performance when gradually adding

Model	Method					LFW	CFP-FP	IJB-A (TAR@FAR)		TinyFace		IJB-S	
	VA	CI	ME	DE	PA	Accuracy	Accuracy	FAR=0.001%	FAR=0.01%	Rank1	Rank5	Rank1	Rank 5
Baseline						99.75	98.16	82.20	93.05	46.75	51.79	37.14	46.75
A	✓					99.70	98.35	82.42	93.86	55.26	59.04	51.27	58.94
B	✓	✓				99.78	98.30	94.70	96.02	57.11	63.09	59.87	66.90
C	✓	✓	✓			99.77	98.50	94.75	96.27	57.30	63.73	59.66	66.30
	✓	✓	✓		✓	99.78	**98.66**	**96.10**	97.29	55.04	60.97	59.71	66.32
D		✓	✓	✓		99.65	97.77	80.06	92.14	34.76	39.86	29.87	40.69
		✓	✓	✓	✓	99.68	98.00	94.37	96.42	35.05	40.13	50.00	56.27
E (all)	✓	✓	✓	✓		99.75	98.30	95.00	96.27	61.32	66.34	60.74	66.59
	✓	✓	✓	✓	✓	**99.78**	98.64	96.00	**97.33**	**63.89**	**68.67**	**61.98**	**67.12**

Table 1: Ablation study over the whole framework. "VA" indicates "Variation Augmentation" (Section 3), "CI" indicates "Confidence-aware Identification loss" (Section 3.1), "ME" indicates "Multiple Embeddings" (Section 3.3), "DE" indicates "Decorrelated Embeddings" (Section 3.3) and "PA" indicates "Probabilistic Aggregation". (Section 3.5). E(all) uses all the proposed modules.

Method	LFW	YTF	CFP-FP	MF1	
				Rank1	Veri.
FaceNet [29]	99.63	95.1	-	-	-
CenterFace [41]	99.28	94.9	-	65.23	76.52
SphereFace [17]	99.42	95.0	-	75.77	89.14
ArcFace [4]	99.83	98.02	98.37	81.03	96.98
CosFace [39]	99.73	97.6	-	77.11	89.88
Ours (Baseline)	99.75	97.16	98.16	80.03	95.54
Ours (Baseline+VA)	99.70	97.10	98.36	78.10	94.31
Ours (all)	99.75	97.68	98.30	79.10	94.92
Ours (all) + PA	99.78	97.92	98.64	78.60	95.04

Table 2: Our method compared to state-of-the-art methods on Type I datasets. The MegaFace verification rates are computed at FAR=0.0001%. "-" indicates that the author did not report the performance on the corresponding protocol.

new variations as in the grid going down from top row to bottom row. In comparison, the proposed method shows improving performance when adding new variations from top to bottom, which highlights the effect of our confidence-aware representation learning and it further allows to add more variations into the framework training.

We also visualize the features with t-SNE projected onto 2D embedding space. Figure 9 shows that for "Baseline" model, with different variation augmentations, the features actually are mixed and thus are erroneous for recognition. While for "Proposed" model, different variation augmentation generated samples are still clustered together to its original samples, which indicates that identity is well preserved. Under the same settings as above, we also show the effect of using different number of groups in Figure 10. At the beginning, splitting the embedding space into more groups increases performance for both TARs. When the size of each sub-embedding becomes too small, the performance starts to drop because of the limited capacity for each sub-embedding.

5.2.2 Ablation on All Modules

We investigate each module's effect by looking into the ablative models in Table 1. Starting from the baseline, model A is trained with variation augmentation. Based on model A, we add confidence-aware identification loss to obtain model B. Model C is further trained by setting up multiple sub-embeddings. In model E, we further added the decorrelation loss. We also compare with a Model D with all the modules except variation augmentation. Model C, D and E, which have multiple embeddings, are tested w/ and w/o probabilistic aggregation (PA). The methods are tested on two type I datasets (LFW and CFP-FP), one type-II dataset (IJB-A) and one type-III dataset (TinyFace).

Shown in Table 1, compared to baseline, adding variation augmentation improves performance on CFP-FP, TinyFace, and IJBA. These datasets present exactly the variations introduced by data augmentation, i.e., pose variation and low resolution. However, the performance on LFW fluctuates from baseline as LFW is mostly good quality images with few variations. In comparison, model B and C are able to reduce the negative impact of hard examples introduced by data augmentation and leads to consistent performance boost across all benchmarks. Meanwhile, we observe that splitting into multiple sub-embeddings alone does not improve (compare B to C first row) significantly, which can be explained by the strongly correlated confidence among the sub-embeddings (see Figure 5). Nevertheless, with the decorrelation loss and probabilistic aggregation, different sub-embeddings are able to learn and combine complementary features to further boost the performance, i.e., the performance in the second row of Model E is consistently better than its first row.

5.3. Evaluation on General Datasets

We compare our method with state-of-the-art methods on general face recognition datasets, i.e., those Type I datasets with limited variation and high quality. Since the testing images are mostly with good quality, there is limited advantage of our method which is designed to deal with larger variations. Even though, shown in Table 2, our method still stands on top being better than most of the methods while slightly worse than ArcFace. Notice that our baseline model already achieves good performance across all the testing sets. It actually verifies that the type I testing sets do not show significant domain gap from the training set, where even without variation augmentation or embedding decorrelation,

Method	IJB-A (Vrf)		IJB-A (Idt)		IJB-C (Vrf)		IJB-C (Idt)		IJB-S (S2B)		
	FAR=0.001%	FAR=0.01%	Rank1	Rank5	FAR=0.001%	FAR=0.01%	Rank1	Rank5	Rank1	Rank5	FPIR=1%
NAN [44]*	-	88.1±1.1	95.8±0.5	98.0±0.5	-	-	-	-	-	-	-
L2-Face [25]*	90.9±0.7	94.3±0.5	97.3±0.5	98.8±0.3	-	-	-	-	-	-	-
DA-GAN [47]*	94.6±0.1	97.3±0.5	**99.0±0.2**	**99.5±0.3**	-	-	-	-	-	-	-
Cao *et al*. [2]	-	92.1±1.4	98.2±0.4	99.3±0.2	76.8	86.2	91.4	95.1	-	-	-
Multicolumn [43]	-	92.0±1.3	-	-	77.1	86.2	-	-	-	-	-
PFE [31]	-	95.3±0.3	-	-	89.6	93.3	-	-	50.16	58.33	31.88
ArcFace [4]+	93.7±1.0	94.2±0.8	97.0±0.6	97.9±0.4	93.5	95.8	95.87	**97.27**	57.36	64.95	41.23
Ours (Baseline)	82.6±8.3	93.3±3.0	95.5±0.7	96.9±0.6	43.9	86.7	89.85	90.86	37.14	46.75	24.75
Ours (Baseline + VA)	82.4±8.1	93.9±3.5	95.8±0.6	97.2±0.5	47.6	90.6	90.16	91.20	51.27	58.94	31.19
Ours (all)	95.0±0.9	96.3±0.6	97.5±0.4	98.4±0.4	91.6	93.7	94.39	96.08	60.74	66.59	37.11
Ours (all) + PA	**96.0±0.8**	**97.3±0.4**	97.5±0.3	98.4±0.3	**95.0**	**96.6**	**96.00**	97.06	**61.98**	**67.12**	**42.73**

Table 3: Our model compared to state-of-the-art methods on IJB-A, IJB-C and IJB-S. "-" indicates that the author did not report the performance on the corresponding protocol. "*" indicates fine-tuning on the target dataset during evaluation on IJB-A benchmark and "+" indicates the testing performance by using the released models from corresponding authors.

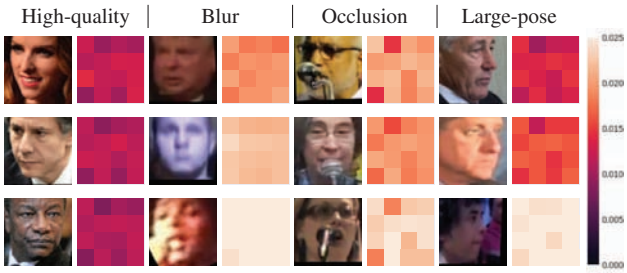


Figure 11: Heatmap visualization of sub-embedding uncertainty on different types of images from IJB-C dataset, shown on the right of each face image. 16 values are arranged in 4×4 grids (no spatial meaning). Brighter color indicates higher uncertainty.

the straight training can lead to good performance.

5.4. Evaluation on Mixed/Low Quality Datasets

When evaluating on more challenging datasets, those state-of-the-art general methods encounter performance drop as the challenging datasets present large variations and thus large domain gap from the good quality training datasets. Table 3 shows the performance on three challenging benchmarks: IJB-A, IJB-C and IJB-S. The proposed model achieves consistently better results than the state-of-the-arts. In particular, simply adding variation augmentation ("Ours (Baseline + VA)") actually leads to a worse performance on IJB-A and IJB-C. When variation augmentation is combined with our proposed modules ("Ours"), significant performance boost is achieved. Further adding PA with "Ours", we achieve even better performance across all datasets and protocols. Notice that IJB-A is a cross-validation protocol. Many works fine-tune on training splits before evaluation (shown with "*"). Even though, our method without fine-tuning still outperforms the state-of-the-art methods with significant margin on IJB-A verification protocol, which suggests that our method indeed learns the representation towards dealing with unseen variations.

Table 3 last column shows the evaluation on IJB-S, which is so far the most challenging benchmark targeting real surveillance scenario with severe poor quality images. We show the Surveillance-to-Booking (S2B) protocol of IJB-S. Other protocol results can be found in supplementary. As IJB-S is recently released, there are few studies that have evaluated on this dataset. To comprehensively evaluate our model, we use the publicly released models from ArcFace [4] for comparison. Our method achieves consistently better performance across Rank-1 and Rank-5 identification protocol. For TinyFace, as in Table 1, we achieve 63.89%, 68.67% rank-1 and rank-5 accuracy, where [3] reports 44.80%, 60.40%, and ArcFace achieves 47.39%, 52.28%. Combining Table 2, our method achieves top level accuracy on general recognition datasets and significantly better accuracy on challenging datasets, which demonstrates the advantage in dealing with extreme or unseen variations.

Uncertainty Visualization Figure 11 shows uncertainty scores for the 16 sub-embeddings reshaped into 4×4 grids. High-quality and low-quality sub-embeddings are shown in dark and light colors respectively. The uncertainty map presents different patterns for different variations.

6. Conclusion

In this work, we propose a universal face representation learning framework, URFace, to recognize faces under all kinds of variations. We firstly introduce three nameable variations into MS-Celeb-1M training set via data augmentation. Traditional methods encounter convergence problem when directly feeding the augmented hard examples into training. We propose a confidence-aware representation learning by partitioning the embedding into multiple sub-embeddings and relaxing the confidence to be sample and sub-embedding specific. Further, the classification and adversarial losses on variations are proposed to decorrelate the sub-embeddings. By formulating the inference with an uncertainty model, the sub-embeddings are aggregated properly. Experimental results show that our method achieves top performance on general benchmarks such as LFW and MegaFace, and significantly better accuracy on challenging benchmarks such as IJB-A, IJB-C and IJB-S.

References

[1] Hakan Bilen and Andrea Vedaldi. Universal representations: The missing link between faces, text, planktons, and cat breeds. *arXiv:1701.07275*, 2017. 2

[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. 8

[3] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *ACCV*, 2018. 1, 6, 8

[4] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CVPR*, 2019. 1, 2, 5, 7, 8

[5] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 5

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 4

[7] 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. 5

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

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2

[10] 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. 1, 6

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

[12] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016. 6

[13] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *ECCV*, 2012. 2

[14] 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, 6

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017. 2

[16] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 2

[17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 1, 2, 7

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[19] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan. Pose-aware face recognition in the wild. In *CVPR*, 2016. 1

[20] Iacopo Masi, Anh Tun Trn, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, 2016. 1, 2

[21] 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. 6

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

[23] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *CIP*, 2014. 6

[24] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017. 1

[25] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv:1703.09507*, 2017. 2, 8

[26] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *NIPS*, 2017. 2

[27] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *CVPR*, 2018. 2

[28] Swami Sankaranarayanan, Azadeh Alavi, Carlos D Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. In *BTAS*, 2016. 1

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 2, 7

[30] 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 *WACV*, 2016. 6

[31] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, 2019. 1, 5, 8

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

[33] Kihyuk Sohn, Sifei Liu, Guangyu Zhong, Xiang Yu, Ming-Hsuan Yang, and Manmohan Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. In *ICCV*, 2017. 1

[34] Kihyuk Sohn, Wendy Shang, Xiang Yu, and Manmohan Chandraker. Unsupervised domain adaptation for distance metric learning. *ICLR*, 2019. 1, 2

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

[36] 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

[37] Youssef Tamaazousti, Hervé Le Borgne, Céline Hudelot, Mohamed El Amine Seddik, and Mohamed Tamaazousti. Learn-

ing more universal representations for transfer-learning. *IEEE trans. on PAMI*, 2019. 2

[38] Feng Wang, Weiyang Liu, Haijun Liu, and Jian Cheng. Additive margin softmax for face verification. *arXiv:1801.05599*, 2018. 2, 5

[39] 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. 1, 2, 3, 5, 7

[40] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *Computer Vision and Pattern Recognition*, 2019. 2

[41] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 1, 2, 7

[42] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011. 6

[43] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. *ECCV*, 2018. 8

[44] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *CVPR*, 2017. 8

[45] Xin Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. *CVPR*, 2019. 1, 2

[46] Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016. 5

[47] Jian Zhao, Lin Xiong, Jianshu Li, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 3d-aided dual-agent gans for unconstrained face recognition. *IEEE trans. on PAMI*, 2018. 8