Learning to Learn across Diverse Data Biases in Deep Face Recognition

Chang Liu[1,2*] Xiang Yu[2] Yi-Hsuan Tsai[2] Masoud Faraki[2] Ramin Moslemi[2]
Manmohan Chandraker[2,3] Yun Fu[1]
[1]Northeastern University [2]NEC Labs America [3]University of California, San Diego
[1]{liu.chang6, yunfu}@ece.neu.edu, [2]{xiangyu, ytsai, mfaraki ,rmoslemi, manu}@nec-labs.com

Abstract

Convolutional Neural Networks have achieved remarkable success in face recognition, in part due to the abundant availability of data. However, the data used for training CNNs is often imbalanced. Prior works largely focus on the long-tailed nature of face datasets in data volume per identity, or focus on single bias variation. In this paper, we show that many bias variations such as ethnicity, head pose, occlusion and blur can jointly affect the accuracy significantly. We propose a sample level weighting approach termed Multi-variation Cosine Margin (MvCoM), to simultaneously consider the multiple variation factors, which orthogonally enhances the face recognition losses to incorporate the importance of training samples. Further, we leverage a learning to learn approach, guided by a held-out meta learning set and use an additive modeling to predict the MvCoM. Extensive experiments on challenging face recognition benchmarks demonstrate the advantages of our method in jointly handling imbalances due to multiple variations.

1. Introduction

Deep face recognition has achieved remarkable progress [4, 6, 27, 39, 42, 50, 56], with strong results on public benchmarks [19, 57]. However, real-world data distributions are usually long-tailed, whereby a method trained with uniform sampling over the imbalanced training data leads to degraded accuracy. Since it is impractical to collect data that sufficiently covers a wide variety of the imbalance factors, there is a pressing need to develop training methods that can mitigate dataset bias along multiple factors of variations.

In current literature, long-tailed or imbalanced data distribution is usually analyzed in terms of per-class data volume, or a single bias factor such as ethnicity [10, 11, 44, 52] or head pose [29, 35, 60, 64]. Previous approaches distinguish long-tailed classes (minority in samples) from head classes (majority in samples) to mitigate the bias. However, we observe that there usually exist more than one bias variation factors. As shown in Fig. 1, several bias factors jointly influence the overall data distribution. We hypothesize that dealing with such *multiple factors of imbalance* results in a feature space that allows better test-time generalization. Moreover, recent methods focus on class-level imbalance, where samples within the same class share the same importance [2, 21, 37]. This is limited in practice, as different images from the same person would likely differ in their importance (e.g., frontal and profile views). Some other methods [20, 54] compensate the loss with the sample hardness, which is general but cannot attribute the hardness to any of the concrete variation factor. Thus, we hypothesize considering *sample-level variation* instead of class-level, and *explicitly* model each of the variation factor into the loss design.

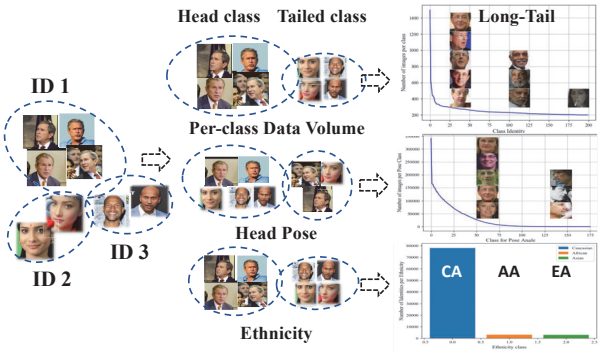

Figure 1. While traditional methods only consider per-class data volume or single bias factor for long-tailed effects, multiple bias factors such as head pose and ethnicity jointly manifest as long-tailed effects in MS-Celeb-1M [14]. Further, samples from the same identity can show different variations – for example, images from ID1 show both frontal and profile poses – indicating that accounting for identity or class-level variation is not sufficient. Hence, our MvCoM aims at *explicitly* modeling the sample-level multiple long-tailed variations jointly for face recognition.

To handle data imbalance, classical methods [2, 16, 41] introduce re-weighted loss functions by assigning higher

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

loss weights to long-tailed classes and lower weights to head classes. Cao et al. [2] mentioned that "label-distribution-aware" re-weighting approaches are advantageous in computational efficiency. However, the assigned weights are usually either fixed based on prior statistics or obtained by sophisticated design choices [5]. We seek a more adaptive re-weighting method that can count for per-sample variation regarding multiple variation factors while potentially sacrificing certain training efficiency. *Meta-learning* [9] is such an adaptive differentiable mechanism to iteratively learn sample-level importance and further contribute to recognition model update. It allows a plug-in mechanism to many recognition losses such as Cosine Loss [50].

Specifically, our proposed framework deals with head pose, ethnicity, blur and occlusion as multiple factors of variation that cause data imbalances, besides per-class data volume. First, we show that the weighted identification loss which is commonly used in re-weighting methods [21, 37], is equivalent to a learnable margin built into the cosine loss (Sec. 3.1). Thereby, we represent each imbalance factor through a corresponding learnable margin. Second, we propose an additive framework to indicate a sample's variation importance using the class volume margin as prior, together with residuals as other variations (Sec. 3.2). With the carefully designed sample-level margin, we orthogonally equip it with cosine loss or its variants, termed Multi-variation Cosine Margin (MvCoM). During training, the proposed MvCoM controls the contribution of each instance in the loss function by assigning its dedicated margin considering all imbalance factors. To realize a meta-learning framework for MvCoM update, we introduce four variation classifiers corresponding to the four variation factors. By hard sample mining over a held-out meta-learning set (no identity overlap with the training set, and its identity data is not used in updating recognition model) to select the samples that are most variation-different from the current training batch, we meta-learn the MvCoM and feedback to the recognition loss update (Sec. 3.2.2). Fig. 2 summarizes our approach.

In the experiments, our method empirically achieve consistently better performance across five challenging datasets highlighting all the long-tailed variations such as occlusion, head pose, blur and ethnicity. Moreover, we find that the proposed MvCoM can be equipped with many backbones such as CosFace and URFace (see Table 3), demonstrating the wide applicability to face recognition platforms. We also visualize the learned sampling importance alongside all those long-tailed variations in Fig. 3 and verify that our MvCoM indeed assigns significant weights to those long-tailed factors which leads to overall smaller loss.

Our technical contributions are thus concluded:

- To our best knowledge, we are the primary several to *explicitly* model multiple long-tailed variation factors, such as ethnicity, pose and occlusion, in an additive formulation within a single framework for face recognition.
- We move beyond class-level imbalance to propose a novel sample-level Multi-variation Cosine Margin (MvCoM) that better compensates distribution imbalance from multiple factors.
- We propose a meta-learning based differentiable mechanism to adaptively learn the proposed MvCoM, enabling an end-to-end unified recognition training scheme.
- Extensive experiments on both controlled and challenging benchmarks show that our method can better mitigate distribution imbalances to outperform prior methods.

2. Related Work

Deep Face Recognition While other face recognition works are related, we only focus on the ones applying CNNs due to their impressive recent gains. Seminal works such as DeepFace, DeepID [45, 47] were among the first to surpass human-level accuracy. A series of recent works [6, 8, 27, 42, 49, 50, 56, 65, 66, 68] design more effective learning losses to further advance the state-of-the-art. Specifically, they focus on designing margins with respect to the angle or the cosine space or a combination of the two. For more complete comparisons, we refer the readers to a survey [51]. We note that these methods either assume the training datasets have balanced distribution or simply remove tail classes from training sets. To better utilize long-tailed data, we propose a comprehensive Multi-variation Cosine Margin (MvCoM) to address data imbalances by considering multiple causative factors such as ethnicity, pose, occlusion and blur.

Imbalanced Data Classification While classification with imbalanced data is a wide direction, we focus on methods specific to face recognition. Early methods directly change the sampling frequency [16, 41]. However, the rebalancing mostly applies empirical rules based on prior statistics, which may lead to sub-optimal training. To adaptively learn the sampling, recent methods exploit hard negative mining [7, 24], metric learning [18, 26, 34, 35, 59, 61, 62] and meta learning [15, 21, 37, 55]. Liu et al. [28] use a dynamic meta-embedding with an associated memory to enhance the representation. AdaptiveFace [25] analyzes the difference between rich and poor classes to propose an adaptive margin. Despite the above advances in addressing the long-tailed problem, those methods only consider the per-class data volume as a cause of imbalance. A recent work of Cao et al. [1] explores other long-tailed factors including ethnicity. But they consider only factors correlated with identity, which excludes other significant factors like pose, occlusion or blur. In contrast, we propose a unified framework to handle a general set of multiple factors, that can be related or unrelated to identity. While Wu et al. [58] share the high-level idea of sampling matters for training, they consider metric space sampling and make a spherical distribution assumption for

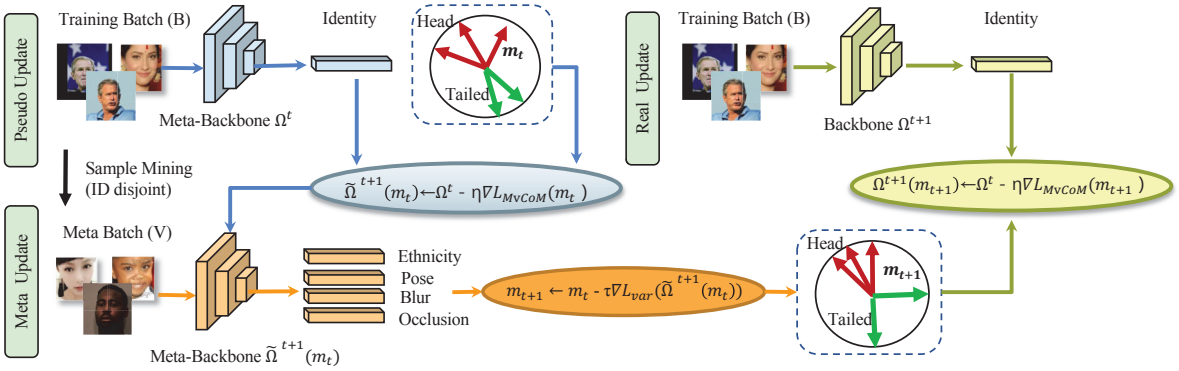

Figure 2. The proposed method flowchart. In Sec. 3.1, we firstly show that traditional re-weighting methods are equivalent to the margin-modulated cosine loss. To jointly tackle multiple variations that cause the long-tailed distribution, we propose the Multi-variation Cosine Margin (MvCoM) (Sec. 3.2). Then, MvCoM is learned via a learning-to-learn scheme specified as three steps (Sec. 3.2.2): (1) recognition model pseudo update. (2) MvCoM meta-update with pseudo recognition model. (3) recognition model real update with updated MvCoM.

it. Our method makes no prior assumption on training data distribution. We instead leverage meta-learning to adaptively generate a balanced training data distribution.

Meta Learning The aim of meta-learning is to train a meta-learner optimized over a set of learning tasks. Each task is typically associated with a dataset. Generally, the approaches are categorized into three categories. (1) *Model based methods* use memory to record the intermediate learned models and incorporate recent model updates with older ones to prevent the forgetting issue [31, 32, 38]. (2) *Metric based methods* learn embedding vectors of input data explicitly and use them to design proper kernel functions, with the prediction usually being a weighted sum over all the kernel functions [43,46,48]. (3) *Optimization based methods* aim to adjust the optimization algorithm so that the model can learn under limited conditions, such as few training samples, data with bias or unseen domain data [9, 13, 33, 36, 63]. Our method lies in the third category. We set up multiple tasks corresponding to the variations that cause long-tailed imbalances. Assuming bias in training data, we seek an optimization method to update the margin, such that our main task of face recognition training is less biased. Note that our focus is on dealing with data bias while [13] emphasizes model generalization to unseen domains.

3. Our Approach

Fig. 2 illustrates our overall framework. We start by explaining traditional re-weighting methods and show their equivalence to optimizing a margin-based identification loss (Sec. 3.1). As the factors causing long-tailed distribution are usually diverse, we propose a sample-level multi-variation cosine margin (MvCoM) as an additive modeling combining all the long-tailed variation factors to enhance a canonical identification objective, i.e., cosine loss [50] (Sec. 3.2). Further, we introduce a three-stage meta-learning approach to dynamically update MvCoM and use the MvCoM for recognition model training (Sec. 3.2.2).

3.1. Interpreting Margin as Sampling Importance

Traditional methods [21, 37] seek to address imbalanced data distributions by introducing a sampling importance weight σ_{y_i} to weigh each sample loss term so as to compensate each sample's imbalance level:

$$\min_{\Omega} \quad \frac{1}{N}\sum_{j=1}^{N} \sigma_{y_j} \mathcal{L}(f(x_j;\Omega), y_j), \tag{1}$$

where N is the number of classes, $\mathcal{L}$ is a general loss function, $\{(x_j, y_j)\}^N$ denotes training set with x_j as sample and y_j as class label. $f(x;\Omega)$ is a convolutional neural network (CNN) backbone generated feature as commonly used in deep face recognition, where Ω stands for the network parameters. The class-level weight σ_{y_j} is designed to compensate for class imbalances. If a class has few samples which is long-tailed, the weight should be large such that its contribution to the overall objective can suitably penalize the model to account for this long-tailed condition.

Without loss of generality, we consider Cosine Loss [50] as the $\mathcal{L}$ in Eqn. 1, which has seen significant recent success in face recognition:

$$\mathcal{L}_{cos} = -log\frac{e^{\mathbf{s}\cdot\cos\theta_{y_j}-\bar{\mathbf{m}}}}{e^{\mathbf{s}\cdot\cos\theta_{y_j}-\bar{\mathbf{m}}} + \sum_{y_k\neq y_j}^{C} e^{\mathbf{s}\cdot\cos\theta_{y_k}}}. \tag{2}$$

In Eqn. 2, $\cos\theta_{y_j}$ is the inner product between the feature vector $f(x_j;\Omega)$ and j-th class template w_{y_j}, that is, $\cos\theta_{y_j} = w_{y_j}^T f(x_j;\Omega)$. The margin $\bar{m}$ is set as a positive constant to squeeze the inner product $\cos\theta_{y_j}$ such that the separating hyper-planes are pushed further away and $\mathbf{s}$ is a scale factor to facilitate training convergence. Combining

Eqn. 2 with Eqn. 1, we obtain:

$$\min_{\Omega} \frac{1}{N} \sum_{y_j=1}^{N} -log \frac{\left[e^{\mathbf{s}\cdot\cos\theta_{y_j}-\bar{\mathbf{m}}}\right]^{\sigma_{y_j}}}{\left[e^{\mathbf{s}\cdot\cos\theta_{y_j}-\bar{\mathbf{m}}} + \sum_{y_k\neq y_j}^{C} e^{\mathbf{s}\cdot\cos\theta_{y_k}}\right]^{\sigma_{y_j}}} \tag{3}$$

When near convergence, the denominator in Eqn. 3 is close to constant $[e^{\mathbf{s}-\bar{\mathbf{m}}} + C - 1]^{\sigma_{y_j}}$ as $\theta_{y_j} \approx 0, \theta_{y_k} \approx \frac{\pi}{2}$. Then, the decisive component is generally the numerator, which is further rearranged as the following:

$$\begin{aligned}\left[e^{\mathbf{s}\cdot\cos\theta_{y_j}-\bar{\mathbf{m}}}\right]^{\sigma_{y_j}} &= e^{\sigma_{y_j}\mathbf{s}\cdot\cos\theta_{y_j}-\sigma_{y_j}\bar{\mathbf{m}}} \\ &= e^{\mathbf{s}'\cdot\cos\theta_{y_j}-\mathbf{m}_{y_j}}\end{aligned} \tag{4}$$

Replacing the numerator of the loss in Eqn. 3 with Eqn. 4, it can be shown that Eqn. 3 is equivalent to a modified Cosine Loss $\mathcal{L}_{cos}$, where $\mathbf{s}' = \sigma_{y_j}\mathbf{s}$ and $\mathbf{m}_{y_j} = \sigma_{y_j}\bar{\mathbf{m}}$ are defined as the new scalar and new margin, respectively. In contrast to cosine loss defined in Eqn. 2, in the new formulation, both the scale and margin are proportional to the class-level sampling weight σ_{y_j}. *Therefore, the importance sampling problem can be interpreted as learning the per-class margin* $\mathbf{m}_{y_j}$, and $\mathbf{s}'$ can be derived as $\mathbf{s}' = \frac{\mathbf{m}_{y_j}}{\bar{\mathbf{m}}}\mathbf{s}$.

3.2. Multi-variation Cosine Margin Loss

Cosine Loss assumes a constant margin that assigns equal importance to all the data, which inevitably pushes the model focus more on the head classes and leads to biased estimation. Meanwhile, class-level importance cannot account for intra-class variations. Prior work has considered a single weight reflected from the recognition loss, to indicate sample importance [20]. However, such weight does not distinguish bias from other factors such as label noise or outlier. In contrast, we search for explicit anchors that correspond with a few known and important causes of distribution bias, namely, *class volume, ethnicity, head pose, blur and occlusion*. Thereby, we train a classifier for each variation to quantify bias corresponding to it.

Thus, we propose the sample-level multi-variation cosine margin (MvCoM) to flexibly capture sample-level variations. Formally, we model our MvCoM $\mathbf{m}_{y_j,j}$ in an additive manner by combining the class-volume margin $\mathbf{m}^{cls}_{y_j}$, and a set of margin residual terms $\mathbf{r}^k_j$ representing the importance of each variation k. The additive assumption stems from each variation equally and independently contributing to the sample importance. With experimental trial, we find that class-volume factor can be stably estimated by statistics prior [2]. Thus, starting with the prior, we accumulate other factors' importance contribution to form the overall margin:

$$\begin{aligned}&\mathbf{m}_{y_j,j} = m^{cls}_{y_j} + \sum_k \lambda_k \mathbf{r}^k_j \\ &k \in \{vol., eth., pose, blur, occ.\}\end{aligned} \tag{5}$$

where $vol.$, eth, $pose$, $blur$ and $occ.$ stand for per-class data volume, ethnicity, head pose, blur level and occlusion variations. Notice that other variations may be similarly considered if necessary. λ_k is a weighting factor for each variation (empirically we set all the λ_k as 1). The overall objective is:

$$\mathcal{L}_{MvCoM} = -log \frac{e^{\mathbf{s}\cdot cos\theta_{y_j}-\mathbf{m}_{y_j,j}}}{e^{\mathbf{s}\cdot cos\theta_{y_j}-\mathbf{m}_{y_j,j}} + \sum_{y_k\neq y_j}^{C} e^{s\cdot cos\theta_{y_k}}} \tag{6}$$

The method effectiveness is highly depended on the MvCoM estimation. Ideally, MvCoM is dynamically updated during training to highlight the samples with variations that are less present in the training distribution. The remainder is to estimate each component of MvCoM: the class-volume margin $\mathbf{m}^{cls}_{y_j}$ and variation-aware margin residual $\mathbf{r}^k_j$.

3.2.1 Estimate Class-volume Margin

Following [2], we use the class-wise statistics as the prior for the class-volume margin:

$$m^{cls}_{y_j} = \frac{\alpha}{n^{1/4}_{y_j}} \tag{7}$$

where j is sample index, α is a hyper-parameter (0.45 used in the experiment) and n_{y_j} is class y_j volume.

3.2.2 Meta-learn Variation-aware Margin Residual

To estimate the residual terms of MvCoM in Eqn. 5, we leverage a learning-to-learn framework [9, 21], by considering each sample's long-tailed factor variations within a training batch $\{(x_j, y_j, \mu^k_j)\}^{|B|}$, where y_j is the class label, μ^k_j the variation k's label and $|B|$ the batch size. This is achieved by introducing variation classifiers to predict per-sample long-tailed factors. Further we introduce a meta-learning face dataset, which is a typical "in-the-wild" distribution and independent from training set distribution. With the online mined samples from this meta-learning set that present largest variation difference to the current training batch, we meta-update the proposed MvCoM and further utilize it to update the face recognition model.

Long-tailed Variation Classification To quantitatively indicate how a training sample is biased alongside each of the pre-defined variations, we introduce the variation classifiers to predict the variation level. Given our choice of the four variations above, we set up four independent classifiers $g(\cdot; v_k)$ as shown in Fig. 2, where v_k indicates the classifier parameter. For example, we label the ethnicity information of our training set MS-Celeb-1M into African American, Caucasian, East Asian and South Asian categories to conduct a 4-way classification[1]. Other variations' labeling

[1] We omit other too limited data volume ethnicities such as Latino in the training set, to guarantee the classifier's unbiasedness.

are explained in Sec. 4 "Variation Augmentation". A cross entropy loss is used to update the variation classifiers:

$$\mathcal{L}^k_{var}(x_j, \Omega, v_k, \mu^k_j) = \sum_j \mathcal{L}_{ce}(g(f(x_j;\Omega); v_k), \mu^k_j) \quad (8)$$

where $\mathcal{L}^k_{var}$ is the cross-entropy loss for variation task k and μ^k_j is the variation label for sample j. The variation classifiers are trained on the same training data as face recognition. The difference is we re-balance the original imbalanced data according to the variation labels, i.e., forcing the volume for each variation level the same by increasing the occurrence of the long-tailed data, denoted as $\hat{T}_k$ in Algorithm 1. This data re-balance cannot be directly applied to face recognition training, because the joint multiple variations' re-balance is not trivial. Notice that the re-balanced $\hat{T}_k$ is based on each single variation k. In this way, we maximally guarantee the variation classifiers are trained balanced. Hence, we make sure in the later meta-learning stage, the imbalance is from the training batch, not from the variation classifiers.

Online Meta-learning Batch Construction We posit that samples that share the similar long-tailed variation result in similar classifier logits $g(f(x_j;\Omega))$. To reshape the training set distribution to be more balanced, we search for the distribution that is *complementary* to the current training distribution. This is achieved by selecting samples from the meta-learning set V that have the largest logit distance from the current training batch. Accordingly, the objective to search for such samples compares the logit distance:

$$x_m : \underset{x_m \in V}{\text{argmax}} \|g(f(x_m;\Omega); v_k) - g(f(x_j;\Omega); v_k)\|_2 \quad (9)$$

where x_j is from training batch B and x_m is from meta-learning batch V. $g(\cdot; v_k)$ are variation k's classifier logits. By mining the meta-learning batches, the original training batch's bias information is fed back to meta-update MvCoM.

Meta-learning Optimization for MvCoM

1) Pseudo Recognition Model Update. At each iteration t, we uniformly sample a batch B from the training data and feed it to update the recognition model parameters Ω with margin $\boldsymbol{m}_{j,t}$:

$$\tilde{\Omega}^{t+1}(\boldsymbol{m}_{j,t}) : \Omega^t - \eta \frac{\partial \sum_{k,j\in T} \mathcal{L}_{MvCoM}(f(x_j;\Omega^t), y_j; \boldsymbol{m}_{j,t})}{\partial \Omega} \quad (10)$$

where sample x_j is from training set T. From this procedure, we see that by adjusting margin $\boldsymbol{m}_{j,t}$, we adjust the overall loss $\mathcal{L}_{MvCoM}$ and it backpropagates to update the model parameter $\tilde{\Omega}^{t+1}$. Thus, $\tilde{\Omega}^{t+1}$ is a function of $\boldsymbol{m}_{j,t}$ while Ω^t and $\boldsymbol{m}_{j,t}$ are independent.

2) Margin Residual Meta-Update. We exploit the online sample mining described by Eqn. 9 to prepare the meta-learning batch from V. Given that the current $\boldsymbol{m}_{j,t}$ is sub-optimal due to the original biased training data, we seek to send the meta-learning batch to the variation classifiers, with the pseudo-updated $\tilde{\Omega}^{t+1}$, to reduce the classifiers prediction error to meta-learn the margin $\boldsymbol{m}_{j,t+1}$. This $\boldsymbol{m}_{j,t+1}$ compensates the previous step data bias to achieve lower variation classification error. Further acknowledging that $\tilde{\Omega}^{t+1}$ is the function of $\boldsymbol{m}_{j,t}$, we meta-update $\boldsymbol{m}_{j,t+1}$ as:

$$\boldsymbol{m}_{j,t+1} : \boldsymbol{m}_{j,t} - \tau \frac{\partial \sum_{k,j\in V} \mathcal{L}^k_{var}(x_j, \tilde{\Omega}^{t+1}(\boldsymbol{m}_{j,t}), v_k, \mu^k_j)}{\partial \boldsymbol{m}_{j,t}} \quad (11)$$

As the class-level margin prior $\mathbf{m}^{cls}_{y_j}$ is unchanged from $\boldsymbol{m}_{j,t}$ to $\boldsymbol{m}_{j,t+1}$, Eqn. 11 is effectively meta-updating the margin residual from $\boldsymbol{r}_{j,t}$ to $\boldsymbol{r}_{j,t+1}$ through Eqn. 5. As a result, the updated margin $\boldsymbol{m}_{j,t+1}$ should be better than the previous update $\boldsymbol{m}_{j,t}$, in the sense that it results in smaller variation-level classification errors on the meta-learning set by balancing the long-tailed training distribution for multiple factors of variation.

3) Real Recognition Model Update. We apply the obtained new importance margin $\boldsymbol{m}_{j,t+1}$ to conduct the update for the actual recognition model:

$$\Omega^{t+1} : \Omega^t - \eta \frac{\partial \sum_{k,j\in T} \mathcal{L}_{MvCoM}(f(x_j;\Omega^t), y_j; \boldsymbol{m}_{j,t+1}))}{\partial \Omega} \quad (12)$$

Algorithm 1 Multi-variation Cosine Margin meta-learning

Require: Training set T, meta-learning set V
Require: Learning rates η and τ, iteration steps t_1 and t_2
  for $t = 1, 2, \cdots, t_1$ **do**
    Sample a mini-batch B from the training set T
    Compute loss $\mathcal{L}_{MvCoM}$ with Eqn. 6
    Update $\Omega \leftarrow \Omega - \eta \nabla_\Omega \mathcal{L}_{MvCoM}(\boldsymbol{m}^{cls}_y)$
  end for
  for $t = t_1 + 1, \cdots, t_1 + t_2$ **do**
    Sample a mini-batch B from the training set T
    Set $r^k_j \leftarrow 0, \forall j \in B$, denote by $\boldsymbol{r}^k := \{r^k_j, j \in B\}$
    Set $\boldsymbol{m}_t \leftarrow \sum_k \boldsymbol{r}^k + \boldsymbol{m}^{cls}_{\boldsymbol{y}}$
    Update $\tilde{\Omega}(\boldsymbol{m}_t) \leftarrow \Omega - \eta \nabla_\Omega \mathcal{L}_{MvCoM}(\boldsymbol{m}_t)$ with Eqn. 10
    for $k = 1 : 4$ **do** ▷ 4 factors of variations
      Sample B_v from V with Eqn. 9.
      $\boldsymbol{r}^k \leftarrow \boldsymbol{r}^k - \tau \nabla_{\boldsymbol{r}^k} \mathcal{L}_{var}(\tilde{\Omega}(\boldsymbol{m}_t))$ with Eqn. 11.
    end for
    Set $\boldsymbol{m}_{t+1} \leftarrow \sum_k \boldsymbol{r}^k + \boldsymbol{m}^{cls}_{\boldsymbol{y}}$
    Update $\Omega \leftarrow \Omega - \eta \nabla_\theta \mathcal{L}_{MvCoM}(\boldsymbol{m}_{t+1})$ with Eqn. 12
  end for
  Update $g(f(.); v_k)$ (Eqn. 8) with variation re-balanced $\hat{T}_k$

The overall procedure is summarized in Algorithm 1. Although our meta-learning shares the high-level structure

as [21], we consider multiple branches for $\boldsymbol{r}_j^k$ to estimate the residuals instead of a single weight. Moreover, [1, 21] consider only the class-level importance weighting, whereas our method considers the finer-scale sample-level importance. Another difference from [21] is that we leverage an independent meta-learning set which has no prior distribution correlation with the training set, while they use a held-out set which shares the same distribution as the training set.

4. Implementation Details

We use MS-Celeb-1M [14] with the clean list from ArcFace [6] for training data. For the meta-training set, we adopt VGGFace2 [3] and exclude the duplicate identities to prevent additional benefit for the training. The baseline models are trained with CosFace loss [50] for 30 epochs with empirically fixed margin $m = 0.35$. After pre-training, we discard the classifier and fine-tune the models with the proposed framework for 18 epochs to ensure convergence.

Variation Augmentation We use mechanical turk to label ethnicity in the training set, including African American, Caucasian, East Asian and South Asian. For head pose, following the pose angle setting in MultiPIE [12], we group every $30°$ as one class and thus obtain 7 classes ranging from $-90°$ to $90°$. For blur, we apply Gaussian kernel with four different kernel sizes $(3, 7, 11, 15)$ to augment the training images. For occlusion, we adopt five different block sizes $(5, 11, 17, 23, 29)$ to randomly black out the training images with the specific size.

Complexity We use the modified 100-layer ResNet [17] as the backbone. All the variation classifiers are linear classifiers. Compared to the CosFace baseline, our framework newly introduces four variation classifiers. But it almost does not increase the network complexity as each variation classifier is less than 10-way. The time complexity for our training is nearly twice longer than the baseline training due to one additional feed forward and a meta-learning step. Since testing only utilizes the recognition model, the runtime for inference is the same as CosFace.

5. Experiments

In this section, we organize the experiments as: **(1)** Extensive ablation study over the five variation factors, and compare to the baseline CosFace [50]. **(2)** Evaluation on challenging benchmarks that are prototypical on variations, i.e., RFW [53] for ethnicity, CFP [40] and CP-LFW [67] for head poses, IJB-A [23] for video blur and OC-LFW for occlusion. **(3)** Evaluation on general face recognition benchmarks LFW [19] and MegaFace [22] **(4)** Visualization of sample images with the predicted margin residuals alongside all the variation factors. **(5)** Further insights on the margin-weighted validation loss, embedding distributions, and the magnitude of the margin residual.

Method	OC-LFW	CFP-FP	RFW					
			CA	AF	EA	IN	Avg ↑	Bias ↓
CosFace*	94.41	98.16	99.01	97.62	97.20	97.96	97.94	0.67
Ours (single)	94.52	98.35	99.06	97.90	**97.83**	98.23	98.25	**0.49**
Ours (all)	**94.83**	**98.41**	**99.16**	**98.06**	97.78	**98.28**	**98.32**	0.51

Table 1. Ablation study on variation-specific benchmarks, OC-LFW for occlusion, CFP-FP for head pose, and RFW for ethnicity where CA, AA, EA and IN are abbreviated for Caucasian, African American, East Asian and Indian respectively. $*$: self-implemented CosFace as baseline. "Ours (single)" means "Ours (occlusion)", "Ours (pose)", "Ours (ethnicity)" respectively for each variation-spcific dataset. "Ours (all)": adding all the proposed variations for MvCoM.

Method	IJB-A (Vrf)	
	FAR@0.01%	FAR@0.001%
CosFace*	97.13	93.22
Ours (ethnicity)	97.24	94.91
Ours (pose)	97.27	95.12
Ours (blur)	97.42	95.58
Ours (occlusion)	97.25	95.21
Ours (ethnicity + pose)	97.20	95.12
Ours (ethnicity + pose + blur)	97.45	95.65
Ours (all)	**97.46**	**95.69**

Table 2. Ablation study on in-the-wild IJB-A dataset with multiple variations. $*$: self-implemented CosFace serves as baseline for all our ablation methods for a fair comparison.

5.1. Studies on Variation-Specific Benchmarks

While the proposed MvCoM complements various recognition losses, in this evaluation, we use CosFace as the baseline. All the ablations are built on top of this baseline for fair comparison. To highlight each component's function, we evaluate on challenging datasets prototypical of specific variations. We use RFW [53] for ethnicity, CFP-FP [40] for head poses, and OC-LFW for occlusion variation. We also evaluate on IJB-A as an in-the-wild dataset that incorporates multiple variations for all our ablation methods.

Benchmark protocols. LFW verification protocol is used for RFW, CFP-FP, IJB-A and OC-LFW. For CFP, we focus on the frontal-profile (FP) protocol.

MvCoM is robust to occlusions. In Table 1, OC-LFW is an occlusion evaluation protocol of LFW [19] that contains more than 13,000 images from 5749 identities. For each verification pair, we randomly set occlusion masks on one of the images, and conduct the same verification protocol as LFW. Although the performance on LFW is saturated, all methods only achieve under 95% accuracy on OC-LFW. We observe that our method with single variation already outperforms the baseline. By adding all variations, the accuracy further increases as more variation factors provide a more complete regularization for representation learning.

MvCoM handles large poses. CFP-FP [40] consists of face image pairs with one image of large pose variation, and most of the image pairs are with high resolution. In Table 1, the single margin ablation outperforms the baseline clearly. While "Ours (all)" is generally better than "Ours (single)". We observe the same trend as in OC-LFW, which consis-

Method	OC-LFW	CP-LFW	CFP-FP	IJB-A (Vrf)		RFW					
				FAR=0.001%	FAR=0.01%	CA	AA	EA	IN	Avg ↑	Bias ↓
ArcFace [6] † (CVPR'19)	94.56	92.08	98.37	93.7	94.2	98.80	97.48	96.80	97.38	97.61	0.84
URFace [42] (CVPR'20)	94.60	92.31	98.30	95.0	96.3	98.35	96.76	96.10	96.63	96.96	0.96
CurricularFace [20] (CVPR'20)	-	**93.13**	98.37	-	-	-	-	-	-	-	-
MagFace [30] (CVPR'21)	-	92.87	98.46	-	-	-	-	-	-	-	-
DebFace [10] (ECCV'20)	-	-	-	-	-	95.95	93.67	94.33	94.78	94.68	0.83
RL-RBN [52] (CVPR'20)	-	-	-	-	-	97.08	94.87	95.57	95.63	95.79	0.93
CIFP [59] (CVPR'21)	-	-	-	-	-	97.08	94.87	95.57	95.63	95.79	0.93
GAC [11] (CVPR'21)	-	-	-	-	-	97.60	97.03	95.65	96.82	96.78	0.82
DAM [26] (ICCV'21)	-	-	-	-	-	96.30	94.51	94.31	95.20	95.08	0.78
CosFace [50]* (CVPR'18)	94.41	92.06	98.16	93.2	97.1	99.01	97.62	97.20	97.96	97.94	0.67
CB-CosFace [37]* (ICML'18)	94.44	92.04	98.24	94.6	97.2	99.03	**98.23**	97.36	97.83	98.10	0.61
LDAM-CosFace [2]* (NeuIPS'19)	94.54	92.05	98.31	94.5	97.2	98.93	97.80	97.23	97.50	97.86	0.65
MetaCW [21]* (CVPR'20)	94.48	92.06	98.28	94.1	97.2	99.13	97.86	97.73	98.11	98.20	0.55
MvCoM-URFace (ours)	**94.92**	92.86	**98.47**	**96.0**	**97.6**	98.85	97.18	97.15	96.98	97.54	0.76
MvCoM-CosFace (ours)	94.83	92.75	98.37	95.7	97.5	**99.16**	98.06	**97.78**	**98.28**	**98.32**	**0.51**

Table 3. Challenging variation-specific face recognition benchmarks comparison. "-": the authors did not report the performance on the corresponding protocol. "*": self-implemented methods. "†" indicates the testing performance by using the released models from corresponding authors. In RFW (BUPT-BalancedFace), CA, AA, EA and IN are abbreviated for Caucasian, African American, East Asian and Indian respectively.

Method	LFW	MF1	
		Rank1	Veri.
CenterFace [56]	99.28	65.23	76.52
SphereFace [27]	99.42	75.77	89.14
ArcFace [6]	**99.83**	81.03	96.98
URFace [42]	99.75	79.10	94.92
CurriculumFace [20]	99.80	**81.26**	**97.26**
DomainBlancing [1]	99.78	-	-
MagFace [30]	**99.83**	-	-
CosFace [50]*	99.73	80.03	95.54
CB-CosFace [37]*	99.81	80.18	95.75
LDAM-CosFace [2] *	99.75	80.73	96.78
MetaCW [21]*	99.78	80.32	96.22
MvCoM-URFace (Ours)	99.78	80.63	96.28
MvCoM-CosFace (Ours)	99.80	**81.30**	**97.22**

Table 4. General face recognition benchmarks comparison. The MegaFace verification rates are computed at FAR=0.0001%. "*": self-implemented methods. "-": the authors did not report the performance on the corresponding protocol. Notice that MegaFace1 is based on uncleaned protocol, of which numbers are lower than the cleaned protocol.

tently demonstrates that by adding the proposed MvCoM, the accuracy is significantly improved.

MvCoM is less biased with respect to ethnicity. RFW consists of four races (Caucasian, East Asian, African American, Indian) data from MS-Celeb-1M to study ethnicity bias in face recognition. We have excluded the identities from RFW that are duplicated in MS-Celeb-1M. In Table 1 RFW column, we find that while both the CosFace baseline and our method achieve strong accuracy, ours is slightly higher. More importantly, following [11], we highlight the *bias*, defined as the standard deviation over the accuracy of four ethnicity subsets. The bias across CA, AA, EA and IN is much smaller for our method, showing the effectiveness of our learned margin that leads to more balanced performance across different ethnicities.

MvCoM is accurate across diverse variations. IJB-A (Vrf) is an in-the-wild dataset with multiple long-tailed variations. In Table 2, we observe that all the single factor ablations are better than the CosFace baseline, indicating that IJB-A contains such long-tailed variations and our method indeed alleviates the issue. Further, we notice that "Ours (blur)" is better compared to other single variation ablations by 0.2%, and "Ours (ethnicity+pose+blur)" is better than "Ours (ethnicity+pose)" by more than 0.2%, which is consistent to the observation that IJB-A is a low-quality surveillance video setting with large blur degradation.

5.2. Evaluation on Challenging Benchmarks

MvCoM captures long-tailed variations well. We compare to both general state-of-the-arts and long-tailed re-weighting specific methods on challenging variation-specific datasets, across the top three sets of rows in Table 3. In general, our method shows consistently better performance over other methods, e.g., 0.3% higher than second best on OC-LFW, 1.0% higher than second best on IJB-A FAR$=0.001\%$. While re-weighting based methods in the third set show strong performance especially on RFW, our method achieves a clearly lower bias of 0.51, defined as standard deviation of accuracy reported across the four ethnicity subsets [11]. In addition to the performance advantages compared to re-weighting methods, our joint consideration of multiple variation factors better represents the long-tailed distribution.

MvCoM is complementary to face recognition backbones. Interestingly, we observe that MvCoM can combine with different recognition architectures such as CosFace and URFace in Table 3. When comparing MvCoM-CosFace and MvCoM-URFace to their baselines, we see clear improvements, which suggests that our MvCoM can complement a variety of recognition frameworks.

5.3. Evaluation on General Benchmarks

MvCoM retains accuracy on more balanced test data. We compare to the state-of-the-arts on general face recognition benchmarks with limited variations, namely LFW [19] and MegaFace [22]. We self-implement CosFace and use it as

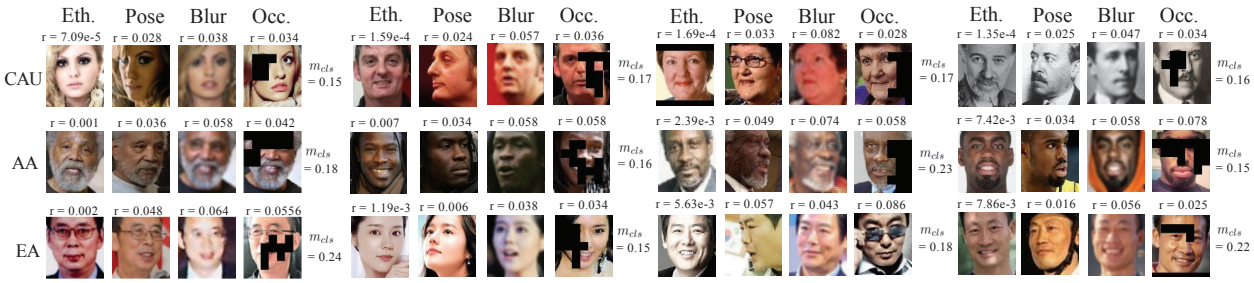

Figure 3. Sample level margin visualization across all the factors. Larger margin corresponds to more tailed class.

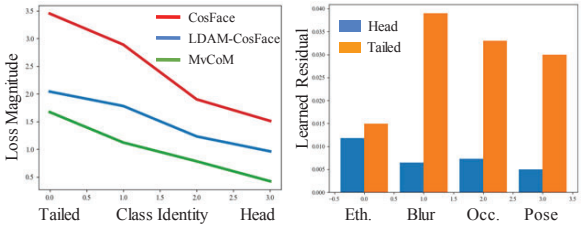

Figure 4. Left: Curve of validation loss magnitude versus the tail to head classes. Our MvCoM (in green) shows significantly lower loss. Right: Histogram of the learned residual magnitude over the long-tailed variations. Tailed classes' MvCoM is clearly larger.

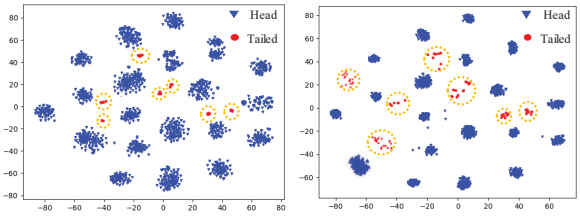

Figure 5. Left: Original head and tailed embedding distribution. Right: MvCoM modulated head and tailed embedding distribution.

the backbone to further implement the Class-Balance CosFace (CB-CosFace) [37], Label-distribution-aware margin loss (LDAM-CosFace) [2], and Meta Conditional Weights (MetaCW) [21]. The main purpose of this evaluation is to show that, our method is consistently among the top, while less imbalanced testing data does not degrade our performance. In Table 4, "Ours" achieves close to best on LFW and the first on MegaFace challenge 1 with uncleaned protocol. Note that while our method uses an additional meta-learning set for training, it is only utilized to feedback the importance weight and no identity information from this auxiliary set is used to train the recognition model.

5.4. Further Insights

MvCoM learns meaningful per-sample margins. We randomly show identities from MS-Celeb-1M in Fig. 3 (more in the supplement). Images of different variations (each column) within the same identity (each row) are presented. We consistently observe that the margin residuals for the head classes are smaller, while those for tailed classes are relatively larger, which suggests the learned MvCoM works as expected to emphasize on the tailed class samples.

Visualization of margin modulation. We verify whether the learned MvCoM can compensate the distribution imbalance and whether the loss with the learned margin drops more significantly. On MS-Celeb-1M, we count the class volume to group the identities and form x-axis of Fig. 4 "Left", ranging from tailed to head. The y-axis is MvCoM loss in Eqn. 6. As expected, our method achieves significantly lower loss compared to LDAM-CosFace [2]. In Fig. 4 "Right", we compare the learned residual between head and tailed classes across all the variations. The residual for tailed classes is consistently higher than head classes across all the variations. Moreover, we randomly pick a set of head and tailed classes and visualize the feature distribution in Fig. 5. Compared to original distribution (left), the feature space with our MvCoM modulation (right) effectively enlarges tailed classes' variance and shrinks head classes'.

6. Discussions and Conclusions

Our models are trained on a public dataset. Consent is obtained by the dataset providers. We will remove any subject image where privacy concern is not properly addressed. Though face recognition may potentially be used for unlawful surveillance or discrimination, our work has the positive benefit of alleviating a critical ethical concern with biases in face recognition, which have been observed to have detrimental consequences in many societal outcomes. The limitation of our work is the training efficiency, yet we trade off the training efficiency for better model efficacy.

In this work, we explicitly handle multiple bias factors in face recognition. This is in contrast to prior works that mostly focus on single bias factor. A learning to learn scheme is proposed to provide the training batch biased distribution feedback in the form of a novel sample-level Multi-variation Cosine Margin (MvCoM), which can be orthogonally equipped with many recognition losses such as Cosine Loss. Empirical results demonstrate our method's top performance on general benchmarks, and clear advantage on challenging variation-specific benchmarks. Avenues for future work include applying the proposed MvCoM for wider range of data bias problems.

References

[1] Dong Cao, Xiangyu Zhu, Xingyu Huang, Jianzhu Guo, and Zhen Lei. Domain balancing: Face recognition on long-tailed domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5671–5679, 2020. 2, 6, 7

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019. 1, 2, 4, 7, 8

[3] 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. 6

[4] Aruni Roy Chowdhury, Xiang Yu, Kihyuk Sohn, Erik Learned-Miller, and Manmohan Chandraker. Improving deep face recognition by clustering unlabeled faces in the wild. In *ECCV*, 2020. 1

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2

[6] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CVPR*, 2019. 1, 2, 6, 7

[7] Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1851–1860, 2017. 2

[8] Masoud Faraki, Xiang Yu, Yi-Hsuan Tsai, Yumin Suh, and Manmohan Chandraker. Cross-domain similarity learning for face recognition in unseen domains. In *CVPR*, 2021. 2

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017. 2, 3, 4

[10] Siyue Gong, Xiaoming Liu, and Anil K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *ECCV*, 2020. 1, 7

[11] Siyue Gong, Xiaoming Liu, and Anil K. Jain. Mitigating face recognition bias via group adaptive classifier. In *CVPR*, 2021. 1, 7

[12] R. Gross, I. Matthew, J.F. Cohn, T. Kanade, and S. Baker. MultiPIE. *Image and Vision Computing*, 2009. 6

[13] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z. Li. Learning meta face recognition in unseen domains. In *CVPR*, 2020. 3

[14] 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, 6

[15] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 2

[16] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 1, 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[18] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. 2

[19] 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, 7

[20] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 4, 7

[21] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *CVPR*, 2020. 1, 2, 3, 4, 6, 7, 8

[22] 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, 7

[23] 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. 6

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[25] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11947–11956, 2019. 2

[26] Jiaheng Liu, Yudong Wu, Yichao Wu, Chuming Li, Xiaolin Hu, Ding Liang, and Mengyu Wang. Dam: Discrepancy alignment metric for face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3814–3823, 2021. 2, 7

[27] 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

[28] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 2

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

[30] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *CVPR*, 2021. 7

[31] Gaurav Mittal, Chang Liu, Nikolaos Karianakis, Victor Fragoso, Mei Chen, and Yun Fu. Hyperstar: Task-aware

hyperparameters for deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8736–8745, 2020. 3

[32] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International Conference on Machine Learning*, 2017. 3

[33] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. In *arXiv preprint arXiv:1803.02999*, 2018. 3

[34] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016. 2

[35] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N. Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017. 1, 2

[36] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017. 3

[37] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018. 1, 2, 3, 7, 8

[38] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of Machine Learning Research*, 2016. 3

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

[40] 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

[41] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer, 2016. 1, 2

[42] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K Jain. Towards universal representation learning for deep face recognition. In *CVPR*, 2020. 1, 2, 7

[43] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems*, 2017. 3

[44] Kihyuk Sohn, Wenling Shang, Xiang Yu, , and Manmohan Chandraker. Unsupervised domain adaptation for distance metric learning. In *ICLR*, 2019. 1

[45] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014. 2

[46] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. 3

[47] 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

[48] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Neural Information Processing Systems*, 2016. 3

[49] Feng Wang, Xiang Xiang, Jian Cheng, and Alan L Yuille. Normface: l_2 hypersphere embedding for face verification. *ACM MM*, 2017. 2

[50] 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, 6, 7

[51] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 2018. 2

[52] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9322–9331, 2020. 1, 7

[53] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 692–702, 2019. 6

[54] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *AAAI*, 2020. 1

[55] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017. 2

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

[57] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011. 1

[58] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017. 2

[59] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 578–586, 2021. 2, 7

[60] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017. 1

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

[62] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017. 2

[63] Yaobin Zhang, Weihong Deng, Yaoyao Zhong, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. Adaptive label noise cleaning with meta-supervision for deep face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 15065–15075, 2021. 3

[64] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, Shuicheng Yan, and Jiashi Feng. Towards pose invariant face recognition in the wild. In *CVPR*, 2018. 1

[65] 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. 2

[66] Kai Zhao, Jingyi Xu, and Ming-Ming Cheng. Regularface: Deep face recognition via exclusive regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1136–1144, 2019. 2

[67] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying crosspose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, pages 18–01, 2018. 6

[68] Yutong Zheng, Dipan K. Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *CVPR*, 2018. 2