LEARNING TO LEARN ACROSS DIVERSE DATA BIASES IN DEEP FACE RECOGNITION

Anonymous authors
Paper under double-blind review

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 with respect to the number of instances per identity. In this paper, we show that besides the imbalanced class volume distribution, other variations such as ethnicity, head pose, occlusion and blur can also significantly affect accuracy. To address the problem, we propose a sample level weighting approach called Multi-variation Cosine Margin (MvCoM) which orthogonally enhances the conventional cosine loss function 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 Schroff et al. (2015); Wen et al. (2016); Liu et al. (2017); Wang et al. (2018); Deng et al. (2019), with strong results on public benchmarks Huang et al. (2007); Lior Wolf & Maoz (2011). However, real-world training data distributions are usually imbalanced or long-tailed, whereby a method trained with uniform sampling of the training data leads to degraded accuracy. Since it is impractical to collect data that sufficiently covers a wide variety of the imbalance factors or variations, 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. Previous approaches distinguish long-tailed classes (minority in samples) from head classes (majority in samples) to mitigate the volume imbalance. However, factors besides volume, such as subject ethnicity and head pose are also well-known contributors to data imbalances, as shown in Figure 1, while imaging artifacts such as blur and occlusion present additional factors of variations. We hypothesize that dealing with such additional *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 Ren et al. (2018); Cao et al. (2019); Jamal et al. (2020). This is limited in practice, as different images from the same person would likely differ in their importance (e.g., frontal and profile views). Thus, we hypothesize considering *sample-level variation* instead of class-level leads to better training.

To handle data imbalance, classical methods He & Garcia (2009); Shen et al. (2016) introduce re-weighted loss functions by assigning higher loss weights to long-tailed classes and lower weights to head classes. However, assigned weights are usually either fixed based on prior statistics or obtained by sophisticated design choices Cui et al. (2019). Active learning Grabinger & Dunlap (1995) or reinforcement learning Mnih et al. (2015) are potential avenues for greater adaptivity. However, a drawback of active learning is that the feedback from the recognition performance to parameter updates is not differentiable. Similarly, reinforcement learning suffers from slow convergence due to sample inefficiency caused by non-differentiability of the forward path and use of the Reinforce technique. As an alternative, this work considers *meta-learning* Finn et al. (2017a) as a differentiable mechanism to iteratively learn sample-level importance weights and use them to update the face

recognition model. This allows a plug-in mechanism which is fully compatible with popular losses such as *cosine loss* Wang et al. (2018).

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 Ren et al. (2018); Jamal et al. (2020) 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). We make a specific choice to apply the margin to the well-known cosine loss or its variants, terming it 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 instance level weighting, we leverage a held-out meta-learning set (no identity overlap with the training set), where hard sample mining is applied to select the samples that are most different from the current training batch. The meta-learner is updated with the hard samples in every iteration to provide the MvCoM that is most effective in regularizing the recognition loss for the current training batch (Sec. 3.2.2). Figure 2 illustrates the proposed approach.

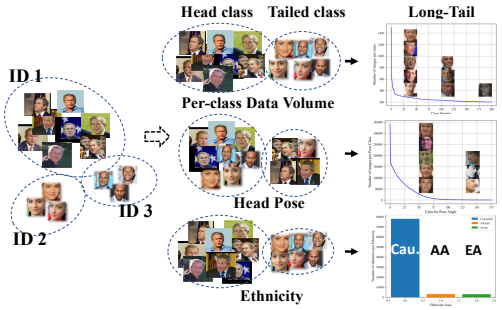


Figure 1: While traditional methods only consider per-class data volume as a factor for long-tailed effects, several other factors such as head pose and ethnicity manifest as long-tailed effects in MS-Celeb-1M Guo et al. (2016). Further, samples from the same identity can show different variations – for example, images from ID1 show both frontal and profile poses – indicating that accounting for identity or class-level variation is not sufficient. Our MvCoM accounts for sample-level variations while jointly considering multiple long-tailed variations in face recognition.

Our technical contributions are summarized as:

- To the best of our knowledge, we are the first to address multiple factors that contribute to distribution imbalance besides data volume, such as ethnicity and head pose, 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 for distribution imbalance.
- We propose a meta-learning based differentiable mechanism to jointly learn the importance margin for each imbalance variation.

Our empirical contributions are summarized as:

- Extensive experiments on five challenging recognition benchmarks show that our method can consistently outperform prior state-of-the-art to mitigate data imbalances across minority ethnicities, non-frontal head poses, blurriness and occlusion.
- The proposed MvCoM complements various backbones such as CosFace and URFace (see Table 3, demonstrating the wide applicability to different face recognition methods.
- Sampling importance analysis in Figure 3 verifies that our MvCoM indeed assigns more weights to long-tail factors, which leads to overall smaller loss magnitudes.

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 Taigman et al. (2014); Sun et al. (2014) were among the first to surpass human-level accuracy. A series of recent works Wen et al. (2016); Wang et al. (2017a); Liu et al. (2017); Wang et al. (2018); Zheng et al. (2018); Deng et al. (2019); Zhao et al. (2019) 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

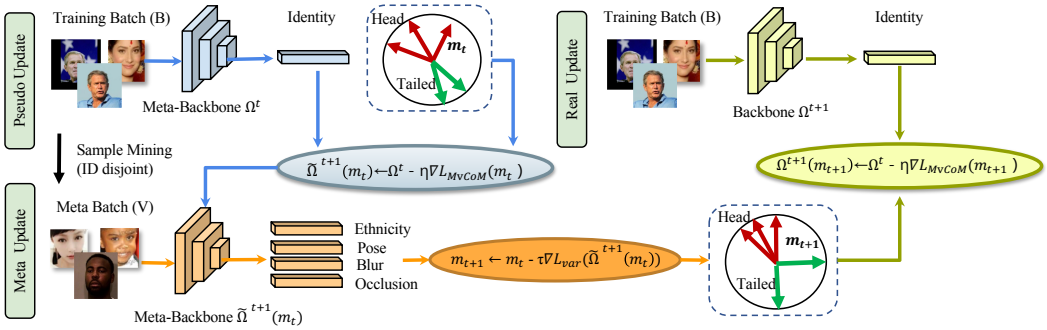


Figure 2: The proposed method flowchart. In Sec. 3.1, we firstly show that traditional re-weighting methods are equivalent to the margin-modulated cosine loss. To jointly tackle multiple variations that cause the long-tailed distribution, we propose the Multi-variation Cosine Margin (MvCoM) (Sec. 3.2). 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.

readers to a survey Wang & Deng (2018). 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 in the presence of imbalanced data is a significant direction, we focus on methods specific to face recognition. Early methods directly change the sampling frequency He & Garcia (2009); Shen et al. (2016). However, the re-balancing 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 Dong et al. (2017); Lin et al. (2017), metric learning Huang et al. (2016); Oh Song et al. (2016); Zhang et al. (2017) and meta learning Ha et al. (2016); Wang et al. (2017b); Ren et al. (2018); Jamal et al. (2020). Liu et al. Liu et al. (2019b) use a dynamic meta-embedding with an associated memory to enhance the representation. AdaptiveFaceLiu et al. (2019a) 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. Cao et al. (2020) 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. Wu et al. (2017) share the high-level idea that sampling matters for training, they consider metric space sampling and make a spherical distribution assumption for it. Our method makes no such assumptions 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 Santoro et al. (2016); Munkhdalai & Yu (2017). (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 Vinyals et al. (2016); Sung et al. (2018); Snell et al. (2017). (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 Ravi & Larochelle (2017); Finn et al. (2017b); Nichol et al. (2018); Guo et al. (2020). 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 Guo et al. (2020) emphasize model generalization to unseen domains.

3 OUR APPROACH

Figure 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) to enhance a canonical identification objective, namely cosine loss Wang et al. (2018) (Sec. 3.2). Further, we formulate MvCoM as an additive modeling combining the class-volume prior with all other factors that cause long-tailed imbalance (Sec. 3.2.2).

3.1 INTERPRETING MARGIN AS SAMPLING IMPORTANCE

Traditional methods Ren et al. (2018); Jamal et al. (2020) 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:

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

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 and y_j as the samples and the label of the y_j-th class. $f(x;\Omega)$ is a convolutional neural network (CNN) backbone 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 and hence 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 Wang et al. (2018) 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 y_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:

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

Usually, the denominators in Eqn. 3 are similar, that is, all close to $[e^{\mathbf{s}-\bar{\mathbf{m}}} + C - 1]^{\sigma_{y_i}}$. Then, the decisive component is generally the numerator, which may be 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 loss 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 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 and its variants assume a constant margin shared across the dataset, which is equivalent to assigning equal sampling importance for all training data. This inevitably pushes the trained model to focus only on the head classes and hence leads to biased estimation. Moreover, if the sampling

importance is modeled only as class-level weighting, intra-class variations among the samples are not considered – for example, two samples with different image qualities from the same class would be assigned the same importance.

Thus, we propose the sample-level multi-variation cosine margin (MvCoM) to flexibly capture sample-level variations, using different importances to contribute to the overall loss. We leverage domain knowledge on face recognition to deem *class volume, ethnicity, head pose, blur and occlusion* as the multiple variations that cause imbalanced distribution, while noting that other variations may be similarly considered if necessary. Formally, we model our MvCoM $\mathbf{m}_{y_j,j}$ in an additive manner by combining the class-volume margin $\mathbf{m}_{y_j}^{cls}$, and a set of margin residual terms $\mathbf{r}_j^k$ representing the importance of each variation k:

$$\mathbf{m}_{y_j,j} = m_{y_j}^{cls} + \sum_k \lambda_k \mathbf{r}_j^k, \tag{5}$$
$$k \in \{vol., eth., pose, blur, occ.\},$$

where $vol.$, eth, $pose$, $blur$ and $occ.$ stand for per-class data volume, ethnicity, head pose, blur level and occlusion variations, respectively and λ_k is a weighting factor for each variation. For further details, please refer to the supplementary material. Thus, 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 effectiveness of our method highly depends on the estimation of MvCoM. Ideally, MvCoM could be dynamically updated during training to highlight the samples with variations (i.e. ethnicity, head pose, etc.) that are less present in the training distribution. The remainder is to estimate each component of MvCoM: the class-volume margin $\mathbf{m}_{y_j}^{cls}$ and variation-aware margin residual $\mathbf{r}_j^k$.

3.2.1 Estimate the class-volume margin

Following Cao et al. (2019), we use the class-wise statistics as the prior for the class-volume magin:

$$m_{y_j}^{cls} = \frac{\alpha}{n_{y_j}^{1/4}}, \tag{7}$$

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

3.2.2 Meta-learn the variation-aware margin residual

To estimate the residual terms of MvCoM in Eqn. 5, we leverage a learning-to-learn framework Finn et al. (2017b); Jamal et al. (2020), by considering each sample's long-tailed factor variations within a training batch $\{(x_j, y_j, \mu_j^k)\}^{|B|}$, where y_j is the class label, μ_j^k 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 an independent meta-learning face dataset. With the online mined samples from this meta-learning set that have complementary variations 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 Figure 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. More details for other variation labels are in the supplementary. A cross entropy loss is used to update the variation classifiers:

$$\mathcal{L}_{var}^k = \sum_j \mathcal{L}_{ce}(g(f(x_j; \Omega); v_k), \mu_j^k), \tag{8}$$

where $\mathcal{L}_{var}^k$ is the cross-entropy loss for variation task k and μ_j^k is the variation label for sample j. Also, only the feature extractor portion of the model Ω is involved in $\mathcal{L}_{var}^k$. 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, 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 towards balanced.

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 may be 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:

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

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 complementary meta-learning batches, the original training batch's bias information is fed back to meta-update the MvCoM. As we maintain the variation classifiers' training to be balanced, the bias information can only from the training batches not the classifiers. Next we illustrate a typical iteration of our meta learning procedure.

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 $\boldsymbol{m}_{j,t}$:

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

where sample x_j is from training set T. Note that update for the model Ω can be rolled back to the previous iteration $t-1$ if the current model Ω does not achieve better performance. 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 such that we meta-learn the more favorable 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:

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

As the class-level magin 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:

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

The overall procedure is summarized in Algorithm 1. Although our meta-learning shares the high-level structure as Jamal et al. (2020), we consider multiple branches for $\boldsymbol{r}^k_j$ to estimate the residuals

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. $*$: self-implemented CosFace baseline. "Ours (single)" means "Ours (occlusion)", "Ours (pose)", or "Ours (ethnicity)" for the respective variation-specific dataset. "Ours (all)": adding all the proposed variations for MvCoM.

instead of a single class conditional weight. Moreover, Jamal et al. (2020); Cao et al. (2020) consider only the class-level importance weighting, whereas our method considers the finer-scale sample-level importance. Another difference from Jamal et al. (2020) 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 Experiments

In this section, we organize the experiments as: **(1)** Extensive ablation study over the five variation margins, and compare to the baseline CosFace Wang et al. (2018). **(2)** Visualization of a set of sample images together with the predicted margin residuals across the factors of ethnicity, head pose, blurriness, and occlusion, to indicate the effectiveness of the learned margins. **(3)** Study of the margin-weighted validation loss as well as the magnitude of the margin residual with respect to head and tail classes. **(4)** Evaluation on challenging benchmarks that are prototypical on variations, such as RFW Wang et al. (2019) for ethnicity, CFP Sengupta et al. (2016) and CP-LFW Zheng & Deng (2018) for head poses, IJB-A Klare et al. (2015) for video blur and OC-LFW for occlusion. Evaluation on general face recognition benchmarks such as LFW Huang et al. (2007), YTF Wolf et al. (2011) and MegaFace Kemelmacher-Shlizerman et al. (2016) is in supplementary due to space limit.

Algorithm 1 Multi-variation Cosine Margin meta-learning

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

4.1 Ablations 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. Specifically, we use RFW Wang et al. (2019) for ethnicity, CFP-FP Sengupta et al. (2016) 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.

Method	IJB-A (Vrf)	
	FAR@0.001%	FAR@0.01%
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.

Method	OC-LFW	CP-LFW	CFP-FP	IJB-A (Vrf)		RFW	
				FAR=0.001%	FAR=0.01%	Avg	Bias ↓
NAN Yang et al. (2017)	-	-	-	-	88.1	-	-
L2-Face Ranjan et al. (2017)	-	-	-	90.9	94.3	-	-
DA-GAN Zhao et al. (2018)	-	-	-	94.6	97.3	-	-
VGGFace2 Cao et al. (2018)	-	-	-	-	92.1	-	-
Multicolumn Xie & Zisserman (2018)	-	-	-	-	92.0	-	-
ArcFace Deng et al. (2019) †	94.56	92.08	98.37	93.7	94.2	97.61	0.84
URFace Shi et al. (2020)	94.60	92.31	98.30	95.0	96.3	96.96	0.96
DomainBlancing Cao et al. (2020)	-	92.63	-	-	-	-	-
CosFace Wang et al. (2018)*	94.41	92.06	98.16	93.2	97.1	97.94	0.67
CB-CosFace Ren et al. (2018)*	94.44	92.04	98.24	94.6	97.2	98.10	0.61
LDAM-CosFace Cao et al. (2019)*	94.54	92.05	98.31	94.5	97.2	97.86	0.65
MetaCW Jamal et al. (2020)*	94.48	92.06	98.28	94.1	97.2	98.20	0.55
MvCoM-CosFace (Ours)	94.83	92.75	98.37	95.7	97.5	**98.32**	**0.51**
MvCoM-URFace (Ours)	**94.92**	**92.86**	**98.47**	**96.0**	**97.6**	97.54	0.76

Table 3: Challenging variation-specific face recognition benchmarks comparison. "-": original work does not report performance on the corresponding protocol. "*": our self-implemented methods. "†": testing performance using models obtained from corresponding authors. In RFW, CA, AA, EA and IN are abbreviations for Caucasian, African American, East Asian and Indian, respectively.

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.

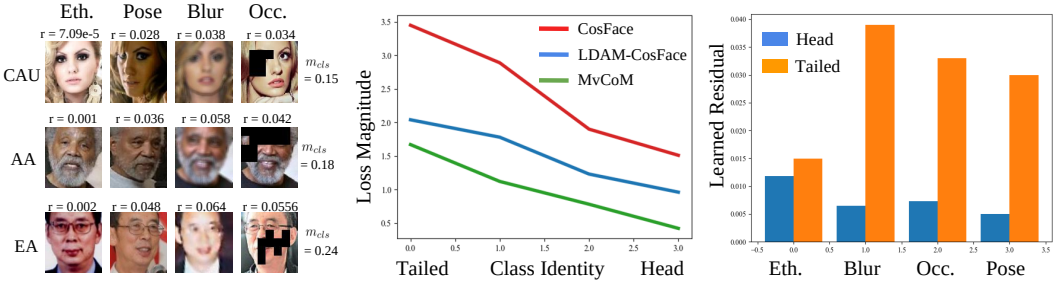
OC-LFW As shown in Table. 1, OC-LFW is an **occlusion** evaluation protocol of LFW Huang et al. (2007), which contains more than 13,000 images from 5749 identities from the internet. For each verification pair, we randomly set occlusion masks on one of the pair images, and conduct the same verification protocol as LFW. Although LFW is saturated, OC-LFW is not saturated as all the methods only achieve under 95% accuracy. We observe that our method with single variation already outperforms the baseline with a clear gain. By adding all variations, the accuracy further increases as more margins provide a more complete regularization for representation learning.

CFP-FP CFP-FP Sengupta et al. (2016) are face image pairs with one image of **large pose variation**, and most of the image pairs are with high resolution. 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 consistently demonstrates that by adding the proposed sample level variation cosine margin, the accuracy is significantly improved.

RFW RFW consists of four race (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, we highlight the *bias*, defined as the standard deviation over the accuracy on those four ethnicity subsets. The bias across CA, AF, EA and IN is much smaller for our method, showing the effectiveness of our learned margin leading to more balanced performance across different ethnicities.

IJB-A (Vrf) We also show an ablation study on IJB-A under the verification protocol, which is an in-the-wild dataset including multiple long-tailed variations. In Table 2, we show the methods with each variation factor, i.e., ethnicity, pose, blur and occlusion. We see 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%, as well "Ours (ethnicity+pose+blur)" is better than "Ours (ethnicity+pose)" by more than 0.2%. This corroborates the intuition that blur is a major distribution imbalance factor in IJB-A whose samples are mostly drawn from low-quality video frames.

Visualization of Per-sample Margin We visualize the learned margin prior and residual to draw the connection with image appearance in Fig. 3 (a), using Eqn. 11. We randomly select three identities from different ethnicities and show the images of different variations (each column) within the same identity (each row). Considering the class volume imbalance, our method initializes the margin as m^{cls} which is larger for tail class (the 3rd row) and smaller for head class (the 1st row). In terms of ethnicity with other factors fixed (the 1st column), variation-aware residual assigns more weight on the minority ethnicity such as African American and East Asian. Similarly for other variations from column two to column four, we organize the tail classes in rows two and three, with the head class in row one. We consistently observe that the margin residuals for the head class are smaller, while those



(a) (b) (c)

Figure 3: (a) Sample level margin visualization across all the factors. Larger margin corresponds to more tailed class. (b) Curve of validation loss magnitude versus the tail to head classes. With our MvCoM (in green), the validation error is significantly reduced (lower) compared to baseline and a re-weighting state-of-the-art LDAM-CosFace Cao et al. (2019). (c) Histogram of the learned residual magnitude over the proposed four long-tailed variations. For tail classes, the learned margin is clearly larger, which expectedly highlights the tail classes in the loss function.

for tail classes (row two and three) are relatively larger. It clearly suggests that the learned MvCoM works as expected to emphasize on samples from the tail class.

Visualization of Margin Histogram We verify whether the learned MvCoM can compensate the distribution imbalance and whether the loss with the learned margin weight is reduced compared to other methods. We group the identities by counting the class volume and form the x-axis of Figure 3(b) as class identity, ranging from tail class to head class. The y-axis is the margin weighted identification loss (Eqn. 6). As expected, our method achieves significantly lower loss on the validation set compared to LDAM-CosFace Cao et al. (2019). In Figure 3(c), we compare the learned residual between head and tail classes across all the variations. The residual for tail classes is consistently higher than head classes across all the factors, again confirming that the learned MvCoM indeed highlights the tail classes.

4.2 Evaluation on Challenging Benchmarks

We compare to both state-of-the-art face recognition methods and long-tailed re-weighting based methods on challenging variation-specific datasets, across the first two sets of rows in Table 3. In general, our method shows consistently better performance over other methods, for example, 0.3% higher than next best on OC-LFW, 1.1% higher than next best on IJB-A FAR$=0.001\%$. While re-weighting based methods in the second row 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. Besides the performance advantages compared to re-weighting methods, our framework jointly considers multiple variation factors, which can better represent the long-tailed distribution. Moreover, our method considers a sample-level learnable cosine margin, which is more flexible than class-level and can be applied to wider range of data distributions. Besides, we observe that MvCoM can combine with different recognition architectures such as CosFace and URFace. 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 Conclusions

In this work, we developed the first method to handle multiple factors that cause long-tailed data distribution in face recognition, namely class volume, ethnicity, head pose, blur and occlusion. This is in contrast to prior works that mostly focus on class volume and also do not consider within-class sample-level variations. A multi-variation meta-learning scheme is proposed to provide complementary feedback on the biased distribution in the form of a novel sample-level Multi-variation Cosine Margin (MvCoM), which is amenable to enhancing popular losses used in face recognition. Empirical results demonstrate that our method achieves top performances on general face recognition benchmarks, with clear advantages on challenging variation-specific benchmarks. Avenues for future work include use of the proposed MvCoM in other practical settings such as few-shot learning.

6 ETHICS STATEMENT

Our work relies on face recognition models trained on public datasets commonly used in computer vision research, where we believe consent has been obtained from all the subjects by the dataset providers. We will remove any subject images where privacy concerns have not been addressed in the provided datasets. We note the concern that face recognition methods may potentially be used for unlawful surveillance or discrimination and support the development of community guidelines that balance innovation in computer vision with regulations on its usage. 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 health, hiring, policing and judicial outcomes. For instance, the proposed method can prevent incorrect face recognition for minority ethnicities and lower the risk of incorrectly recognizing faces due to low quality images, large head poses or heavy occlusions.

7 REPRODUCIBILITY

We highlight several components within this submission that effectively support the reproducibility of this work. **1)** Implementation details are provided in Appendix A.3, where the training dataset, backbone architecture, training schemes, varaition label preparation and the running complexity are carefully analyzed. **2)** A more detailed theoretical analysis regarding our methodology is presented in Appendix A.1 and A.2. **3)** A better understanding of our method, that is, additional evaluations and comparisons on more general face datasets and further visualizations on the variational margin residuals, are provided in Appendix A.4 and A.5.

REFERENCES

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*, pp. 5671–5679, 2020. 3, 7, 8, 15

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*, pp. 1567–1578, 2019. 1, 5, 8, 9, 15

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, 14

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 1

Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CVPR*, 2019. 1, 2, 8, 14, 15

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*, pp. 1851–1860, 2017. 3

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 2017a. 1

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017b. 3, 5

Scott Grabinger and Joanna Dunlap. Rich environments for active learning: a definition. In *Research in Learning Technology*, 1995. 1

R. Gross, I. Matthew, J.F. Cohn, T. Kanade, and S. Baker. MultiPIE. *Image and Vision Computing*, 2009. 15

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

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. 2, 14

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 3

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 1, 3

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*, pp. 5375–5384, 2016. 3

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, 7, 8, 15

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, 5, 6, 7, 8, 15

Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016. 7, 15

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. 7

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*, pp. 2980–2988, 2017. 3

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

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*, pp. 11947–11956, 2019a. 3

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, 15

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*, pp. 2537–2546, 2019b. 3

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. In *Nature*, 2015. 1

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

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

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*, pp. 4004–4012, 2016. 3

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

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

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018. 1, 2, 3, 4, 8, 15

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

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

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. 7, 8, 15

Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pp. 467–482. Springer, 2016. 1, 3

Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6817–6826, 2020. 8, 15

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

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*, pp. 1891–1898, 2014. 2

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

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

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

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

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, 4, 7, 8, 14, 15

Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 2018. 3

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*, pp. 692–702, 2019. 7

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pp. 7029–7039, 2017b. 3

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 1, 2, 15

Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011. 7

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*, pp. 2840–2848, 2017. 3, 15

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

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

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*, pp. 5409–5418, 2017. 3

Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *CVPR*, 2019. 13

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

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*, pp. 1136–1144, 2019. 2

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*, pp. 18–01, 2018. 7

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

A APPENDIX

In this material, we firstly show in Sec. A.1 that the re-weighting method Eqn. 1 can be interpreted as the CosFace loss with new scale $\sigma_{y_i}\boldsymbol{s}$ and new margin $\sigma_{y_i}\bar{\boldsymbol{m}}$. Then, we illustrate our proposed meta-updating module in Sec A.2 with more details. A complement implementation detail to our main submission experiment section, is in Sec A.3 to provide more information for the reproduction purpose. Due to the main submission space limit, we add the general face recognition datasets evaluation in Sec A.4. Finally, in Sec. A.5, we visualize the instance-level variation-aware margin with more samples as an extension of our main submission Fig. 3, which is expected to provide a more complete view of the visualization.

A.1 SAMPLE IMPORTANCE INTERPRETED AS COSINE LOSS

In our main submission Sec. 3.1, we show that the sample importance from the re-weighting method (Eqn. 13) is equivalent to the Cosine Loss. In this section, we provide a more detailed proof showing that actually the re-weighting objective (Eqn. 13) can be numerically approximated by the Cosine Loss.

Connecting to the main submission, we introduce the importance weight σ_{y_i} to re-weight each sample loss as the following:

$$\min_{\Omega} \quad \frac{1}{N}\sum_{i=1}^{N}\sigma_{y_i}\mathcal{L}_{cos}(f(x_i;\Omega),y_i)). \tag{13}$$

Combining Eqn. 1 with Cosine Loss, we obtain:

$$\min_{\Omega}\frac{1}{N}\sum_{i=1}^{N}-log\frac{\left[e^{\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}}}\right]^{\sigma_{y_i}}}{\left[e^{\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}}}+\sum_{k\neq y_i}^{C}e^{\mathbf{s}\cdot cos\theta_{i,k}}\right]^{\sigma_{y_i}}}. \tag{14}$$

As the denominator is non-negative, we obtain the following for the denominator:

$$\left[e^{\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}}}+\sum_{k\neq y_i}^{C}e^{\mathbf{s}\cdot cos\theta_{i,k}}\right]^{\sigma_{y_i}} \tag{15}$$

$$\begin{aligned}&= e^{\sigma_{y_i}(\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}})}+\sum_{k\neq y_i}^{C}e^{\sigma_{y_i}\mathbf{s}\cdot cos\theta_{i,k}}\\&+\frac{e^{\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}}}}{\left[\sum_{k\neq y_i}^{C}e^{\sigma_{y_i}\mathbf{s}\cdot cos\theta_{i,k}}\right]^{(1-\sigma_{y_i})}}+\frac{\sum_{k\neq y_i}^{C}e^{\mathbf{s}\cdot cos\theta_{i,k}}}{\left[e^{\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}}}\right]^{(1-\sigma_{y_i})}}\end{aligned} \tag{16}$$

$$\approx e^{\sigma_{y_i}(\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}})}+\sum_{k\neq y_i}^{C}e^{\sigma_{y_i}\mathbf{s}\cdot cos\theta_{i,k}} \tag{17}$$

In Eqn. 16, we only provide the two major terms and two first-order terms. As indicated in Zhang et al. (2019), $\boldsymbol{s}$ empirically can choose $\sqrt{2}log(C-1)$. Assume $cos\theta_{i,y_i}\approx 1$ and take $\bar{\boldsymbol{m}}=0$ for simplicity, $e^{\boldsymbol{s}cos\theta_{i,y_i}-\bar{\boldsymbol{m}}}=(C-1)^{\sqrt{2}}$, and $\left[\sum_{k\neq y_i}^{C}e^{\sigma_{y_i}\mathbf{s}\cdot cos\theta_{i,k}}\right]^{(1-\sigma_{y_i})}\approx(C-1)^{(1-\sigma_{y_i})}$. The last two terms in Eqn. 16 are $O(1)$. Thus, comparing to the two major terms, they can be omitted.

Replacing the denominator of Eqn. 14 with its approximate counterpart from Eqn. 17, we show in Eqn. 18 that the original sampling importance re-weighting objective is equivalent to a Cosine Loss with new scale $\sigma_{y_i}\boldsymbol{s}$ and new margin $\sigma_{y_i}\bar{\boldsymbol{m}}$.

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}-log\frac{\left[e^{\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}}}\right]^{\sigma_{y_i}}}{\left[e^{\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}}}+\sum_{k\neq y_i}^{C}e^{\mathbf{s}\cdot cos\theta_{i,k}}\right]^{\sigma_{y_i}}}\\&\approx\frac{1}{N}\sum_{i=1}^{N}-log\frac{\left[e^{\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}}}\right]^{\sigma_{y_i}}}{\left[e^{\mathbf{s}cos\theta_{i,y_i}-\bar{\mathbf{m}}}\right]^{\sigma_{y_i}}+\sum_{k\neq y_i}^{C}e^{\sigma_{y_i}\mathbf{s}\cdot cos\theta_{i,k}}}\end{aligned} \tag{18}$$

A.2 META-LEARNING UPDATE DETAILS

As shown in Fig.1, there are mainly three updating steps in our meta-updating procedure. We hereby specify one typical iteration of the proposed meta learning framework from the gradient perspective as the following:

$$\tilde{\Omega}^{t+1}(\boldsymbol{r}^t) \leftarrow \Omega^t - \eta \frac{\partial \sum_{i \in T} L_{MvCoM}(f(x_i; \Omega^t), y_i; m_{y_i} + \boldsymbol{r_j^t})}{\partial \Omega} \tag{19}$$

$$\boldsymbol{r}^{t+1} \leftarrow \boldsymbol{r}^t - \tau \frac{\partial \sum_{k, i \in V} L_{var}^k(f(x_i; \tilde{\Omega}^{t+1}(\boldsymbol{r}^t)), \mu_j^k)}{\partial \boldsymbol{r}} \tag{20}$$

$$\Omega^{t+1} \leftarrow \Omega^t - \eta \frac{\partial \sum_{i \in T} L_{MvCoM}(f(x_i; \Omega^t), y_i; m_{y_i} + \boldsymbol{r_j^{t+1}})}{\partial \Omega}. \tag{21}$$

It corresponds to three steps of the update in Sec. 3.3.2 of the main submission. First, we feed a training batch into the deep face recognition model to compute identification loss with a prior class-aware margin where margin residual is initialized as zero and update the model in Eqn. 19 to a temporary model which is termed as "Pseudo updated" model in the main submission. Then, following our hard sampling rule, we sample a meta-learning batch which is the most distinct from the training batch, feed it into the pseudo updated model, compute the classification loss for 4 variation tasks and update the margin residual in Eqn. 20. It is worth noting that the meta-updated margin residual carries the compensating information to each long-tailed variation factors compared to the previous margin. Last, we add the meta-updated margin residual on top of previous margin prior, go back to the first step to compute the identification loss and update the original pre-updated model with newly meta-learned margin in Eqn. 21.

In a nutshell, the purpose of meta-learning framework is to learn the best margin (or margin residual $\boldsymbol{r}$) in the sense that it results in smaller classification errors in variation tasks on meta-learning set by balancing multiple variation factors.

A.3 IMPLEMENTATION DETAILS

The whole code base is implemented with Pytorch v1.1. We use the clean list from ArcFace Deng et al. (2019) for MS-Celeb-1M Guo et al. (2016) as the training data. For the meta-training set, we adopt VGGFace2 Cao et al. (2018) and exclude the duplicate identities since it contains multiple variation factors that can potentially benefit the training dataset. The baseline models in the experiments are trained with CosFace loss Wang et al. (2018) 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.

Pre-training Specifically, we adopt the 100-layer ResNet proposed in Wang et al. (2018) as our embedding network Ω. As introduced in the main paper, our training of the face recognition engine is divided into two steps. For the first step, we pre-train a CosFace-like engine backbone. The training data is the cleaned MS-Celeb-1M Guo et al. (2016) with 84K identities and around $4.8M$ images. We train the CosFace model with 30 epochs with initial learning rate 0.1 Then, the learning rate is multiplied by 0.1 at the 14^{th}, 20^{th} and 23^{th} epoch each time. The momentum is set as 0.9 with weight decay as $5e^{-4}$. We use SGD as the optimizer. Batch size is set to 512. The overall training is conducted on a 8-core Titan-X gpu with pytorch parallel model training.

Alternative Meta-Optimization After pre-training, we keep the whole backbone engine from the first step. For the identification classifier (a fully-connected layer with 512×84K dimension), we also keep it for our main task identification. Meanwhile, as introduced in the main paper, we introduce 4 variation task classifiers, namely a pose classifier (a fully-connected layer with 512×7 dimension), an occlusion classifier (a fully-connected layer with 512×5 dimension), a blur classifier (a fully-connected layer with 512×4 dimension) and an ethnicity classifier (a fully-connected layer with 512×4 dimension). There are typically two rounds of forward pass and backward pass in one iteration of the fine-tuning. For the first round, a training batch is fed to the recognition backbone and calculate the identification loss via identification classifier, which is known as "Pseudo Update" explained in the main paper method part. Then, a backward pass is conducted to update the network

parameter Ω. After the first round, we meta-update the residual of the margin for each instance in the training batch. For the second round, the updated margin is contributed to calculate the loss and the recognition model is updated. Such process lasts 18 epochs to ensure convergence. We initialize the learning rate as 0.001 and reduce by 0.1 at every 8^{th}, 14^{th} and 16^{th} epochs. The momentum is set as 0.9 with weight decay as $5e^{-4}$. We use SGD as the optimizer. Batch size is set to 512, the same as the pre-training step.

Variation Label We apply mechanical turk to label the ethnicity of the training set, including African American, Caucasian, East Asian and South Asian. For head pose, following the pose angle setting in MultiPIE Gross et al. (2009), 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 with blur effect. For occlusion, we adopt five different block sizes $(5, 11, 17, 23, 29)$ to randomly black out the training images with the specific size. All the above mentioned labels will be released after submission.

Complexity All the variation classifiers are linear classifiers. Compared to CosFace, our framework newly introduce four variation classifiers. But it almost does not increase the network complexity, because the variation classifiers are less than 10-way, while the recognition classifier in the baseline is around 80000-way. Regarding the training convergence, for one iteration, since our framework require twice the recognition forward and backward, and once the meta-learning forward and backward, the time complexity for our training is nearly two times longer than the baseline training. Since testing only utilizes the recognition model, the runtime for inference is the same as CosFace.

A.4 Evaluation on General Benchmarks

We compare our method against the state-of-the-arts on general face recognition benchmarks with limited variations, namely LFW Huang et al. (2007), CFP Sengupta et al. (2016), and MegaFace Kemelmacher-Shlizerman et al. (2016). We self-implement the CosFace and use it as the backbone to further implement the Class-Balance CosFace (CB-CosFace) Ren et al. (2018), Label-distribution-aware margin loss (LDAM-CosFace) Cao et al. (2019), and Meta Conditional Weights (MetaCW) Jamal et al. (2020) as comparison to the state-of-the-art long-tailed methods. Some datasets like LFW have reported saturated results due to its limited variation. The main purpose of this evaluation is to show that our method is consistently among the top level of the cutting-edge methods, where the less imbalanced testing data distribution does not degrade our method's performance. In Table 4, "Ours" achieves the second best on LFW, the first on CFP-FP, and MegaFace challenge 1 with uncleaned protocol. Furthur, 1) Our MvCoM is consistently better than the previous long-tailed re-weighting methods on all datasets. 2) The MvCoM is modular to many backbones such as CosFace and URFace, demonstrating the wide applicability to different face recognition platforms.

Method	LFW	CFP-FP	MF1	
			Rank1	Veri.
Wu et al. Wu et al. (2017)	98.37	-	-	
CenterFace Wen et al. (2016)	99.28	-	65.23	76.52
SphereFace Liu et al. (2017)	99.42	-	75.77	89.14
ArcFace Deng et al. (2019)	**99.83**	98.37	81.03	96.98
DomainBlancing Cao et al. (2020)	99.78	-	-	
URFace Shi et al. (2020)	99.75	98.30	79.10	94.92
CosFace Wang et al. (2018)*	99.73	98.16	80.03	95.54
CB-CosFace Ren et al. (2018)*	99.81	98.24	80.18	95.75
LDAM-CosFace Cao et al. (2019) *	99.75	98.31	80.73	96.78
MetaCW Jamal et al. (2020)*	99.78	98.28	80.32	96.22
MvCoM-CosFace (Ours)	99.80	98.37	**81.30**	**97.22**
MvCoM-URFace (Ours)	99.78	**98.47**	80.63	96.28

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.

A.5 Per-Sample Margin Visualization

In this section, we visualize the per-sample margin for more instances to verify our learned margin is capable of compensating the distribution imbalance in terms of multiple variation factors. In Fig. 4,

Figure 4: Multi-variation Cosine Margin (MvCoM) visualization across all the factors,i.e. ethnicity, quality, pose, blur and occlusion.

we observe that our model consistently assigns larger margin weight to the samples from the tailed class of each variation such as non-Caucasian, large pose, blur and occluded images.