CoReFace: Sample-Guided Contrastive Regularization for Deep Face Recognition

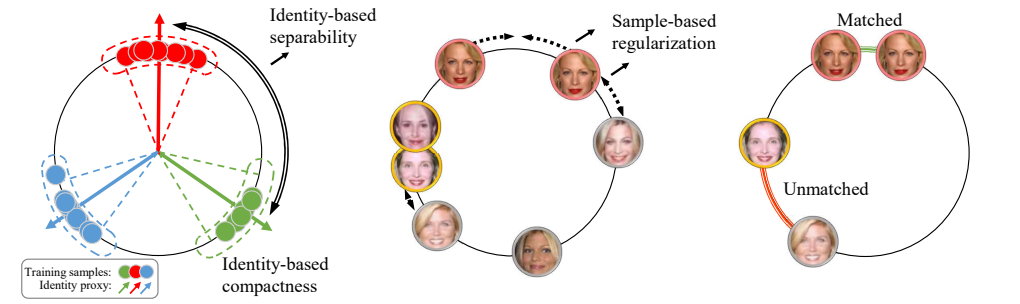
Youzhe Song[a], Feng Wang[a,*]

[a]*Shanghai Key Laboratory of Multidimensional Information Processing, School of Computer Science and Technology, East China Normal University, 3663 North Zhongshan Road, 200062, Shanghai, China.*

Abstract

The discriminability of the feature representation is crucial for face recognition. However, previous methods rely solely on the learnable weights of the classification layer, which represent the identities. This reliance could be problematic as the evaluation process depends on the similarity between pairs of face images and requires minimal identity information learned during training. As a result, there is an inconsistency between the training and evaluation processes, which can confuse the feature encoder and hinder the effectiveness of identity-based methods. To address this problem, we propose a novel approach namely Contrastive Regularization for Face Recognition (CoReFace), which applies sample-level regularization in feature representation learning. Specifically, we employ sample-guided contrastive learning to directly regularize the training based on the sample-sample relationship and thus align it with the evaluation process. To avoid image quality degradation, we augment the embeddings instead of the images in order to integrate contrastive learning into face recognition. Additionally, we introduce a new contrastive loss function for the regularization of representation distribution. This function incorporates an adaptive margin and a supervised contrastive mask to ensure stable loss values and prevent interference with the identity supervision signals. Finally, we explore new pair-coupling protocols in order to overcome the problem of semantically repetitive signals in contrastive learning. Extensive experiments demonstrate the efficacy and efficiency of our CoReFace approach, which achieves competitive results compared to state-of-the-art methods.

*Corresponding author. Email: fwang@cs.ecnu.edu.cn.



(a) Identity-based training process (b) Sample-based training process (c) Sample-based evaluation process

Figure 1: (a) Current identity-based methods aim at intra-class compactness and inter-class separability in training. However, the identity-based training pays little attention to the sample-sample relationship which is the foundation of the evaluation process. The points with gray borderlines represent face images from distinct identities, and other points with the same borderline color come from the same identity. (b) Our CoReFace takes contrastive learning as regularization to directly constrain the sample-sample relationship during training and improve discrimination in evaluation as illustrated in (c).

Keywords: Face recognition, discriminative representations, contrastive regularization, contrastive learning

1. Introduction

Face recognition (FR) is a well-established task that holds significant importance in various applications. FR evaluation scenarios can be broadly categorized into two types: verification and identification. In both cases, the similarity between the face images forms the basis for the comparison. To better adapt to the real-world situations, the training dataset used in face recognition excludes those identities for evaluation [1].

Recent state-of-the-art (SoTA) FR methods are based on identity, where face images with labels are used to train a classifier that discriminates between different identities. However, during evaluation, the classifier is typically discarded as the identity information learned during training is of little use for the sample-sample comparison in the evaluation process.

Many existing methods aim to enhance the performance in some specific subtasks of face recognition, such as low-quality images [2] and variant head poses [3]. However, these methods are all built upon the foundation of general face recognition methods. In order to improve face recognition in the

general situation, a series of identity-based methods using margin [4, 5, 6] have been proposed to improve the intra-class compactness.However, these identity-based methods overlook the holistic feature space [7]. Meanwhile, several other approaches emphasize inter-class separability as a key factor in achieving feature discriminability by utilizing various loss functions for regularization [8, 9, 10, 7]. These *identity-based* methods leverage the classifier's weight as a proxy for identities to explore sample-identity or identity-identity relationships during training. However, during the evaluation which is *sample-based*, the identity information in the classifier becomes irrelevant and the differentiation is based solely on the chosen samples.

As illustrated in Figure 1, the feature distribution in identity-based training could achieve high intra-class compactness and inter-class separability with the help of identity proxy features. In the sample-based evaluation, the classifier is removed and the sample pairs are then used for verification. Additionally, the face images used in the evaluation are from different identities than those used in training. Consequently, the discriminative characteristics of the evaluation distribution may not be as comparable as those in training.

To address the aforementioned problem, this paper introduces a novel method called **Co**ntrastive **Re**gularization for **Face** Recognition (CoReFace). This approach uses contrastive learning to regulate the training process and ensure that the sample-sample relationship aligns with the evaluation goal, thus improving the performance of face recognition. Contrastive learning works by bringing semantically similar samples closer and pushing dissimilar samples away from each other [11]. In the field of face recognition, class-guided contrastive learning has been explored, where positive pairs are composed of samples from the *same identity*. For instance, triplet loss is applied either solely [12] or jointly [13, 14] with the identity-based classifying methods. However, with the recent development of identity-based methods, these approaches could interfere while joint training with other identity-based methods [15, 6]. On the other hand, sample-guided contrastive learning has demonstrated promising progress in unsupervised learning [16, 17, 18]. In this case, positive pairs are formed by applying stochastic data augmentation on the same image. In our approach, we leverage sample-guided contrastive learning as a regularization approach to adjust the relationship between samples so as to learn a more semantic and consistent feature distribution during training and evaluation.

However, integrating sample-guided contrastive learning with identity-based methods is non-trivial. First, general face recognition requires a large

number of high-quality images to learn the differences between identities. The commonly used data augmentations in contrastive learning can hinder the convergence of FR models [19, 20]. To make sample-guided contrastive learning applicable to FR, we propose a new pipeline that uses feature augmentation instead of data augmentation to generate positive pairs. Second, sample-guided contrastive learning is typically designed to be applied exclusively. When jointly trained with the identity-based methods, its effectiveness becomes insignificant. To address this issue, we make several improvements to the contrastive loss function to enable the effective regularization. Third, the scale of the negative sample pool plays a crucial role in contrastive learning [18, 21, 22, 23]. We further discovere a *Semantically Repetitive Signal* (SRS) problem, where certain combinations of samples repeatedly contribute to the optimization and push the related part of the distribution with an inappropriate magnitude. To alleviate this problem, we explore new strategies of pair coupling in contrastive learning. The main contributions of this paper are summarized as follows:

- We propose a novel framework that applies regularization in FR using contrastive learning. Unlike previous methods adjusting feature distribution with sample-identity pairs, our approach utilizes sample-sample relationships that are consistent between training and evaluation.
- We propose a contrastive loss function that performs effective regularization by incorporating an adaptive margin to strengthen the contrastive supervision signal and a supervised contrastive mask to avoid collisions in joint training.
- We investigate the SRS problem in contrastive learning in situations with limited negative samples and explore new pair-coupling protocols to alleviate this problem.
- We perform extensive experiments on widely-used benchmarks to show the superiority of our proposed framework over existing approaches.

2. Related Works

2.1. Identity-based methods with margin

In recent years, there has been an emerging trend in identity-based methods using extra margin for face recognition (FR) [4, 5, 6, 24, 25]. They are

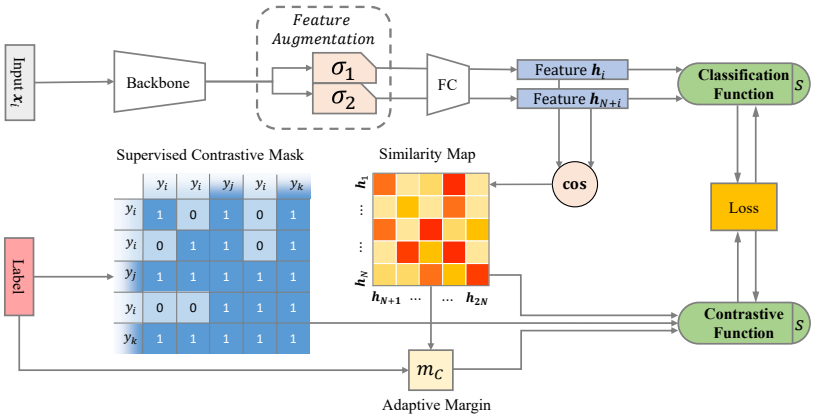


Figure 2: Illustration of the framework of our CoReFace. To address the problem of image quality degradation, we introduce a feature augmentation module between the backbone and the FC layer. Our contrastive loss function consists of an adaptive-margin loss and a supervised contrastive mask. To avoid the semantically repetitive signals, we employ a new pair-coupling protocol in the similarity computation for contrastive learning. The contrastive loss regularize the training process based on sample-sample relationships.

also the foundation of multi-modal face recognition [26, 27]. These methods normalize the representation embeddings of images and classes (or identities) before their multiplication [28, 29], which degrades the product to the cosine value of the angle between the two vectors. During training, a margin parameter is used to increase the distance between the matched sample-identity pairs and unrelated ones. This promotes the intraclass compactness by minimizing the distances between the representations of the images from the same identity. The normalization alleviates the misguidance of the feature norm in the Softmax loss by projecting the features onto a hypersphere, while the margin imposes a strong constraint on the sample-identity feature pairs on this hypersphere [28]. Although these methods achieve high intra-class compactness, they fail to fully exploit the holistic feature space [7]. To align the training with the evaluation process, our CoReFace applies constraints of the sample-sample relationships during training.

2.2. Feature Regularization in Face Recognition

Feature distribution is the foundation of face recognition evaluation, as both the two subtasks (verification and identification) rely on feature similarity between face images [4, 8]. To improve the performance of face recognition, some methods introduce additional constraints to adjust the feature

distribution in a holistic manner. These constraints include restricting the magnitude of representation features [30] and controlling the Euclidean distance between the representations and the identity weights [8]. Since the identity weights act as class proxies, previous research suggests that they can facilitate a holistic feature distribution [31, 9, 7, 10]. By constraining the energy function, Euclidean distances, or angles between the identity weights, better feature distributions can be achieved.

However, the methods mentioned above indirectly adjust sample similarities using identity information specific to the training process. Their effectiveness is primarily observed during training with little assurance of generalization to the evaluation process where the identities are unknown. In this paper, we propose a novel contrastive regularization approach by introducing a new contrastive loss within a novel framework. In contrast to existing methods, our approach directly adjusts the relationship between image features so as to make the training consistent with the evaluation process and improve the performance of face recognition.

2.3. Contrastive Learning for Face Recognition

Contrastive learning aims to cluster semantic neighbors as distribution neighbors in the representation space [11]. Class-guided contrastive learning [12, 28] has been applied to face recognition [12]. It considers samples from the same class as semantic neighbors. However, it has been shown to hinder performance in joint training and underperform compared to the identity-based methods [6, 15]. On the other hand, sample-guided contrastive methods use data augmentation results to form positive pairs. These methods typically construct a large negative sample pool for comparison [18, 21, 22, 23]. With extensive datasets and sufficient training, they exhibit promising unsupervised learning performance. However, applying sample-guided contrastive learning in face recognition is challenging due to the semantic damage introduced by commonly used data augmentation techniques [19]. Additionally, avoiding conflicts between the two supervision signals during training is also a key issue. In this paper, we introduce a new framework namely CoReFace to address the image quality degradation problem and maintain the effectiveness of regularization during training. CoReFace incorporates feature augmentation to mitigate the semantic damage caused by data augmentation. Moreover, our contrastive loss utilizes an adaptive margin to supervise well-performing identity-based methods and incorporates a supervised contrastive mask to prevent conflicts during joint training. We

also identify and address the semantically repetitive signal problem by exploring new pair-coupling protocols in contrastive learning.

3. Methodology

Figure 2 illustrates our CoReFace framework. Sample-guided contrastive learning is used to regulate the training process with the sample-sample relationships. To tackle the problem of image quality degradation, feature augmentation is employed instead of data augmentation for positive pair composition. A novel contrastive loss is proposed by integrating an adaptive margin and a supervised contrastive mask (SCM). Furthermore, new pairing strategies are developed to handle the issue of "Semantically Repetitive Signal" (SRS) in contrastive learning, which distorts the feature distribution and disrupts the similarity calculation.

Algorithm 1 Pseudo code of our CoReFace loss on Pytorch.

Input: Image features (H_1, H_2) from augmentation channel (σ_1, σ_2), identity labels Y, identity matrix I, momentum parameter α, scale parameter s, adaptive margin in the last step m_C
Output: CoReFace loss value $\mathcal{L}_C$
$H_1, H_2 = \text{normalize}(H_1), \text{normalize}(H_2)$
$SM = (H_1 \cdot H_2^\top).\text{clamp(-1, 1)}$
$Y_C = \text{range}(\text{len}(H_1))$
$SM' = SM.\text{clone}().\text{detach}()$
$pos = SM' \cdot I$
for i in Y_C:
    $SM'[i, Y = Y[i]] = 0$
$neg = SM'.\text{max(dim=1)}$
$m = (pos - neg).\text{mean}()$
$m_C = \alpha \cdot m + (1 - \alpha) \cdot m_C$
$SM = SM - m_C \cdot I$
$\mathcal{L}_C = \text{CrossEntropyLoss}(SM \cdot s, Y_C)$
return $\mathcal{L}_C$

3.1. Feature Augmentation

The image quality degradation caused by data augmentation cannot be ignored [19, 20]. To address this issue and make sample-guided contrastive

learning applicable to FR, we propose augmenting the features instead of the images for positive pair composition. As depicted in Figure 2, we pass the hidden embedding after the backbone through two dropout channels σ_1 and σ_2 with distinct masks. Dropout [32] randomly disables certain parts of the input with a certain probability. It can be viewed as a form of augmentation between adjacent layers [33, 23]. In our approach, the dropout masks are randomly generated in every mini-batch and applied to all input samples. Random noise and linear transformations are another two candidate for feature augmentation. But they need proper design and both of them have two variables which make it more difficult to apply them than dropout. What's more, dropout does not modify the original input data. This prevents introducing artificial bias that may occur when adding noise or linear transformations.

By using feature augmentation, we can compose positive pairs for contrastive learning while avoiding image quality degradation. Moreover, compared to data augmentation, which is applied to the input sample and passes the augmented samples through the entire model twice, our feature augmentation operates on the features and reduces computation by almost half.

3.2. CoReFace Loss Function

The identity-based supervision signal could dominate the training as it takes advantage from the labels, while aggressive regularization that conflicts with other supervision signals is also ineffective. Adapting to different datasets makes these problems even more difficult. To overcome the these issues, we propose a new contrastive loss function that is both adaptive and harmonious during joint training. The pseudo code is presented in Algorithm 1. Our CoReFace loss function produces consistent loss values and considers the identity labels to prevent conflicts with the identity supervision signal.

Both the sample-guided contrastive loss functions and the identity loss functions in face recognition are based on the cross-entropy loss function. The common forms of these two types of losses are as follows:

$$\mathcal{L}_{Cla} = -\log \frac{e^{s \cdot P(\boldsymbol{h}_i, \boldsymbol{W}_{y_i})}}{e^{s \cdot P(\boldsymbol{h}_i, \boldsymbol{W}_{y_i})} + \sum_{j=1, j \neq y_i}^{n} e^{s \cdot Q(\boldsymbol{h}_i, \boldsymbol{W}_j)}}, \tag{1}$$

$$\mathcal{L}_{Con} = -\log \frac{e^{\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_{N+i})/\tau}}{\sum_{j=1}^{2N} \mathbb{1}_{[j \neq i]} e^{\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)/\tau}}, \tag{2}$$

where $P(\boldsymbol{h}_i, \boldsymbol{W}_{y_i})$ and $Q(\boldsymbol{h}_i, \boldsymbol{W}_j)$ are two different functions to modulate the positive and the negative pair production of the feature $\boldsymbol{h} \in \mathbb{R}^d$, $\boldsymbol{W} \in \mathbb{R}^{d \times n}$ is the weight of the classifier with d being the feature dimension and n being the number of classes, h_i and h_{N+i} are obtained from one feature augmented by different augmentation channels, $\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j) = \frac{\boldsymbol{h}_i^\top \boldsymbol{h}_j}{\|\boldsymbol{h}_i\| \|\boldsymbol{h}_j\|}$ is the cosine similarity, s and τ are two scale parameters used in the identity loss function and the contrastive loss function respectively, and $\mathbb{1}_{[j \neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $j \neq i$.

Adaptive Margin. As the most similar negative pair and the positive pair influence the decision boundary the most, our contrastive loss updates the margin m by taking into consideration the difference between their similarities. This margin guarantees that the magnitudes of the exponentiated numerator and the denominator in the softmax function are similar, thus maintaining a steady loss value. In order to handle extreme data that may introduce noise, we employ the Exponential Moving Average (EMA) [25]. Specifically, let $m_C^{(k)}$ be the average of the margin of the k-th iteration with $m_C^0 = 0$, and let α be the momentum parameter that is empirically set to 0.99. For a pair $(\boldsymbol{h}_i, \boldsymbol{h}_j)$ where $i < j$, $m_C^{(k)}$ is updated as:

$$m_C^{(k)} = \alpha m^{(k)} + (1 - \alpha) m_C^{(k-1)}, \tag{3}$$

$$m^{(k)} = \frac{1}{N} \sum_{i=1}^{N} \left(\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_{N+i}) - Maxneg_i \right), \tag{4}$$

$$Maxneg_i = \max(\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)), j \in [1, 2N], j \neq N + i, \tag{5}$$

where N is the number of samples. Taking m as the difference between angles as in ArcFace [6] is also a candidate approach. However, it changes the angle of the vector pairs directly, which needs to include the triangle function and increases the complexity of the derivation. This results in a nan value when being used as the contrastive loss. To sum up, the contrastive loss with adaptive margin can be formulated as

$$\mathcal{L}_C = -\log \frac{e^{s(\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_{N+i}) - m_C)}}{e^{s(\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_{N+i}) - m_C)} + \sum_{j=1, j \neq i, j \neq N+i}^{n} e^{s \cdot \text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)}}. \tag{6}$$

Supervised Contrastive Mask. In supervised learning, contrastive methods may contradict the fact that some samples are from the same class in FR, i.e. $y_i = y_j$, whose features should be similar. When cooperating with

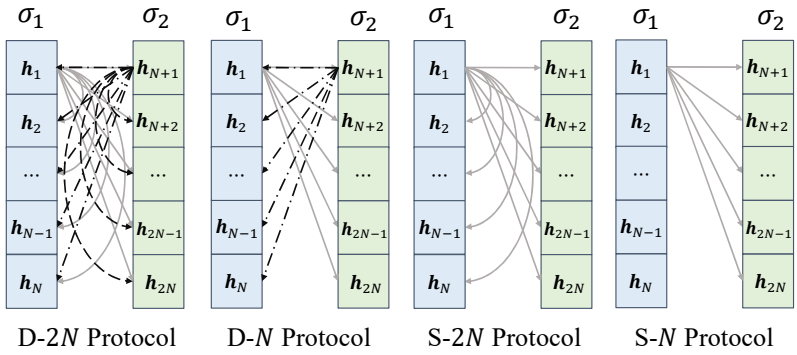


Figure 3: Four types of pair-coupling protocols. Every combination of two augmentation channels σ_1 and σ_2 represents the feature combinations of the images in a mini-batch. If there are multiple augmentation channel combinations, the feature pairs are duplicated.

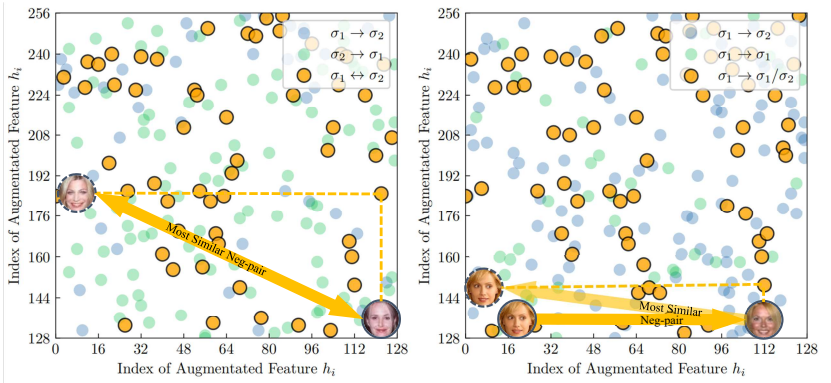
the identity loss, the two methods could disturb each other in the interpretation of the supervision signals.

To avoid this conflict between the contrastive regularization and the identity-based training, we introduce Supervised Contrastive Mask (SCM). With the help of the labels, we create SCM to exclude the distraction of samples from the same class. Specifically, we set the similarity score $\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)$ to 0 for feature pairs where $i < j$, $j \neq N + i$, and $y_i = y_j$.

3.3. Pair-Coupling Protocol for SRS Problem

We uncover a problem known as Semantically Repetitive Signal (SRS), where some specific negative pairs are excessively emphasized, resulting in a distorted distribution and abnormal drawing and pulling of the affected features. To gain a better understanding of this problem, we have investigated pair-coupling protocols, which determine how positive and negative pairs are formed. Figure 3 illustrates four distinct protocols. Let $(\sigma_i \rightarrow \sigma_j)$ represent a pair where the first and the second features are from the i-th and j-th mask channels respectively, where $i, j \in \{1, 2\}$. The number of mask channels in the first position determines whether the pairs are formed in a *single* or *double* way, while the number of mask channels in the second position affects the number of negative samples, either N or $2N$.

Figure 4 illustrates the repetition of key negative pairs in a well-trained identity-based model. Here, the coordinates of a point represent the indexes of a feature and its most similar negative counterpart within a batch. Points



(a) Most similar neg-pair in D-N protocol (b) Most similar neg-pair in S-2N protocol

Figure 4: The coordinates of a point are the indexes of a feature and its most similar negative feature in a mini-batch. σ_1 and σ_2 denote the augmentation channel. When two ordered pairs are mirrored, the points overlap and are painted yellow. The blossom yellow points in (a) and (b) demonstrate the symmetric problem in the ways and the number of negative samples separately. We further illustrate the SRS problem by including a descriptive example in each of the two figures. The y-index of $(\sigma_1 \to \sigma_1)$ points are increased by 128.

are colored blue or green according to the feature channels of a pair, while the overlapping points are colored yellow. For relation $(\sigma_1 \leftrightarrow \sigma_2)$, the most similar negative feature of h_i and h_j could be each other. While for $(\sigma_1 \to \sigma_1/\sigma_2)$, the most similar negative pair could be (h_i, h_j) and (h_j, h_i') when h_i' is the corresponding augmented version of h_i. It is obvious that many feature pairs are mirrored as many points are in yellow. The contrastive loss function results in partially duplicated loss values, which leads to a unexpected example mining strategy and inappropriate back-propagation throughout the training process.

To address this problem, we propose a *Single-way N Protocol* that reduces the symmetry in the pair-coupling process. Specifically, we only compute the similarity of $(\sigma_1 \to \sigma_2)$ within a batch and disregard the other three compositions to eliminate the calculation of extra repeated losses. This may seem contradictory to the common contrastive learning setting that requires more negative samples for comparison [18, 21]. However, these methods are typically supported by complex data augmentations and a large comparison pool. The stochastic nature of the augmentations and the availability of abundant candidates provide more possibilities for a given feature. While in

(a) CASIA-WebFace Dataset

(b) MS1MV2 Dataset

Figure 5: Several sample images of training set. (a) CASIA-WebFace dataset for training. (b) MS1MV2 dataset for training. Each row represents an identity. These images are the face crops after face alignment and cropping. MS1MV2 is a semi-automatically refined version of the MS-Celeb-1M dataset proposed in [6].

FR, data augmentation is destructive and not employed, and a large batch size (e.g., 8,192 [18]) is generally not applicable.

Upon the integration of the supervised contrastive mask and the single-way N protocol, we modify $Maxneg_i$ in equation (5) and the contrastive loss function from equation (6) to equations (7) and (8), respectively. The identity loss involves processing features h_i and h_{N+i} through distinct augmentation channels. The entire supervision signal, represented by equation (9), incorporates both the identity part and the equation (8).

$$Maxneg_i = \max(\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)), j \in [N{+}1, 2N], y_i \neq y_j, \tag{7}$$

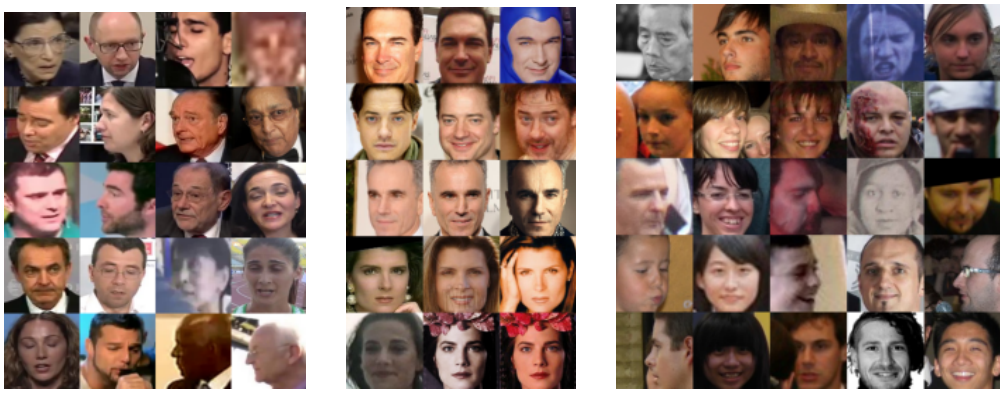
$$\mathcal{L}_{CoRe} = -\log \frac{e^{s(\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_{N+i}) - m_C)}}{e^{s(\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_{N+i}) - m_C)} + \sum_{j=N+1, y_i \neq y_j}^{2N} e^{s\ \text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)}}, \tag{8}$$

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{Cla}(\boldsymbol{h}_i) + \mathcal{L}_{Cla}(\boldsymbol{h}_{N+i})\right) + \lambda \mathcal{L}_{CoRe}(\boldsymbol{h}_i, \boldsymbol{h}_{N+i}). \tag{9}$$

4. Experiments

4.1. Datasets

1) Training Data: We use CASIA-WebFace [34] and MS1MV2 [6] for model training. Specifically, the CASIA-WebFace dataset is applied with



(a) IJB-C Dataset (b) FaceScrub Dataset (c) MegaFace Dataset

Figure 6: Several sample images of the evaluation datasets. (a) IJB-C dataset for evaluation. (b) FaceScrub dataset for evaluation. (c) MegaFace dataset for evaluation.

(a) Cross-Age LFW (b) Cross-Pose LFW

Figure 7: Several sample images of the evaluation datasets. (a) Cross-Age LFW dataset for face verification. (b) Cross-Pose LFW dataset for face verification. The matched and unmatched pairs are bounded with green and red respectively. Each row in the colored boxes represents a verification pair. These two datasets are challenging due to their variations of ages and poses.

Methods (%)	Venue	LFW	AgeDB	CFP-FP	CALFW	CPLFW
CosFace[5]	CVPR 2018	99.81	98.11	98.12	95.76	92.28
ArcFace [6]	CVPR 2019	**99.83**	98.28	98.27	95.45	92.08
MV-Softmax [24]	AAAI 2020	99.80	97.95	98.28	96.10	92.83
CurricularFace [25]	CVPR 2020	99.80	98.32	98.37	**96.20**	93.13
SCF-ArcFace [43]	CVPR 2021	99.82	98.30	98.40	96.12	93.16
MagFace [44]	CVPR 2021	**99.83**	98.17	98.46	96.15	92.87
AdaFace [20]	CVPR 2022	99.82	98.05	98.49	96.08	**93.53**
DDC [45]	TPAMI 2023	99.8	98.0	98.3	96.0	92.3
FaceT-B [46]	TIFS 2023	99.82	98.18	98.23	95.68	92.62
CoReFace	Ours	**99.83**	**98.37**	**98.60**	**96.20**	93.27

Table 1: Verification accuracy (%) on LFW, AgeDB, CFP-FP, CALFW, and CPLFW. The **Best** results are emphasized in bold.

ResNet50 for ablation studies while the MS1MV2 dataset is applied for comparison with other SOTA methods. CASIA-WebFace contains about 0.5M face images of 10K individuals, which are originally collected from IMDb by searching the celebrities' names. MS1MV2 contains about 5.8M face images of 85K individuals. The noises and the potential outliers are removed within each class from the MS-Celeb-1M dataset by a semi-automatic refinement method. Figure 5(a) and Figure 5(b) show some samples randomly selected from these two datasets where each row represents an identity.

2) Testing Data: We extensively evaluate our approach on eight benchmarks, including LFW [1], AgeDB [35], CFP-FP [36], CPLFW [37], CALFW [38], IJB-B [39], IJB-C [40], and MegaFace [41]. Some examples of CALFW [38] and CPLFW [37] are shown in Figure 7 where the positive pairs and the negative pairs are bounded with green and red boxes respectively. IJB-B [39] and IJB-C [40] are used for reliability evaluation. These two datasets provide tons of face pairs, and most of them are unmatched. They ask for high accuracies for positive pairs (TAR) when the ratio of mistakes on negative pairs (FAR) are controlled. We show some examples of IJB-C in Figure 6(a). The variations on the poses, illuminations, and qualities make these two IJB datasets quite challenging. MegaFace [41] takes FaceScrub [42] as a probe set and provides a gallery set that contains 1M images, which makes it a difficult large-scale dataset. Figure 6(b) and Figure 6(c) show some examples of the probe set and the gallery set respectively.

Methods(%)	IJB-B(TAR@FAR)			IJB-C(TAR@FAR)		
	1e-6	1e-5	1e-4	1e-6	1e-5	1e-4
Softmax	46.73	75.17	90.06	64.07	83.68	92.40
SphereFace [4]	39.40	73.58	89.19	68.86	83.33	91.77
CosFace [5]	40.41	89.25	94.01	87.96	92.68	95.56
ArcFace [6]	38.68	88.50	94.09	85.65	92.69	95.74
SCF-ArcFace [43]	-	90.68	94.74	-	94.04	96.09
Magface [44]	42.32	90.36	94.51	**90.24**	94.08	95.97
QMagFace [51]	-	-	94.70	-	-	96.19
SphereFace-R [52]	45.64	86.55	94.51	80.19	93.01	95.96
DDC [45]	-	-	94.7	-	-	96.1
FaceT-B [46]	-	-	94.37	-	-	95.72
CoReFace	**47.02**	**91.33**	**95.09**	89.34	**94.73**	**96.43**

Table 2: 1:1 verification on IJB-B and IJB-C. The **Best** results are emphasized in bold.

4.2. Implementation Details

We follow the settings commonly used in recent works [24, 20, 47, 25, 44, 6] to ensure the fairness of comparison. Multi-task Cascaded Convolutional Networks (MTCNN) [48] is taken to mark the five face landmarks for face alignment. The face images are then cropped and resized to 112×112. Image pixels are normalized by subtracting 127.5 and then dividing by 128.

We employ ResNet50 and ResNet100 [49] as backbones for CASIA-WebFace and MS1MV2 respectively. ArcFace is used as the identity loss. Our framework is implemented with Pytorch [50]. We train the models on 4 NVIDIA A100 GPUs with a batch size of 512. All models are trained using SGD algorithm with an initial learning rate of 0.1. We set the momentum to 0.9 and the weight decay to 5×10^{-4}. On CASIA-WebFace, the training finishes after 40 epochs, and the learning rate is divided by 10 at the 22nd and the 30th epochs respectively. In MS1MV2, we divide the learning rate by 10 at the 8th, the 14th, and the 20th epochs, and stop training after 24 epochs. We set the scale parameter s to 64 for both the identity loss and our contrastive loss, and set λ to 0.05. The possibility of dropout is set to 0.4, following AdaFace. For a fair comparison of the evaluation results, all methods without specifications are implemented with ResNet100 and MS1MV2.

Methods (%)	Id	Ver
CosFace [5]	97.91	97.91
ArcFace [6]	98.35	98.48
MV-Softmax [24]	97.76	97.80
CurricularFace [25]	**98.71**	98.64
BroadFace [47]	98.70	98.95
CircleLoss [53]	98.50	98.73
FaceT-B [46]	97.99	97.92
CoReFace	98.69	**99.06**

Table 3: Face identification and verification on MegaFace Challenge using FaceScrub as the probe set. Id refers to the rank-1 face identification accuracy with 1M distractors, and Ver refers to the face verification **TAR** (@FAR=$1e-6$).

4.3. Experiment Results

1) Results on LFW, CFP-FP, AgeDB, CALFW and CPLFW: Table 1 compares our CoReFace with other recent SOTA approaches on diverse benchmarks. Compared to the original ArcFace, our CoReFace outperforms it on four out of the five datasets with remarkable margins and achieves the same performance on the last one. This improvement is due to the incorporation of contrastive regularization in CoReFace, which successfully addresses the inconsistency problem between identity-based training and sample-based evaluation, which is ignored in previous approaches. Among all the approaches, AdaFace considers image quality during training. This may explain its superior performance on CPLFW, where different poses may cause occlusions on faces and result in lower accuracies. Our method achieves the highest accuracies on the other four datasets. Notably, while our CoReFace has the same performance as ArcFace and CurricularFace on LFW and CALFW, it significantly outperforms them on the other datasets.

2) Results on IJB-B and IJB-C: The IJB-B dataset contains 1,845 subjects with 21.8K still images and 55K frames from 7,011 videos. The IJB-C dataset expands IJB-B, and contains about 3,500 identities with a total of 31.3K images and 117.5K unconstrained video frames. Table 2 shows the performances of different methods for 1:1 verification on IJB-B and IJB-C. Our method achieves the highest True Acceptance Rates (TARs) for nearly all False Acceptance Rates (FARs) on these two datasets. As IJB-B has fewer matches, it becomes the most challenging situation when the FAR is set to 10^{-6}, and only about 8 negative matches are allowed to be wrong. Com-

Setting Groups	Methods	Average
Single Supervision	Identity-only	93.60
	Triplet-only	91.03
Contrastive Only	NT-Xent	63.61
	SupCon	67.87
	CoReFace	86.68
Data Augmentation	NT-Xent	92.78
	SupCon	91.49
Feature Augmentation	NT-Xent	93.60
	SupCon	93.60
	Cos m=0.7	93.60
	CoReFace	**93.75**

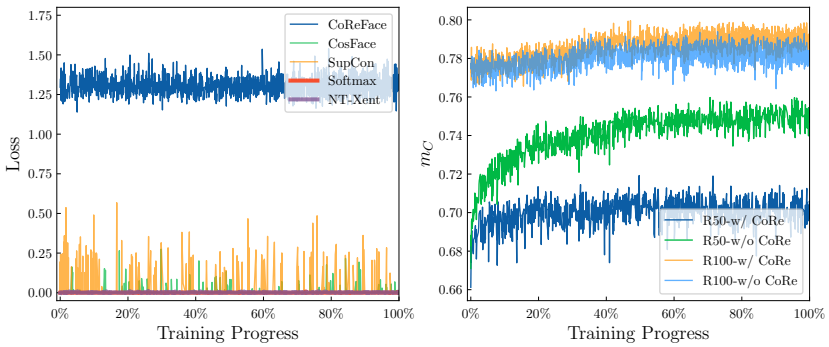
Table 4: Average verification performance (%) of different methods. All experiments are based on a pre-trained ResNet50 ArcFace model with 90.45% average performance. To avoid the influence of the hyper-parameter, $\lambda = 1$ is set for all experiments.

pared with other methods whose TARs are lower than Softmax, our model is more competitive under such an extreme circumstance. Furthermore, when there is a higher FAR bound (e.g., 10^{-4}) or the evaluation dataset is larger, CoReFace still outperforms the competitors.

3) Results on MegaFace: Finally, we demonstrate the efficacy of our method on the MegaFace Challenge. The gallery set of MegaFace contains 1M images of 690K subjects. We follow [6] to remove the face images with wrong labels and evaluate our method on the refined dataset. Table 3 compares the performances of different methods. For the identification task, CoReFace achieves competitive performance which is only 0.02% lower compared to the highest one CurricularFace [25]. For the verification task, CoReFace outperforms all the other approaches with a clear margin, which is also the only one that achieves a TAR higher than 99%. Without the need for complex structure reformations, CoReFace implements an sample-sample regularization to improve the feature distribution and boost the performance of large-scale face recognition.

4.4. Ablation Studies

In this section, we conduct detailed ablation studies from four aspects to demonstrate the effectiveness of our method, including the effects of our framework and the pair-coupling protocol, the speed, the hyper-parameters,



(a) Loss of different contrastive loss functions (b) m_C in different training settings

Figure 8: (a) The loss variations of different contrastive methods in joint training with R100. (b) The adaptive margin (m_C) variations caused by CoReFace on different backbone models. Some methods keep their loss values nearly 0 and fail to supervise in training.

and the feature distribution. As the performance on LFW is almost saturated (with an accuracy of about 99.8% using the ResNet100 model), we report the performances on AgeDB, CFP-FP, CALFW, CPLFW, and their average in our ablation studies.

1) Effects of our framework: In Figure 9, three different frameworks are evaluated: *Original, w/o* L_C, and *w/* L_C. *Original* refers to the traditional identity-based framework, while the latter two settings adopt our framework. Experiments on the identity-based methods, CosFace, and ArcFace demonstrate the effectiveness of our CoReFace loss. Dropout also has a positive influence on the average performance.

Table 4 compares the effectiveness of our contrastive loss function with other alternative methods in various settings. It is observed that the *Contrastive Only* group performs inferiorly compared to the identity-based methods, indicating the importance of incorporating the identity loss into face recognition tasks. The most commonly used contrastive loss function NT-Xent [18] and the label-guided SupCon [54] are taken for comparison. The performance degradation of them in the *Data Augmentation* group, compared to the *Identity-only* method, confirms the semantic damage caused by commonly used image augmentations. The *Feature Augmentation* group, which follows our framework and utilizes CoReFace, shows significant improvement, while the other methods have little effect. Notably, we set $\lambda = 1$ to prevent hyperparameter disturbance, and the improvement of our method becomes

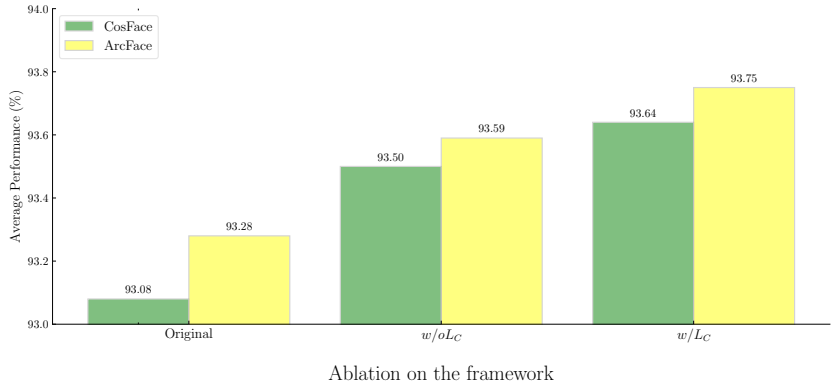


Figure 9: Average verification performance (%) of CosFace and ArcFace. The Original implementation takes no dropout or contrastive loss. *w/o* L_C takes the same setting as our framework but our contrastive loss.

more significant after parameter selection.

Figure 8 further visualizes the contrastive loss values and the adaptive margin m_C during joint training, demonstrating that our method generates stable and reasonable loss values with the aid of the adaptive margin. Figure 8 shows how other contrastive methods fail to provide consistent supervision. The change in m_C with different backbones confirms the adaptability of our method, relieving the need for tedious hyperparameter tuning for different model scales and the ability of our contrastive loss to effectively enhance the difference between the similarities of positive and negative pairs.

The efficacy of our supervised contrastive mask (SMC) is illustrated in Figure 10. The findings suggest that the masked version consistently outperforms the unmasked version. Despite some conflicting situations caused by the stochastic sampling during training, the mask effectively resolves the supervision conflicts between the two signals.

Figure 10 also demonstrates the impact of various pair-coupling protocols. Among the four protocols, the D-2N protocol, D-N protocol, and S-2N protocol include repeated key negative pairs. The D-2N protocol exhibits lower average performance due to a higher number of negative pairs. The single-way N protocol performs better than the others. This supports our assumption that symmetry in pair coupling hinders the performance.

2) Ablation on speed: Figure 11 provides a comparison of the training

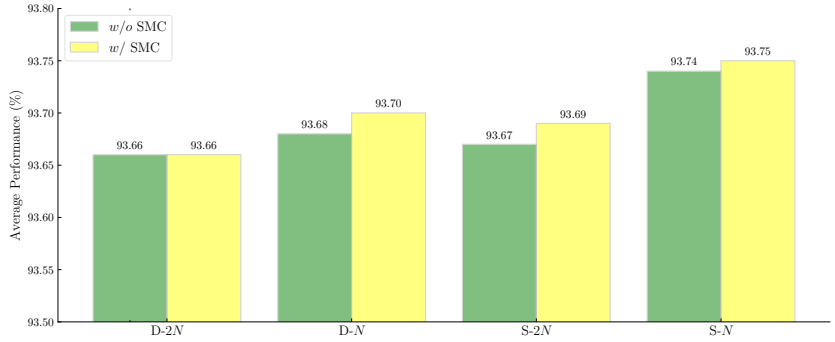
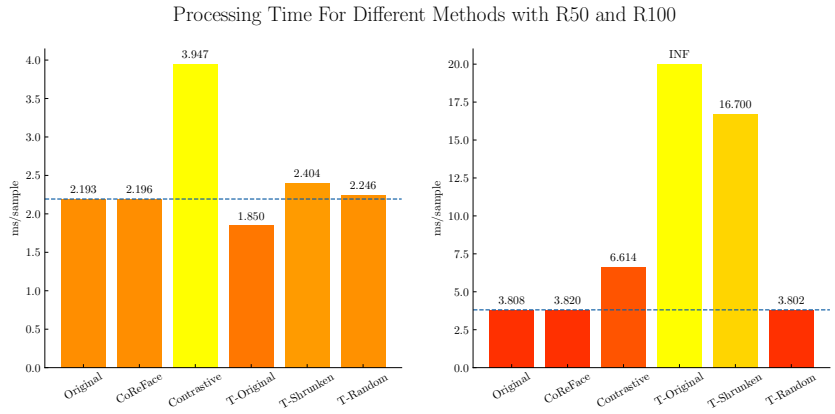


Figure 10: Average verification performances (%) of different pair-coupling protocols and the supervised contrastive mask (SMC). S and D mean *single-way* and *double-way* respectively. N and $2N$ represent the number of candidates for a given sample.



(a) R50-CASIA (b) R100-MS1MV2

Figure 11: Average processing time for a sample with each method on one NVIDIA A100 GPU. (a) Processing time using CASIA-WebFace as training set with ResNet50 as the backbone. (b) Processing time using MS1MV2 as training set with ResNet100 as the backbone. "T-" means triplet loss variations. R100 with original triplet loss needs more than 40GB video memory and fails to be trained. Our CoReFace shows considerable efficiency that is almost the same as the original structure.

Augmentation Type	AgeDB	CFP-FP	CALFW	CPLFW	Average
no	95.08	**95.67**	93.72	89.93	93.60
random noise (0.005)	94.90	95.38	93.78	90.23	93.58
random noise (0.01)	94.98	95.53	93.92	90.25	93.67
random noise (0.03)	94.85	95.36	93.82	90.12	93.54
scaling (0.01)	94.77	95.41	**93.95**	90.17	93.57
dropout (0.4)	**95.03**	95.50	93.92	**90.53**	**93.75**

Table 5: The effects of different feature augmentations. The numbers in the parentheses represent the standard deviation for random noise, scaling factor for scaling, and the probability of discard for dropout.

speed among different frameworks: the original identity-based framework, our feature-augmentation-based CoReFace framework, the data-augmentation-based Contrastive framework, and the Triplet frameworks. The speed is reported for a batch size of 128 for Original, CoReFace, and Contrastive frameworks. We also consider three situations for triplet loss, which calculates one loss value using three images. To make the samples in a batch generally equal, we shrink the batch size of triplet loss to 42 and it possesses a similar number of samples compared with other methods, namely Triplet-Shrunken. We then randomly take the negative samples from the whole dataset with a Triplet-Shrunken batch and name it Triplet-Random.

As shown in Figure 11, our CoReFace framework only incurs negligible extra time, i.e. 1.4‰ for R50 and 3.3‰ for R100, compared to the original identity-based method. In contrast, the common Contrastive framework nearly doubles the processing time. Triplet-Original, which uses a batch size three times larger than Original, is faster for a single sample with R50 and CASIA-WebFace, but cannot be applied on a GPU with 40GB memory for R100 and MS1MV2. Triplet-Shrunken addresses the memory explosion issue, but still faces challenges with the negative sample selection which is time-consuming. Triplet-Random does not have these problems, but lacks control over the selection of negative samples. It performs similar to a degraded version of the Contrastive method, contributing positive pairs and potentially wrong negative pairs.

3) Ablation on different types of feature augmentation: Table 5 shows the effects of different feature augmentation methods. To simplify the design, random noise is a tensor generated from a normal distribution $\mathcal{N}_1$ that is the

λ	AgeDB	CFP-FP	CALFW	CPLFW	Average
0.01	98.23	98.51	96.20	93.12	96.52
0.03	**98.37**	98.60	96.20	93.17	96.58
0.05	**98.37**	98.60	96.20	**93.27**	**96.61**
0.07	98.17	98.59	96.17	93.12	96.51
0.1	98.32	**98.63**	**96.25**	93.17	96.59
0.2	98.30	98.49	96.20	93.10	96.52

Table 6: Verification performances (%) with different λ.

Model		CALFW			CPLFW		
		ArcFace	CoReFace	Diff.	ArcFace	CoReFace	Diff.
R50	+	1322.0	1456.0	↑10.1%	961.4	1172.4	↑21.9%
	−	99.4	93.0	↓6.6%	73.7	78.6	↑6.6%
R100	+	1769.1	1744.2	↓1.4%	1417.7	1380.5	↓0.4%
	−	27.0	26.9	↓2.6%	13.4	11.9	↓11.2%

Table 7: The similarity scores of ArcFace and CoReFace on CALFW and CPLFW with different backbones. + and − denote the positive pairs and the negative pairs respectively.

same size as the features. Linear transformation is achieved by randomly scaling according to the scale from normal distribution $\mathcal{N}_2$. The means are both 0. The standard deviation are controlled by parameters. The generated random noise is scaled by the magnitude of the feature $\|\boldsymbol{h}\|$ and added to the original feature $\boldsymbol{h}$, while the scaling scale factor is multiplied by the original feature $\boldsymbol{h}$ and then added. As shown in the Table 5, when the standard deviation scale is set to a small value (0.01), the average performance is lower than the original method. For random noise, it performs better than the original method within a small range of standard deviation, requiring careful hyperparameter tuning to achieve good results. Meanwhile, dropout outperforms these two feature augmentation methods. It is noticeable that the probability 0.4 for dropout is not from hyperparameter selection, but rather directly inherited from AdaFace. Dropout not only has an impact on the data but also acts as a regularization on the model structure. It is simple and effective without changing the input.

4) Ablation on the effects of hyper-parameters: A larger λ value results in stronger regularization, and forces the features to be more separated. However, this may contradict the intrinsic similarity of face images. Table 6

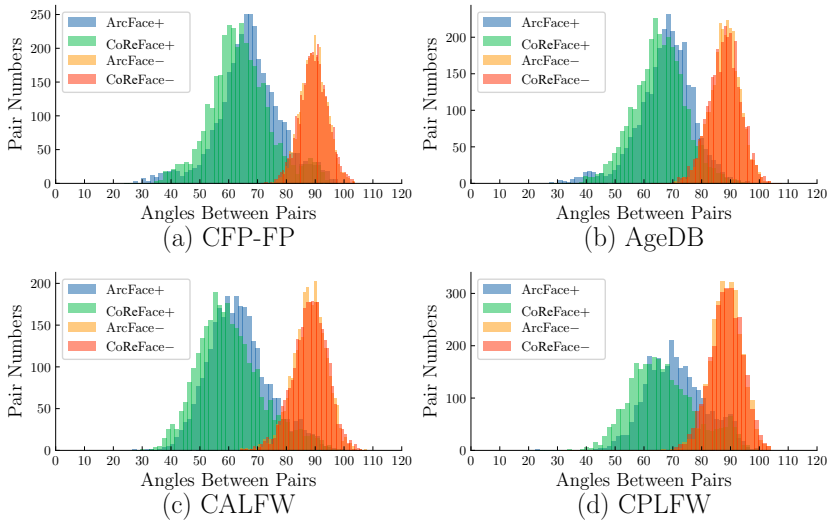
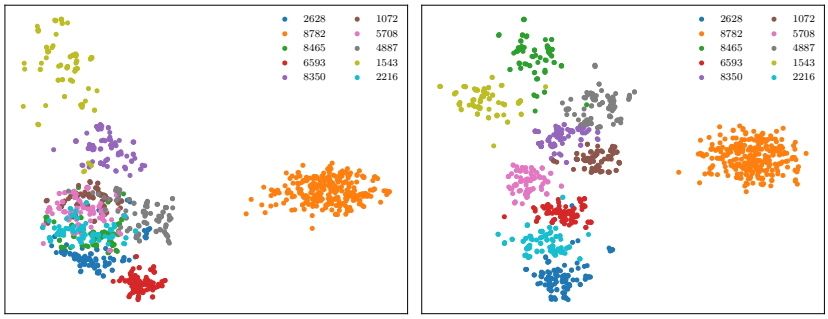


Figure 12: The angle distributions of ArcFace and CoReFace on four datasets training with CASIA-WebFace and R50. + and − denote the positive pairs and the negative pairs respectively.

presents the verification performances with different values of λ. The average performance remains relatively steady within the selected range. This indicates that our method is insensitive to the hyper-parameters. Based on the experimental results, we choose $\lambda = 0.05$ as our final training setting.

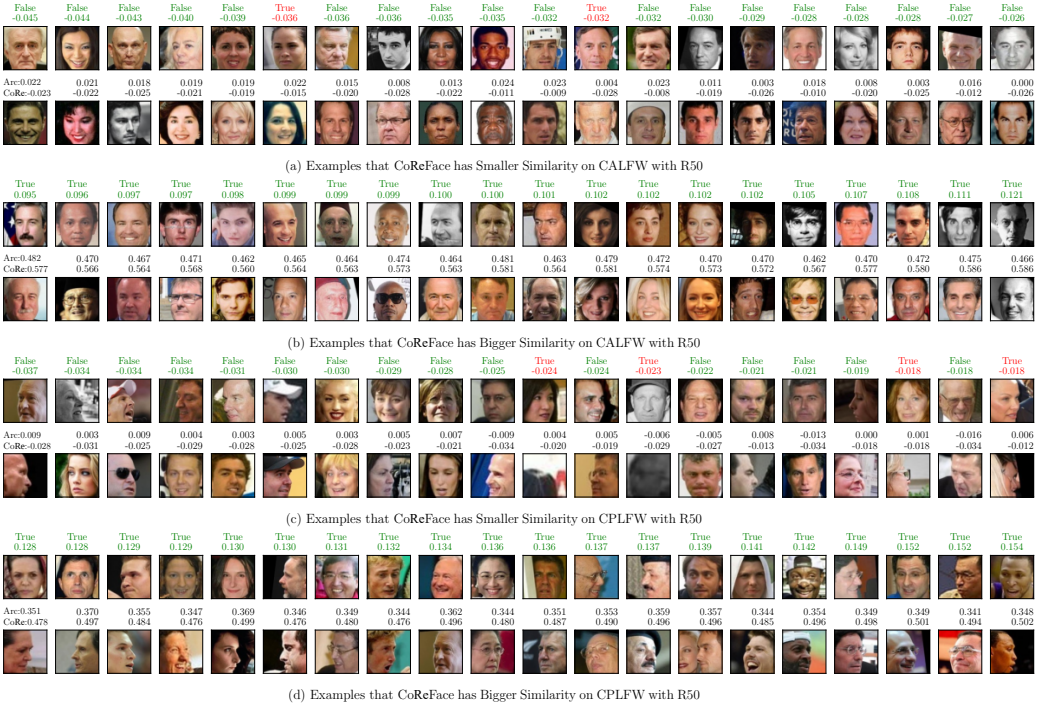
5) Effects on feature distribution: To better understand the impact of our approach on feature space, we visualize the distribution of similarities between the positive and the negative pairs in the evaluation datasets. As seen in Figure 12, the angles of the positive pairs in CoReFace are closer to 0 compared to ArcFace, maintaining a clear margin across datasets with age and pose variations. Furthermore, we show the feature distribution of some randomly selected unseen identities with t-SNE [55] in Figure 13. It illustrates that the feature clusters generated by our CoReFace are more semantically separated. This demonstrates the ability of our method to adjust the feature distribution.

We further investigate how CoReFace changes the similarity distribution by summing up the similarity scores of two methods on datasets with different backbones. Each of CALFW and CPLFW contains 3,000 positive pairs and 3,000 negative pairs respectively. We find that CoReFace changes the



(a) ArcFace (b) Our CoreFace

Figure 13: Visualization for the feature distribution with t-SNE. We show the indexes of the randomly selected unseen identities.



(a) Examples that CoReFace has Smaller Similarity on CALFW with R50

(b) Examples that CoReFace has Bigger Similarity on CALFW with R50

(c) Examples that CoReFace has Smaller Similarity on CPLFW with R50

(d) Examples that CoReFace has Bigger Similarity on CPLFW with R50

Figure 14: Examples of the difference between ArcFace and CoReFace with R50 as backbone. The label and the similarity subtraction are placed on the top of a pair. The similarity scores of the two methods are put in the middle of a pair. If CoReFace has a proper change in similarity on the basis of the pair label, it will be painted green. Otherwise, we show it in red. The samples in (a) and (c) are from the intersection of 15%∼25% of the similarity distributions of the two methods, while those in (b) and (d) are from the intersection of 75%∼85% of the similarity distributions of the two methods.

similarity distribution differently on R50 and R100. We observe different changes in the similarity distribution for R50 and R100 as shown in Table 7. Compared with ArcFace, the subtraction of the similarity score summation on the positive pairs and the negative pairs is increased by more than 15% on CALFW and CPLFW with R50 as the backbone. This demonstrates the effectiveness of re-distribution with our method on a relatively small model. For R100, the model's representation ability is already strong enough, but our method still imposes greater penalties on the negative pairs.

Figure 14 displays the face image pairs and the similarity scores of them computed by ArcFace and CoReFace. The images are from the intersection of the same similarity distribution segments of the two methods, with the 20 pairs showing the largest absolute values of the subtraction. For the samples where CoReFace has smaller scores, both methods have low accuracies. The main difference is that our method produces more and lower negative scores. However, for another group of samples where CoReFace has larger scores, our approach shows an angular margin on all the pairs. This implies that our method shows better discrimination ability. Generally, CoReFace demonstrates its superiority in dealing with pose and age variations.

5. Conclusion

We have presented our CoReFace to regulate the feature distribution based on the sample-sample relationship and align the training with the evaluation in face recognition. This is achieved by integrating sample-guided contrastive learning in our framework. To address the degradation caused by the commonly-used data augmentations, we augment the embeddings instead of images for positive pair composition in contrastive learning. By incorporating an adaptive margin and a supervised contrastive mask, our contrastive loss generates steady loss values and avoids collision with the identity supervision signals. Additionally, the new pair-coupling protocol alleviates the similarity problem caused by the symmetry of pairs. Extensive experiments on the popular face recognition benchmarks and ablations demonstrate the effectiveness and efficiency of our proposed approach and highlight the great potential of contrastive learning for regularization in face recognition. Our concise framework allows for an easy application to existing FR methods.

Limitations and future work: Although using dropout at the end of the CNN backbone for feature augmentation reduces forward propagation time by nearly half, it also limits the diversity compared to data augmentation.

This limitation restricts the potential of contrastive learning. Furthermore, if we could learn augmented images, face recognition in low-quality scenarios would greatly benefit. A new backbone-only approach that uses augmented images and contrastive learning to improve low-quality face recognition is worth exploring.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61375016) and ECNU Multifunctional Platform for Innovation (001) Center for High Performance Computing.

References

[1] G. B. Huang, M. A. Mattar, T. L. Berg, E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, in: Tech. Rep., 2007.

[2] K. Zhang, D. Zheng, J. Li, X. Gao, J. Lu, Coupled discriminative manifold alignment for low-resolution face recognition, Pattern Recognition 147 (2024) 110049.

[3] J. Yang, Z. Wang, B. Huang, J. Xiao, C. Liang, Z. Han, H. Zou, Headpose-softmax: Head pose adaptive curriculum learning loss for deep face recognition, Pattern Recognition 140 (2023) 109552.

[4] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: Proc. of CVPR, 2017, pp. 6738–6746.

[5] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: Large margin cosine loss for deep face recognition, in: Proc. of CVPR, 2018, pp. 5265–5274.

[6] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proc. of CVPR, 2019, pp. 4690–4699.

[7] Y. Duan, J. Lu, J. Zhou, Uniformface: Learning deep equidistributed representation for face recognition, in: Proc. of CVPR, 2019, pp. 3415–3424.

[8] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: Proc. of ECCV, 2016.

[9] K. Zhao, J. Xu, M. Cheng, Regularface: Deep face recognition via exclusive regularization, in: Proc. of CVPR, 2019, pp. 1136–1144.

[10] S.-M. Yang, W. Deng, M. Wang, J. Du, J. Hu, Orthogonality loss: Learning discriminative representations for face recognition, IEEE Transactions on Circuits and Systems for Video Technology (2021). doi:10.1109/TCSVT.2020.3021128.

[11] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: CVPR, 2006.

[12] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proc. of CVPR, 2015, pp. 815–823.

[13] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Proc. of NeurIPS, 2014, pp. 1988–1996.

[14] Y. Sun, X. Wang, X. Tang, Deeply learned face representations are sparse, selective, and robust, in: Proc. of CVPR, 2015, pp. 2892–2900.

[15] S. Horiguchi, D. Ikami, K. Aizawa, Significance of softmax-based features in comparison to distance metric learning-based features, IEEE Trans. Pattern Anal. Mach. Intell. (2020).

[16] M. Ye, X. Zhang, P. C. Yuen, S. Chang, Unsupervised embedding learning via invariant and spreading instance feature, in: Proc. of CVPR, 2019, pp. 6210–6219.

[17] Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proc. of CVPR, 2018, pp. 3733–3742.

[18] T. Chen, S. Kornblith, M. Norouzi, G. E. Hinton, A simple framework for contrastive learning of visual representations, in: Proc. of ICML, Vol. 119, 2020, pp. 1597–1607.

[19] Y. Shi, X. Yu, K. Sohn, M. Chandraker, A. K. Jain, Towards universal representation learning for deep face recognition, in: Proc. of CVPR, 2020, pp. 6816–6825.

[20] M. Kim, A. K. Jain, X. Liu, AdaFace: Quality adaptive margin for face recognition, in: CVPR, 2022.

[21] K. He, H. Fan, Y. Wu, S. Xie, R. B. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proc. of CVPR, 2020, pp. 9726–9735.

[22] X. Chen, K. He, Exploring simple siamese representation learning, in: Proc. of CVPR, 2021.

[23] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: EMNLP, 2021.

[24] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, T. Mei, Mis-classified vector guided softmax loss for face recognition, in: Proc. of AAAI, 2020.

[25] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, F. Huang, Curricularface: Adaptive curriculum learning loss for deep face recognition, in: Proc. of CVPR, 2020, pp. 5900–5909.

[26] W. Hu, H. Hu, Orthogonal modality disentanglement and representation alignment network for nir-vis face recognition, IEEE Transactions on Circuits and Systems for Video Technology 32 (6) (2022) 3630–3643. doi:10.1109/TCSVT.2021.3105411.

[27] W. Hu, H. Hu, Dual adversarial disentanglement and deep representation decorrelation for nir-vis face recognition, IEEE Transactions on Information Forensics and Security 16 (2021) 70–85. doi:10.1109/TIFS.2020.3005314.

[28] F. Wang, X. Xiang, J. Cheng, A. L. Yuille, Normface: L_2 hypersphere embedding for face verification, in: Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017, 2017, pp. 1041–1049.

[29] R. R. an, L2-constrained softmax loss for discriminative face verification, ArXiv preprint abs/1703.09507 (2017).

[30] Y. Zheng, D. K. Pal, M. Savvides, Ring loss: Convex feature normalization for face recognition, in: Proc. of CVPR, 2018, pp. 5089–5097.

[31] X. Zhang, Z. Fang, Y. Wen, Z. Li, Y. Qiao, Range loss for deep face recognition with long-tailed training data, in: Proc. of ICCV, 2017, pp. 5419–5428.

[32] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. (2014).

[33] K. R. K. an, Dropout as data augmentation, ArXiv preprint abs/1506.08700 (2015).

[34] D. Yi, Z. Lei, S. Liao, S. Z. Li, Learning face representation from scratch, ArXiv preprint abs/1610.02915 (2016).

[35] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, S. Zafeiriou, Agedb: The first manually collected, in-the-wild age database, in: CVPR, 2017.

[36] S. Sengupta, J. Chen, C. D. Castillo, V. M. Patel, R. Chellappa, D. W. Jacobs, Frontal to profile face verification in the wild, in: WACV, 2016.

[37] T. Zheng, W. Deng, Cross-pose lfw : A database for studying cross-pose face recognition in unconstrained environments, in: Tech. Rep., 2018.

[38] T. Zheng, W. Deng, J. Hu, Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments, arXiv:1708.08197 (2017).

[39] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. C. Adams, T. Miller, N. D. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, P. Grother, IARPA janus benchmark-b face dataset, in: CVPR, 2017.

[40] B. Maze, J. C. Adams, J. A. Duncan, N. D. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, P. Grother, IARPA janus benchmark - C: face dataset and protocol, in: ICB, 2018.

[41] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, E. Brossard, The megaface benchmark: 1 million faces for recognition at scale, in: Proc. of CVPR, 2016, pp. 4873–4882.

[42] H. Ng, S. Winkler, A data-driven approach to cleaning large face datasets, in: ICIP, IEEE, 2014, pp. 343–347.

[43] S. Li, J. Xu, X. Xu, P. Shen, S. Li, B. Hooi, Spherical confidence learning for face recognition, in: CVPR, 2021.

[44] Q. Meng, S. Zhao, Z. Huang, F. Zhou, MagFace: A universal representation for face recognition and quality assessment, in: CVPR, 2021.

[45] B. Uzun, H. Cevikalp, H. Saribas, Deep discriminative feature models (ddfms) for set based face recognition and distance metric learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (5) (2023) 5594–5608.

[46] Y. Zhu, M. Ren, H. Jing, L. Dai, Z. Sun, P. Li, Joint holistic and masked face recognition, IEEE Transactions on Information Forensics and Security 18 (2023) 3388–3400.

[47] Y. Kim, W. Park, J. Shin, BroadFace: Looking at tens of thousands of people at once for face recognition, in: Proc. of ECCV, 2020.

[48] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (10) (2016) 1499–1503.

[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. of CVPR, 2016, pp. 770–778.

[50] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, in: NeurIPS Workshop, 2017.

[51] P. Terhörst, M. Ihlefeld, M. Huber, N. Damer, F. Kirchbuchner, K. B. Raja, A. Kuijper, Qmagface: Simple and accurate quality-aware face recognition, in: WACV, 2023, pp. 3473–3483.

[52] W. Liu, Y. Wen, B. Raj, R. Singh, A. Weller, Sphereface revived: Unifying hyperspherical face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2) (2023) 2458–2474.

[53] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, Y. Wei, Circle loss: A unified perspective of pair similarity optimization, in: Proc. of CVPR, 2020, pp. 6397–6406.

[54] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: Proc. of NeurIPS, 2020.

[55] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.

Youzhe Song received the B.E. degree in School of Computer Science and Technology from the DongHua University (DHU), Shanghai, China, in 2021. He is currently pursuing the M.E. degree at the School of Computer Science and Technology from the East China Normal University (ECNU), Shanghai, China. His research interests include face recognition and open-set retrieval.

Feng Wang received his PhD in Computer Science from the Hong Kong University of Science and Technology in 2007 and BsC from Fudan University, China, in 2001 respectively. Before joining East China Normal University as an associate professor in the Department of Computer Science and Technology, he was a research fellow in the City University of Hong Kong and the Institute Eurecom, France. His research interests include multimedia information retrieval, pattern recognition, and IT in education.