QAMFACE: QUADRATIC ADDITIVE ANGULAR MARGIN LOSS FOR FACE RECOGNITION

He Zhao, Yongjie Shi, Xin Tong, Xianghua Ying, Hongbin Zha

Key Lab of Machine Perception (MOE)
Peking University

ABSTRACT

The angular-based softmax losses and their variants achieve great success in face recognition based on deep learning. ArcFace [1] which directly maximize decision boundary in angular space is one of the most popular and effective loss function. In this paper, we analyze the inherent limitations of ArcFace, including the non-monotonic logit and gradient curve, and inappropriate trend of loss value. To address these problems, we propose a novel loss function named the Quadratic Additive Angular Margin Loss (QAMFace). It takes the value of the angle through a quadratic function rather than cosine function as the target logit. Our QAMFace is easy to implement and only adds negligible computational overhead. Experiments on several relevant benchmarks show that QAMFace performs better in convergence on feature embedding, and consistently outperforms the state-of-the-art face recognition methods. Our codes will be released soon.[1]

Index Terms— Face Recognition, Loss Function, Margin

1. INTRODUCTION

Face recognition (FR) is one of the most widely studied topics in computer vision. Recent years have witnessed the great success of convolutional neural networks (CNNs) in face recognition [1, 2, 3]. The deep convolutional networks map the face image, typically after a pose normalization step, into an embedding feature vector. Those features of the same person are close to each other, while features of different individuals have considerable distances.

The successes of deep face CNNs can be mainly credited to three attributes. The first is the training data. In the past three decades, many face databases have been constructed with a clear tendency from small-scale to large-scale, from single-source to diverse-sources. Three databases with over one million images, namely MS-Celeb-1M [4], VGGface2 [5], and Megaface [6] lead breakthrough on several benchmarks. The second attribute is the network architecture. High capacity CNNs, such as ResNet [7] and Inception-ResNet [8] can obtain better performance compared to VGGNet [9] and GoogLeNet [10]. The third attribute is the design of the loss functions, which play critical roles in learning to extract discriminative face features.

We divide the design of loss functions for face recognition into two main streams. Softmax-based methods [11, 12] train a multi-class classifier that can separate different identities of the training set, while other methods directly learn an embedding, such as the triplet loss [13]. As the combinatorial explosion in the number of face triplets for the triplet loss and the training difficulty induced by semi-hard mining, we focus on softmax-based loss in this paper. Recently, several variants have been proposed to enhance the discriminative power of the softmax loss. Center Loss [14] improves the intra-class compactness by imposing additional loss term that penalizes the Euclidean distance between samples and their representation centers. Liu *et al.* proposed the A-Softmax loss in SphereFace [3] that imposes a multiplicative angular margin penalty to enforce extra intra-class compactness. In order to stabilize training, ArcFace [1] directly adds additive margin penalty and obtains state-of-the-art performance.

Although the efficiency of ArcFace has been proved by several previous works [15, 16, 17], it still has two limitations. First, the gradient curve of ArcFace is not monotonic, and the geometric decision margins of ArcFace have a part of overlap, which increase the convergence difficulty. Second, under the same inter-class margin, ArcFace loss increases while the target angle is decreasing, which is contrary to the motivation to narrow the intra-class distance. Considering these problems, we carefully design a novel loss function named the Quadratic Additive Angular Margin Loss (QAMFace). QAMFace takes a quadratic function rather than cosine function as target logit, which directly optimizes the geodesic distance margin and performs better in convergence. The main contributions of this paper are summarized as follows:

- We analyze the limitations of ArcFace, the state-of-the-art loss function of face recognition and provide reasonable theoretical interpretation.
- Considering these problems, we propose a novel loss function called QAMFace, to improve the discriminative power of the face features and stabilize the training process.
- QAMFace is easy to implement and only adds negligible computational complexity. We achieve state-of-the-art performance on several public benchmarks, including pose, age variant datasets and large-scale unconstrained datasets.

[1] https://github.com/MccreeZhao/QAMFace

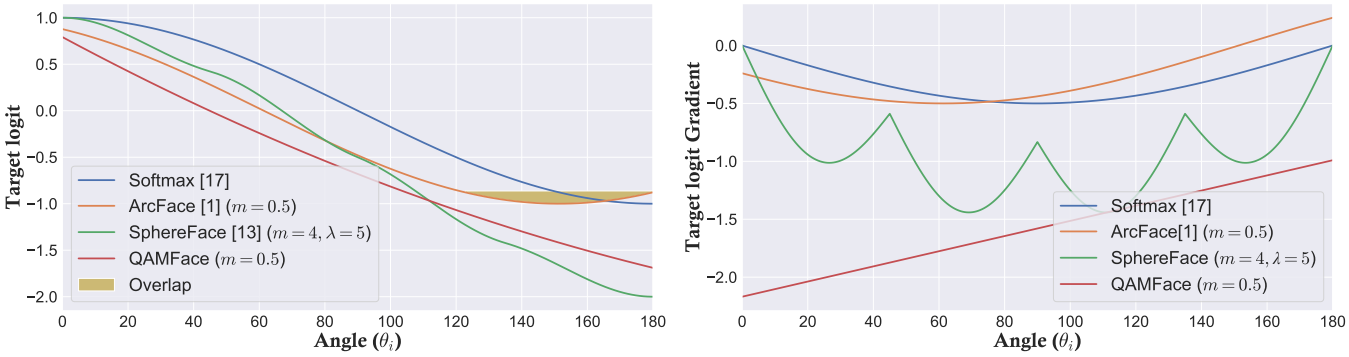

Fig. 1. Curves of target logits (Left) and the target logits gradient (Right) with respect to θ_i, the angle between the target feature vector and the class weight. The target logit of ArcFace has a part of overlap area, and its gradient curve is not monotonic. Our QAMFace avoids these problems and perform better in convergence. (To better display the target logit, we add a bias term on QAMFace which do not affect the calculation of loss and gradient.)

2. RELATED WORKS

Softmax is the most widely used classification loss funcion [12, 11], which is presented as:

$$\mathcal{L}_s = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}, \tag{1}$$

where $x_i \in \mathbb{R}^d$ denotes the deep feature of the i-th sample, belonging to the y_i-th class. The dimension of the embedding feature is set as d. $W_j \in \mathbb{R}^d$ denotes the j-th column of the weight $W \in \mathbb{R}^{d\times n}$, and $b_j \in \mathbb{R}$ is the bias term of the final fully connected layer. The $W_{y_i}^T x_i + b_{y_i}$ is also called the target logit. The batch size and the class number are represented with N and n, respectively. The traditional softmax loss could only separate each sample but could not explicitly learn discriminative features to minimize inter-class similarity and intra-class distance.

In **SphereFace** [3] and **NormFace** [18], the bias term is being removed and the target logit is transformed as: $W_j^T x_i = \|W_j\|\|x_i\|\cos\theta_j$. Then, they fix $\|W_j\| = 1$ and $\|x_i\| = s$ by ℓ_2 normalization, which make the predictions only depend on the angle between the feature vector and the weight. Therefore, the distance metric of training and test process is unified. The modified loss function can be formulated as :

$$\mathcal{L}_{ns} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cdot\cos\theta_{y_i}}}{\sum_{j=1}^{n} e^{s\cdot\cos\theta_j}} \tag{2}$$

As the embedding features are distributed around each feature center on the hypersphere, **ArcFace** [1] adds an additive angular margin penalty m between x_i and W_{y_i} to simultaneously enhance the intra-class compactness and inter-class discrepancy. ArcFace loss is presented as:

$$\mathcal{L}_{Arc} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{3}$$

3. PROPOSED METHOD

In this section, we first introduce QAMFace, our novel loss function. Then, we analysis the improvement of QAMFace compared with previous methods.

3.1. QAMFace

Most of the previous softmax-based loss functions take cosine function and its variants as the target logit. Different from them, our QAMFace takes the quadratic function as the logit. We first constructed a quadratic function: $f(x) = (2\pi - x)^2$. Then, an additive angular margin m is added to minimize the intra-class distances and the target logit can be formulated as : $f(\theta) = s\left(2\pi - (\theta + m)\right)^2$. In the same way, we remove the bias term and fix the length of W_j and x_i by ℓ_2 normalization. Finally, the proposed QAMFace is defined as:

$$\mathcal{L}_{QAM} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\left(2\pi-\left(\theta_{y_i}+m\right)\right)^2}}{e^{s\left(2\pi-\left(\theta_{y_i}+m\right)\right)^2} + \sum\limits_{j=1,j\neq y_i}^{n} e^{s(2\pi-\theta_j)^2}}, \tag{4}$$

subjects to

$$W_j = Wj/\left\|W_j\right\|, x_i = x_i/\left\|x_i\right\|, \theta = \arccos(W_j^T x_i).$$

3.2. Comparison on Different Loss Functions

There are two main advantages of using the quadratic function to replace the cosine function.

Firstly, the target logit and logit gradient curves of QAMFace is monotonic when the angle between the feature vector and the weight belongs to $[0, \pi]$. Therefore, QAMFace has a stronger convergence ability than previous methods. As shown in Figure 1, the logit curve of ArcFace has a part of overlap when the θ_i deviates too much from the center W_{y_i}, which may mislead the convergence direction of the network. SphereFace [3] defines a monotone decreasing function $\psi\left(\theta_{y_i,i}\right) = (-1)^k\cos\left(m\theta_{y_i,i}\right) - 2k$ to replace the original cosine function. However, the training of SphereFace is not stable enough. To avoid divergence at the beginning

of training, they employ an annealing optimization strategy and introduces new hyperparameters that need to be tuned carefully. Different from them, our quadratic logit function is naturally monotone decreasing without any overlap area. Besides, the gradient curve of our logit is linear and monotonic. During the training process, the loss value decreases gradually with the target angle θ_i. At the same time, the absolute value of gradient in ArcFace and Softmax also decrease, which will decelerate the convergence speed. On the contrary, the absolute gradient of our QAMFace is monotonically increasing, which continuously strengthens the supervision.

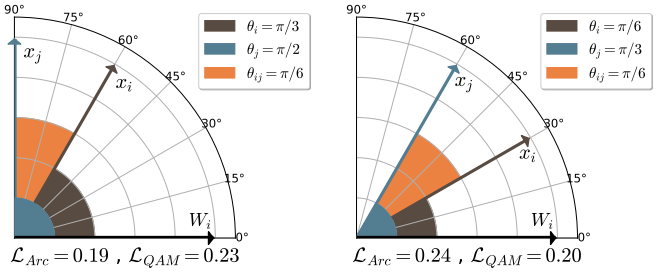

Fig. 2. Two binary classification cases which share the same inter-class distance while have different intra-class compactness.

Secondly, our QAMFace is more accordant with the target of the face recognition algorithm, which is minimizing the intra-class distance. To simplify and better visualize this problem, we analyze the binary classification case and provide an example. In Figure 2, x_i and x_j denote two deep features of samples belong to class y_i and y_j. W_i denotes the i-th column of the weight W, which can be seen as the class center of class y_i. The purpose of our network is to minimize θ_i, the intra-class distance between x_i and W_i, and maximize θ_{ij}, the inter-class distance. The two training phases shown in Figure 2 share the same inter-class margin ($\theta_{ij} = \pi/6$), and the θ_i in the right side is smaller, which represents higher intra-class compactness. Therefore, We should naturally assign a larger penalty to the left case. However, the value of ArcFace loss on the left side (0.19) is smaller than the right (0.24), which is not conducive to the convergence. Different from that, the value of our QAMFace decreases from 0.23 to 0.20 with the target angle θ, which is proceeding as expected. In addition, we present the complete loss curve of different loss functions under a fixed θ_{ij}. As shown in Figure 3, our QAMFace is monotonically increasing with the intra-class distance, which leads to better convergence.

4. EXPERIMENT

4.1. Implementation Details

Data Preparation We employ MS-Celeb-1M dataset [4] as our training data, which contains about 3.8M images of 85k unique identities. During the test, besides LFW [19], the most widely used benchmark, we also report the performance of QAMFace on the recent large-pose and large-age datasets

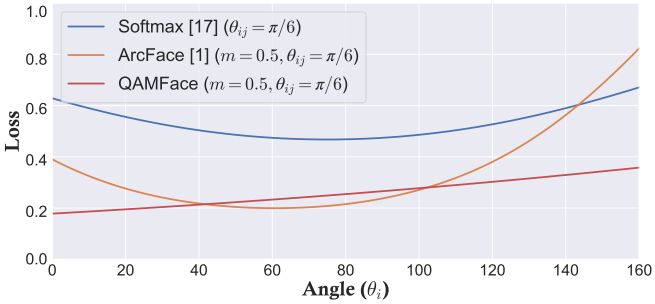

Fig. 3. The loss curve of QAMFace and other loss functions where the inter-class distance θ_{ij} is fixed to $\pi/6$.

CPLFW [15] and CALFW [16]. In addition, we test the proposed QAMFace on large-scale dataset IJB-C [20] and video face dataset YTF [21].

Training Details For data prepossessing, we employ a facial landmark detection algorithm MTCNN [22] to detect and align all faces in training and testing. Specifically, each face is cropped from the detected region and aligned to 112×112 images with similarity transformation based on the detected five facial landmarks. During the training, the random horizontal flip is applied as data augmentation. For the embedding network, we employ the same CNN architectures, modified ResNet50 as ArcFace [1]. It produces 512-dimension discriminative features for each image.

The feature scale s, and the angular margin m is set to 6 and 0.5. We set the batch size to 256, momentum to 0.9, and weight decay to $5e{-}4$. The network parameters are optimized by stochastic gradient descent. The learning rate starts from $2e-1$ and is divided by 10 at 80K, 200K, 360K iterations. We finish the training process at 400K iterations.

Table 1. Verification accuracy (%) of QAMFace on LFW, CPLFW, and CALFW, compared with other state-of-the-arts. Human-I/F represent individual and fusion results.

Method	LFW(%)	CPLFW(%)	CALFW(%)
Center Loss [14]	98.75	77.48	85.48
SphereFace [3]	99.27	81.40	90.30
VGGFace2 [5]	99.43	84.00	90.57
ArcFace [1]	99.82	**92.08**	**95.87**
Human-I [15, 16]	97.27	81.21	82.32
Human-F [15, 16]	**99.85**	85.24	86.50
QAMFace	99.82	**92.85**	**96.10**

4.2. Evaluation on LFW, CPLFW and CALFW

First, we conduct experiments on three commonly used benchmarks LFW [19], CPLFW [15], and CALFW [16]. LFW is one of the most widely utilized benchmarks for unconstrained face verification. We report the mean verification accuracy via ten-fold cross-validation on its 6,000 pairs of faces. Due to big data-driven deep learning methods, the performance of LFW gradually becomes saturated (over 99%).

To further confirm the efficiency of our QAMFace, we conducted additional experiments on Cross-Pose LFW (CPLFW) and Cross-Age LFW (CALFW). They focus on pose and age variation, two main challenges of face recognition. Compared to LFW, the accuracy of previous methods on these two benchmarks drops about 10% - 20% under the same verification protocols.

We compare QAMFace loss against many existing state-of-the-art face recognition methods in Table 1. From the results, we can see that the proposed QAMFace outperforms the CenterLoss, SphereFace, and VGGFace2 by a significant margin and is comparable with ArcFace and Human-Fusion results in LFW. Furthermore, our QAMFace not only exceeds human-fusion results but also reduces the error of CPLFW by 9.7% and CALFW by 5.6% compared with ArcFace, the previous state-of-the-art method. Furthermore, we plot the angle distribution histogram of both positive and negative pairs on CPLFW and CALFW in Figure 4. As can be seen, QAMFace has a closer positive pair distance (left shift) and enlarge the inter-class discrepancy (right shift). The results indicate that the proposed QAMFace further enhances the discriminative power of deeply learned face features.

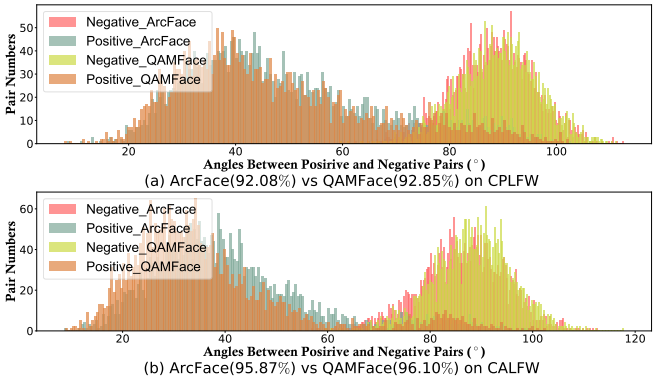

Fig. 4. Angle distributions of positive and negative pairs on CPLFW and CALFW.

4.3. Evaluation on YTF and IJB-C

We then evaluate our method on video face dataset YTF (The YouTube Face) and large-scale template face dataset IJB-C. These two benchmarks are closer to real-world surveillance scenarios and environments. YTF contains 3,425 videos of 1,595 different subjects. We follow the *unrestricted with labeled outside data protocol* to report the average accuracy with cross-validation on ten folds of 5,000 video pairs. IJB-C is a challenging unconstrained benchmark that contains wide pose and resolution variations. It consists of 31,334 images and 11,779 videos from 3,531 subjects. Here, we report the true-accept rates (TAR) vs. false-accept rates (FAR) under *1:1 face verification protocol*.

The verification accuracy of our QAMFace and other methods on YTF are presented in Table 2. It can be seen that our QAMFace outperforms most of the state-of-the-art loss functions by a large margin and achieves a mean accuracy of 98.02%. The performance of QAMFace is comparable with ArcFace [1], which utilizes a deeper network ResNet-100 as its backbone. The bridged gap of computational complexity proves the effectiveness of our method.

Table 2. Verification accuracy (%) of QAMFace on YTF, compared with other state-of-the-arts.

Method	Backbone	Data	Accuracy (%)
FaceNet [13]	GoogLeNet	200 M	95.52
CenterFace [14]	-	0.7M	94.90
NormFace [18]	ResNet-28	0.5M	94.72
CosFace [23]	ResNet-64	5 M	97.60
ArcFace [1]	ResNet-100	3.8M	**98.02**
RegularFace [24]	ResNet-20	3.2M	96.70
AFRN [25]	ResNet-100	3.2M	97.10
QAMFace	ResNet-50	3.8M	**98.02**

Table 3. Verification accuracy (%) of QAMFace on IJB-C compared with other state-of-the-arts (TAR@FAR). The results of previous methods are from [17].

Method	1:1 Verification TAR		
	FAR=1e-5	FAR=1e-4	FAR=1e-3
FaceNet [13]	33.30	48.69	66.45
Crystal Loss [26]	87.35	92.29	95.63
CosFace [23]	86.94	91.82	95.37
ArcFace [1]	87.28	92.13	95.55
AdaCos [17]	88.03	92.40	95.65
AFRN [25]	**88.30**	**93.00**	**96.30**
QAMFace	**91.02**	**94.27**	**96.26**

Table 3 presents the numerical results of different methods on IJB-C. Our QAMFace consistently achieves better performance compared with the state-of-the-art methods, especially on the low FAR cases. Although the performance of AFRN [25] is better than QAMFace at high FAR, it is noteworthy that the backbone of AFRN is a modified ResNet-101, which is much deeper than our ResNet-50. Besides, it performs low-rank bi-linear pooling on the feature maps, which is time-consuming, while our method is simple yet effective.

5. CONCLUSION

In this paper, we propose a Quadratic Additive Angular Margin Loss function, which can effectively enhance the discriminative power of embedded features for face recognition. We provided well-formed geometrical and theoretical interpretations to verify the effectiveness of our QAMFace. Comprehensive experiments on LFW, CPLFW, CALFW, YTF, and IJB-C demonstrate that our method consistently outperforms other state-of-the-art loss functions. The code and details will be released under the MIT license.

6. REFERENCES

[1] Jiankang Deng, Jia Guo, Niannan Xue, et al., "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.

[2] Wang Feng, Weiyang Liu, Haijun Liu, and Cheng Jian, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. PP, no. 99, pp. 1–1, 2018.

[3] Weiyang Liu, Yandong Wen, Zhiding Yu, et al., "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017, pp. 212–220.

[4] Yandong Guo, Lei Zhang, Yuxiao Hu, et al., "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*. Springer, 2016, pp. 87–102.

[5] Qiong Cao, Li Shen, Weidi Xie, et al., "Vggface2: A dataset for recognising faces across pose and age," in *FG*. IEEE, 2018, pp. 67–74.

[6] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *CVPR*, 2016, pp. 4873–4882.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[8] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[9] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] Christian Szegedy, Wei Liu, Yangqing Jia, et al., "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[11] Yaniv Taigman, Ming Yang, et al., "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014, pp. 1701–1708.

[12] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al., "Deep face recognition.," in *BMVC*, 2015, vol. 1, p. 6.

[13] Florian Schroff, Kalenichenko, et al., "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[14] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*. Springer, 2016, pp. 499–515.

[15] T. Zheng and W. Deng, "Cross-pose LFW:: A database for studying cross-pose face recognition in unconstrained environments," Tech. Rep. 18-01, Beijing University of Posts and Telecommunications, February 2018.

[16] Tianyue Zheng, Weihong Deng, and Jiani Hu, "Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments," *CoRR*, vol. abs/1708.08197, 2017.

[17] Xiao Zhang, Rui Zhao, Yu Qiao, et al., "Adacos: Adaptively scaling cosine logits for effectively learning deep face representations," in *CVPR*, 2019, pp. 10823–10832.

[18] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille, "Normface: l 2 hypersphere embedding for face verification," in *ACM MM*. ACM, 2017, pp. 1041–1049.

[19] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[20] Brianna Maze, Jocelyn Adams, James A Duncan, et al., "Iarpa janus benchmark-c: Face dataset and protocol," in *ICB*. IEEE, 2018, pp. 158–165.

[21] Lior Wolf, Tal Hassner, and Itay Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR*, 2011, pp. 529–534.

[22] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[23] Hao Wang, Yitong Wang, Zhou, et al., "Cosface: Large margin cosine loss for deep face recognition," in *CVPR*, 2018, pp. 5265–5274.

[24] Kai Zhao, Jingyi Xu, and Ming-Ming Cheng, "Regularface: Deep face recognition via exclusive regularization," in *CVPR*, 2019, pp. 1136–1144.

[25] Bong-Nam Kang, Yonghyun Kim, Bongjin Jun, and Daijin Kim, "Attentional feature-pair relation networks for accurate face recognition," in *ICCV*, October 2019.

[26] Rajeev Ranjan, Ankan Bansal, Hongyu Xu, et al., "Crystal loss and quality pooling for unconstrained face verification and recognition," *arXiv preprint arXiv:1804.01159*, 2018.