Ada^{2}NPT: An Adaptive Nearest Proxies Triplet Loss for Attribute-Aware Face Recognition with Adaptively Compacted Feature Learning

Lei Ju LJ00434@SURREY.AC.UK
Zhenhua Feng Z.FENG@SURREY.AC.UK
Muhammad Awais MUHAMMAD.AWAIS@SURREY.AC.UK
Josef Kittler J.KITTLER@SURREY.AC.UK
School of Computer Science and Electronic Engineering, University of Surrey, UK

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Attribute-aware face recognition has gained increasing attention in recent years due to its potential to improve the robustness of face recognition systems. However, this may raise concerns about potential biases and privacy issues. To alleviate this, some studies involve complex designs to obtain independent ID and attribute features and fuse them based on the application scenario (for better accuracy or fairness). In this paper, we obviate their complex design and demonstrate that the Nearest neighbours Proxy Triplet (NPT) loss has an intrinsic capability for feature disentanglement. To further enhance the effectiveness of NPT, we propose a novel margin-based loss, namely Adaptive-rank NPT, which naturally separates the identity and attribute features. While a margin-based loss ensures inter-class separability, it imposes no constraints on intra-class compactness. The samples that meet the inter-class margin will not contribute to network training. To mitigate this issue, we propose an adaptive distance measurement to promote the compactness of the learned features, resulting in the final Ada^{2}NPT loss. The experimental results obtained on several benchmarks demonstrate the superiority and merits of the proposed loss function over the state-of-the-art losses in terms of accuracy and fairness.

Keywords: Attribute-Aware Face Recognition, Triplet-based Loss, Face Recognition Fairness

1. Introduction

Most modern face recognition systems Deng et al. (2021); An et al. (2022) use Convolutional Neural Networks (CNNs) to extract identity-discriminative features from facial images. However, these methods may overlook essential attributes, such as gender, ethnicity, etc., that could further improve the performance of a face recognition system. Attribute-aware face recognition Howard et al. (2019); Grother et al. (2019) aims to integrate additional attributes into the recognition process, offering enhanced capabilities of a trained network for practical applications with diverse populations and varying scenarios.

There are two main approaches for handling attribute features: entangling attribute features with identity or disentangling them. Typically, the entanglement method yields better recognition accuracy as it includes rich information. However, addressing potential

biases and privacy concerns becomes critical with this method, as attribute and identity features may be mutually correlated, thus reflecting the biases inherent in the training database. To mitigate this issue, disentangling attribute features from identity features is essential for developing fair and unbiased recognition systems. By focusing on identity-related information, we can mitigate the risk of biased decision-making and protect the individual's privacy. Moreover, disentangled features allow for more flexible models, making it easier for researchers to choose related attributes to develop more fair/accurate algorithms for different scenarios.

To extract low-correlated attribute and ID features, the choice of loss function is important. A CNN-based model usually uses cross-entropy Deng et al. (2019a) or triplet loss. Cross-entropy primarily focuses on the decision boundary and learns a sophisticated classifier, resulting in a less structured feature space where attribute and identity features are inadequately disentangled. To mitigate this issue, it is a common practice to introduce other losses or regularisation terms to specify the desired solution more precisely. For instance, Debface Gong et al. (2020) constructs classifiers for each attribute and applies adversarial losses among them to disentangle attribute features from identity. Specifically, an attribute-specific classifier aims to predict the attribute accurately by giving the attribute features but is incapable of estimating other attributes by giving other attribute features. With the above constraint, the loss design is complicated when more attributes are involved.

In contrast, a triplet-based loss Schroff et al. (2015) focuses on learning discriminative embeddings by constraining the distance between anchor examples and positive examples to be smaller than the distance between anchor examples and negative examples. It encourages the model to learn embeddings that make images of the same class closer and those of different classes farther apart. As a result, triplet loss is more adept at learning independent embeddings. Thus, we opt for a triplet-based loss in our design. Specifically, we employ Nearest Proxies Triplet Loss (NPT Loss) Khalid et al. (2022) as our baseline. Compared with the conventional triplet loss, NPT does not require numerous triplets across all the classes, which is time-consuming. Instead, it only considers the closest negative class during training. Consequently, NPT inherently possesses hard (closest) class mining capability, leading to more discriminative identity features. In this work, we further investigate NPT to uncover its innate capacity for feature disentanglement. Note that NPT has a low convergence rate because it only considers the closest negative class in each iteration. A naive way to boost the speed is to update the network with losses from more negative classes, but this will fail to mine hard classes and result in a compromised performance for challenging samples. To address this issue, we propose adaptive-rank NPT ($AdaNPT$) which adaptively adjusts the number of negative classes (the rank) during the whole training process based on the amplitude of the loss.

As a margin-based loss, $AdaNPT$ focuses on discriminative feature learning but ignores the distribution of positive samples when they satisfy the margin constraint. Consequently, the identity features can have considerable intra-class variations, which can compromise the identification or verification performance. Some works attempt to optimise the Euclidean distance between the samples and their corresponding class centres when located inside the margin hinge Wen et al. (2016). However, the Euclidean distance is sensitive to outliers. To address this, we propose to use a novel distance metric to measure the within-class distance adaptively based on the sample vicinity to its class weight vector. This controls the network

training and focuses on the challenging learnable samples by amplifying the magnitude of their gradients. Moreover, it ignores the well-learned samples while mitigating the effects of the noisy/low-quality samples.

In summary, the main contributions of this paper include:

- We investigate the NPT loss for attribute-aware face recognition and demonstrate that NPT achieves automated feature disengagement, avoiding complicated network designs.
- We propose an adaptive-rank annealing mechanism to effectively control the network training process with NPT. The adapted rank guides the network to disambiguate the classes from coarse to fine by initially considering all negative cohorts and gradually shifting the focus to the most challenging one.
- We use a new distance metric to compact the intra-class representations. This novel distance metric is robust to outliers while condensing the intra-class samples more forcefully.
- We evaluate the proposed method on several well-known benchmarks and demonstrate the effectiveness of the proposed loss. Moreover, we show that our solution achieves better performance in terms of fairness across different gender and ethnicity cohorts.

2. Related Work

2.1. Attribute-Aware Face Recognition

The performance of modern face recognition systems has significantly improved in the last decade, but they still suffer from diverse appearance variations of human faces. To alleviate this, researchers tend to utilise attributes, *e.g.,* gender and ethnicity, to extract contextual and structural information to be more tolerant of these variations. Hu et al. (2017) thoroughly examined the issue of fusing identity features with attribute features by recasting feature fusion as a gated two-stream neural network. Given the assumption that attributes could share low-level features from a representation learning standpoint, some research delves into multi-task learning Rudd et al. (2016); Ranjan et al. (2017) for high-performance face recognition.

While the performance of recognition systems improved, researchers noted that identity features preserving sensitive attributes, such as ethnicity, could introduce bias. Thus, more studies investigated facial attributes for fairer performance. Quadrianto et al. (2019) proposed a reprocessing method to obtain balanced data for network training. In a different approach, Zhang and Sang (2020) employed data augmentation with adversarial samples to balance training data with varying sensitive attributes. Ramaswamy et al. (2021) enhanced this approach by generating data pairs, such as individuals with and without glasses, to ensure fairness.

However, Wang et al. (2019b) claimed that having a balanced training data set is insufficient to improve fairness. Therefore, they designed a metric to measure a model's bias to re-scale the training set and a method to mask the gender-sensitive information in the representation. Liu et al. (2019) thought applying a constant margin to all cohorts is sub-optimal. Therefore, they proposed a fair loss, assigning different margins to various classes.

To simultaneously handle fairness among several attributes, Gong et al. (2020) decouple the learned features into four independent embeddings and set adversarial losses among them to reduce the feature correlation. Xu et al. (2021) introduced an inequality penalty to their network. They defined an instance FPR as the ratio between the number of non-target similarities above a unified threshold and the total number of non-target similarities. By adopting the penalty based on the instance FPR, they gain consistent FPR, *i.e.* fairness, across various attributes. These methods require efforts in training data or designing complex loss functions which are not flexible in real applications. In this paper, we provide a more straightforward solution by investigating triplet-based losses.

2.2. Triplet-based Loss

Schroff et al. (2015) proposed the triplet loss, which minimises the distance between the anchor point and positive examples within the class and maximises the distance between the anchor point and negative examples without the class. However, easy triplets may dominate the training progress by random sampling, resulting in poor performance for challenging unseen samples. To address this issue, Hermans et al. (2017) performed hard-sample mining within each training batch. For natural hard-sample mining, the Nearest Proxies Triplet (NPT) loss Khalid et al. (2022) was proposed to optimise the closest negative class instead of all categories and thus inherently mines challenging classes. In this paper, we first investigate and reveal that NPT possesses the capacity to attain independent attribute features. However, NPT still suffers from slow convergence due to the optimisation of only the closest negative class. This inflexibility hinders NPT's applicability to large training datasets. To boost the training speed, we introduce an adaptive-rank annealing mechanism that considers all negative ranks in the initial stage and progressively anneals the rank as the network training goes on.

2.3. Compacted Feature Learning

A triplet-based loss aims to ensure that the distance between the features of the same individual is smaller than the distance between the features of different individuals by a predefined margin. Different from the triplet loss, another loss category deals with face recognition as a multi-class classification task. Therefore, softmax and cross-entropy are used to learn an identity classifier to determine the class boundaries. However, cross-entropy may not always yield highly discriminative features, as they solely concentrate on the boundary. To address this limitation, Wen et al. (2016) proposed the centre loss, which establishes a centre for each class and minimises the distance between each sample in a mini-batch and its corresponding class centre, thereby reducing intra-class distance. However, Euclidean distance is sensitive to outliers, which can misdirect the optimisation process. Alternatively, L-softmax Liu et al. (2016) modified the classifier output by incorporating the magnitudes of classifier weights, identity embeddings, and the cosine of angles. It uses an additional positive scalar to the angle, which further constrains the softmax due to the cosine function's monotonically decreasing nature. Subsequently, Liu et al. (2017) normalised the classifier weights making the network focus on the angle compactness of the embedding. In contrast, AM-softmax Wang et al. (2018a) normalises the embeddings to eliminate the effects of embedding magnitudes and uses a margin to the cosine value to

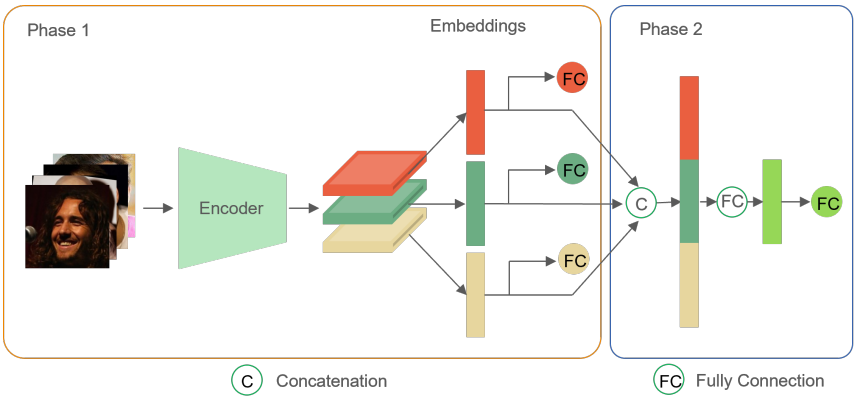


Figure 1: The overall network architecture of the proposed method with a two-phase training strategy. First, the encoder generates embeddings for gender (red), identity (green), and ethnicity (yellow) simultaneously. The identity embeddings from this stage can be used for unbiased face recognition. Second, these embeddings are fused as a final embedding for high-performance face recognition accuracy.

learn compact and discriminative representations. Later, arcface Deng et al. (2019a) further enhanced AM-softmax by replacing the cosine margin with the angle margin.

Note that all these methods tend to impose stricter Euclidean, cosine, or angle margins on the classifier for compacted feature learning. However, they fail to consider the samples located within the margins. The challenging samples will dominate the training progress in the later stage and make the network sensitive to outliers. To alleviate these concerns, we use a novel distance measurement that adaptively determines well-trained samples' contributions based on their distance to their class centres.

3. The Proposed Method

In this section, we first introduce the overall network architecture and the NPT loss. We also analyse the ability of NPT to promote feature disentanglement. Then we present our adaptive-rank annealing mechanism and introduce the adaptive distance measurement for our final Ada^{2}NPT loss.

We use a modified ResNet-50 model as our backbone Wang et al. (2018b) and employ a two-phase training strategy, as shown in Fig. 1. In the first phase, we learn identity, gender, and ethnicity attributes through three separate branches. Following the multi-task network training, we use the fused embedding to obtain the final identity classification results in the second phase, while the phase 1 network remains frozen during this stage. Note that, both the embeddings obtained by the first and second phases can be used for face recognition, depending on the requirement of fairness or accuracy in a specific task.

3.1. From Triplet to NPT

The triplet loss Schroff et al. (2015) was proposed to train face recognition systems on large-scale datasets. It calculates the similarity between two image embeddings. The aim

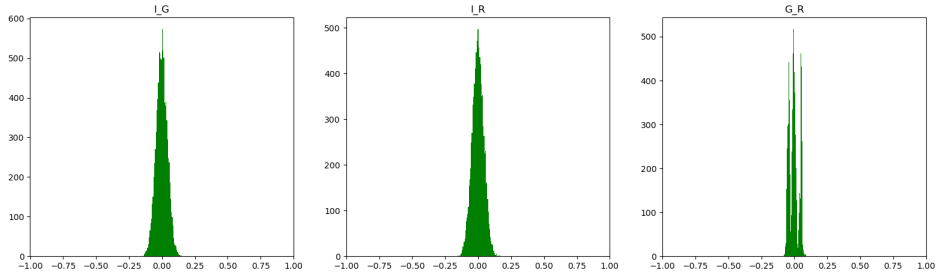


Figure 2: The distribution of the cosine similarities among different attribute embeddings. By applying the NPT loss, the cosine similarities identity and gender (I_G), identity and ethnicity (I_R), and gender and ethnicity (G_R) are close to 0.

is to maximise the distance between a pair of images from the same class and minimise the distance between two images from different classes. In other words, the goal is to maximise the inter-class distance and minimise the intra-class distance. In the training phase, a batch updates the network based on the loss of triplets $(\mathcal{I}_i^a, \mathcal{I}_i^p, \mathcal{I}_i^n)$, where $\mathcal{I} \in \mathbb{R}^{W \times H \times 3}$, W and H are the width and height of an image. $\mathcal{I}_i^a$ is the ith anchor, $\mathcal{I}_i^p$ is the positive sample with the same identity, and $\mathcal{I}_i^n$ is the negative sample. Each triplet needs to meet:

$$\|e(\mathcal{I}_i^a) - e(\mathcal{I}_i^p)\|_2^2 - \|e(\mathcal{I}_i^a) - e(\mathcal{I}_i^n)\|_2^2 + m < 0, \tag{1}$$

where $e : \mathbb{R}^{W \times H \times 3} \rightarrow \mathbb{R}^E$ is a mapping that obtains the feature embedding of the input image. E is the dimensionality of the output feature embedding vector.

The triplet loss can learn discriminative feature embeddings by constructing anchor, positive and negative pairs inside a batch. Instead of using actual samples as positive and negative examples, NPT uses proxy vectors representing class centres. The proxy vectors are learned during training and updated to encourage the distance between the anchor and its positive proxy centre to be smaller than the distance between the anchor and the proxy centres of other classes. However, most anchor and negative centre pairs are easy to learn, and this limits the performance of challenging samples. Instead of a complicated hard-sample mining strategy, the NPT loss focuses on the separation between an image embedding and its nearest-neighbour negative proxy/class-weight vector:

$$NPT = \frac{1}{N} \sum_{i=1}^{N} [\|e(\mathcal{I}_i) - \mathbf{w}_{pi})\|_2^2 - \|e(\mathcal{I}_i) - \mathbf{w}_{ni})\|_2^2 + m]_+ \tag{2}$$

where N is the batch size and m is the margin. $\mathbf{w}_{pi} \in \mathbb{R}^E$ and $\mathbf{w}_{ni} \in \mathbb{R}^E$ are the weight vectors (proxies) of the positive and nearest negative classes to $\mathcal{I}_i$. All the weight and embedding vectors are normalised to unit vectors. The separation from the closest neighbour implicitly guarantees separation from the other classes.

3.2. The disentanglement ability of NPT

In this paper, we investigate and demonstrate the capability of NPT in feature disentanglement. We build our network as shown in Fig. 1 and train it using multiple losses,

Feature	Accuracy			
	LFW	CFP-FP	CPLFW	AgeDB
Gender	65.82	65.67	52.72	67.83
Ethnicity	69.47	68.57	55.13	67.65

Table 1: The 1:1 face verification accuracy (%) when using the gender/ethnicity embeddings.

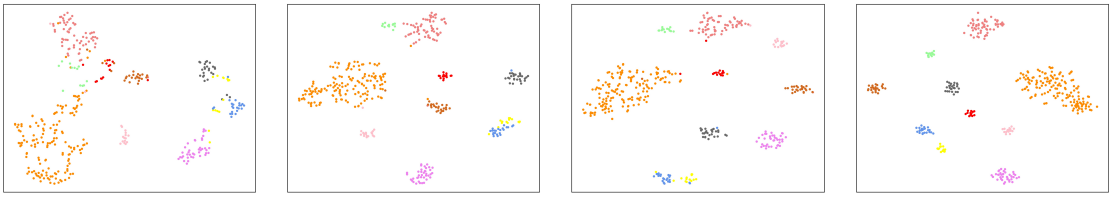
characterising the complementary objectives. To analyse the properties of the embeddings outputted by the backbone encoder, we normalise the outputs from the gender, ethnicity, and identity branches separately. We then calculate the cosine similarity between identity and gender (I_G), identity and ethnicity (I_R), and gender and ethnicity (G_R). We show the distribution of the cosine similarity among the attributes in Fig. 2. We can see that most of the cosine similarity scores are close to 0 which means that embeddings from different attributes are almost orthogonal to each other, exhibiting low correlations among each other. Besides, we conduct 1:1 face verification on several popular benchmarks with gender/ethnicity embeddings as shown in Table 1. It is clear that gender/ethnicity embeddings contain less identity information and can only result in performance close to random guessing. These observations demonstrate that NPT has the innate ability to disentangle the embeddings of various attribute pairs.

The intuitive interpretation of this property is that the NPT loss is designed to minimise the intra-class distances of embeddings. When applied to gender classification, embeddings from different male identities are constrained to be close to each other, as they share the same gender attribute. Conversely, for identity classification, these individuals' embeddings are expected to be distant from each other due to their unique identities. To fulfill both gender and identity classification requirements, the gender and identity embeddings should exhibit low correlation. This principle applies to all attribute pairs, demonstrating the effectiveness of the NPT loss in feature disentanglement.

3.3. Adaptive-rank Annealing

NPT has a natural mechanism for hard sample mining. However, human faces include heavy appearance variations, the distance to many other negative classes can be pretty close to the nearest one. By restricting the network update to the nearest negative proxy, each sample can only contribute to separating two classes in one iteration. Its focus on the top-one negative category results in the need for more iterations for network optimisation.

A solution to this problem is to relax the rank number so that the training progress can benefit from several pairs in each iteration. However, using multiple negative proxies tends to blur the class boundaries, and makes it more difficult to learn generally good embeddings within the rank hinge. As a result, the accelerated learning algorithm cannot achieve a good quality discrimination in the latent space and its solution to the face recognition problem is sub-optimal. To alleviate this problem, we propose Adaptive-rank NPT that learns discriminative embeddings with fast convergence. As the larger rank can introduce more contributors to the training process, we consider all negative pairs in the initial training to help the network learn general feature representations. Then we adaptively reduce the rank



(a) rank:4000 (b) rank:1500 (c) rank:200 (d) rank:1

Figure 3: Progress of *AdaNPT* training. From the 2D visualisation, it is evident that the network is generally able to learn separable embeddings, starting with a high number of proxies, and gradually focusing on the challenging negative classes as the rank is annealed. Ultimately, *AdaNPT* finds a discriminative latent space.

in the training progress to make the learned features more discriminative until the rank reaches 1. With the rank annealing parameter, our *AdaNPT* is formulated as follows:

$$AdaNPT_t = \frac{1}{N}\sum_{i=1}^{N}[\|e(\mathcal{I}_i) - \mathbf{w}_{pi})\|_2^2 - \frac{1}{r_t}\sum_{j=1}^{r_t}\|e(\mathcal{I}_i) - \mathbf{w}_{ni})\|_2^2 + m]_+, \tag{3}$$

where r_t is the rank at stage $t \in \{1, 2, ..., T\}$. When $r_t = 1$, the *AdaNPT* loss degrades to the basic NPT loss. For r_t, we use a Gaussian-style function as our rank annealing strategy. Based on the amplitude of the loss, the annealed rank r_t can be represented as:

$$r_t = r_{t-1} - (\frac{f * M}{\sqrt{2 * \pi}}e^{-\alpha g^2} + \frac{M}{T}), \tag{4}$$

where M is the number of classes of the training set, f and α are hyper-parameters to control the shape of the function and T is the maximal times of rank updates. g is the absolute amplitude value defined as $|(l_c - l_p)/l_p|$, where l_c and l_p are the average loss values of current and previous epochs.

For stable training progress, we keep the rank unchanged after each update until the loss enters a local plateau. If the loss is not optimised after 2 epochs, then we consider it enters the local plateau, and the rank should be adjusted. To boost the progress of training, we only focus on significant changes, which means that the loss is recognised as optimised if it is 10% less than the previous best loss. Therefore, the largest amplitude value that can trigger our *AdaNPT* loss will be 0.1. To meet this, $r_{t-1} - r_t$ should be close to M/T when $g = 0.1$. By holding this rule, we approximately set α as 426. regarding the other parameters, we set the maximal times of rank updates as $T = 5$ to guarantee that the rank will anneal to 1 within 5 adaption steps. We set the upper boundary of $r_{t-1} - r_t$ as $2 * M/5$. When $g = 0$, $\frac{f*M}{\sqrt{2*\pi}} = \frac{M}{5}$, by solving this, we roughly set f as 0.5.

With these settings, our *AdaNPT* loss can learn discriminative embeddings in a coarse-to-fine manner. For better understanding, we randomly select ten identities and project their embeddings to 2D space by T-SNE. As shown in 3, the embeddings become more and more discriminative when the rank anneals to 1.

3.4. Adaptive Distance Measurement for Compacted Feature Learning

Incorporating adaptive-rank annealing into the network enables learning informative embeddings through hard-class mining. However, akin to other margin-based loss functions, our *AdaNPT* loss can only ensure that a sample's distance to its corresponding class centre is further from other negative classes by a margin. However, samples located within the hinge will not contribute to the training process, even if they remain distant from the centre. Consequently, *AdaNPT* loss fails to enhance the compactness of each cohort and obtains less discriminative features.

To address this limitation, we aim to ensure that all samples continue to contribute to the training process, persistently pushing them toward their corresponding class centres, even after meeting the margin conditions. To achieve this, we could resort to the centre loss Wen et al. (2016), which minimises the Euclidean distance between each sample and its corresponding centre, explicitly constraining the intra-class compactness. However, Euclidean distance is sensitive to outliers. The network may focus on hard noisy samples but neglect high-quality challenging samples.

Throughout the training process, the network is optimised based on the direction (gradient) and stride (loss value) derived from the training objectives. Since outliers generate larger loss values, they dominate the network training in later stages, leading to over-fitting and diminished generalisation capabilities. Motivated by these concerns, we propose a novel distance measurement technique to balance the contributions of easy and challenging samples. To achieve this, we incorporate Rectified Wing (RWing) loss Feng et al. (2020) with our *AdaNPT* loss. RWing was originally designed for regression tasks to ensure easy samples have a larger gradient magnitude while difficult samples exhibit the opposite effect. In our work, we adopt it as a distance metric to measure the contribution of the samples located within the margin. In our work, we define the measurement as:

$$D(x) = \begin{cases} 0 & \text{if } |x| < b \\ hln(1 + (|x| - b))/\mu & \text{if } b \leq |x| < h \\ |x| - C & \text{Otherwise,} \end{cases} \tag{5}$$

where $|x|$ represents the Euclidean distance between a sample and its class centre, b is the lower boundary defining well-trained samples, and h is the upper boundary. The network should concentrate on samples with losses in the interval $[b, h)$. If the loss exceeds h, the loss becomes a linear function. C is set equal to $hln(1 + (h - b))/\mu$ to guarantee the continuity of the function and the hyper-parameter μ is set to 1.

3.5. The Final loss

We obtain our final Ada^{2}NPT loss function by combining adaptive-rank annealing and adaptive distance measurement into the NPT loss:

$$Ada^2NPT = \begin{cases} AdaNPT & \text{if } AdaNPT > 0 \\ D(d_p) & \text{Otherwise ,} \end{cases} \tag{6}$$

where $d_p = ||e(\mathcal{I}_i) - \mathbf{w}_{pi})||_2^2$ denotes the distance between the sample embedding and its corresponding class centre. The Ada^{2}NPT loss merges the benefits of adaptive-rank annealing, concentrating on hard class mining, with an adaptive distance measure. This addresses

the shortcomings of margin-based loss functions regarding their propensity to achieve compactness within each group. In our experiments, we apply the Ada^{2}NPT on ID classification and set the margin to 1. The hyper-parameters d and w to 0.1 and 0.9, respectively. For gender and ethnicity classification, we adopt the NPT loss for training. The whole network loss is a combination of three tasks:

$$Loss = L_{ID} + L_{ethnicity} + L_{gender}, \tag{7}$$

where L_{ID}, $L_{ethnicity}$ and L_{gender} are losses of ID, ethnicity and gender respectively.

4. Experiments

We first compare the proposed method with the state-of-the-art methods. Then, we demonstrate the capability of the proposed method in unbiased face recognition across various demographics. Last, we analyse the contribution of each adaptive mechanism in the ablation study. To perform a comprehensive analysis, we train our model on both tiny (number of images $<$ 0.5M) and large-scale datasets (number of images $>$ 0.5M) to demonstrate the effectiveness of our method. To be specific, we use the **CASIA-WebFace** Yi et al. (2014) dataset, which contains 494,414 images of 10,575 real identities collected from the web, as a tiny benchmark. We use the **MS1M-V2** Deng et al. (2019a) and **MS1M-V3** Deng et al. (2019b) datasets, containing 5.8M images of 85K identities and 5.1M faces of 91k identities, respectively, as large-scale benchmarks.

We train our model, with Resnet-50, Resnet-100 and Resnet-100 as the backbone, on CASIA-WebFace, MS1M-V2, and MS1M-V3 for 32, 24, and 24 epochs, respectively, on 2 RTX3090 GPU cards with a batch size of 512. We optimise the network with SGD and set the momentum and weight decay to 0.9 and 5e-4, respectively. The learning rate starts from 1e-1 and decreases by a factor of 10 at 10/10/10, 20/18/18, and 28/22/22 epochs. For pre-processing, we crop the input image to 112×112 and align faces with five landmarks, following Wang et al. (2018b).

4.1. Comparison with State-of-The-Art

We first compare our method with the state-of-the-art approaches on several popular benchmarks, including LFW Huang et al. (2008), CFP-FP Sengupta et al. (2016), CPLFW Zheng and Deng (2018), AgeDB Moschoglou et al. (2017), IJB-B Whitelam et al. (2017), and IJB-C Maze et al. (2018). We present the 1:1 face verification performance of models trained with MS1M-V2, MS1M-V3, and CASIA-WebFace in Tables 2, 3, and 4, respectively.

According to the tables, our method consistently achieves the best results on LFW, AgeDB, and IJB-C. It also attains the second-best performance on CPLFW and IJB-B, which contain abundant faces with large poses. This indicates that our method may not fully address pose variations when compared with AdaFace. The specific focus of AdaFace on challenging samples might cause this disparity, while our approach also aims to maintain fairness across samples. The average performance indicates our superior overall face verification results on various occasions. The superior performance of our method on various benchmarks demonstrates the effectiveness of the proposed Ada^{2}NPT loss in learning discriminative embeddings for face recognition. By incorporating adaptive-rank annealing

Method	Benchmark						
	LFW	CFP-FP	CPLFW	AgeDB	Average	IJB-B	IJB-C
CosFace Wang et al. (2018b)	99.81	98.12	92.28	98.11	97.08	94.80	96.37
ArcFace Deng et al. (2019a)	99.83	98.27	92.08	98.28	97.12	94.25	96.03
CurricularFace Huang et al. (2020)	99.80	98.37	93.13	98.32	97.41	94.80	96.10
BroadFace Kim et al. (2020)	99.85	98.63	93.17	98.38	97.51	94.97	96.3
MagFace Meng et al. (2021)	99.83	98.46	92.87	98.17	97.33	94.51	95.97
SCF-ArcFace Li et al. (2021)	99.82	98.40	93.16	98.30	97.42	94.74	96.09
AdaFace Kim et al. (2022)	99.82	98.49	93.53	98.05	97.47	95.67	96.89
Ours	99.86	98.42	93.43	98.39	97.53	95.59	96.91

Table 2: A comparison with the state-of-the-art methods trained on MS1M-V2 in terms of 1:1 verification accuracy (%), with ResNet-100 as the backbone. For IJB-B and IJB-C, the face verification TAR ($@FAR = 1e-4$) is reported.

Method	Benchmark						
	LFW	CFP-FP	CPLFW	AgeDB	Average	IJB-B	IJB-C
VPL-ArcFace Deng et al. (2021)	99.83	99.11	93.45	98.60	97.75	95.56	96.76
AdaFace Kim et al. (2022)	99.83	99.03	93.93	98.17	97.74	95.84	97.09
Ours	99.85	99.01	93.55	98.62	93.76	95.60	97.11

Table 3: A comparison with the state-of-the-art methods trained on MS1M-V3 in terms of 1:1 verification accuracy (%), with ResNet-100 as the backbone. For IJB-B and IJB-C, the face verification TAR ($@FAR = 1e-4$) is reported.

Method	Benchmark			
	LFW	CFP-FP	CPLFW	AgeDB
ArcFace Deng et al. (2019a)	99.30	95.30	89.85	94.23
CurricularFace Huang et al. (2020)	99.36	95.61	89.88	94.18
AdaFace Kim et al. (2022)	99.42	96.41	89.97	94.38
Ours	99.55	96.40	89.92	94.96

Table 4: A comparison with the state-of-the-art methods trained on CASIA-WebFace in terms of 1:1 verification accuracy (%), with ResNet-50 as the backbone.

and adaptive distance measurement, we successfully improve the compactness and discriminative capacity of the network, outperforming the state-of-the-art methods in numerous scenarios.

4.2. Fairness on Gender and Ethnicity

For a fair comparison with Debface, we train our models on the BUPT-Balancedface dataset Wang and Deng (2020), which contains 1.3M images from 28K celebrities and is approximately ethnicity-balanced, with 7K identities per ethnicity (Caucasian, Indian, Asian, and African). We test the performance on the RFW dataset Wang et al. (2019a). As RFW only provides ethnicity labels, following Khalid et al. (2021), we train a ResNet-18 model with the combination of IMDB Rothe et al. (2018), UTKFaces Zhang et al. (2017), AgeDB Moschoglou et al. (2017), AFAD Niu et al. (2016), and AAF Cheng et al. (2019) datasets to estimate the gender labels.

Method	Accuracy ↑				Bias ↓	Accuracy ↑		Bias ↓
	Caucasian	African	Asian	Indian		Female	Male	
DebFace	95.95	93.67	94.33	94.78	0.83	94.86	95.43	0.29
Ours	**95.96**	**93.82**	**94.42**	**94.80**	**0.78**	**95.01**	**95.45**	**0.22**

Table 5: A comparison of the ethnicity and gender fairness between the proposed method and DebFace in terms of 1:1 verification accuracy (%).

Following the methodology in DebFace Gong et al. (2020), we use the standard deviation of the 1:1 verification accuracy among the ethnicity/gender cohorts to assess the fairness of the trained model. As reported in Table 5, the embeddings obtained through our model benefit from the intrinsic disentanglement ability of the Ada^{2}NPT loss, resulting in better fairness than DebFace across both ethnicity and gender cohorts. This improved fairness demonstrates the effectiveness of the Ada^{2}NPT loss in addressing the challenges of attribute-aware face recognition, particularly in terms of achieving a more balanced performance across diverse populations. By incorporating adaptive-rank annealing and adaptive distance measurement into the loss function, our method not only enhances the overall recognition performance but also ensures a more equitable treatment of different demographic groups, thereby contributing to the development of more reliable and fair face recognition systems.

4.3. Ablation Study

To evaluate the contribution of each proposed component, we conduct an ablation study based on the model trained with the CASIA-Webface dataset. As shown in Table 6, the model effectively learns the latent space from coarse to fine using adaptive-rank annealing. The feature embeddings are further compacted by incorporating the adaptive distance measurement, which leads to enhanced discrimination.

Interestingly, the performance decreases when jointly training the identity with gender and ethnicity. This outcome is expected since our Ada^{2}NPT loss disentangles the attribute features from the identity-related features, preserving some sensitive features in the process. The absence of such information can impact the face recognition results. To address this issue, we construct an additional identity classifier capable of fusing the features of identity, gender, and ethnicity. By integrating a comprehensive set of information, we achieve the best results on LFW and AgeDB. The ablation study underscores the effectiveness of each component in our method and their collective influence on overall performance. By adjusting rank annealing and distance measurement adaptively, our model attains enhanced recognition results while preserving fairness across demographic groups. Additionally, the ablation study emphasises the importance of incorporating all available information, such as identity, gender, and ethnicity features, to optimise the performance of attribute-aware face recognition systems.

5. Conclusion and Future Work

In this paper, we first investigated and empirically validated the intrinsic ability of the NPT loss for feature disentanglement. Then we designed the adaptive-rank annealing mechanism to enhance the efficiency of NPT loss when training with large-scale datasets. We also

Method	Benchmark			
	LFW	CFP-FP	CPLFW	AgeDB
NPT	99.36	96.25	89.79	94.88
AdaNPT	99.45	96.30	89.84	94.91
Ada^{2}NPT	99.52	96.38	89.87	94.93
Ada^{2}NPT + G	99.51	96.36	89.85	94.90
Ada^{2}NPT + R	99.48	96.35	89.82	94.89
Ada^{2}NPT + G + R	99.44	96.22	89.80	94.85
Final Fusion	**99.55**	**96.40**	**89.92**	**94.96**

Table 6: Ablation study of the proposed method in terms of 1:1 face verification accuracy (%). G and R represent gender and ethnicity, respectively.

tackled the problem of large intra-class distances that result in sub-optimal latent spaces. Despite satisfying inter-class margins, we used an innovative adaptive distance measurement metric to enhance intra-class compactness explicitly. By incorporating these two adaptive mechanisms, we transformed the NPT loss into Ada^{2}NPT loss. The new loss function achieved superior fairness and comparable recognition accuracy across popular benchmarks.

While our proposed adaptive mechanisms for rank annealing and distance measurement demonstrate promising results, there remain some hyper-parameters within the model. Future work could explore assigning adaptive margins to training samples based on their quality to improve performance further and simplify the loss function. Besides, inspired from Wang and Isola (2020), we may add extra terms to the loss for better intra-class compactness. Overall, our research contributes to the development of more efficient, unbiased, and attribute-aware face recognition algorithms, laying the groundwork for future advancements in this domain.

Acknowledgments

This work was supported in part by the EPSRC grants MVSE (EP/V002856/1).

References

Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *CVPR*, pages 4042–4051, 2022.

Jingchun Cheng, Yali Li, Jilong Wang, Le Yu, and Shengjin Wang. Exploiting effective facial patches for robust gender recognition. *Tsinghua Science and Technology*, 24(3): 333–345, 2019.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019a.

Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *ICCV Workshops*, pages 0–0, 2019b.

Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas, and Stefanos Zafeiriou. Variational prototype learning for deep face recognition. In *CVPR*, pages 11906–11915, 2021.

Zhen-Hua Feng, Josef Kittler, Muhammad Awais, and Xiao-Jun Wu. Rectified wing loss for efficient and robust facial landmark localisation with convolutional neural networks. *International Journal of Computer Vision*, 128(1):2126–2145, 2020.

Sixue Gong, Xiaoming Liu, and Anil K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, pages 330–347, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58526-6.

Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

John J Howard, Yevgeniy B Sirotin, and Arun R Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In *BTAS*, pages 1–8. IEEE, 2019.

Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, Zheng Lu, Sankha S. Mukherjee, Timothy M. Hospedales, Neil M. Robertson, and Yongxin Yang. Attribute-enhanced face recognition with neural tensor fusion networks. In *ICCV*, Oct 2017.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *CVPR*, pages 5901–5910, 2020.

Syed Safwan Khalid, Muhammad Awais, Chi-Ho Chan, Zhenhua Feng, Ammarah Farooq, Ali Akbari, and Josef Kittler. Npt-loss: A metric loss with implicit mining for face recognition, 2021. URL https://arxiv.org/abs/2103.03503.

Syed Safwan Khalid, Muhammad Awais, Zhenhua Feng, Chi Ho Chan, Ammarah Farooq, Ali Akbari, and Josef Kittler. Npt-loss: Demystifying face recognition losses with nearest proxies triplet. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *CVPR*, pages 18750–18759, 2022.

Yonghyun Kim, Wonpyo Park, and Jongju Shin. Broadface: Looking at tens of thousands of people at once for face recognition. In *ECCV*, pages 536–552. Springer, 2020.

Shen Li, Jianqing Xu, Xiaqing Xu, Pengcheng Shen, Shaoxin Li, and Bryan Hooi. Spherical confidence learning for face recognition. In *CVPR*, pages 15629–15637, 2021.

Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *ICCV*, pages 10052–10061, 2019.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017.

Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *ICB*, pages 158–165. IEEE, 2018.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *CVPR*, pages 14225–14234, 2021.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPR workshops*, pages 51–59, 2017.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, pages 4920–4928, 2016.

Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *CVPR*, pages 8227–8236, 2019.

Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *CVPR*, pages 9301–9310, 2021.

Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *FG*, pages 17–24, 2017. doi: 10.1109/FG.2017.137.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018.

Ethan M. Rudd, Manuel Günther, and Terrance E. Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, pages 19–35, Cham, 2016. Springer International Publishing.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

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*, pages 1–9. IEEE, 2016.

Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018a.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, June 2018b.

Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *CVPR*, pages 9322–9331, 2020.

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 *ICCV*, October 2019a.

T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pages 9929–9939, 2020.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, pages 5310–5319, 2019b.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016.

Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *CVPR workshops*, pages 90–98, 2017.

Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *CVPR*, pages 578–586, 2021.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Yi Zhang and Jitao Sang. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In *ACM*, pages 4346–4354, 2020.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, pages 5810–5818, 2017.

Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7), 2018.