PPL: Pairwise Prototype Learning for Masked Face Recognition

Minsoo Kim[12]
kim1102@ust.ac.kr
Gi Pyo Nam[1]
gpnam@kist.re.kr
Yu-Jin Hong[3]
yjhong@hoseo.edu
Ig-Jae Kim*[12]
drjay@kist.re.kr

[1] Korea Institute of Science and Technology, South Korea
[2] University of Science and Technology, South Korea
[3] Hoseo university, South Korea

Abstract

The occlusion caused by a facemask emerged as a new challenge in face recognition as the pandemic of COVID-19 made wearing a facemask an everyday practice. The recognition performance of the previous approaches degraded in recognising a face with a facemask since the models were trained to extract features from the overall face and not many face samples with a facemask were available. Although the previously proposed method of using both face data without facemasks and those with synthesised facemasks for training improved the recognition performance, a certain drop in the recognition performance for faces without facemasks was observed. Thus, we propose an approach that can achieve robust recognition of faces with facemasks without compromising that of faces without facemasks. This study broke free from using a single prototype and designed Pairwise Prototype Learning (PPL) which separated the prototype depending on the facemask condition of the face data. Models trained with the proposed PPL method outperformed those trained with previously suggested methods in recognising both faces with and without facemasks. On top of presenting a new MFW+ dataset for masked face recognition benchmark, our study found a simple and intuitive way to improve recognition performance on all benchmarks, overcoming the limitation of using a single prototype in face recognition for faces with facemasks. All codes of PPL are available at https://github.com/kim1102/PPL-MFR.

1 Introduction

With the construction of large-scale face data and the advancement of convolutional neural networks (CNN), the performance of face recognition has been dramatically improved. However, in the outbreak of the unprecedented pandemic of COVID-19, a new challenge of recognising masked faces arose. On top of the majority of face recognition data being based on faces not wearing facemasks, the occlusion by a facemask made it hard to recognise faces with facemasks, and it brought limitations to using the existing face recognisers

* indicates the corresponding author.

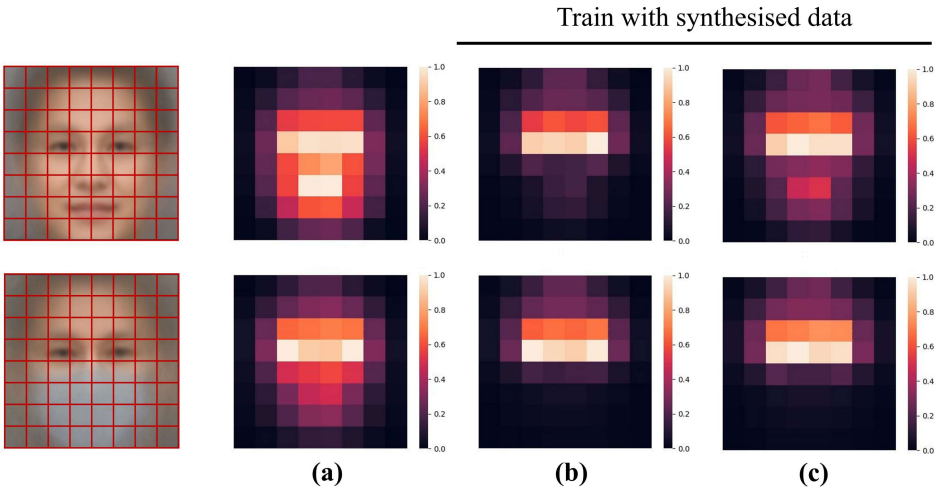

Figure 1: The contribution of each patch to the final feature extraction. Value of each patch were calculated by averaging the cosine similarities between the original images and modified ones in which the pixel value of each patch was filled with 0. The images used for the calculation were about 2,000 frontal face images sliced into the size of 8×8. (a) A model trained without synthesised data, (b) A model trained with synthesised data, and (c) A model trained with the suggested PPL.

without causing accuracy degradation [12]. Models trained with datasets whose samples are mostly unmasked faces extract features from the overall face. Thus, the extracted features from the lower part of faces when a face is wearing a facemask can serve as a component to degrade the accuracy, being useless for calculating feature distance as shown in Fig. 1 (a).

Several breakthroughs were suggested for masked face recognition, such as restoring masked region[5], [11], [13] or making models ignore the masked region [30], [22]. Among the many strategies, data augmentation utilising virtual masked faces has enabled the model to achieve significant performance improvements in masked face recognition [15]. However, these methods have a limitation compromising the recognition performance under an unmasked face environment, despite the enhancement in terms of accuracy for masked face recognition [10]. The reason behind this problem is that masked faces and unmasked faces both contribute to generating only one representative feature, which is known as a prototype, in a linear layer. Although the information delivered by a face with a face mask differs from that delivered by faces without a face mask, the training phase based on a single prototype requires one prototype that covers samples under both conditions. It leads to ignoring the lower part of a face and extracting the features in the upper face intensively, such as an eye area, due to the characteristic of the deep learning model that extracts and trains common features within each class. Therefore, the models trained with synthesised masked face data extract features, ignoring the information from the lower face, as shown in Fig. 1 (b). And this caused degradation in unmasked face recognition performance due to the absence of information from the lower face.

To resolve these problems and overcome the limitations, we propose Pairwise Prototype Learning (PPL) which could embed the features in consideration of both unmasked and masked face images without any information loss by using two prototypes in the training procedure shown as Fig. 1 (c). The proposed method trains each prototype by classifying the samples in the same class according to mask conditions in order to break free from the constraint of the single prototype approach. Using two prototypes that correspond to both face conditions makes it possible for the model to compare the feature embeddings and pro-

Datasets	# of subjects	# of images (videos)	Mask	Usage
CFP-FP [28]	500	7,000	-	Test
AgeDB-30 [25]	568	16,488	-	Test
IJB-B [35]	1,845	11,754 (7,011)	-	Test
IJB-C [24]	3,531	31,334 (11,779)	-	Test
LFW [16]	5,749	13,233	-	Test
CALFW [37]	5,749	12,174	-	Test
CPLFW [36]	5,749	11,652	-	Test
VGGFace2 [4]	9,131	3.31M	-	Train/Test
CASIA [34]	10K	494K	-	Train
MS-Celeb-1M V3 [8]	93K	5.1M	-	Train
MegaFace [19]	690K	4.7M	-	Train/Test
WebFace260M [38]	4M	260M	-	Train
MFR2 [2]	53	269	Yes	Test
MFW [15]	300	3,000	Yes	Test
RMFRD [33]	525	14,000	Yes	Test

Table 1: Datasets widely in use for face recognition. Datasets currently available for face recognition are mostly based on faces without a facemask.

totypes under the same mask condition in the training stage and release the strict similarity constraint between masked and unmasked faces. The contribution of this study can be summarised as follows:

1) We propose PPL (Pairwise Prototype Learning) as an approach that overcomes the limitation of using a single prototype in masked face recognition and is easy to apply to the previously proposed softmax-based methods.

2) We demonstrate that the models trained with PPL achieve better performance in recognising both masked and unmasked faces compared to the models trained without PPL.

3) We present a new dataset of MFW+ which is an extension of MFW [15], the published benchmark dataset for measuring the performance of face recognition for masked faces.

2 Related Work

Many proposed methods suggested for face recognition can be classified into two types: softmax-based and non-softmax-based. The method proposed by [6] is one of the non-softmax-based methods that utilise siamese networks and contrastive loss for learning similarity metrics. Another non-softmax-based approach, Triplet loss [27] directly reduces feature distances extracted from the same person while keeping the distance of feature extracted from different person far. However, these methods have a problem in that pair selection is time-consuming, and the pair configuration greatly affects the recognition performance [23]. The softmax based methods [31], [34], [17], [20], [23], [32], on the other hand, is free from pair construction because they compare the similarity between the feature embeddings extracted from the input images and prototypes which represent the identities that are included in dataset. For this reason, softmax and its variants have become the general methods for face recognition.

High-performance face recognition requires not only well-designed loss functions but also large training face datasets under various environmental conditions. For this reason, datasets involving large-scale facial data under various environments were collected and published by several research groups [14], [1], [38]. However, the collected datasets were mostly based on faces not wearing facemasks and which made models hard to recognise masked faces. Inspiringly, datasets such as RMFRD [33], MFW [15], and MFR2 [2] were

collected for masked faces but the subject number of all these three datasets barely reach a thousand which means not suitable for training. Table. 1 shows the lack of masked face datasets compared to the other face datasets. To overcome the challenges related to the small number of masked face samples, synthesising a virtual mask on an unmasked face image was proposed [2], [15]. By including samples with face masks, models can extract features excluding the areas where occlusion occurs [29]. For this reason, mask synthesis enabled models to achieve better performance in masked face recognition compared to a model trained without synthesised data.

Even though models trained with masked face data yielded remarkable performance gains in masked face recognition, they incurred performance degradation on unmasked faces [10]. This result is related to the characteristic of softmax-based learning using a single prototype: many suggested softmax and its variant methods including [7] used in report [10] gather feature embeddings towards a single prototype at the training stage. These features had a negative impact on the recognition performance when applied to the mask synthesis method. A single prototype generated from both unmasked and masked face samples makes models extract most of the features from the upper face and leads models to discard the rich identity information contained in the lower face. On the other hand, there was an attempt of using multiple prototypes [9] in a competition for masked face recognition [3]. However, the approach that is proposed in [9] for label-noised data removal does not guarantee the consistency of feature embedding and prototype comparisons depending on mask conditions. With this intuition, we propose a separate prototype containing representative features with different mask conditions for masked face recognition.

3 Proposed method

Compared to the conventional approaches using softmax-variant loss, the proposed PPL has two major features; 1) composed of separate linear layers and forwards the feature embeddings to different linear layers depending on the face mask conditions, and 2) embedding similarity loss is employed to maintain the similarity between unmasked and masked face feature embeddings.

3.1 Logits of PPL

To use softmax-variant loss for learning, the model consists of a backbone network that extracts feature embeddings and a linear layer that allows the embeddings to be projected to an appropriate place to have proper discrimination. The posterior probability of the model having a backbone network f and a linear layer W can be described as follows:

$$p(x) = W(f(x)) \tag{1}$$

The key idea of PPL is not to use a single prototype but forwards feature embeddings to separate prototypes depending on the facemask conditions. To store two different prototypes, we designed two linear layers of the same size. In training the model using PPL, the backbone network forward is performed on all input images and they are forwarded to two different linear layers depending on the facemask conditions. The posterior probability $p'(x)$ inferred

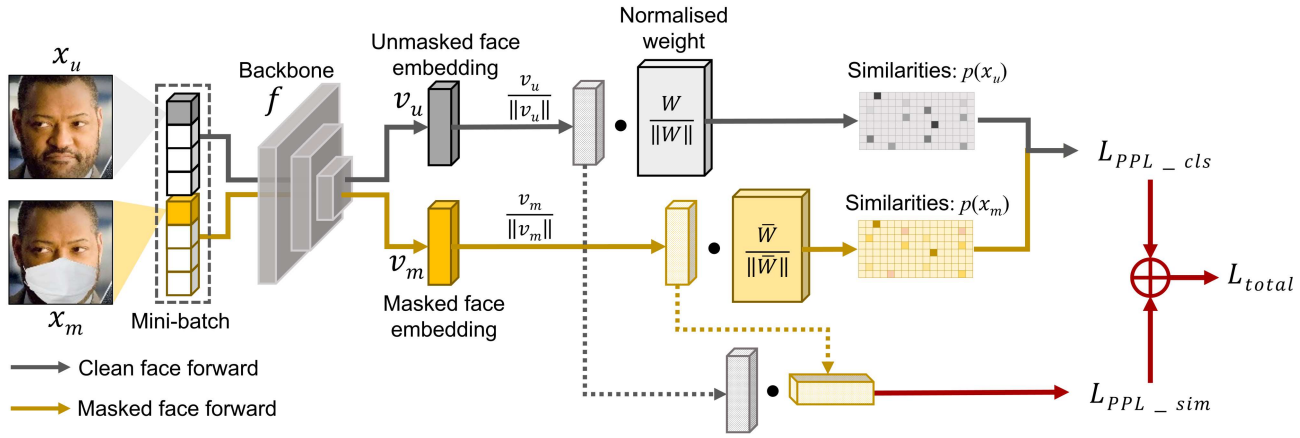

Figure 2: Overview of the proposed method of using PPL. PPL uses two separate linear layers to store separate prototypes depending on the facemask conditions. In the training process, losses are calculated using the feature embedding and prototype of the corresponding facemask condition. The feature embeddings achieved from unmasked and masked face images are brought closer directly by PPL-SIM loss.

from the model using PPL can be described as follows:

$$p'(x) = \begin{cases} W(f(x)) & \text{if } x = x_u \\ \bar{W}(f(x)) & \text{if } x = x_m \end{cases} \tag{2}$$

Since the masked faces were synthesised, we could determine whether an input image x is wearing a face mask (x_m) or not (x_u) in the training process. Each prototype being able to store representative features depending on the facemask condition enables the model to learn the optimal extraction methods under varying conditions. This approach of using PPL can be applied to any type of softmax variant loss previously proposed and it can provide a new posterior probability. Angular margin loss, one of the representative softmax based losses for training face recognition networks, can be described as follows:

$$L_m = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(m_1\theta_{y_i}+m_2)-m_3)}}{e^{s(cos(m_1\theta_{y_i}+m_2)-m_3)}+\sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}, \tag{3}$$

$$cos\,\theta_j = \frac{{W_j}^T f(x_i)}{\| W_j \| \| f(x_i) \|} \tag{4}$$

whereas θ_j is the angle between prototype W_j and feature embedding $f(x_i)$, N is the number of image samples in mini-batch. m_1, m_2, m_3 are margin hyper parameters suggest by [23], [7], [32] and s is scaling parameter. Angular margin loss in which PPL is applied can be described as follows:

$$L_{ppl-cls} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(m_1\theta'_{y_i}+m_2)-m_3)}}{e^{s(cos(m_1\theta'_{y_i}+m_2)-m_3)}+\sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta'_j}}, \tag{5}$$

$$cos\,\theta'_j = \begin{cases} \frac{{W_j}^T f(x_i)}{\|W_j\|\|f(x_i)\|} & \text{if } x = x_u \\ \frac{{\bar{W}_j}^T f(x_i)}{\|\bar{W}_j\|\|f(x_i)\|} & \text{if } x = x_m \end{cases} \tag{6}$$

Dataset		# of genuine pairs	# of imposter pairs
MFW	unmask-mask	3,000	3,000
	mask-mask	3,000	3,000
MFW+	unmask-mask	13,865	8,247,553
	mask-mask	5,425	8,040,556

Table 2: Details of the extended MFW+ benchmark compare to the original MFW benchmark.

3.2 Embedding similarity loss

As seen in Eq. 5 and Eq. 6, there is no similarity constraint between feature embeddings extracted from unmasked and masked face. This can lead to low similarity between masked and unmasked faces and low intra-class compactness. PPL-similarity loss $L_{ppl-sim}$ is added to this requirement, making masked and unmasked face embeddings directly closer. Similarity loss $L_{ppl-sim}$ can be calculated as follows:

$$L_{ppl-sim} = \frac{1}{C}\sum_{i=1}^{C} max\left(1 - \left(\frac{f(x_u^i)}{||f(x_u^i)||}\right)^T \left(\frac{f(x_m^i)}{||f(x_m^i)||}\right), 0\right) \tag{7}$$

C is the number of unmasked and masked face pairs existing in a mini batch. Since PPL-SIM loss requires unmasked and masked face embeddings as inputs, we designed a mini-batch to include the original unmasked face sample when the synthesised masked face image is included in the mini-batch. Finally, the PPL applied to the angular margin loss is described as the combination of modified angular margin loss Eq. 5 and embedding similarity loss Eq. 7:

$$L_{ppl} = L_{ppl-cls} + L_{ppl-sim} \tag{8}$$

The overall process of model training using PPL can be seen in Fig. 2.

4 Experiments

4.1 MFW+ dataset

The original MFW, published for the masked face recognition benchmark is composed of 300 IDs and 3,000 images. With two duplicate IDs found in MFW, MFW actually contains 298 IDs and 2,980 images. To evaluate models under variant mask conditions and environments, we gathered supplement data manually from the web. The refined and extended MFW, which we named MFW+, contains 606 IDs, 2,911 unmasked face images, and 2,838 masked face images. We used this MFW+ dataset to construct genuine pairs and imposter pairs according to the two facemask conditions (i.e., unmasked-masked faces and masked-masked faces). All possible face pair combinations were used to construct genuine pairs and imposter pairs. Table. 2 describes the details of the MFW+ dataset and comparison with the original MFW dataset. More details of the MFW+ dataset are included in supplementary material section.

4.2 Implmentation details

For training, we generated masked face data from the MS-CELEB-1M dataset using the face mask synthesis method proposed by [15]. The resulting Masked-msceleb dataset consisted of 93K IDs with 4.6M masked face images paired with 4.6M unmasked face images. Images used for training and testing were aligned to a size of 112 × 112 following the method used

Method	Clean face benchmark			Masked face benchmark	
	LFW	IJB-B	IJB-C	MFW+ (U-M)	MFW+ (M-M)
(1) CosFace, without synthesis(m=0.35) [32]	99.33	93.89	95.22	62.86	59.28
(2) CosFace(m=0.35)	99.18	92.59	94.03	80.73	75.93
(3) ArcFace, without synthesis(m=0.50) [7]	99.38	94.29	95.68	70.27	64.79
(4) ArcFace(m=0.50)	99.20	92.81	94.52	83.03	77.70

Table 3: Performance degradation on unmasked face recognition caused by masked face data. The 1:1 verification accuracy(%) is reported on LFW benchmark. We also report the verification accuracy of IJB-B, IJB-C and MFW+ on TAR@FAR = 1e-4. Written U in the table represents an unmasked face while M represents a masked face.

Method	Clean face benchmark							Masked face benchmark	
	LFW	CFP-FP	CPLFW	AgeDB	CALFW	IJB-B	IJB-C	MFW+ (U-M)	MFW+ (M-M)
(1) CosFace(m=0.35) [32]	99.18	93.30	88.15	97.55	95.08	92.59	94.03	80.73	75.93
(2) ArcFace(m=0.50) [7]	99.20	92.51	88.68	97.62	95.47	92.81	94.52	83.03	77.70
(3) CurricularFace [17]	99.23	93.04	87.78	97.32	95.22	91.57	93.46	78.36	74.80
(4) BroadFace [21]	99.27	92.43	88.57	97.45	95.35	92.69	94.43	83.18	77.88
(5) AdaFace [20]	99.30	92.87	89.27	97.53	95.53	93.29	95.02	83.54	78.27
(6) FocusFace (ArcFace) [26]	99.27	92.43	88.78	97.6	95.53	93.05	94.69	83.40	77.77
(7) MaskInv-HG (ArcFace) [18]	99.3	93.46	89.13	97.82	95.6	93.40	94.92	83.82	78.36
(8) CosFace + PPL	99.3	**94.99**	88.9	97.98	95.63	93.65	94.99	80.50	75.59
(9) ArcFace + PPL	99.33	94.1	89.65	**98.07**	**95.7**	93.59	95.25	83.69	**78.80**
(10) AdaFace + PPL	**99.35**	94.74	**89.97**	98.03	**95.7**	**94.14**	**95.58**	**84.27**	78.27

Table 4: Benchmark results by different methods. We compared the models trained with the methods previously published with the model trained with suggested PPL. The 1:1 verification accuracy(%) of LFW, CFP-FP, CPLFW, AgeDB-30, and CALFW benchmarks is reported. We also report the verification accuracy of IJB-B, IJB-C, MFW+(U-M) and MFW+(M-M) on TAR@FAR = 1e-4.

in [23], and normalised to an average value of 0.5 and a standard deviation of 0.5 so that the pixel values were 0 to 1. The IR-SE50 proposed by [7] was used as the backbone and features were extracted with a size of 512 dimensions. The batch configuration consisted of 256 unmasked face samples and 256 masked face samples, which made the size of the mini-batch 512 in total. Stochastic gradient descent (SGD) was used as the optimizer, and weight decay was set to 5e-4 and momentum to 0.9. The learning rate started at 0.1 for initial learning and was scheduled to be multiplied by 0.1 at 10, 16, and 22 epochs. The training ended at 25 epochs.

4.3 Model evaluation

To test the recognition performance of the proposed approach, the recognition performance for unmasked and masked faces was measured using various data sets. To verify the recognition performance for unmasked faces, LFW [16], CFP-FP [28], CPLFW [36], AgeDB-30 [25], and CALFW [37] were used along with IJB-B [35] and IJB-C [24] benchmarks. As for verifying the recognition performance for masked faces, the MFW+ benchmark was used. We also evaluated models on IFRT [10] challenge, which provides both unmasked and masked face recognition benchmarks.

Degradation in unmasked face recognition. To observe the degradation in unmasked face recognition caused by masked face data, the models in Table. 3 (1) and Table. 3 (3) were trained using only the original unmasked face data from Masked-msceleb. As shown in Table. 3, training a model with masked face data leads to performance degradation in unmasked face recognition. Models trained with ArcFace and synthesised masked face data

Method	Childeren	African	Caucasian	South Asian	East Asian	All	Mask
(1) CosFace	42.95	46.56	63.14	50.65	31.51	51.75	73.86
(2) ArcFace	49.33	54.40	71.57	59.66	37.73	60.18	79.99
(3) CurricularFace	36.39	41.79	61.18	47.50	24.56	44.63	73.57
(4) BroadFace	49.32	54.87	71.55	59.37	40.02	61.79	80.20
(5) AdaFace	51.01	57.32	73.86	62.06	41.03	62.71	80.82
(6) FocusFace (ArcFace)	52.08	57.90	74.19	61.95	41.68	64.01	79.87
(7) MaskInv-HG (ArcFace)	52.35	61.08	76.54	67.29	47.68	68.92	78.78
(8) CosFace +PPL	50.14	59.28	74.43	64.47	45.16	65.49	74.96
(9) ArcFace +PPL	54.37	66.77	80.49	71.75	52.35	73.32	78.99
(10) AdaFace +PPL	**56.97**	**68.97**	**82.41**	**75.40**	**54.81**	**75.55**	**80.90**

Table 5: Benchmark results by different methods on IFRT.

suffer 0.18% degradation in LFW, compared to models trained without synthesised data. For IJB-B and IJB-C, performance degradations of 1.48% and 1.16% were recorded. However, training with synthesised masked face data resulted in a 12.76% performance improvement in the MFW+(U-M) benchmark and a 12.91% performance improvement in the MFW+(M-M) benchmark. Similarly, the model trained with CosFace performed better on masked face recognition, although the performance degraded in unmasked face recognition when the synthesized masked face data was included in the training dataset.

Effects of PPL. Although the proposed PPL could not completely prevent the performance degradation of unmasked face recognition caused by masked face data, PPL made it possible to obtain remarkable performance elevation for both unmasked and masked faces compared to the model trained without PPL. As shown in Table. 4, model trained using ArcFace with PPL showed a performance improvement of over 0.13% on the LFW dataset, 1.59% on CFP-FP, and over 0.45% on AgeDB-30 compare to model using ArcFace only. For IJB-B and IJB-C datasets, accuracy improvements of 0.78% and 0.73% were observed. In masked face recognition, the model trained using ArcFace with PPL showed a performance improvement of over 0.66% in the MFW+(U-M) benchmark and over 1.1% in the MFW+(M-M) benchmark compared to the model trained with ArcFace without PPL. Even for other approaches, such as cosface and adaface, models trained using PPL achieve remarkable performance improvements in unmasked face recognition compared to models trained only with cosface and adaface. Interestingly, the performance gap between models with and without PPL in mask face recognition is not large, but similar. This means that models trained with single prototype mainly use the upper part of the face to extract features.

Comparison with other methods for masked face recognition. To confirm the effectiveness of PPL, we compared our method with other approaches for masked face recognition. For a fair comparison, both FocusFace [26] and Mask-Inv [18] were trained using ArcFace loss. All methods were equally trained with the scratch manner as described in 4.2. And in the case of Mask-Inv, a pure ArcFace model trained only on the unmasked face data reported in Table. 3 (3) was used as the teacher model. As can be seen from the Table. 4 and Table. 5, PPL+ArcFace achieved comparable performance with the FousFace and MaskInv in masked face recognition. However, model trained with PPL achieved a significant performance gap in the clean face benchmark compared to other approaches. The model trained with PPL showed a performance gap of 9.31% compared to FocusFace and 4.4% compared to maskInv-HG in IFRT-All. These results show that prototype separation is a very effective and simple way to protect models from performance degradation caused by masked face data.

Method	Consistency	Utilisation	IJB-B	IJB-C	MFW+ (U-M)	MFW+ (M-M)
(Baseline) Single prototype	-	-	92.81	94.52	83.03	77.70
(a) Random forward (-)	0.000	1.000	92.70	94.28	81.84	75.91
(b) Random forward	0.000	1.000	92.74	94.37	83.66	78.65
(c) Max pooling	0.993	0.004	92.48	94.16	82.38	76.81
(d) PPL (-)	1.000	1.000	93.52	95.17	81.74	76.59
(e) PPL	1.000	1.000	**93.59**	**95.25**	**83.69**	**78.80**

Table 6: Recognition performance by linear layer configurations. TAR@FAR=1e-4 is reported on IJB-B, IJB-C, MFW+. (-) in the table indicates the trained model without embedding similarity loss.

5 Discussion

5.1 Effect of prototype separation

In order to verify that forwarding the embeddings depending on the face mask condition is important, we compared and analysed six models trained with different linear layer configurations. The models in Table. 6 (a) and Table. 6 (b) are the model which forwards the embeddings to linear layers regardless of the mask conditions. The model in Table. 6 (c) is a model that forwards embeddings to closer prototypes using max pooling proposed by [9]. The models in Table. 6 (d) and Table. 6 (e) are the model which forwards the embeddings to separate linear layers depending on the mask conditions, as proposed in PPL. All three conditions used ArcFace loss for classification loss. To evaluate how the prototypes were constructed according to each of the three conditions, two values were measured at the end of the training stage: prototype consistency and prototype utility. Prototype consistency was measured to determine how consistently an embedding was multiplied by a specific prototype. The prototype consistency was calculated as follows:

$$Consistency = \frac{1}{C}\sum_{i=1}^{C}\left(1+\left(\frac{p_i}{p_i+\overline{p}_i}log_2\left(\frac{p_i}{p_i+\overline{p}_i}\right)+\frac{\overline{p}_i}{p_i+\overline{p}_i}log_2\left(\frac{\overline{p}_i}{p_i+\overline{p}_i}\right)\right)\right) \tag{9}$$

We also measured prototype utility to determine whether the two separate linear layers were fully utilised. Prototype utilisation was calculated as follows:

$$Utilisation = 1-\frac{1}{C}\sum_{i=1}^{C}\frac{|n_i-\overline{n}_i|}{n_i+\overline{n}_i} \tag{10}$$

In the equation, C is the total number of classes. p_i and n_i are the number of the unmasked face embeddings and the number of total embeddings in class i, which are to be multiplied with prototype W_i. $\overline{p}_i$ and $\overline{n}_i$ are the number of clean face embeddings and the number of total embeddings in class i, which are to be multiplied with prototype $\overline{W}_i$. High prototype consistency indicates that each prototype consists of a consistent images based on mask conditions. High prototype utility indicates that the model uses both prototypes equally in the training phase. Table. 6 shows each model's prototype consistency, utility, and bench results. The model trained using PPL was the only model with high consistency and utility and showed the best performance compared to other models. As seen in Table. 6, maxpooling cannot utilise both prototypes equally. This is because once the embeddings are passed to the linear layer, one prototype is fixed as identity center, preventing other prototypes from being composed of another images. Therefore, maxpooling may suitable for training label noise data but not for masked face data.

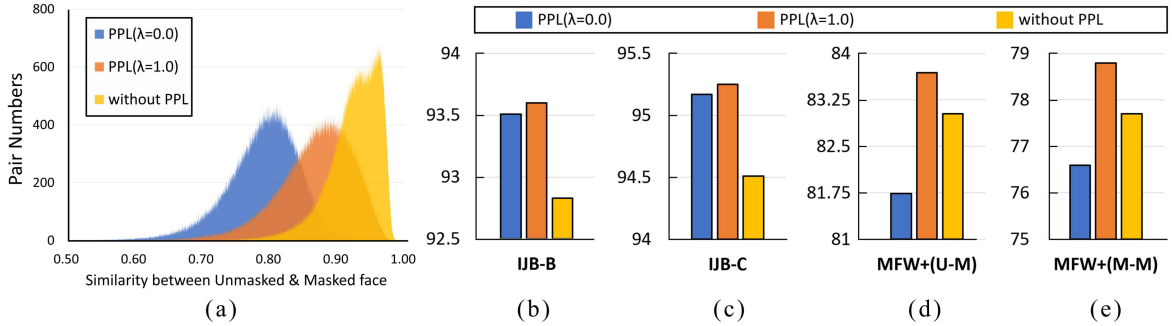

Figure 3: Effect of PPL-SIM loss in training. (a) Cosine similarities between unmasked and masked face embeddings, (b) TAR@FAR=1e-4 on IJB-B, (c) TAR@FAR=1e-4 on IJB-C, (d) TAR@FAR=1e-4 on MFW+(U-M), and (e) TAR@FAR=1e-4 on MFW+(M-M).

5.2 Analysis of PPL-SIM loss

To evaluate how PPL-SIM loss effects on the training using PPL, we modified Eq. 8 as follows:

$$L_{ppl'} = L_{ppl-cls} + \lambda L_{ppl-sim} \tag{11}$$

As the equation shows, the hyperparameter λ allows the PPL to decide whether to use the PPL-SIM loss. As shown in Fig 3, the model trained using PPL with PPL-SIM loss ($\lambda = 1.0$) had a higher similarity between masked and unmasked faces and better recognition performance compared to the model trained without PPL-SIM loss ($\lambda = 0.0$). However, this does not mean that achieving high similarity between masked and unmasked faces achieves high performance. The model trained without PPL recorded lower recognition performance compared to the model trained using PPL, although it had higher similarity between masked and unmasked face embeddings. Interestingly, achieving excessive similarity between unmasked and masked faces does not guarantee high performance in both unmasked and masked face recognition. Again, the key of PPL is the separation of prototypes based on mask conditions, which allows the model to compare training samples to appropriate prototypes.

6 Conclusion

We proposed PPL that can achieve robust recognition of faces with and without facemasks. This study found that using a single prototype is not reasonable when training model with synthesised masked face data and designed an approach of using separated prototypes in training. The proposed method PPL (Pairwise Prototype Learning) can prevent the performance degradation in unmasked face recognition and enable models to achieve promising accuracy on masked face recognition. This approach can be regarded as a strategy to actively overcome occlusion, which is considered as a major obstacle in face recognition and will inspire future research by publishing a meaningful dataset of masked faces.

7 Acknowledgement

This research was supported by R&D program for Advanced Integrated-intelligence for IDentification (AIID) through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT (2018M3E3A1057288), and the KIST Institutional Program "Multimodal Visual Intelligence for Cognitive Enhancement of AI Robots"(2E31582).

References

[1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021.

[2] Aqeel Anwar and Arijit Raychowdhury. Masked face recognition for secure authentication. *arXiv preprint arXiv:2008.11104*, 2020.

[3] Fadi Boutros, Naser Damer, Jan Niklas Kolf, Kiran Raja, Florian Kirchbuchner, Raghavendra Ramachandra, Arjan Kuijper, Pengcheng Fang, Chao Zhang, Fei Wang, et al. Mfr 2021: Masked face recognition competition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2021.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[5] Lele Cheng, Jinjun Wang, Yihong Gong, and Qiqi Hou. Robust deep auto-encoder for occluded face recognition. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1099–1102, 2015.

[6] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[8] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[9] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, pages 741–757. Springer, 2020.

[10] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444, 2021.

[11] Nizam Ud Din, Kamran Javed, Seho Bae, and Juneho Yi. A novel gan-based network for unmasking of masked face. *IEEE Access*, 8:44276–44287, 2020.

[12] Mustafa Ekrem Erakın, Uğur Demir, and Hazım Kemal Ekenel. On recognizing occluded faces in the wild. *arXiv preprint arXiv:2109.03672*, 2021.

[13] Shiming Ge, Chenyu Li, Shengwei Zhao, and Dan Zeng. Occluded face recognition in the wild by identity-diversity inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3387–3397, 2020.

[14] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[15] Je Hyeong Hong, Hanjo Kim, Minsoo Kim, Gi Pyo Nam, Junghyun Cho, Hyeong-Seok Ko, and Ig-Jae Kim. A 3d model-based approach for fitting masks to faces in the wild. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 235–239. IEEE, 2021.

[16] 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.

[17] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[18] M. Huber et al. Mask-invariant face recognition through template-level knowledge distillation. In *FG 2021*, pages 1–8, 2021.

[19] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016.

[20] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18750–18759, 2022.

[21] Yonghyun Kim, Wonpyo Park, and Jongju Shin. Broadface: Looking at tens of thousands of people at once for face recognition. In *European Conference on Computer Vision*, pages 536–552. Springer, 2020.

[22] Yande Li, Kun Guo, Yonggang Lu, and Li Liu. Cropping and attention based approach for masked face recognition. *Applied Intelligence*, 51(5):3012–3025, 2021.

[23] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[24] 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 *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.

[25] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.

[26] Pedro C Neto, Fadi Boutros, João Ribeiro Pinto, Naser Darner, Ana F Sequeira, and Jaime S Cardoso. Focusface: Multi-task contrastive learning for masked face recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[28] 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 *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

[29] Krishna Kumar Singh, Hao Yu, Aron Sarmasi, Gautam Pradeep, and Yong Jae Lee. Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond. *arXiv preprint arXiv:1811.02545*, 2018.

[30] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 773–782, 2019.

[31] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015.

[32] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[33] Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Qi Hong, Hao Wu, Peng Yi, Kui Jiang, Nanxi Wang, Yingjiao Pei, et al. Masked face recognition dataset and application. *arXiv preprint arXiv:2003.09093*, 2020.

[34] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[35] 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 *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98, 2017.

[36] 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.

[37] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[38] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.