Localization using Multi-Focal Spatial Attention for Masked Face Recognition

Yooshin Cho[1]   Hanbyel Cho[1]   Hyeong Gwon Hong[2]   Jaesung Ahn[2]

Dongmin Cho[3]   JungWoo Chang[3]   Junmo Kim[1,2]

[1]School of Electrical Engineering, KAIST, South Korea
[2]Kim Jaechul Graduate School of AI, KAIST, South Korea
[3]Alchera Inc

Abstract— Since the beginning of world-wide COVID-19 pandemic, facial masks have been recommended to limit the spread of the disease. However, these masks hide certain facial attributes. Hence, it has become difficult for existing face recognition systems to perform identity verification on masked faces. In this context, it is necessary to develop masked Face Recognition (MFR) for contactless biometric recognition systems. Thus, in this paper, we propose Complementary Attention Learning and Multi-Focal Spatial Attention that precisely removes masked region by training complementary spatial attention to focus on two distinct regions: masked regions and backgrounds. In our method, standard spatial attention and networks focus on unmasked regions, and extract mask-invariant features while minimizing the loss of the conventional Face Recognition (FR) performance. For conventional FR, we evaluate the performance on the IJB-C, Age-DB, CALFW, and CPLFW datasets. We evaluate the MFR performance on the ICCV2021-MFR/Insightface track, and demonstrate the improved performance on the both MFR and FR datasets. Additionally, we empirically verify that spatial attention of proposed method is more precisely activated in unmasked regions.

I. INTRODUCTION

With the advent of deep neural networks, the accuracy of Face Recognition (FR) has become more than over 99% in controlled environments [6], [23], [1]. Accordingly, identity authentication systems that employ FR have been widely used in our daily life (e.g., airports, companies, and smartphones). However, owing to the recent global COVID-19 pandemic, it has become mandatory to wear facial masks to protect public health. These facial masks cover the nose and mouth, and hence, they significantly decrease the performance of previous face recognition systems. Thus, the necessity of Masked Face Recognition (MFR) has been highlighted, and various MFR studies have been proposed during the pandemic.

Previous studies on MFR have focused on the effective extraction of mask-invariant features [27], [11], [21], [19], [4]. The method proposed in [27] discards the features obtained from the lower half region of images by utilizing the prior knowledge that a mask is located in the lower half of a face. In addition, MMD loss, MSE loss, or adversarial loss with an auxiliary mask-usage classification branch has been adopted to reduce the distance between the features obtained from

(a) Original   (b) Blue   (c) Green   (d) Black   (e) White

Fig. 1: Examples of the synthetic masked face, which is generated by the online masked face generation function of the ICCV2021/Insightface track.

masked and unmasked images [19], [14]. These approaches have successfully improved the MFR performance by forcing networks to neglect the facial attributes (nose and mouth) that are typically behind a mask. However, this conflicts with the goal of conventional FR and leads to performance degradation on FR datasets.

To optimize the trade-off between the performances of conventional FR and MFR, we propose the method that more precisely separates masked regions from unmasked region by adopting a spatial attention module [3]. Conventionally, spatial attention modules, such as the Convolutional Block Attention Module (CBAM) [25] have been adopted to make networks focus on foreground objects, and have demonstrated the localization capability that separates foreground regions from background regions. In this paper, we propose Complementary Attention Learning (CAL) that adversarially utilizes the complementary spatial attention to enhance the localization capability of CBAM inspired by recent studies on unbiased visual recognition [24]. We train complementary spatial attention to learn undesirable information (e.g., mask-usage classification) to prevent standard spatial attention from focusing on the undesirable region (e.g., masked region). We describe details in Section II-B

Additionally, we propose Multi-Focal Spatial Attention (MFSA), which separates an image into three regions (i.e., unmasked regions, masked regions, and background regions). Previously, *sigmoid* function has been generally adopted to normalize spatial attention of CBAM, and divided the elements into binary (e.g., foreground and background). Thus, if we train complementary attention to focus on masked region, background region is more likely to be activated at standard attention. To alleviate the issue, we employ *softmax* function to classify elements into N-way. Also, instead of channel attention module of CBAM, we employ the convolution

(a) Convolutional Block Attention Module (CBMA)
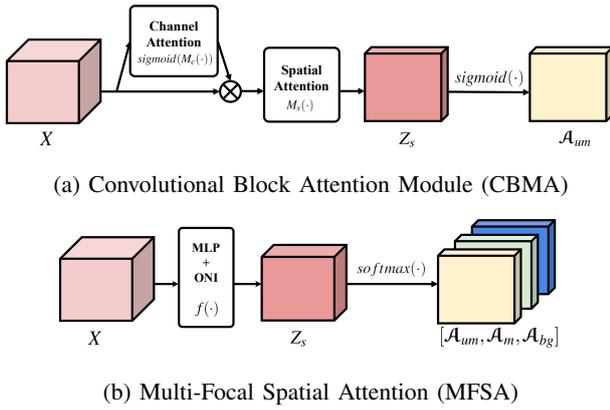

(b) Multi-Focal Spatial Attention (MFSA)

Fig. 2: **Schematic of the Convolutional Block Attention Module (CBAM) and Multi-Focal Spatial Attention (MFSA).** We replace the channel attention of CBAM and the last *sigmoid* function with the convolutional layers and *softmax* function, respectively. Additionally, we adopt the Orthogonalization by Newton's Iteration (ONI) to the weight of convolutional layers to disentangle the representations.

layers that followed by batch normalization and ReLU. To disentangle the representation of N-way attention, we adopt the Orthogonalization by Newton's Iteration (ONI) [12], and it orthogonalizes the weight of convolution layers. Weight orthogonalization enforces each output channel to depend on a different channel of input channel, and it reduces causality and correlation between output channels. With the modification, we increase the model representation capacity, and empirically demonstrate improved performance on both MFR and FR datasets.

We follow the experimental setup of the baseline models presented at the Insightface track of the ICCV2021-MFR challenge [4]. We train networks on the CASIA-Webface [26] and MS1MV3 datasets [6], [5], [9] with masked face augmentation that synthesizes masked face with a given probability. To implement masked face augmentation, we employ the online masked face generation function provided by Insightface. We visualize the examples of synthetic masked face in the Fig 1. We train networks with randomly synthesized masked face, and evaluate on the ICCV2021-MFR/Insightface track, which is a **real masked face** dataset. To analyze the trade-off between MFR and FR performances, we evaluate the FR performance on IJB-C, Age-DB, CALFW, and CPLFW datasets [28], [17], [29], [16]. From the results, we empirically verify that proposed method achieves the superior performance on both FR and MFR datasets.

As wearing masks have been highly recommended, many disciplines that require to recognize and interact with human need to develop masked face recognition algorithms (e.g., cyber-security, transportation, public health, human-computer interaction, and smart technologies). Especially, our method demonstrates the masked region by spatial attention maps, so our method is explainable to human, which is

an essential behavior for human-computer interaction. Also, our method can be applied to other type of facial occlusions (e.g., glasses and hats), if synthetic image generator is exist. Therefore, we expect our algorithm can be ubiquitously applied to other disciplines.

II. LOCALIZATION USING MULTI-FOCAL SPATIAL ATTENTION

We propose the method that precisely localizes the unmasked region of a face. First, we propose the Complementary Attention Learning (CAL), which prevents spatial attention from being activated in an undesirable area, such as the masked region of a face. Second, we propose the Multi-Focal Spatial Attention (MFSA), which does not only divide the region into binary (e.g., foreground and background), but divides into 3-way (e.g., masked region, unmasked region, and background region). Details of the proposed method are described in the following sections.

A. Preliminary: CBAM

Before introducing our method, we briefly describe Convolutional Block Attention Module (CBAM) [25]. CBAM is a representative attention module that sequentially applies channel attention and spatial attention modules. It can be expressed using the following formulas:

$$\mathcal{A}_c = sigmoid(M_c(X)), \qquad X_c = \mathcal{A}_c \otimes X, \qquad (1)$$
$$\mathcal{A}_s = sigmoid(M_s(X_c)), \qquad Y = \mathcal{A}_s \otimes X_c, \qquad (2)$$

where $X, Y \in \mathbb{R}^{C \times H \times W}$ are the input and output features of the CBAM, respectively. $\otimes$ denotes element-wise multiplication, and $M_c(\cdot)$ and $M_s(\cdot)$ are the channel and spatial attention modules, respectively. $X_c \in \mathbb{R}^{C \times H \times W}$ is an intermediate feature refined with channel attention. The channel attention module $M_c(\cdot)$ is composed of max-pooling, average-pooling along spatial dimensions, and multi-layer perceptrons (MLPs). The spatial attention module $M_s(\cdot)$ is composed of max-pooling, average-pooling along a channel dimension, and convolution layers. $M_c(X) \in \mathbb{R}^{C \times 1 \times 1}$, $M_s(X_c) \in \mathbb{R}^{1 \times H \times W}$ are normalized by *sigmoid* function to compute the channel and spatial attention, respectively. Channel attention $\mathcal{A}_c \in \mathbb{R}^{C \times 1 \times 1}$ and spatial attention $\mathcal{A}_s \in \mathbb{R}^{1 \times H \times W}$ represent the importance of channel and spatial locations, respectively. Empirically, the CBAM demonstrates the localization capability, and $\mathcal{A}_s$ focuses on the foreground objects. In this work, we enhance the localization capability to make the network precisely focus on the unmasked region of a face.

B. Complementary Attention Learning

To extract mask-invariant features while minimizing the loss of the FR performance, we propose the method that precisely separates a masked region by employing and enhancing the localization capability of the CBAM. First, we generate two attention maps, $\mathcal{A}_{um}, \mathcal{A}_m$ and two features, X_{um}, X_m, using the following equations:

$$Z_s = M_s(X_c), \quad \mathcal{A}_{um} = sigmoid(Z_s), \quad \mathcal{A}_m = 1 - \mathcal{A}_{um}, \quad (3)$$
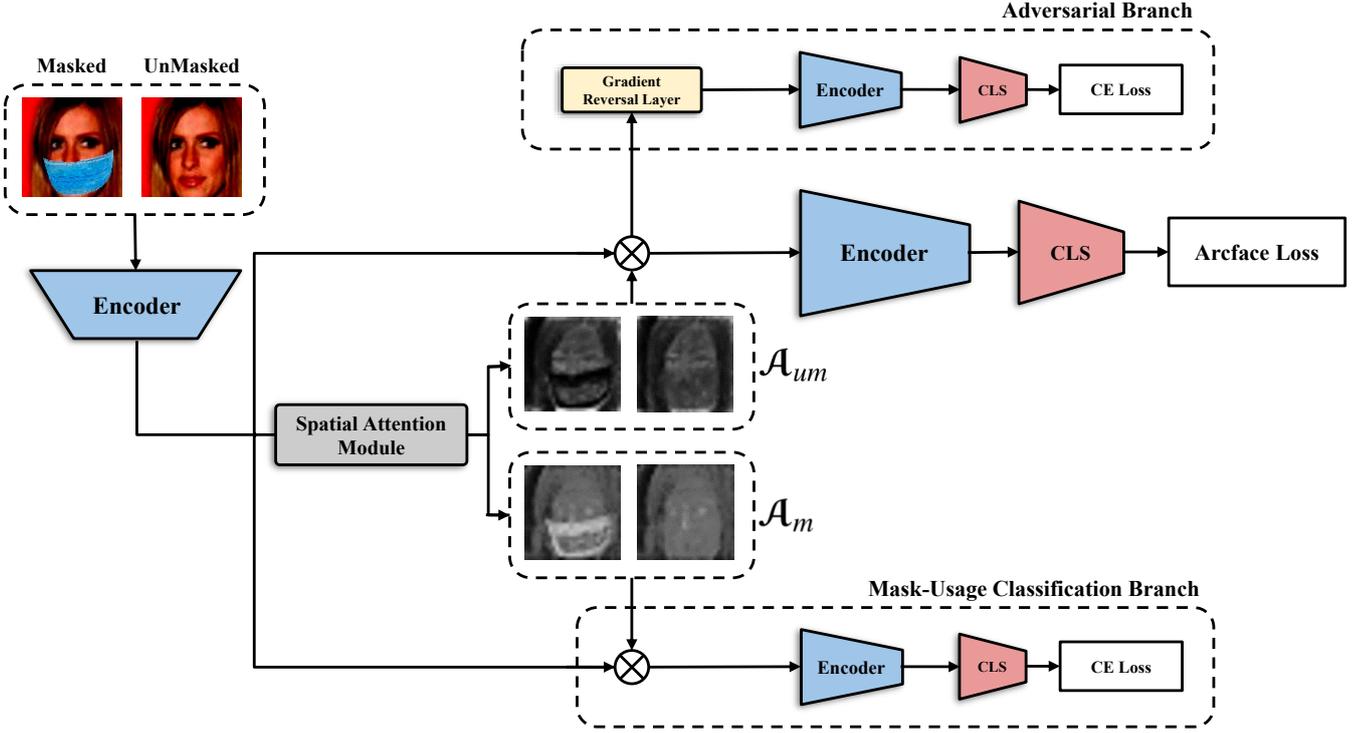
Fig. 3: **Overview of proposed localization method.** Spatial attention module generates attention maps from input images to separate masked and unmasked regions. A_m, A_{um} are masked attention and unmasked attention, respectively. To localize masked and unmasked regions without annotation, A_m is used to learn face recognition with arcface loss and, A_{um} is used to learn mask-usage classification with cross-entropy loss. We empirically verify that masked and unmasked regions are successfully localized, and the performances on both MFR and FR are improved with CAL

$$X_{um} = \mathcal{A}_{um} \otimes X_c, \qquad X_m = \mathcal{A}_m \otimes X_c, \qquad (4)$$

where, X_c is the intermediate feature computed using eq 1, and M_s is the spatial attention module of the CBAM. We illustrate the procedure in the Fig 2a. $\mathcal{A}_m \in \mathbb{R}^{1 \times H \times W}$ is the complementary attention of $\mathcal{A}_{um} \in \mathbb{R}^{1 \times H \times W}$; hence, they activate mutually exclusively. Then, two distinct features X_{um}, X_m are generated by mutually exclusive attention $\mathcal{A}_{um}, \mathcal{A}_m$, respectively. We use X_{um} for training face recognition, and it makes $\mathcal{A}_{um}$ to localize the face region in an image. To prevent $\mathcal{A}_{um}$ focus on the masked region of a face, we use X_m to train mask-usage classification. X_m is optimized to get information related to mask-usage. This makes $\mathcal{A}_m$ focus on masked region, and prevents $\mathcal{A}_{um}$ from being activated in the masked region owing to their complementary property. This method is named "Complementary Attention Learning" (CAL), and compare the performance of baseline with an auxiliary adversarial learning branch. We visualize the overall procedure of CAL and adversarial branch in the Fig 3. Adversarial branch is the same structure with the mask-usage branch, excepts there is the gradient reversal layer [7] at the beginning of the branch. For comparison, we use X_{um} for training the adversarial branch, and it makes X_{um} invariant to mask-usage.

We train FR with arcface loss [6], and train the mask-usage classification branch and adversarial branch with the cross-entropy loss [8], as expressed by the following equations:

$$\mathcal{L}_{arc} = -\frac{1}{N_b} \sum_{i=1}^{N_b} log \left(\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum_{j=1, j \neq y_i}^{N_c} e^{s(cos(\theta_j))}} \right),$$
$$(5)$$

$$\theta_j = cos^{-1} \left(\frac{W_j^\top \cdot z_i}{\|W_j\| \|z_i\|} \right) \qquad (6)$$

$$\mathcal{L}_{CE} = -\frac{1}{N_b} \sum_{i=1}^{N_b} log \left(\frac{e^{z_{y_i}}}{\sum_{j=1}^{N_c} e^{z_{y_j}}} \right), \qquad (7)$$

where N_b and N_c are mini-batch size and the number of the classes, respectively. y_i, z_i is the target label index and logit of x_i, respectively. $W \in \mathbb{R}^{C \times N_c}$ is the weight of last linear classifier, where C is the dimension size of the z. s and m are the scale and margin of arcface loss, and θ_j is the angle between features z_i and weight W_j.

C. Multi-Focal Spatial Attention

In the CBAM, the spatial attention module divides the region into binary regions, such the foreground and background. Then, with CAL, we train complementary attention $\mathcal{A}_m$ to localize a masked region and standard attention $\mathcal{A}_{um}$ to localize a unmasked region. In this case, it is ambiguous to classify a background region that does not belong to the masked and unmasked regions. Therefore, we propose

Train Datasets	Models	Test Datasets					
		MFR	MR-ALL	IJB-C	Age-DB	CALFW	CPLFW
	Baseline	18.49	23.65	69.63	94.07	94.07	88.97
	Baseline + MA=0.1	31.23	26.35	**58.54**	93.96	93.27	88.7
	Baseline + Adv + MA=0.1	33.36	26.64	31.63	93.95	93.15	88.7
	CBAM + CAL + MA=0.1	35.35	**28.58**	50.36	93.92	**93.35**	**89.2**
	MFSA + CAL + MA=0.1	**35.76**	27.61	48.21	**94.03**	93.22	88.73
CASIA	Baseline + MA=0.3	40.37	26.55	**47.82**	93.87	**93.32**	88.82
	Baseline + Adv + MA=0.3	41.73	24.31	30.62	93.33	93	88.37
	CBAM + CAL + MA=0.3	42.1	26.14	35.13	93.8	92.8	88.53
	MFSA + CAL + MA=0.3	**43.44**	**28.94**	35.99	**93.95**	93.07	**88.92**
	Baseline + MA=0.5	42.83	21.80	**18.34**	93.07	92.7	87.68
	Baseline + Adv + MA=0.5	43.15	**22.36**	8.92	92.95	92.6	**87.98**
	CBAM + CAL + MA=0.5	44.4	20.74	11.45	92.68	92.52	87.9
	MFSA + CAL + MA=0.5	**45.2**	21.87	11.86	92.9	**92.83**	87.97
	Baseline	65.86	80.53	94.80	98.30	96.17	**92.90**
MS1MV3	Baseline + MA=0.5	78.25	69.41	**93.68**	97.90	96.03	92.50
	Baseline + Adv + MA=0.5	78.48	68.71	93.54	97.03	95.13	92.30
	CBAM + CAL + MA=0.5	78.45	69.30	93.57	**98.03**	**96.13**	92.72
	MFSA + CAL + MA=0.5	**78.70**	**69.64**	93.62	97.90	96.08	92.70

TABLE I: **Open-sourced face recognition datasets verification performances.** We report 1:1 verification TAR (@FAR=1e-5) on the IJB-C dataset, and verification performance (%) of Age-DB, CALFW and CPLFW. "MFR" and "MR-ALL" denote TAR (@FAR=1e-4) on the masked test set and TAR (@FAR=1e-6) on the multi-racial test set of the ICCV2021-MFR/Insightface track, respectively. "MA" means the masked face augmentation probability. Best in bold, second-best underlined.

the Multi-Focal Spatial Attention (MFSA) to apply N-way classification to a region. There are three classes for MFR (unmasked, masked, and background regions); hence, we use MFSA with 3-way classification. Then, MFSA can be expressed by the following formulas:

$$Z_s = f(X), \qquad [\mathcal{A}_{um}, \mathcal{A}_m, \mathcal{A}_{bg}] = softmax(Z_s) \quad (8)$$

$$X_{um} = \mathcal{A}_{um} \otimes X, \quad X_m = \mathcal{A}_m \otimes X, \quad X_{bg} = \mathcal{A}_{bg} \otimes X, \quad (9)$$

where $f(\cdot)$ is a network composed of pointwise convolution layers, batch normalization layers, and ReLU [15], [18]. To enhance the discriminative capability of $f(\cdot)$, we adopt Orthogonalization by Newton's Iteration (ONI) [12]. It orthogonalizes the weight of the pointwise convolution layers, and disentangle the attention representations. Orthogonalization is a popular technique, which is well-conditioning the network training behavior [2], [13]. Also, weight orthogonalization enforces each channel of Z_s to depend on a different input channel, so it reduces causality and correlation between attentions. To compute attention, $Z_s \in \mathbb{R}^{3 \times H \times W}$ is normalized by $softmax$ function along the channel dimension. Finally, $\mathcal{A}_{um}, \mathcal{A}_m, \mathcal{A}_{bg} \in \mathbb{R}^{1 \times H \times W}$ are element-wise multiplied to the input feature X to generate three features X_{um}, X_m, X_{bg}, respectively. We visualize the procedure of MFSA in the Fig 2b. We utilize the X_{um}, X_m by following the CAL described in Section II-B. X_{bg} is not explicitly utilized during training, but it alleviates the ambiguity of the background region.

III. EXPERIMENTS

A. Training details

We adopt ResNet-50 [10] as the backbone architecture. We train networks using the standard data augmentation (i.e., flipping, translation, cropping), and mask augmentation using the tools introduced in the ICCV2021-MFR/Insightface track [4]. We train the CASIA-Webface and MS1MV3 datasets [26], [5], [6], [9] by employing SGD with a mini-batch size of 512. Momentum and weight decay are set to 0.9 and 5e-4, respectively. We set initial learning rate to 0.2, and employ the polynomial learning rate decay scheduler [20], [22] with 2 epochs of warm restart. We finish the training at 25 epochs and 34 epochs for MS1MV3 and CASIA-Webface datasets, respectively. Following the setup of [6], we set the scale s to 64 and the margin m to 0.5 for arcface loss.

B. Evaluation details

We evaluate the conventional FR performance on the four benchmark FR datasets: IJB-C, Age-DB, CALFW, CPLFW [28], [17], [29], [16]. Additionally, we report the performance on the multi-racial face dataset (MR-ALL) and masked face dataset (MFR) provided by the ICCV2021-MFR/Insightface track [4].
Masked Test Set of ICCV2021-MFR/Insightface track (MFR): It contains 6,964 real-world masked facial images and 13,928 unmasked facial images. There are 20,892 images of 6,964 identities. We evaluate the 1:1 face verification TAR (@FAR=1e-4).
Multi-Racial Test Set of ICCV2021-MFR/Insightface track (MR-ALL): It consists of four demographic groups
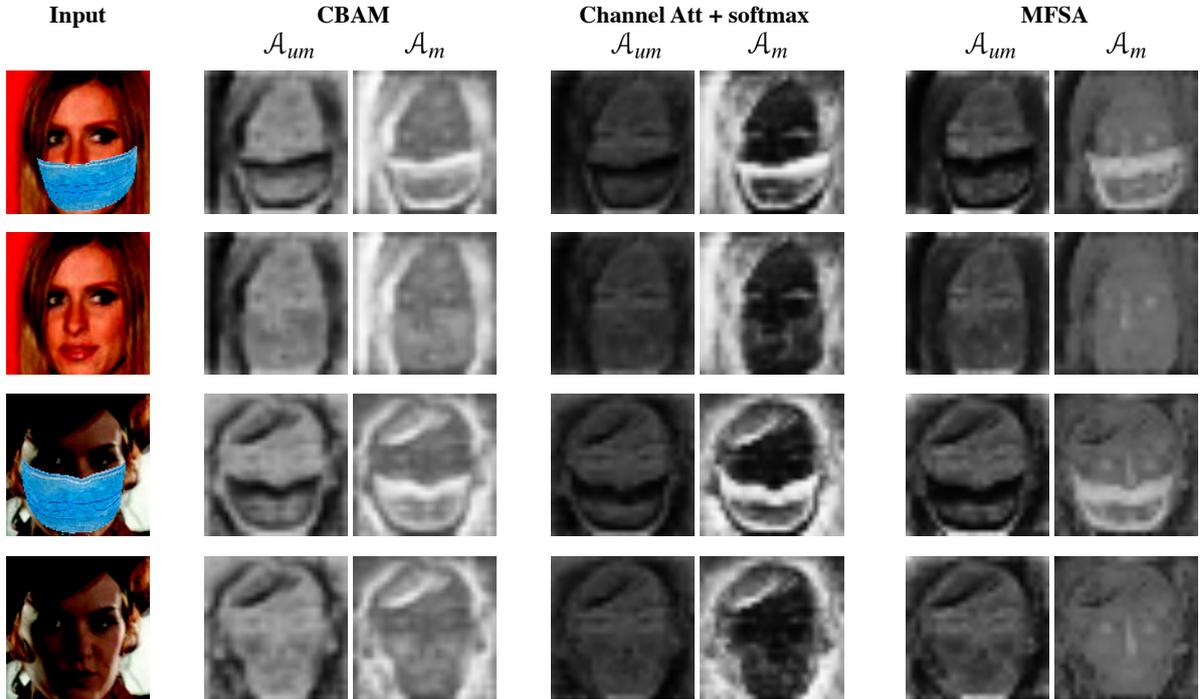
Fig. 4: **Visualization of the attention maps.** From input images, spatial attention modules generates attention maps $\mathcal{A}_{um}$, $\mathcal{A}_m$. We visualize the attention maps as varying the attention modules. "Channel Att $+$ $softmax$" denotes the intermediate attention module that replace $sigmoid$ function of CBAM with $softmax$ function.

(African, Caucasian, South Asia, and East Asia), and contains 1.6M images of 242K identities. We evaluate the 1:1 face verification TAR (@FAR=1e-6).

IJB-C: It is large-scale face recognition dataset that contains 148.8K images of 3,531 identities. We evaluate the 1:1 face verification TAR (@FAR=1e-5).

To investigate efficacy of proposed method, we evaluate the performance of FR and MFR as varying the spatial attention modules (CBAM, MFSA) and training methods (Adv, CAL). "Baseline" denotes the performance of ResNet-50 trained with arcface loss. "MA" means the probability of the masked face augmentation. "Adv" denotes the adversarial learning with auxiliary adversarial branch to make features invariant to mask-usage.

From the results shown in Table I, we verify that Multi-Focal Spatial Attention (MFSA) with Complementary Attention Learning (CAL) demonstrates the best performance on MFR regardless of the MA probability and train datasets. Additionally, CAL shows smaller IJB-C performance drop than for the Adv. For IJB-C, Baseline+MA obtains the best performance, but CAL+MA obtains the second best performance. It indicates CAL successfully enhances the localization capability of spatial attention modules and more precisely extracts mask-invariant features. Therefore, the performances of conventional FR datasets are less degenerated.

C. Qualitative Results

We visualize two attention maps $\mathcal{A}_{um}$ and $\mathcal{A}_m$ as varying the spatial attention modules in the Fig 4. As we expected, in the CBAM, background regions are ambiguous to be

classified. Therefore, background regions are not clearly removed on $\mathcal{A}_{um}$, and it is not desirable phenomenon. "Channel Att $+$ $softmax$" denotes the attention module that replace $sigmoid$ function of CBAM with $softmax$ function. In the "Channel Att $+$ $softmax$", the background regions are not ambiguous to classified, but unmasked regions are not activated at $\mathcal{A}_{um}$. Also, background regions are activated at $\mathcal{A}_m$. We suspect that the representation capacity should be increased to alleviate the problem. Therefore, we propose MFSA that replace channel attention with convolution layers with ONI, and obtain the desirable results. $\mathcal{A}_{um}$ of MFSA is relatively invariant to the mask-usage, and unmasked, masked, and background regions are more clearly divided.

D. Ablations

Models	MFR	MR-ALL	IJB-C
Baseline	18.49	23.65	**69.63**
Baseline + MA=0.1	**31.23**	**26.35**	58.54
Baseline + Adv + MA=0.1	33.36	26.64	31.63
CBAM + MA=0.1	32.71	27.24	37.20
CBAM + Adv + MA=0.1	34.04	27.47	33.66
CBAM + CAL + Adv + MA=0.1	33.89	27.148	40.02
CBAM + CAL + MA=0.1	**35.35**	**28.5**8	**50.36**

TABLE II: Comparisons of 1:1 verification performance (%) on MFR, MR-ALL, and IJB-C. CAL shows the best performance on all test sets. Best in bold.

To investigate the efficacy of CAL, we conduct ablation studies as varying the training methods. As shown in Table II.

We verify that MA improves the MFR performance, but degenerates the MR-ALL and IJB-C performance. With Adv, the performance of MFR is improved by 2.13%, but the performance of IJB-C is significantly decreased by 16.91%. By contrast, CBAM with CAL improves the MFR performance by 3.12%, and decreased the FR performance by 8.18%. Notably CAL + Adv shows the worse performance than CAL. It indicates that CAL is more effective training method to extract mask-invariant features than adversarial learning.

IV. CONCLUSION

In this paper, we propose the method to extract the mask-invariant features by employing and enhancing the localization capability of the CBAM, which is a representative attention module. First, we propose the Complementary Attention Learning (CAL) that adversarially utilizes the complementary attention to prevent the standard attention is being activated on the undesirable area. From the ablation studies, we empirically demonstrate that CAL is more efficient to extract mask-invariant feature than simple adversarial learning. Second, we propose the Multi-Focal Spatial Attention (MFSA) that divides a image into N-way. It alleviates the ambiguity of the background regions classification. From the visualized attention maps of CBAM and MFSA, we verify that MFSA successfully neglects the background regions and extracts mask-invariant features. Additionally, MFSA with CAL gets the best MFR performance regardless of the train dataset and MA probability with relatively smaller FR performance degeneration.

REFERENCES

[1] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1578–1587, 2022.

[2] Y. Cho, H. Cho, Y. Kim, and J. Kim. Improving generalization of batch whitening by convolutional unit optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5321–5329, 2021.

[3] Y. Cho, Y. Kim, H. Cho, J. Ahn, H. G. Hong, and J. Kim. Rethinking efficacy of softmax for lightweight non-local neural networks. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1031–1035. IEEE, 2022.

[4] J. Deng, J. Guo, X. An, Z. Zhu, and S. Zafeiriou. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444, 2021.

[5] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[6] J. Deng, J. Guo, N. Xue, and S. 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.

[7] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[8] I. J. Good. Rational decisions. In *Breakthroughs in statistics*, pages 365–377. Springer, 1992.

[9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] S. Hemathilaka and A. Aponso. A comprehensive study on occlusion invariant face recognition under face mask occlusion. *arXiv preprint arXiv:2201.09089*, 2022.

[12] L. Huang, L. Liu, F. Zhu, D. Wan, Z. Yuan, B. Li, and L. Shao. Controllable orthogonalization in training dnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6429–6438, 2020.

[13] L. Huang, Y. Zhou, F. Zhu, L. Liu, and L. Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4874–4883, 2019.

[14] M. Huber, F. Boutros, F. Kirchbuchner, and N. Damer. Mask-invariant face recognition through template-level knowledge distillation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[16] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.

[17] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, page 5, 2017.

[18] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[19] P. C. Neto, F. Boutros, J. R. Pinto, N. Darner, A. F. Sequeira, and J. 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.

[20] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[21] D. Qi, K. Hu, W. Tan, Q. Yao, and J. Liu. Balanced masked and standard face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1497–1502, 2021.

[22] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[23] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. 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.

[24] T. Wang, C. Zhou, Q. Sun, and H. Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[26] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[27] Y. Zhang, X. Wang, M. S. Shakeel, H. Wan, and W. Kang. Learning upper patch attention using dual-branch training strategy for masked face recognition. *Pattern Recognition*, 126:108522, 2022.

[28] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, February 2018.

[29] T. Zheng, W. Deng, and J. Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017.