Exploring Masked Image Modeling for Face Anti-spoofing

Xuetao Ma, Jun Zhang, Yunfei Zhang, and Daoxiang Zhou(✉)

College of Data Science, Taiyuan University of Technology, Taiyuan 030024, China
{maxuetao1427,zhangjun3398,zhangyunfei4062}@link.tyut.edu.cn,
zhoudaoxiang@tyut.edu.cn

Abstract. Face anti-spoofing (FAS) is an indispensable step in face recognition systems. In order to distinguish spoofing faces from genuine ones, existing methods always require sophisticated handcrafted features or well-designed supervised networks to learn discriminative representation. In this paper, a novel generative self-supervised learning inspired FAS approach is proposed, which has three merits: no need for massive labeled images, excellent discriminative ability, and the learned features have good transferability. Firstly, in the pretext task, the masked image modeling strategy is exploited to learn general fine-grained features via image patches reconstruction in an unsupervised encoder-decoder structure. Secondly, the encoder knowledge is transferred into the downstream FAS task. Finally, the entire network parameters are fine-tuned using only binary labels. Extensive experiments on three standard benchmarks demonstrate that our method can be exceedingly close to the state-of-the-art in FAS, which indicates that masked image modeling is able to learn discriminative face detail features that are beneficial to FAS.

Keywords: Face anti-spoofing · Self-supervised learning · Masked image modeling · Transformer network

1 Introduction

Face recognition has entered the commercial era and is widely used in various scenarios. However, there are many places in the face recognition system that may be attacked.

The most common form of attack is presentation attack, for instance photo print and video replay, which greatly threatens the reliability and security of face recognition systems and makes face anti-spoofing (FAS) a challenging problem.

Over the past ten years, considerable FAS approaches have been put forward successively, which can be grouped into handcrafted methods and convolutional neural network (CNN) based methods. Although they have shown promising

This work was supported by the National Natural Science Foundation of China (62101376) and Natural Science Foundation of Shanxi Province of China (201901D211078).

S. Yu et al. (Eds.): PRCV 2022, LNCS 13534, pp. 814–826, 2022.
https://doi.org/10.1007/978-3-031-18907-4_62

FAS performance, their discriminative and generalization capability still needs to be improved. Firstly, the huge number of labeled face images are required. The performance of these methods relies heavily on the supervision signals, like binary labels, remote photoplethysmography (rPPG) [31] and depth maps [16]. The accuracy may degenerate once the supervision information has some errors. Secondly, the convolution operation acts in a local manner, and therefore it cannot capture the long-range visual context that plays a crucial role in visual pattern recognition. Thirdly, the transfer capability of the learned feature is not encouraging in discriminating unknown types of presentation attacks.

In recent years, self-supervised learning (SSL) has emerged as the most prominent technology to overcome the shortage of supervised learning that require massive labeled data in computer vision. The core idea behind SSL is to learn general features via pretext task, then the learned knowledge is transferred to a specific downstream task, such as recognition, segmentation and detection. It should be pointed out that the pretext task uses a large-scale unlabeled dataset for pre-training, and then uses another relatively small labeled dataset for fine-tuning. SSL is superior to supervised learning in pre-training tasks, and the pretext task does not require labels that makes the model free from massive and complex label information, such as depth maps and rPPG. To sum up, there are two kinds of SSL models: generative and contrastive. The pretext task of contrastive SSL methods seeks to learn image level general semantic features [6,12]. Inspired by BERT [8], masked image modeling (MIM) as a generative self-supervised method has been extensively studied in the past two years. With the help of self-attention mechanism [24] in the transformer models, the two generative SSL methods dubbed masked autoencoders (MAE) [11] and simple masked image modeling (SimMIM) [29] achieved outstanding performance and even surpassed the supervised learning baselines on some image processing tasks. The MIM learns the general image features via masking random patches of the original image and reconstructing the missing pixels. It has the following four advantages: (i) pretext task does not require image label information (ii) can learn general image detail features (iii) the learned general features have excellent transfer capability (iv) can capture the long-range global relationship of features because of the self-attention mechanism in the transformer encoder.

Generally speaking, the pixel details or global spatial structure of an image will be changed for spoofing faces, such as pixel blurring in printed photos and image warping in hand-held photos. In other words, the key discrepancies between spoofing faces and genuine faces come from the image fine-grained information [23] and the global correlation between the features at different regions.

Because the MIM can reconstruct image pixels perfectly even though most regions of the image are masked, which reveals that MIM is capable of learning image detail information and capturing image spatial structure. Accordingly, our initial motivation for this work is to learn detailed features for faces through MIM, which is helpful for detecting presentation attacks. What is more, the transformer encoder network can learn the global correlation between visual

features, which is an important clue to distinguish between genuine and spoofing faces.

From the above analysis, in order to address the aforementioned issues of existing FAS methods, this paper proposes a novel and simple method to learn general and discriminative features for FAS under the SSL framework. The overall pipeline of our method is illustrated in Fig. 1. In the pretext task stage, the MIM is exploited to learn general face detail features in an unsupervised fashion under transformer encoder-decoder architecture. Afterward, the trained encoder knowledge is utilized to initialize the encoder of our downstream FAS task. Since we consider FAS as an image classification problem, and therefore the encoder is followed by a simple network only with global average pooling (GAP) and fully connected (FC) layers instead of the decoder. The main contributions of this paper are threefold:

- To our knowledge, this work is the first attempt to exploit generative SSL for FAS. The SSL strategy renders our method can achieve better results than supervised learning methods on the premise of using a large amount of unlabeled images for pre-training, which effectively reduces the cost of labeling.
- We explore the effectiveness of two different MIM models in learning general face detail features that have superior discriminative ability and transfer advantages.
- We conduct extensive FAS experiments on three popular datasets. The results show that our method offers competitive performance compared with other FAS methods.

2 Related Work

2.1 Face Anti-spoofing

The majority of FAS methods are based on supervised learning. From the early period of handcrafted feature methods, such as LBP [21], etc., these methods require at least binary label as supervised information. With the rise of deep learning, there are more types of clues that have been proven to be discriminative to distinguish spoofing faces. In [1], depth maps are introduced into the FAS task firstly. In addition, [16] leverages depth maps and rPPG signal as supervision. Besides, reflection maps and binary mask are respectively introduced by [13] and [17]. In the past two years, the Vision Transformer (ViT) structure has achieved success in vision tasks. Some researchers have applied ViT to FAS. Although the new architecture further improves the indicators of FAS, these works still require various types of supervision. For example, ViTranZFAS [10] needs binary labels, and TransRPPG [31] needs rPPG as supervision.

Various types of supervision information seriously increase the cost of labeling, and the quality of labels also greatly affects the performance of models. Therefore, some researches begun to explore the FAS methods based on contrastive SSL [15,20]. These works not only get rid of constraint of labels, but

also achieve better performance than supervised learning. Unlike these methods, this paper adopts generative SSL method.

2.2 Masked Image Modeling

Masked image modeling is a generative self-supervised method. The work in [25] proposes denoising autoencoders (DAE), which corrupts the input signal and learns to reconstruct the original input. Further, the work of [26] takes masking as a noise type in DAE. They randomly set some values in the input data to zero with a certain probability, then train the encoder to reconstruct these values.

DAE first achieved great success in the field of NLP. Transformer [24] and BERT [8] are the most representative architectures. Specifically, a self-attention mechanism is proposed in Transformer to capture the relationship between different tokens. Further, a special token [MASK] is introduced to BERT. The [MASK] will replace some tokens in training phase, then the network predicts the original words in this position. After the masked model has achieved such great achievements in NLP area, a natural question is how to apply this model to computer vision tasks.

Some pioneering works in the recent years has explored the potential of MIM. iGPT [5] reshapes the raw images to a 1D sequence of pixels and predicts unknown pixels. The BEiT [2] proposes a pre-training task called MIM, and also introduces the definition of MIM firstly. In BEiT, the image is represented as discrete tokens, and these tokens will be treated as the construct target of masked patches. Most recently, MAE [11] and SimMIM [29] almost simultaneously obtain state-of-the-art on computer vision tasks. They propose a pre-training paradigm based on MIM, that is, the patches of images are randomly masked with a high probability (usually greater than 50%), then the self-attention mechanism is used in the encoder to learn the relationship between patches, and finally the masked patches is reconstructed in the decoder.

3 Methodology

3.1 Intuition and Motivation

Spoofing faces are very similar in appearance to genuine faces. Their main differences are the image pixel details (blur and color) and the overall image structure (deformation and specular reflection). Learning discriminative cues from numerous labeled samples via CNN is a common way, but it is hard to learn general features, so the generalization ability needs to be improved, and the cost of producing labeled samples is expensive. So how to learn the general discriminative features that can distinguish spoofing faces from genuine ones on small amount labeled faces are the main challenge of FAS.

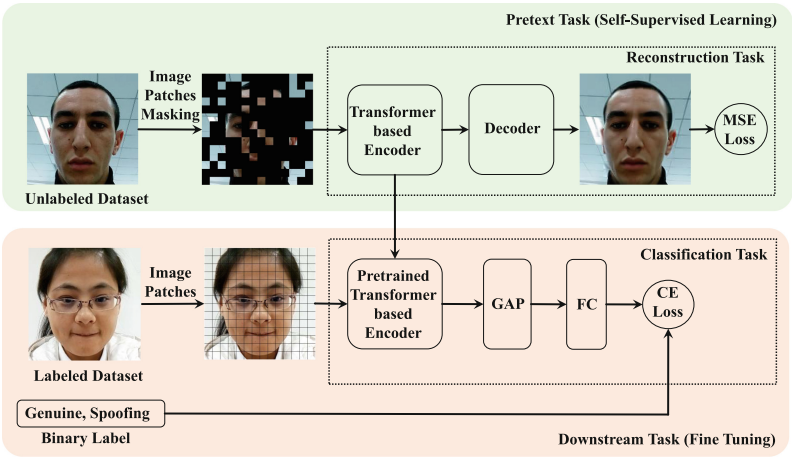

Fig. 1. Overall architecture of our proposed face anti-spoofing with masked image modeling.

3.2 The Proposed Method

Pretext Task Stage. SSL has been recognized as an effective way to remedy the shortcoming of the appetite for a large amount of labeled data. Due to the strong power of MIM in reconstructing image pixels, we argue that it can capture face detail visual features and the image structure via position embedding. Moreover, the global features of face image can be characterized by the self-attention in the transformer. Consequently, the general discriminative face visual cues with good transfer ability can be learned by MIM in an unsupervised manner.

In this paper, we mainly consider two newly proposed MIM methods: MAE [11] and SimMIM [29]. The ViT [9] and swin transformer [18] are adopted as the encoder backbone of MAE and SimMIM respectively. Meanwhile, the experiments of MAE and SimMIM both prove that random mask is more effective, so this paper also adopts the random mask. Concretely, we first divide a face image into several non-overlapping patches and randomly mask a large portion of the patches according to the mask ratio. For MAE, the encoder network with multiple transformer blocks are called to learn latent representations from the remaining unmasked patches. For SimMIM, both unmasked patches and mask tokens are fed into the encoder. All the tokens composed of encoded visible patches and mask tokens are fed into a lightweight decoder that is responsible for regressing the raw pixel values of masked area under mean squared error or l_1 loss.

Downstream Task Stage. Having obtained the knowledge from the trained pretext task, we directly apply the encoder to our downstream FAS task and discard the decoder. For the purpose of recognition, a binary classification network with GAP and FC layers is added after the encoder, and the cross-entropy

loss is employed in this stage. We choose fine-tuning instead of linear probing to conduct supervised training to evaluate the face feature representations.

4 Experiments

4.1 Datasets and Evaluation Metrics

To evaluate the effectiveness of our method, extensive experiments are carried out on three representative datasets. OULU-NPU [3] contains 4950 high-resolution videos from 55 individuals. CASIA-FASD [36] comprises 600 videos from 50 subjects under three types of attacks. Repay-Attack [7] has 1200 videos from 50 persons with 24 videos per person under three kinds of attacks. Three widely used metrics are adopted [32]: attack presentation classification error rate, APCER = FP/(TN+FP), bona fide presentation classification error rate, BPCER = FN/(TP+FN), average classification error rate, ACER = (APCER+BPCER)/2, and equal error rate (EER). The lower scores signify better performance.

4.2 Implementation Details

Our method is implemented via Pytorch on an Ubuntu system with NVIDIA Tesla V100 and 32 GB graphics memory. The input images of pretext task and downstream task are of size 224×224, and each image is into regular non-overlapping patches of size 16×16. It should be pointed out that we did not use any additional datasets such as ImageNet. The epochs of pretext task and fine-tuning for ours MAE (SimMIM) are 1600 (1000) and 100 (100) respectively. The fine-tuning process for the downstream classification task is performed on each dataset or its protocol. Following [35], the frame-level image is used in this paper instead of the entire video. For simplicity, the first 20 frames of each spoofing video from the training set are selected. In order to alleviate the data imbalance problem, we select more frames for the genuine video so that the ratio between positive and negative samples is 1:1. In the testing phase, 20 trials of each video from the test set are conducted, and the average results are reported, for the i-th trail, the i-th frame for each test video is utilized.

4.3 Experimental Results and Analysis

Effect of Mask Ratio. The mask ratio of MIM is an important factor that has an obvious effect on the performance of visual recognition. To assess the impact of the mask ratio on the FAS task, three mask ratios $\{0.50, 0.60, 0.75\}$ are evaluated for both MAE and SimMIM. Several experiments are carried out on the four protocols of OULU-NPU. The results are shown in Fig. 2(a).

For MAE, mask ratio and ACER scores basically show negative correlation. For SimMIM, the performance of the three mask rates in protocol 2 and protocol 3 is very similar. At the same time, the performance of 0.75 mask rate in protocol

1 and protocol 4 is significantly better than other mask rates. These experimental results show that different mask ratios and the choices of MIM models have a great impact on FAS.

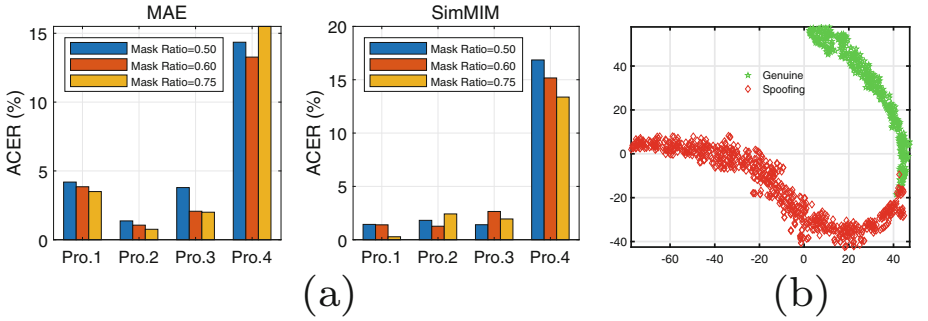

Fig. 2. (a) ACER (%) versus mask ratio under MAE and SimMIM on the four protocols of OULU-NPU dataset. (b) Feature distribution visualization for all 1080 testing videos from OULU-NPU protocol 2 via t-SNE.

Transfer Ability of Pretext Task. When superior performance is shown on a single dataset, one natural question is how well the transfer ability of the MIM pretext task is. To answer this, we conducted six experiments. We first train our MIM pretext task on the training set of OULU-NPU, Replay-Attack, and CASIA-FASD. After knowledge transferring, the fine-tuning of downstream tasks are conducted on the training set of CASIA-FASD and Replay-Attack. All the ACER scores are enumerated in Table 1, we can get the following observations: (1) Even though the pretext tasks are trained on different datasets, the downstream task still has good performance, which reveals the generalization ability of the MIM pretext task is excellent. (2) On the Replay-Attack, the ACER scores for all three cases are 0. (3) The training videos of CASIA-FASD are only 240 and are less than that of OULU-NPU and Replay-Attack. Ours model (SimMIM) achieves better results when the pretext task is performed on a large training dataset than on a small one. Such phenomenon is consistent with the founding in transformer models, i.e., the more training data, the better the performance.

Table 1. ACER (%) of different cases of knowledge transferring. O, C and R denotes OULU-NPU, CASIA-FASD and Replay-Attack.

Pretext Task On	O	R	O	C	C	R
Fine-Tune On	C		R		C	R
SimMIM-0.75	0.343	0.111	0.00	0.00	0.52	0.00
MAE-0.75	0.10	0.47	0.00	0.00	0.06	0.00

Comparison with State-of-the-Art Methods. In what follows, we compare the performance of our approach on OULU-NPU with several classical methods, including three CNN based methods: attention-based two-stream CNN (ATS-CNN) [4], central difference convolutional networks (CDCN) [34] and neural architecture search (NAS) for FAS [33]. Three transformer based methods: temporal transformer network with spatial parts (TTN-S) [28], video transformer based PAD (ViTransPAD) [19] and two-stream vision transformers framework (TSViT) [22]. One SSL-based method: Temporal Sequence Sampling (TSS) [20]. All the comparison results on the four protocols are tabulated in Table 2.

Table 2. Results on OULU-NPU dataset. architecture C and T denotes CNN and transformer. {M, S}-{0.50, 0.60, 0.75} stands for SimMIM and MAE under the mask ratio respectively. Bold values are the best results in each case.

Prot.	Method	APCER (%)	BPCER (%)	ACER (%)	Arc.	Notes
1	ATS-CNN [4]	5.1	6.7	5.9	C	20 TIFS
	CDCN [34]	0.4	1.7	1.0	C	20 CVPR
	NAS-FAS [33]	0.4	**0.0**	0.2	C	21 TPAMI
	TTN-S [28]	0.4	**0.0**	0.2	T	22 TIFS
	ViTransPAD [19]	0.4	0.2	0.3	T	22 arXiv
	TSViT [22]	1.7	0.0	0.9	T	22 JCVR
	TSS [20]	**0.0**	0.2	**0.1**	C	22 PRL
	Ours (S-0.75)	0.44	0.13	0.28	T	–
	Ours (M-0.75)	4.91	2.08	3.5	T	–
2	ATS-CNN [4]	7.6	2.2	4.9	C	20 TIFS
	CDCN [34]	1.5	1.4	1.5	C	20 CVPR
	NAS-FAS [33]	1.5	0.8	1.2	C	21 TPAMI
	TTN-S [28]	**0.4**	0.8	**0.6**	T	22 TIFS
	ViTransPAD [19]	2.0	0.4	1.2	T	22 arXiv
	TSViT [22]	0.8	1.3	1.1	T	22 JCVR
	TSS [20]	**0.4**	0.8	**0.6**	C	22 PRL
	Ours (S-0.60)	1.45	1.08	1.27	T	–
	Ours (M-0.75)	1.18	**0.34**	0.76	T	–
3	ATS-CNN [4]	3.9 ± 2.8	7.3 ± 1.1	5.6 ± 1.6	C	20 TIFS
	CDCN [34]	2.4 ± 1.3	2.2 ± 2.0	2.3 ± 1.4	C	20 CVPR
	NAS-FAS [33]	2.1 ± 1.3	1.4 ± 1.1	1.7 ± 0.6	C	21 TPAMI
	TTN-S [28]	$\mathbf{1.0 \pm 1.1}$	0.8 ± 1.3	$\mathbf{0.9 \pm 0.7}$	T	22 TIFS
	ViTransPAD [19]	3.1 ± 3.0	1.0 ± 1.3	2.0 ± 1.5	T	22 arXiv
	TSViT [22]	2.4 ± 2.6	1.4 ± 2.2	1.9 ± 1.3	T	22 JCVR
	TSS [20]	2.5 ± 1.8	$\mathbf{0.5 \pm 0.6}$	1.5 ± 0.8	C	22 PRL
	Ours (S-0.50)	1.01 ± 0.80	1.81 ± 2.72	1.41 ± 1.18	T	–
	Ours (M-0.75)	1.57 ± 1.40	2.44 ± 3.43	2.00 ± 1.55	T	–
4	ATS-CNN [4]	11.3 ± 3.9	9.7 ± 4.8	9.8 ± 4.2	C	20 TIFS
	CDCN [34]	4.6 ± 4.6	9.2 ± 8.0	6.9 ± 2.9	C	20 CVPR
	NAS-FAS [33]	4.2 ± 5.3	1.7 ± 2.6	2.9 ± 2.8	C	21 TPAMI
	TTN-S [28]	$\mathbf{3.3 \pm 2.8}$	2.5 ± 2.0	2.9 ± 1.4	T	22 TIFS
	ViTransPAD [19]	4.4 ± 4.8	$\mathbf{0.2 \pm 0.6}$	$\mathbf{2.3 \pm 2.4}$	T	22 arXiv
	TSViT [22]	7.4 ± 5.0	1.2 ± 2.2	4.3 ± 1.9	T	22 JCVR
	TSS [20]	4.7 ± 10.5	9.2 ± 10.4	7.1 ± 5.3	C	22 PRL
	Ours (S-0.75)	19.06 ± 17.70	7.69 ± 9.70	13.38 ± 5.92	T	–
	Ours (M-0.50)	14.35 ± 16.36	12.20 ± 12.59	13.27 ± 5.96	T	–

Compared with these state-of-the-art methods, our method does not achieve best performance, especially in protocol 4. Nonetheless, our method still gets competitive results, for examples, the best BPCER in protocol 2, the second best ACER in protocol 2 and APCER in protocol 3. The reason why these methods outperform our method is that they design ingenious but complex models, which increase the consumption of computational resources. It should be noted that our models are relatively simple and do not require complex label information and structure design. This means that our method has great potential ability. For example, the architecture of TTN-S [28] is complex because it combines temporal difference attention, pyramid temporal aggregation and transformer. ViTransPAD [19] has high computation burden since it captures local spatial details with short attention and long-range temporal dependencies over frames. The architecture of TSViT [22] is also complex since it leverages transformer to learn complementary features simultaneously from RGB color space and multi-scale Retinex with color restoration space.

To sum up, the reasons for the excellent performance of our proposed approach are originated from two aspects: (i) masking and reconstruction strategy are well in learning face detail features. (ii) the self-attention of transformer is able to extract image global information.

To investigate our approach more comprehensively, we compare our method with several models on Replay-Attack and CASIA-FASD. All the testing videos of Replay-Attack are recognized correctly, and our EER score is the lowest for CASIA-FASD, which can evidently verify the superiority of our method again.

Table 3. Results on CASIA-FASD and Replay-Attack Datasets. Bold values are the best results in each case.

Methods	CASIA-FASD	Repaly-Attack		Notes
	EER (%)	EER (%)	ACER (%)	
LBP [7]	18.2	13.9	13.8	12 BIOSIG
CNN [30]	4.64	4.46	–	14 arXiv
3D-CNN [14]	1.40	0.30	1.20	18 TIFS
ATS-CNN [4]	3.14	0.13	0.25	20 TIFS
DTN [27]	1.34	0.06	0.02	21 TIFS
Ours (S-0.75)	0.33	**0.00**	**0.00**	–
Ours (M-0.75)	**0.06**	**0.00**	**0.00**	–

Ablation Study. We perform an ablation study on protocol 1 and protocol 2 of OULU-NPU to show that the experimental results not only benefit from ViTs structure but also benefit from MIM. We train the downstream tasks without pretext-task, which is using a pure ViT to train the FAS task separately. All ACER scores are enumerated in Table 4, and we can get the following observations: The pretext task plays a crucial role in the performance of our model. The ACER results of our method are significantly better than pure ViT on the two protocols. Such experimental results sufficiently prove the necessity of MIM.

Table 4. Ablation experimental results on OULU-NPU dataset.

Prot.	Method	APCER (%)	BPCER (%)	ACER (%)
1	ViT (w/o mim)	0.25	20.50	10.38
	M-0.75 (w/ mim)	4.91	**2.08**	**3.50**
2	ViT (w/o mim)	3.32	2.14	2.73
	M-0.75 (w/ mim)	**1.18**	**0.34**	**0.76**

4.4 Visualization

Feature Distribution. To visualize the distribution of our learned features based on MAE, the 1080 testing videos in protocol 2 of OULU-NPU are used, and the GAP processed feature matrix with the dimensions of 768×1080 are fed into the t-SNE algorithm. From Fig. 2(b), it can be seen that the genuine videos and spoofing videos are very distinguishable, which obviously implies that our learned features possess the powerful discriminative capability.

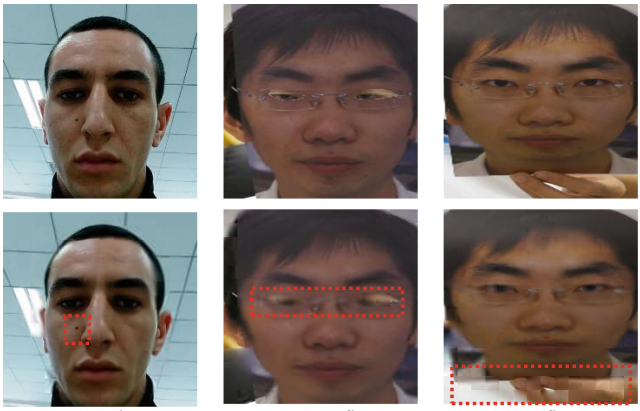

Fig. 3. Reconstruction details marked by red boxes for genuine and spoofing faces.

Reconstruction Details. To further illustrate the effectiveness of our method, we display the reconstruction details for different type of face images, as shown in Fig. 3. Columns 1–3 represents genuine face, eye-cut photo attack and hand-held photo attacks. For FAS task, the differences between spoofing and genuine faces often lies in the pixel details. One can notice that the MIM focuses on the perfect reconstruction of the face area. Among them, for the image in column 2, the reconstruction quality of the eye-cut region is unpromising, for the image in column 3, the reconstruction quality of the hand-held region is incorrect. These parts that cannot be reconstructed well are all non-face areas. This discovery directly prove that our method pays attention to the learning of detailed facial features and autonomously discovers the visual cues of spoofing faces.

5 Conclusion

This paper proposes a novel FAS method under the SSL framework. In the pretext task stage, the MIM strategy is employed to learn general face detail features under an encoder-decoder structure. In the downstream task stage, the knowledge in the encoder is directly transferred, followed by a simple classification network only with GAP and FC layers. Extensive experiments on three standard benchmarks show that our method gets competitive results, which demonstrates the MIM pretext task is effective to learn general and discriminative face features that are beneficial to FAS.

References

1. Atoum, Y., Liu, Y., Jourabloo, A., Liu, X.: Face anti-spoofing using patch and depth-based cnns. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 319–328. IEEE (2017)
2. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
3. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: Oulu-npu: a mobile face presentation attack database with real-world variations. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 612–618 (2017)
4. Chen, H., Hu, G., Lei, Z., Chen, Y., Robertson, N.M., Li, S.Z.: Attention-based two-stream convolutional networks for face spoofing detection. IEEE Trans. Inf. Forensics Secur. **15**, 578–593 (2020)
5. Chen, M., et al.: Generative pretraining from pixels. In: International Conference on Machine Learning, pp. 1691–1703. PMLR (2020)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning (ICML), vol. 119, pp. 1597–1607 (2020)
7. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG), pp. 1–7 (2012)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), pp. 4171–4186 (2019)
9. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations (ICLR), pp. 1–21 (2021)
10. George, A., Marcel, S.: On the effectiveness of vision transformers for zero-shot face anti-spoofing. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE (2021)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738 (2020)

13. Kim, T., Kim, Y., Kim, I., Kim, D.: Basn: enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 494–503 (2019)
14. Li, H., He, P., Wang, S., Rocha, A., Jiang, X., Kot, A.C.: Learning generalized deep feature representation for face anti-spoofing. IEEE Trans. Inf. Forensics Secur. **13**(10), 2639–2652 (2018)
15. Liu, H., Kong, Z., Ramachandra, R., Liu, F., Shen, L., Busch, C.: Taming self-supervised learning for presentation attack detection: In-image de-folding and out-of-image de-mixing. arXiv preprint arXiv:2109.04100v1 (2021)
16. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: binary or auxiliary supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 389–398 (2018)
17. Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X.: Deep tree learning for zero-shot face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4680–4689 (2019)
18. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021)
19. Ming, Z., Yu, Z., Al-Ghadi, M., Visani, M., MuzzamilLuqman, M., Burie, J.C.: Vitranspad: video transformer using convolution and self-attention for face presentation attack detection. arXiv preprint arXiv:2203.01562 (2022)
20. Muhammad, U., Yu, Z., Komulainen, J.: Self-supervised 2d face presentation attack detection via temporal sequence sampling. Pattern Recogn. Lett. **156**, 15–22 (2022)
21. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Proceedings of 12th International Conference on Pattern Recognition, vol. 1, pp. 582–585. IEEE (1994)
22. Peng, F., Meng, S., Long, M.: Presentation attack detection based on two-stream vision transformers with self-attention fusion. J. Vis. Commun. Image Representation **85**, 103518 (2022)
23. Shao, R., Lan, X., Yuen, P.C.: Regularized fine-grained meta face anti-spoofing. In: Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI), pp. 11974–11981 (2020)
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
25. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103 (2008)
26. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., Bottou, L.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**(12), 3371–3408 (2010)
27. Wang, Y., Song, X., Xu, T., Feng, Z., Wu, X.J.: From rgb to depth: Domain transfer network for face anti-spoofing. IEEE Trans. Inf. Forensics Secur. **16**, 4280–4290 (2021)
28. Wang, Z., Wang, Q., Deng, W., Guo, G.: Learning multi-granularity temporal characteristics for face anti-spoofing. IEEE Trans. Inf. Forensics Sec. **17**, 1254–1269 (2022)
29. Xie, Z., et al.: Simmim: a simple framework for masked image modeling. arXiv preprint arXiv:2111.09886 (2021)
30. Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601 (2014)

31. Yu, Z., Li, X., Wang, P., Zhao, G.: Transrppg: remote photoplethysmography transformer for 3d mask face presentation attack detection. IEEE Signal Process. Lett. **28**, 1290–1294 (2021)
32. Yu, Z., et al.: Multi-modal face anti-spoofing based on central difference networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2766–2774 (2020)
33. Yu, Z., Wan, J., Qin, Y., Li, X., Li, S.Z., Zhao, G.: Nas-fas: static-dynamic central difference network search for face anti-spoofing. IEEE Trans. Pattern Anal. Mach. Intell. **43**(9), 3005–3023 (2021)
34. Yu, Z., et al.: Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5295–5305 (2020)
35. Zhang, L.B., Peng, F., Qin, L., Long, M.: Face spoofing detection based on color texture markov feature and support vector machine recursive feature elimination. J. Vis. Commun. Image Represent. **51**, 56–69 (2018)
36. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: 2012 5th IAPR international conference on Biometrics (ICB), pp. 26–31 (2012)