Pattern Recognition Letters 171 (2023) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A simple and effective patch-Based method for frame-level face anti-spoofing

Shengjie Chen[a], Gang Wu[a], Yujiu Yang[b], Zhenhua Guo[b,*]

[a] Tsinghua University, Beijing, China
[b] Tsinghua Shenzhen International Graduate School, Shenzhen, China

ARTICLE INFO

Article history:
Received 6 April 2022
Revised 18 April 2023
Accepted 22 April 2023
Available online 25 April 2023

Edited by: Jiwen Lu

Keywords:
Face anti-spoofing
Liveness detection
Image-level
Attention mechanism
Patch sampling

ABSTRACT

With the wide applications of face recognition, face anti-spoofing has become a major challenge for reliable face recognition. Thus, it is necessary to perform face liveness detection. Most existing methods rely on a whole face image for training and testing and are thus susceptible to the overfitting problem because of limited training samples; meanwhile, liveness information is not fully explored. To address these issues, we propose a simple and effective patch-based approach. There are two main contributions: 1) different patch sampling strategies are applied to a training set and a testing set to overcome the overfitting problem, and 2) an attention mechanism is applied to explore more significant information for liveness detection. We evaluate the proposed approach on four popular and challenging databases: the CASIA-SURF, OULU-NPU, CASIA-FASD and REPLAY-ATTACK databases. The proposed method could obtain very promising liveness detection performance. For example, the average classification error rate (ACER) on the CASIA-SURF database (using RGB images only) was 1.6%, which is the lowest reported error rate to the best of our knowledge.

1. Introduction

In recent years, face recognition has been widely used in many applications, such as time attendance systems, mobile phone unlocking and payment verification. However, due to the poor privacy of face, criminals can easily obtain them through social media or other means. They may use these photos or videos to carry out various present attack (PA) [1] on face recognition systems, such as print attack, replay attack and realistic 3D mask made with skin-like materials. Accordingly, face liveness detection is very important for face recognition.

In recent decades, many approaches have been developed to detect different PAs. The first category is hardware-based methods, which generate multimodal information by adding extra light sources [2] or sensors [3]. However, this strategy will also increase the acquisition cost and may not suitable for some applications. Another idea of liveness detection is based on human-computer interactions. Some face verification systems require users to open their mouths, shake their heads, or blink to prove that they are not pictures. These methods are not friendly to users and require more time to complete verification. In addition, some corresponding attack methods have been proposed, such as printing photos without a nose, eyes, a mouth or some combination of these features [4]. Therefore, an efficient and robust face recognition system requires a more accurate, cheap, fast and user-friendly liveness detection algorithm.

Due to the poor accuracy and generality of face liveness detection methods based on handcrafted features, researchers focused more attention on deep learning methods. Although artificial neural networks have strong feature representation and semantic reasoning capabilities, the limited size of existing face anti-spoofing databases makes it easy to suffer severe overfitting when actually training a discriminative network. For example, the CASIA-SURF database [4], which is currently the largest face liveness detection database, only contains 1000 subjects. We trained a standard ResNet18 [5] for liveness detection using this database. The accuracy of the network on the training set and validation set is shown in Fig. 1. The former gradually converged to about 93% after three epochs, while the latter continued to oscillate at lower values.

To overcome the overfitting problem in network training, researchers [1,6,7] proposed to add auxiliary supervision information, such as depth and temporal information. However, auxiliary supervision information requires additional acquisition or calculation, while temporal information is only available in video-based face recognition systems. We notice that the discriminative information

* Corresponding author.
E-mail addresses: chensjt@mail.tsinghua.edu.cn (S. Chen), wg18@tsinghua.org.cn (G. Wu), yang.yujiu@sz.tsinghua.edu.cn (Y. Yang), zhenhua.guo@sz.tsinghua.edu.cn (Z. Guo).

https://doi.org/10.1016/j.patrec.2023.04.011
0167-8655/

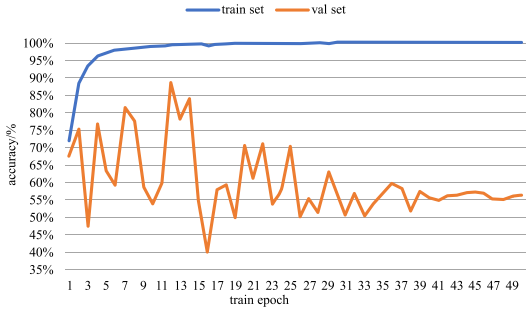


Fig. 1. The accuracy of training and validation sets for the CASIA-SURF database.

in a full face image is rich, including skin texture, color, shadow, local shape, etc., and some information may only be obvious in the current training set. If trained with full images, the network may achieve good performance on the training set based on arbitrary cues. Once the dataset is changed, the performance of the trained network tends to decrease significantly. Therefore, we decide to enhance the generalization ability of the network by reducing the amount of information fed into the network during training. In this paper, we propose a simple and effective patch-based frame-level face liveness detection method, which can achieve considerable performance with only a single RGB image. Specifically, the full face images will be randomly cropped into fixed-size patches as input to the network. In order to avoid overfitting in the training stage and make the network more fully utilize the information of the image in the inference stage, we design different patch sizes for the two stages. The patch size in the inference stage is larger than that in the training stage. In addition, Yang et al. [8] point out that the attention mechanism can help a model choose some important patches. However, patch attention is difficult to learn. Therefore, channel attention and spatial attention may be more suitable for face anti-spoofing since they can help describe fine-grained information in face images.

The main contributions of this work are summarized as follows:

- We propose a frame-level method that is trained with randomly cropped patches from face images. Meanwhile, we propose to use different crop strategies in training and inference, which has not been explored in previous studies.
- We use spatial and channel attention to obtain important information for face liveness detection. Experimental results show that both attention mechanisms are effective.
- Without any auxiliary or outside information, the proposed method achieves the best performance on the CASIA-SURF database when only RGB images are utilized. Comparable results can be obtained on the OULU-NPU database, especially in the three difficult test protocols.

2. Related works

2.1. CNN-Based frame-level face anti-spoofing

Modern convolutional neural network (CNN) with powerful feature representation and semantic reasoning capabilities have been widely exploited in frame-level face anti-spoofing (FAS). Yang et al. [9] propose to use CNN to learn features for FAS. Li et al. [10] believe that RGB and HSV color spaces are not suited for FAS task, and thus propose a deep learning net named CompactNet to learn a new compact space. Yu et al. [6] design a new central difference convolution (CDC), which can capture intrinsic detailed patterns by aggregating both intensity and gradient information. Yu et al. [11] proposed to identify the discriminative salient parts of image using deep reinforcement learning. They extracted and fused high-level and local features of images to assist the discriminative network in identifying fake faces more robustly. Wang et al. [12] developed a learnable gradient operator (LGO) to adaptively learn gradient fine-grained information. The model used the generalization of the existing gradient operator to obtain efficient high-frequency information calculation capabilities, which is conducive to capturing more detailed discriminative clues from the original pixels.The input of these CNN-based models are all full face images, which makes the training suffer from overfitting. In this study, we propose to use patch training, which can improve the generalization ability of the model.

2.2. Patch training

Due to the relatively small scale of face anti-spoofing datasets, some researchers use patches instead of whole face images to train their models. Atoum et al. [13] propose a two-stream CNN-based approach for present attack detection. The patch-based CNN extracts local features. The final detection score of one image is the average score of all patches. To make full use of multimodal data, Shen et al. [14] propose a multistream CNN model, named FaceBagNet, for multi-modal face anti-spoofing. Yang et al. [8] design a region attention module to obtain vital local regions rather than patches randomly cropped from images. However, the abovementioned methods use the same size for training and testing patches. In this study, to avoid the overfitting problem and to make full use of the test samples, we propose to use different sampling schemes in the training and testing stage. To the best of our knowledge, this is the first attempt at a cropping scheme.

2.3. Attention mechanism

Many researchers work on attention mechanisms in computer vision, such as SENet [15], CBAM [16], and nonlocal neural networks [17]. Some researchers have proposed attention-based methods for face anti-spoofing. For example, Yang et al. [8] utilize the RAM structure, which can learn a position offset to crop important patches. Chen et al. [18] separately feed RGB and MSR (converted from an original RGB image in an offline way) images to two CNNs and then combined two CNN features through attention. However, attention is not well studied in patch-level methods. In this paper, we propose to use spatial and channel attention simultaneously to improve performance.

3. The proposed method

Fig. 2 shows the overall structure of our patch-based frame-level face anti-spoofing method. The basic framework of our network references the ResNet18 [5], which has been proven to be effective in multiple vision tasks. The input to the network is patches rather than full face images, and we design different sampling schemes for the training and inference stages. To obtain larger receptive field, the kernel size of the first convolutional layer is 5×5, and the other layers is 3×3 considering the amount of computation. Every ResBlock has two BasicBlocks in which spatial and channel attention are utilized. To make the network compatible with different input sizes (the testing size is different from the training size), we apply a global average pooling (GAP) layer after ResBlock4.

In this section, we first introduce our patch-based sampling scheme. Then, we present spatial and channel attention for face anti-spoofing.

3.1. Patch-based sampling scheme

There are many reasons to utilize a patch-based approach. First, the existing face liveness detection databases are not large. The

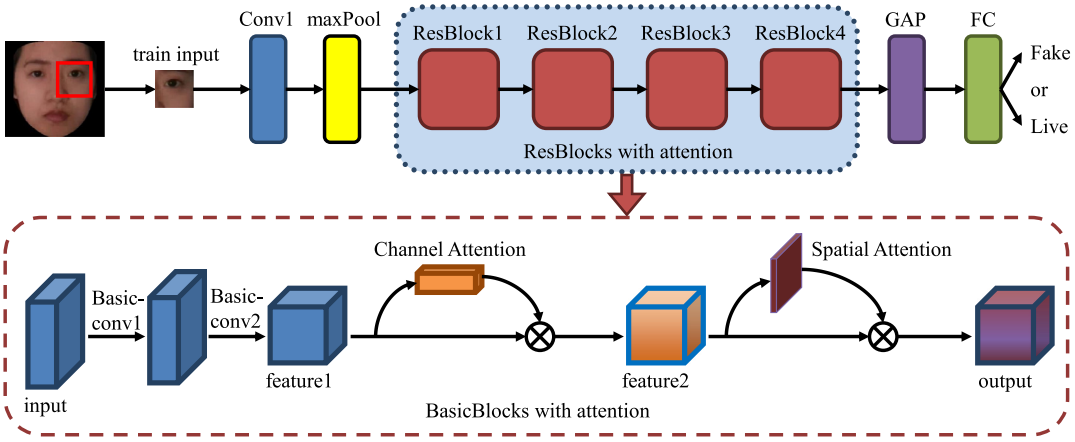
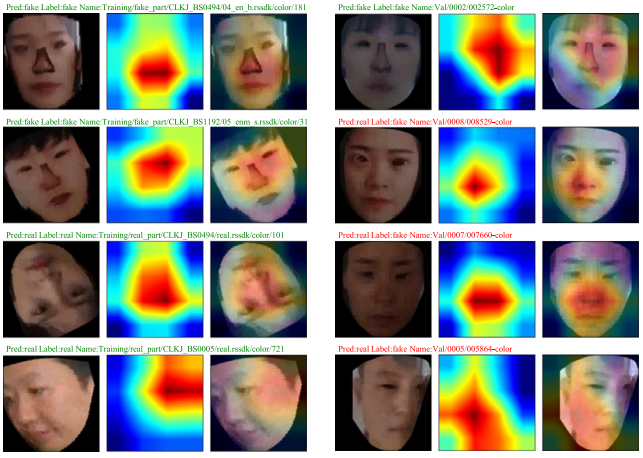


Fig. 2. The proposed network architecture.



(a) Samples from the training set (b) Samples from the validation set

Fig. 3. Some heatmaps from ResNet18 trained on CASIA-SURF using full image. The discriminative cues are obvious in the nose region in (a) training set, making the model focus more on the nose region. However, some samples (the second and third row) in the (b) validation set have no discriminative cue in the nose region, leading to incorrect predictions by the overfitting model. Furthermore, the model is not robust to face poses (the fourth row).

frames extracted from each video are highly similar, which may cause deep learning methods to easily overfit, as shown in Fig. 1. Second, a deep neural network may extract arbitrary discriminative features for known attacks during the training phase since the entire face contains more discriminative information. However, the network may learn some bias in training samples, so it cannot perform well for unknown attacks. For example, the network may only focus on the nose region when it sees many training photos with nose cuts, while it may not perform well with fake test faces without nose cuts, as shown in Fig. 3. Furthermore, we notice that the focus of some samples (the fourth row of Fig.3) are not in nose region. This is mainly because their poses are quite different from most samples in the training set. If we train the model with patches in a more challenging learning task, the model can extract more discriminative features from many regions, not only the nose area.

To increase the diversity of training data, we randomly cut out a patch of size $S_t \times S_t$ from every face image in training stage. The starting (top left) point of the patch is (x_t, y_t). Specifically, x_t and y_t are repectively randomly sampled from $(0, w - S_t - 1)$ and $(0, h - S_t - 1)$, where w and h are the width and height of the full face image.

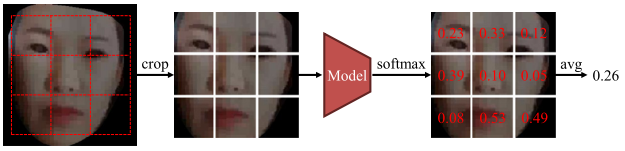


Fig. 4. Schematic diagram of the inference.

In past studies, the input size of the testing sample was commonly the same as that of the training sample. However, the larger the patch is, the more discriminative information it contains. Hence, to avoid the overfitting problem, we train a model with a smaller patch size. To fully use the information of the test samples, we use a large size during the inference stage. This simple scheme has not been studied before, and we empirically measure its effectiveness in face anti-spoofing.

During inference, for each face image, we cut out 9 patches with a size of $S_i \times S_i$. The procedure of obtaining starting points is shown in Algorithm 1. After prediction and Softmax, we calculate

Algorithm 1 Patch starting points in inference.

```
x_c = (w − S_i)//2, y_c = (h − S_i)//2;
starts = [];
for x in [x_c, x_c − S_i, x_c + S_i] do
    x = clip(x, 0, w − S_i − 1);
    for y in [y_c, y_c − S_i, y_c + S_i] do
        y = clip(y, 0, h − S_i − 1);
        starts.append((x, y));
    end for
end for
```

the average of the classification probabilities of these 9 patches. The average value is the prediction score of the image, as shown in Fig. 4.

3.2. Attention-based feature extraction

As shown in Fig. 3, the importance of discriminative features in different regions is not the same. Hence, the spatial attention mechanism should be useful for face liveness detection. As each channel of a feature map can be considered a feature detector [19], channel attention can help to distinguish real faces from fake ones. Thus, we propose combining channel and spatial attention in patch classification.

The channel and spatial attention module follow from CBAM [16]. The computation process of each attention module is shown

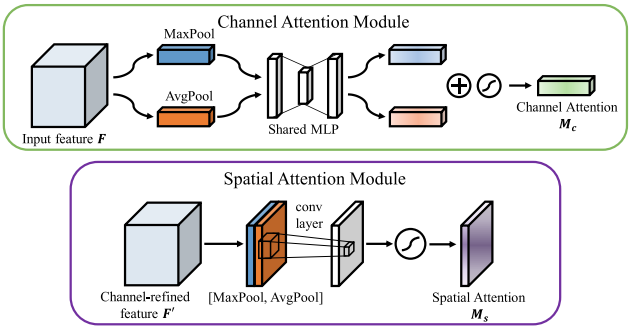


Fig. 5. Diagram of each attention module [16].

Table 1
Summary of four face anti-spoofing databases.

Database	Year	Subjects	Videos
CASIA-SURF [4]	2020	1000	21000
OULU-NPU [20]	2017	55	4950
CASIA-FASD [21]	2012	50	600
REPLAY-ATTACK [22]	2012	50	1200

in Fig. 5. They are computed as:

$$M_C(F) = \sigma\left(MLP(AvgPool(F)) + MLP(MaxPool(F))\right) \tag{1}$$

$$M_S(F') = \sigma\left(f^{7\times7}\left(\left[AvgPool(F'); MaxPool(F')\right]\right)\right) \tag{2}$$

where σ denotes the sigmoid function and MLP is a multilayer perceptron with one hidden layer. The hidden activation size is set to $\mathbb{R}^{(C/r\times1\times1)}$, and r is the reduction ratio. In this paper, r is set to 16. $f^{7\times7}$ represents a convolutional layer, and the size of the convolution kernel is 7×7.

4. Experimental results

4.1. Databases

We adopt four popular and challenging databases for the face anti-spoofing task: the CASIA-SURF database [4], OULU-NPU database [20], CASIA-FASD database [21] and REPLAY-ATTACK database [22]. The details of the four databases are summarized in Table 1.

The CASIA-SURF database, which has 6 types of photo-attacks, is mainly used to study photoprint attacks with different styles. We only utilize RGB images in this study. The training set has 29,266 images, the validation set has 9608 images, and the testing set has 57,710 images. The OULU-NPU database uses 6 types of mobile phones to record videos in 3 different backgrounds. Each mobile phone records 5 videos (2 photo attacks, 2 replay attacks and 1 live face). The purpose of this database is to evaluate the generalization of PA detection methods in realistic mobile authentication scenarios. The Videos in the CASIA-FASD database are recorded with 3 different-resolution cameras (640×480, 480×640, 1920×1080). The CASIA-FASD database and REPLAY-ATTACK database have the same two types of attacks, namely photoprint attacks and video-replay attacks. Since the last three databases only provide videos, we first randomly select 10 frames from each video and then use the MTCNN algorithm [23] to detect faces in every frame. Finally, we crop the face region.

4.2. Experimental setting and evaluation protocol

All experiments are conducted with a NVIDIA TITAN X GPU. The deep learning framework is PyTorch. We adopt cross entropy as the loss function. The SGD optimizer with weight decay (0.0001) and momentum (0.9) is used to optimize our proposed network. The initial learning rate is 0.07. The learning scheduler is CosineAnnealingLRwithRestart [24]. The scheduler restarts every 50 epochs. We train 500 epochs on all the data-bases in this paper. We set the training batch size to 256. We first resize all the images to 256×256. In general, data augmentation methods include rotation by an angle in $(-35^\circ, 35^\circ)$ and flipping.

To compare the experimental results fairly, official test protocols and metrics are employed for the CASIA-SURF and OULU-NPU databases. The metric is the average classification error rate (ACER), which is the average of the attack presentation classification error rate (APCER) and the bona fide presentation classification error rate (BPCER). According to ISO/IEC 30107-3 [25], the formulas of the metrics are shown as follows:

$$APCER_{PAI} = \frac{1}{N_{PAI}}\sum_{i=1}^{N_{PAI}}(1 - Res_i) \tag{3}$$

$$APCER = \max_{PAI=1\cdots S}(APCER_{PAI}) \tag{4}$$

$$BPCER = \frac{1}{N_{BF}}\sum_{i=1}^{N_{BF}} Res_i \tag{5}$$

where N_{PAI} is the number of attack presentations for the given PAI species (print or video in OULU-NPU); N_{BF} is the number of bona fide presentations. $Res_i = 1$ if the i^{th} presentation is classified as an attack presentation, and $Res_i = 0$ if it is classified as a bona fide presentation. APCER is from the most successful PAI species.

We adopt the half total error rate (HTER) as the metric in the cross-testing between the CASIA-FASD and REPLAY-ATTACK databases. HTER is the average the false acceptance rate (FAR) and false rejection rate (FRR). We calculate the equal error rate (EER) for the validation set and then use the threshold determined by the EER to compute the FAR and FRR for the testing set.

Table 2
Comparison with state-of-the-art methods on the CASIA-SURF Dataset*.

Method	Test ACER(%)
FaceBagNet [14]	3.2
Wang et al. [26]	3.9
Huang et al. [27]	1.6‡
Ours	**1.6**

* Results are copied from original publications.
‡ ACER value on the validation set.

4.3. Comparison with state-of-the-art methods

Results on the CASIA-SURF database. The comparison with state-of-the-art methods is shown in Table 2. Note that our method is compared with methods that only use RGB images. Our method can achieve the best result. To the best of our knowledge, our method has the lowest error for the CASIA-SURF database for this test protocol. The test ACER drops by approximately 50%.

Results on the OULU-NPU database. The results are shown in Table 3. In each protocol, the results of the methods without auxiliary information are above the middle line. Compared with results without auxiliary information, our method does not obtain the best result on the first test protocol, but achieves the lowest error on the other three test protocols, which are more difficult and challenging. Therefore, we conclude that our method has a better generalization ability.

Fusion results. Furthermore, as mentioned above, there are some methods that use auxiliary or outside information and ob-

Table 3
Comparison with state-of-the-art methods on the OULU-NPU database. Results are copied from original publications.

Pro	AInf	Method	APCER (%)	BPCER (%)	ACER (%)
1	-	STASN [8]	1.2	2.5	1.9
		DeepPixBis [28]	0.8	0.0	**0.4**
		Ours	1.7	2.5	2.1
	√	Auxiliary [1]	1.6	1.6	1.6
		FAS-SGTD [7]	2.0	0.0	1.0
		CDCN+ [6]	0.4	0.0	0.2
2	-	STASN [8]	4.2	0.3	2.2
		DeepPixBis [28]	11.4	0.6	6.0
		Ours	1.1	2.5	**1.8**
	√	Auxiliary [1]	2.7	2.7	2.7
		FAS-SGTD [7]	2.5	1.3	1.9
		CDCN+ [6]	1.8	0.8	1.3
3	-	STASN [8]	4.7±3.9	0.9±1.2	2.8±1.6
		DeepPixBis [28]	11.7±19.6	10.6±14.1	11.1±9.4
		Ours	1.7±1.1	3.1±4.0	**2.4±2.0**
	√	Auxiliary [1]	2.7±1.3	3.1±1.7	2.9±1.5
		FAS-SGTD [7]	3.2±2.0	2.2±1.4	2.7±0.6
		CDCN+ [6]	1.7±1.5	2.0±1.2	1.8±0.7
4	-	STASN [8]	6.7±10.6	8.3±8.4	7.5±4.7
		DeepPixBis [28]	36.7±29.7	13.3±14.1	25.0±12.7
		Ours	5.0±7.1	6.7±6.9	**5.8±4.9**
	√	Auxiliary [1]	9.3±5.6	10.4±6.0	9.5±6.0
		FAS-SGTD [7]	6.7±7.5	3.3±4.1	5.0±2.2
		CDCN+ [6]	4.2±3.4	5.8±4.9	5.0±2.9

Table 4
Fusion with CDCN++ on OULU-NPU database. We reimplement CDCN++ model.

Protocol	Method	APCER (%)	BPCER (%)	ACER (%)
1	CNCN+	0.4	0.8	0.6
	Ours	1.7	2.5	2.1
	Fusion	0.0	0.8	**0.4**
2	CNCN+	3.9	0.3	2.1
	Ours	1.1	2.5	1.8
	Fusion	0.6	1.9	**1.3**
3	CNCN+	1.0±1.2	5.6±6.1	3.3±2.8
	Ours	1.7±1.1	3.1±4.0	2.4±2.0
	Fusion	0.7±0.8	1.4±1.8	**1.0±0.8**
4	CNCN+	10.8±10.6	4.2±1.9	7.5±5.6
	Ours	5.0±7.1	6.7±6.9	5.8±4.9
	Fusion	6.7±8.0	2.5±2.5	**4.6±4.9**

tain very good performance. For example, CDCN++ [6] is a state-of-the-art method for the OULU-NPU database, which uses auxiliary depth trained from outside information. Since we do not use depth estimation, we can combine our method with CDCN++ to further improve performance. Experimental results are shown in Table 4. Some gaps are seen in our implementation. There are two possible reasons. First, we do not use NAS to search the network for a fair comparison with our method; second, we calculate the APCER according to ISO/IEC 30107-3, which means that the largest $APCER_{PAI}$ is the maximum value among all PAIs rather than their average. As shown in Table 4, the fusion results are better than the results of both of the other methods, indicating that these two methods can complement each other. In our future work, we will study how to effectively fuse depth information for more robust face liveness detection.

Intertesting. Similarly to previous methods [29,30], we conduct intertesting experiments. As shown in Table 5, the proposed method achieves the lowest HTER when the training database is CASIA-FASD and the testing database is REPLAY-ATTACK. The model trained with the high-resolution dataset (CASIA-FASD) can find the difference between fake faces and real faces for the low-resolution dataset (REPLAY-ATTACK). However, the model trained with the low-resolution dataset does not perform well on the high-resolution dataset, possibly because the proposed method can make full use of discriminative information, which may not exist in low-resolution images.

Table 5
Comparison with state-of-the-art methods for intertesting. Results are copied from original publications. The metric is HTER (%).

Method	Train CASIA-FASD	Test REPLAY-ATTACK	Train REPLAY-ATTACK	Test CASIA-FASD
Boulkenafet [29]	30.3		37.7	
Wang [30]	26.1		39.2	
STASN [8]	31.5		**30.9**	
Ours	**25.6**		43.8	

Fig. 6. Experimental results for different patch size on CASIA-SURF.

4.4. Ablation study

In this section, experiments are conducted to explore the influence of patch size and attention mechanism.

4.4.1. Patch size

We train the proposed model without attention modules. The patch size in training and inference stage are both set to 64×64, 80×80, 96×96, 128×128, 160×160 and 192×192 in turn. The results are shown in Fig. 6.

Here, we first discuss the effect of different training patch sizes. As shown in Fig. 6, in general, the larger the training patch size is, the worse the result. We believe that a large image contains so much information that the network fails to explore the information fully with less training samples. Then, we explore the influence of different testing patch sizes. Contrary to the training patch size, the larger the testing patch size is, the better the overall result is. If the testing patches are larger than the training patches, the testing patches can provide more discriminative information so that the model can detect liveness information easily.

4.4.2. Attention mechanism

First, we use patches with different sizes to train the proposed model. The training patch size is set to 64×64, 80×80, and 96×96 since the results are not accurate when the models are trained with larger patches, as shown in 6. The results are shown in Fig. 7. Fig. 6 and Fig. 7 show that the attention mechanism is generally effective for patches of different sizes.

As shown in Fig. 7, the result is not accurate when the training patch size is too large, which is similar to previous results.

Fig. 7. Experimental results for different patch size on CASIA-SURF (with attention).

Table 6
Ablation results of attention mechanism on CASIA-SURF(Training patch size is 80 × 80, testing patch size is 160 × 160).

Model	Channel Attention	Spatial Attention	Val ACER(%)	Test ACER(%)
Model-1	-	-	2.0	2.9
Model-2	√	-	1.3	1.8
Model-3	-	√	1.6	2.5
Ours	√	√	**0.8**	**1.6**

However, in contrast to Fig. 7, testing patches that are too large or too small will drop the ACER. The reason for this is that the proposed attention modules are more sensitive to patch size. The testing patches are so large that there is a large gap between the testing set and training set. We choose the best configuration corresponding to the minimum ACER on the validation set, although a better result (ACER = 1.3%) could be obtained. Hence, we can conclude that 80 × 80 is a good choice for training patches, while 160 × 160 is a suitable choice for testing patches.

We conduct two other experiments to study the influence of channel attention and spatial attention. We train models with different attention modules for the optimal configuration. Table 6 shows that both channel attention and spatial attention contribute to face liveness detection. Specifically, the test ACER is greatly improved when the model combines channel attention and spatial attention simultaneously.

5. Conclusion

In this study, we propose a simple yet effective face liveness detection method based on patch sampling and the attention scheme. We achieve better results for unseen attacks for both RGB and IR images. Our method obtains the lowest ACER on the CASIA-SURF database and performs well on the OULU-NPU database. The proposed method not only can reduce the cost of data collection in practical applications by making better use of images but also is more suitable for some scenarios with IR images only.

The experimental results also prove two characteristics for face liveness detection. First, discriminative cues exist in an entire image, so a larger patch contains more cues. Second, the importance of spoofing cues is not identical for every part of an image. These findings can be helpful for future face anti-spoofing research. In the future, we will study how to combine our method with other information, such as depth, for more robust face liveness detection.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment

This work was partially supported by the Natural Sci-ence Foundation of China (NSFC) (No. 61772296) and the Shenzhen Fundamental Research Fund (No. JCYJ20170412170438636).

References

[1] Y. Liu, A. Jourabloo, X. Liu, Learning deep models for face anti-spoofing: Binary or auxiliary supervision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 389–398.
[2] P.P.K. Chan, W. Liu, D. Chen, D.S. Yeung, F. Zhang, X. Wang, C.-C. Hsu, Face liveness detection using a flash against 2d spoofing attack, IEEE Trans. Inf. Forensics Secur. 13 (2) (2017) 521–534.
[3] G. Heusch, A. George, D. Geissbühler, Z. Mostaani, S. Marcel, Deep models and shortwave infrared information to detect face presentation attacks, IEEE Trans. Biometric. Behav. Ident. Sci. 2 (4) (2020) 399–409.
[4] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H.J. Escalante, S.Z. Li, Casia-surf: a large-scale multi-modal benchmark for face anti-spoofing, IEEE Trans. Biometric. Behav. Ident. Sci. 2 (2) (2020) 182–193.
[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[6] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, G. Zhao, Searching central difference convolutional networks for face anti-spoofing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5295–5305.
[7] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, Z. Lei, Deep spatial gradient and temporal depth learning for face anti-spoofing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5042–5051.
[8] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, W. Liu, Face anti-spoofing: model matters, so does data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3507–3516.
[9] J. Yang, Z. Lei, S.Z. Li, Learn convolutional neural network for face anti-spoofing, arXiv preprint arXiv:1408.5601 (2014).
[10] L. Li, Z. Xia, X. Jiang, F. Roli, X. Feng, Compactnet: learning a compact space for face presentation attack detection, Neurocomputing 409 (2020) 191–207.
[11] B. Yu, J. Lu, X. Li, J. Zhou, Salience-aware face presentation attack detection via deep reinforcement learning, IEEE Trans. Inf. Forensics Secur. 17 (2021) 413–427.
[12] C. Wang, B. Yu, J. Zhou, A learnable gradient operator for face presentation attack detection, Pattern Recognit. 135 (2023) 109146.
[13] Y. Atoum, Y. Liu, A. Jourabloo, X. Liu, Face anti-spoofing using patch and depth-based CNNs, in: 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2017, pp. 319–328.
[14] T. Shen, Y. Huang, Z. Tong, Facebagnet: bag-of-local-features model for multi–modal face anti-spoofing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019. 0–0
[15] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[16] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
[17] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
[18] H. Chen, G. Hu, Z. Lei, Y. Chen, N.M. Robertson, S.Z. Li, Attention-based two-stream convolutional networks for face spoofing detection, IEEE Trans. Inf. Forensics Secur. 15 (2019) 578–593.
[19] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.
[20] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, A. Hadid, 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), IEEE, 2017, pp. 612–618.
[21] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, S.Z. Li, A face antispoofing database with diverse attacks, in: 2012 5th IAPR international conference on Biometrics (ICB), IEEE, 2012, pp. 26–31.

[22] I. Chingovska, A. Anjos, S. Marcel, 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), IEEE, 2012, pp. 1–7.
[23] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.
[24] I. Loshchilov, F. Hutter, Sgdr: stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983 (2016).
[25] I. Standard, Information technology–biometric presentation attack detection—part 3: testing and reporting, Int. Org. Standard.: Geneva, Switzerland (2017).
[26] G. Wang, C. Lan, H. Han, S. Shan, X. Chen, Multi-modal face presentation attack detection via spatial and channel attentions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019. 0–0
[27] Y. Huang, W. Zhang, J. Wang, Deep frequent spatial temporal learning for face anti-spoofing, arXiv preprint arXiv:2002.03723 (2020).
[28] A. George, S. Marcel, Deep pixel-wise binary supervision for face presentation attack detection, in: 2019 International Conference on Biometrics (ICB), IEEE, 2019, pp. 1–8.
[29] Z. Boulkenafet, J. Komulainen, A. Hadid, Face spoofing detection using colour texture analysis, IEEE Trans. Inf. Forensics Secur. 11 (8) (2016) 1818–1830.
[30] G. Wang, H. Han, S. Shan, X. Chen, Cross-domain face presentation attack detection via multi-domain disentangled representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6678–6687.