Attended-Auxiliary Supervision Representation for Face Anti-spoofing

Son Minh Nguyen, Linh Duy Tran, and Masayuki Arai

Graduate School of Science and Engineering
Teikyo University, Utsunomiya, Tochigi 320-8551, Japan
{nguyenminhson1110,duylinh161287}@gmail.com
arai@ics.teikyo-u.ac.jp

Abstract. Recent face anti-spoofing methods have achieved impressive performance in recognizing the subtle discrepancies between live and spoof faces. However, due to directly holistic extraction and the resulting ineffective clues used for the models' perception, the previous methods are still subject to setbacks of not being generalizable to the diversity of presentation attacks. In this paper, we present an attended-auxiliary supervision approach for radical exploitation, which automatically concentrates on the most important regions of the input, that is, those that make significant contributions towards distinguishing the spoof cases from live faces. Through a multi-task learning approach, the proposed network is able to locate the most relevant/attended/highly selective regions more accurately than previous methods, leading to notable improvements in performance. We also suggest that introducing spatial attention mechanisms can greatly enhance our model's perception of the important information, partly intensifying the resilience of our model against diverse types of face anti-spoofing attacks. We carried out extensive experiments on publicly available face anti-spoofing datasets, showing that our approach and hypothesis converge to some extent and demonstrating state-of-the-art performance.

Keywords: Face anti-spoofing · Multi-task learning · Feature Extraction · Self-attention

1 Introduction

Face recognition expertise is currently one of the most prominent subjects attracting a lot of research attention and large collaborations due to its potential for convenient effectiveness in biometric-based security applications. Due to organizations' endeavors, current state-of-the-art modalities have been achieving tremendous accuracy beyond that of human ability. However, the related applications have been generally turned into objects appealing to illegal access of the so-called facial spoofing attacks, due to their incompetence in identifying imposters. As a result, these systems are really vulnerable to such impersonating attacks, *e.g.*, replayed videos or photographs, which are perfectly recorded by sophisticated devices. Face anti-spoofing (FAS) techniques are thus being urgently developed for assurances of face recognition system operation.

There have been many attempts at enhancing effective solutions for preventing facial impostures in recent years. At first, the previous methods[1] used image distortion analysis to capture optical textural differences between presentation attacks (PAs) and live faces. Other methods[2, 3] focused on local binary patterns to identify distinguishable features for further improvements, while some techniques[4, 5] attained some success. However, the nature of the handcrafted features that these methods exploited are not generalized traits, the resulting models hence showed limitations on practical environments, and are prone to dramatically degraded performance in peculiar media.

Most recent studies have obtained significant achievements based on convolutional neural networks (CNNs) along with supplemental dependencies. Luo *et al.*[6] adopted a CNN long short-term memory (LSTM) architecture to learn spatiotemporal patterns in sequential frames for distinguishing genuine cases from spoof activities, whereas LSTM was previously used almost for sequence-related tasks. Atoum *et al.*[7] used an auxiliary supervision associated with a patch-scored modality, and then fed the result into a support vector machine-based classifier for PA detection. Compared with previous work, an algorithm[8] that relied on the extra remote photoplethysmography (rPPG) pattern in addition to the auxiliary supervision[7] made an advance to some extent. Recently, Yang *et al.* attempted to provide enhanced spatiotemporal patterns with the help of a pretrained CNN-LSTM model augmented by an attention mechanism and their own synthetic data[9].

However, it seems that these techniques might not radically capture essential characteristics for genuine and counterfeit case analysis. For instance, the models in [8, 10] put a lot of effort into the full scope of auxiliary supervision, namely facial depth maps and rPPG signals, without any highly selective approaches. In other words, such learning holistic representations of input images might lead to the excessive exploitation and thus guide the models' focus onto redundant/irrelevant information, interfering with the models' perception. Similarly to auxiliary cases, the binary supervision-based networks[9, 2] did not enable the networks to capture sufficient distinct features needed to discern spoof faces. On the contrary, that induces the networks to have a great tendency towards inconsistent features, raising the high possibility of being overfitted and poor performance in most test cases.

To tackle these problems, we developed the idea of radically capturing the most relevant locations, but still making the best use of auxiliary supervision because of its undeniable contribution in prior methods. Hence, we propose an attended-auxiliary supervision (AAS) architecture in which the attended-auxiliary information is comprehensively used in both the highly selective region proposal and inference stages. We represent this AAS/patch-driven facial depth map supervision for live and spoof images. To invigorate the adoption of the AAS, the patches used to estimate these partial depth maps are intended to be sampled from highly selective regions of input images based on their contribution to model's decision. In order to make these regions identifiable, we built a two-module model. The first module uses a pretrained network integrated with

squeeze-and-excitation network (SENet) blocks[11], which are useful for describing the nature of channels in feature blocks, to extract the more distinguishable spatial properties. Then, the employed LSTM network converts the resulting embeddings into temporal information for sequential frame-driven classification. The second module is processed within reciprocal stages. The *Region Proposal Stage* (RPS) is in charge of best aligning initial advice given by the first module with the model's situation, and then proposing the ensuing highly selective regions to the *Attended-Auxiliary Execution Stage* (AAES). In the AAES, the highly selective regions are fully exploited through the corresponding patch-driven depth map regression, which stimulates the network to explore sufficient generalizable representations of both spoof and live images.

Our main contributions are summarized as follows:

• We present a practical solution to radically exploit the most relevant regions of input images based on the patch-driven depth map supervision in both channel and region attention scenarios for 2D PAs. Such AAS adoption has the purpose of partially influencing the network to learn adequate salient information of input images, thereby alleviating the performance degradation caused by irrelevant information.

• To this end, we designed the AAS framework towards a multi-task learning fashion to further advance the model's perception of highly selective regions.

• We demonstrate the reliability of our model and arguments by evaluating our framework on publicly available FAS datasets. Our experimental results show the FAS community another potential concept to resolve ongoing issues in FAS.

2 Prior Work

Traditional approaches. Distinguishable features are the most fundamental keys to recognize spoof cases. Many prior patterns have been, to a degree, flourishingly built upon such ideas since several years ago. However, these patterns are still essentially handcrafted feature representations, such as local binary patterns[12–14], HOG[15, 16], and SIFT[17], which produce modest outcomes via conventional classifiers, namely support vector machines and linear discriminant analysis. The handcrafted-feature-based methods thus showed limitations regarding generalizable representation in PA detection. To cope with these difficulties, researchers have approached these problems in another way that maps input images onto other domains, namely HSV, YCbCr[18, 19], temporal domains[20, 21], and Fourier spectra[22].

In lieu of using a mere single frame, researchers have attempted to exploit the traits of facial motions in several consecutive frames, such as determining "facial liveness" with eye-blinking detection[5, 23] and mouth and lip motions[4].

Deep learning-based approaches. During the deep learning era, the FAS modalities have thrived dramatically on prior difficulties. Some of the modern CNNs[24, 25] have been used as feature extractors to discriminate live and spoof images. In [26], Jourabloo *et al.* tackled anti-spoofing in a particular way that treats PAs as decomposition issues, inversely decomposes a spoof face into spoof

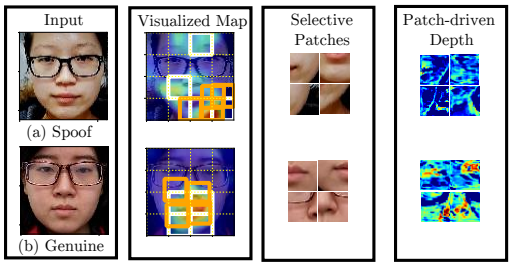


Fig. 1. Translation offsets of the orange regions, given the white initial locations of the AM, are thoroughly regularized based not only on binary supervision, but also on the attended auxiliary supervision in a multi-task learning manner.

noise and a live face, and then uses the spoof noise for classification. At that point, Liu *et al.*[8] attempted to holistically estimate depth maps and rPPG signals from live and spoof faces through a recurrent neural network-CNN model based on ground truths predefined with dense face alignment[27]. Inspired by auxiliary supervision[8], Kim *et al.*[10] introduced an additional auxiliary supervision by extending the reflection map into PAs, which intensified the resilience of their model on PA types. Additionally, some works have proposed spatiotemporal features. In particular, Yang *et al.*[9] adopted spatiotemporal features extracted from discriminative regions for PA detection using a CNN-LSTM structure, which was already pretrained on their own synthetic data.

Motivated by [9, 28], we partly used a CNN-LSTM model, but with advanced subnets in order to precisely interpret temporal features, as an *Advisor Module* (AM), supporting the rest of network in determining where to look first. Furthermore, we imposed extra essential constraints towards a multitask-learning scheme on the proposed model, so that the representation regularization of refined subtle offsets, mentioned in Subsec.3.2, should be more vigorous than its predecessor[9]. As shown in Fig. 1, the AM and the rest of the model share similar considerations in genuine cases, but spoof activities cause a large displacement.

3 Proposed method

The main objective of our approach is to direct the network to autonomously centralize regions of input images possessing the model's decision, rather than digesting completely unprocessed full scope of inputs, which avoids redundant information causing noise and detrimental effects on performance. In contrast with previous auxiliary supervision-based methods, the most striking feature of our approach, as aforementioned, is that we meticulously exploit the patch-driven facial depth map supervision for both live and spoof faces on the basis of the patches. As shown in Fig. 2, the proposed architecture comprises two main interdependent attending modules. The AM bases the fabric of a CNN-LSTM incorporated with SENets on performing the classification task itself and

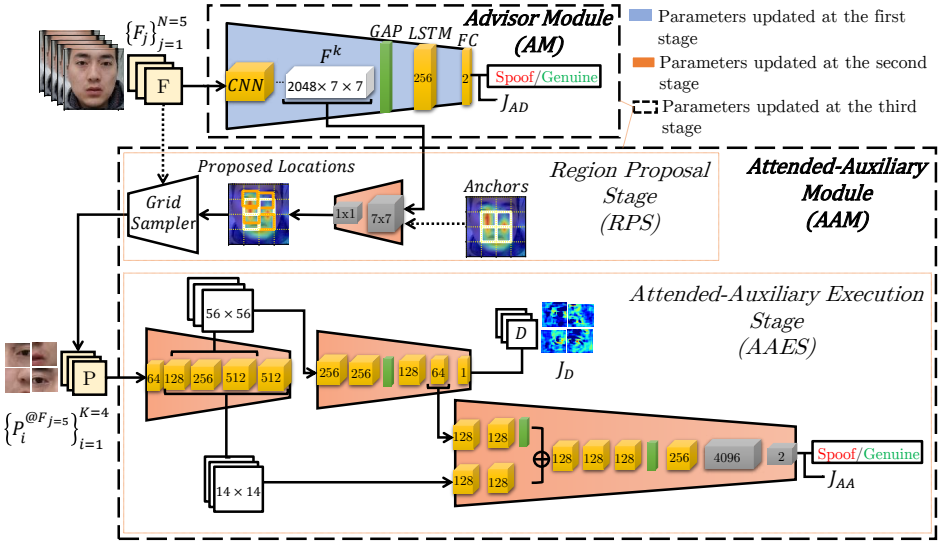


Fig. 2. Proposed framework constituted by interdependent attending modules. The AM plays a main role in providing the AAM with its experience in discerning spoof from genuine images. The AAM, in turn, draws attention to the received advice and sequentially processes images with reciprocal stages. For the architecture interpretation, the white, yellow, and green blocks represent feature blocks, sequential filters followed by a BatchNorm, and a rectified linear unit (RELU) activation function, respectively, each with a given size of 3×3 and pooling layers. Apart from that, the gray blocks are depth-wise convolutions with size 7×7 and point-wise convolutions with size 1×1. At the AAES, the patches proposed by the RPS are used to extract and to synthesize coarse to fine features with the help of the patch-driven depth map supervision D for PA detection. The figured is best viewed online in color.

conveying its initial advice to the *Attended-Auxiliary Module* (AAM) for spotting which part of the input is plausible to look into (hereafter referred to as initial regions). The AAM with reciprocal stages thereafter build its attention to the most relevant regions upon the initial regions.

3.1 Auxiliary Supervision

Depth map. The fundamental disparity between basic PAs, such as print and replay attacks, and live faces is basically manifested in the depth, the distance from points of an object to a capturing device. In practice, such spoof cases, which were previously recorded for several times by diverse devices, are often presented in even and flat facial surfaces, e.g electronic displays and print materials. However, live faces, which are captured directly in front of the camera, are not completely analogous to PA properties, with differences in the color distribution and illumination reconstructions, but more importantly, the depth due to the irregular geometry of live faces. According to this point, we consider

the basic PAs and genuine faces as flat images and depth maps, respectively. However, due to a shortage of related ground truth labels for the depth maps in spoof datasets, PRNET[29] was used to produce the corresponding ground truth labels by estimating the depth maps for genuine images. Specifically, the ground truth depth maps D estimated at facial regions are defined as follows:

$$D(I|y) = \begin{cases} 0, & \text{if y is spoof,} \\ \frac{1}{|D|}d(I), & \text{if y is genuine,} \end{cases} \tag{1}$$

where I is a given input image with label y of either 0 or 1. The depth map distance $d(I)$ is normalized from 0 to 1 by the intensity maximum $|D|$, where values of 0 and 1 represent the farthest and closest points to the camera, respectively. Based on the Eq.1, the spoof cases' supervision is also regarded as depth maps, but with all distance values of zero.

3.2 Network Architecture

The previous model[9] used a simple CNN-LSTM architecture for initial advice providing, and weight-shared CNNs worked under the binary classification for attended region mining, resulting in the exploitation of weak features and the inefficient utilization. To resolve this weakness, our architecture is further advanced by an effective combination of two interdependent attending modules with the use of the patch-driven depth map supervision, which enables the whole network to exploit enough generalized clues that are supposed to strongly boost the model perception. Apart from that, we are still able to achieve robust performance without the use of bipartite auxiliary supervision, as in [10], which requires very time-consuming preparation. Typically, we process the input as video classification by feeding N sequential frames directly into the AM to teach it spatiotemporal information, largely helping the module to accurately converge on key regions.

Advisor Module (AM). Let $\{V_i, y_i\}_{i=1}^{k}$ describe the set of training data, where V_i is the i-th training video among k training videos and y_i is the corresponding label where 0 or 1 represents the attribute of V_i, namely genuine or spoof, respectively. As aforementioned, the CNN presented in Fig. 2 takes as input N sequential frames from $\{V_i\} = \{F_i\}_{j=1}^{m}$, where F_j^i denotes the j-th frame among m frames extracted from the i-th video. We are aware of the main role of the AM, which must be powerful to correctly provide the AAM with the initial regions. Accordingly, we use a pretrained model, a 101-layer residual network (Resnet) pretrained on the ImageNet dataset for the CNN, into which minor SENet subnets were integrated for the channel attention.

SENets. With the SENet functionality, the informative channels of each feature map are remarkably intensified, whereas the less useful ones are greatly suppressed, as opposed to being equally treated by conventional models. Each feature map from each residual block particularly undergoes the squeeze-and-excitation process for channel analysis and enhancement. As illustrated in Fig. 3, a feature map X is firstly shrunk to a C-sized vector by the global average

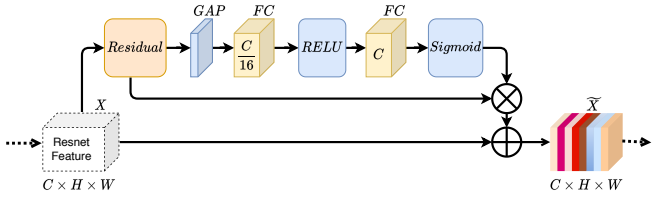


Fig. 3. Fundamental structure of a squeeze-and-excitation block, attached to each of the 101 residual blocks as a channel-wise enhancement submodule.

pooling (GAP) layer at the squeeze stage for the purpose of channel analysis. The shrunk feature is then transferred to the excitation stage, where the channel enhancement is represented by two fully connected (FC) layers, before sending it back to X, with channel-wise weighting thereafter to obtain the augmented feature map $\tilde{X}$. Taking advantage of the subnets increases the confidence in avoiding overfitting caused by unnecessary channels. In addition, the aforesaid LSTM introduced behind the GAP is accountable for interpreting spatial information in the CNN as temporal patterns, which aims to make discrete spatial information more mutually dependent and have a closer relationship between sequential frames.

Attended-Auxiliary Module (AAM). The workings of this module consist in the region attention mechanism with reciprocal functions, both of which are observed in the RPS and AAES.

Region Proposal Stage (RPS). Motivated by [28], which used a spatial transformer network to be sensitive to spatial changes of input, the AAM is thus offered the differentiable attention to any spatial transformation, and it learns to align the initial regions with locations of the most relevant regions via learnable parameters. Because of the difficulties in directly determining these locations, we thus let the model be automated to adjust the learnable parameters as translation offsets, which are subsequently associated with the anchors in order to locate the most relevant regions. To enable the model to optimize the offset parameters during the training period, these were set up as the affine transformation M, whose resulting patches are then manipulated and collected by a grid sampler in the RPS as follows:

$$M = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix}, \tag{2}$$

where s_x, s_y must be constant scaling ratios of an image in the x and y-axes, respectively, in order to achieve K constant-sized patches. In contrast, t_x and t_y along the corresponding axes are adaptive translation offsets for every epoch. As indicated in Fig. 2, these offsets are achieved by attaching two extra filters, namely a 7×7 depth-wise layer followed by a 1×1 point-wise layer, to the $2,048\times 7\times 7$ feature map F_k behind the GAP layer. Deriving the $K\times 2$ output of $[t_x, t_y]$ from these filters is, however, far more efficient in reducing the complexity of

the number of parameters and operations than that from FC layers[28], thereby easing degradation in performance more greatly.

The anchors, which are locations of the initial regions given by the AM, are recommended to the RPS for subsequent proposed patches through the effect of Gradcam[30]. In effect, we used Gradcam to enhance the analytical capability of the AM to interpret its concentration on specific regions of the input image, with regard to the considered label, for the provision of the anchors. According to Gradcam, F^k with the predicted label-related score y_c is extracted for the gradient α_k^c in terms of F^k.

$$\alpha_k^c = \frac{1}{z} \sum_i \sum_j \frac{\delta y_c}{\delta F_{ij}^k}, \tag{3}$$

where z is a normalizing constant. After that, a 7×7 Gradcam map S_k is obtained by applying a RELU layer to the weighted feature map,

$$S_k = RELU(\sum_k \alpha_c^k F^k). \tag{4}$$

The 7×7 Gradcam map shows the important regions, with respect to y_c, which are of interest to the AM's performance. To enhance visualization, we turn the Gradcam map into a score map whose size is upsampled to be similar to that of the input, mentioned in Subsec. 4.2, using bilinear interpolation. An averaged pooling layer overlaid on the score map results in a 4×4 grid, which is used to identify the K highest scored regions in the grid cells. Finally, the anchors extracted from the K regions are transferred to the RPS for the above-mentioned alignment.

In fact, although the RPS is implemented with all the above initial inspirations[9, 28, 30], the offset-regularizing effects of the surroundings, namely the AM and the AAES, on the RPS are far more intensive and extensive than that of the prior approaches in channel interaction and supervision, respectively.

Attended-Auxiliary Execution Stage (AAES). The proposed patches from the RPS are fetched by the AAES coupled with bypass connections, which aim to conflate both low-level and high-level features collectively. Given the obvious benefits of the RPS for the AAES, the partial depth map regression with the corresponding attended patches of a face is much less demanding than the straightforward generation of a full depth map from the face. In addition, based on closely related constraints of regression and classification in a multi-task learning manner, the AAES is fully able to extract enough related features and to ignore the less useful context, helping the RPS quickly converge on optimal offsets. To be more effective, the 14×14 low-level features at the first layers are channel-wise associated with the high-level features to preserve the depth of information because some distinct information existing at the low-level stage might be significantly attenuated or virtually vanish after further processing. The inferred scores behind the softmax layer of the AM that need to receive a greater emphasis by an exponential transformation are eventually fused with probabilistic results of the AAM for the final decision.

Table 1. Ablation study using Protocol 1 of the OULU-NPU dataset.

Model	Channel Attention	Region Attention	AAS	APCER (%)	BPCER (%)	ACER (%)
Model A		✓	✓	2.0	0.7	1.4
Model B	✓			1.6	0.2	0.9
Model C	✓	✓		1.6	**0.0**	0.8
Model D	✓	✓	✓	**1.1**	0.1	**0.6**

Objective Functions. To enable the AAS, we employed a depth loss to supervise regression of the partial depth maps from the corresponding patches.

$$J_D = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{K} || \, CNN_{AAM}(P_i^j, \Theta_D) - D_i^j \, ||_1^2, \tag{5}$$

where J_D depicts the regression loss between the partial depth maps estimated from the corresponding patches P and the ground truth label D, while Θ_D, K, and M are CNN_{AAM} parameters at the AAES, the number of patches, and batch size of training images, respectively. In addition, the losses of both the AM and the AAM trained in a multi-task learning manner need to be attached to different levels of importance.

$$J_{overall} = \lambda_1 J_D + \lambda_2 J_{AD} + \lambda_3 J_{AA}, \tag{6}$$

where $J_{overall}$, J_{AA}, and J_{AD} refer to the overall loss, the softmax cross-entropy losses designed for the AAM, and the AM. During each training stage, we accommodate the magnitudes of weights λ to the stage so as to balance the involved losses.

4 Experiments

4.1 Datasets and Metrics

Datasets. We evaluated the performance of our model by using two publicly available datasets. Specifically, the assessments were conducted on only the SiW[8] and OULU[31] datasets, which are new datasets with very high resolution and a wide variety of practical protocols covering subject, pose, illumination, medium, and attack variations. These protocols can rigorously verify quality and efficiency of a model through cross-testing in media, pose, and unknown attacks.
Evaluation Metrics. To make comparisons with prior works, we report our results with the following metrics: attack presentation classification error rate (APCER)[32]; bona fide presentation classification error rate (BPCER)[32]; and average classification error rate (ACER), which equals (APCER+BPCER)/2[32].

4.2 Implementation Details

The input fed to the network was sequential frames ($N = 5$) of 224×224 pixels. Four patches K with size of 56×56 pixels and s_x, s_y of 0.25 were selected in

Table 2. Intra-testing results on three SiW protocols.

Protocol	Method	APCER (%)	BPCER (%)	ACER (%)
1	Auxiliary[8]	3.58	–	3.51
	STASN[9]	–	–	1.00
	DepthFreq[33]	0.80	0.50	0.67
	BASN[10]	–	–	0.37
	Ours	**0.21**	**0.03**	**0.12**
2	Auxiliary[8]	0.57 ± 0.69	0.57 ± 0.69	0.57 ± 0.69
	DepthFreq[33]	$\mathbf{0.00 \pm 0.00}$	0.75 ± 0.96	0.38 ± 0.48
	STASN[9]	–	–	0.28 ± 0.05
	BASN[10]	–	–	0.12 ± 0.03
	Ours	0.03 ± 0.02	$\mathbf{0.10 \pm 0.10}$	$\mathbf{0.07 \pm 0.06}$
3	STASN[9]	–	–	12.10 ± 1.50
	Auxiliary[8]	8.31 ± 3.81	8.31 ± 3.80	8.31 ± 3.81
	DepthFreq[33]	9.50 ± 1.20	5.30 ± 2.10	7.40 ± 2.90
	BASN[10]	–	–	6.45 ± 1.80
	Ours	$\mathbf{3.35 \pm 2.62}$	$\mathbf{4.15 \pm 1.50}$	$\mathbf{3.75 \pm 2.05}$

our experiments. Our architecture was trained on a single GeForce GTX 1080 GPU and implemented in the PyTorch framework with a learning rate of $5e-5$ for the first two training stages, and $5e-6$ for the last. In addition, λ_1, λ_2, and λ_3 were set as 0.5, 10, and 10 in turn.

Multi-step Approach for Training. In this case, we used a training procedure similar to that used in [9] to divide the progression of training into three main stages. During the first stage, the AM was trained with a learning rate of $5e-5$ over 10 epochs. Likewise, the AAM was trained with the same learning rate for another 10 epochs while the AM's parameters were not updated. Both of them were then trained together over 6 epochs with the learning rate reduced by 10 times to ensure that they fit with each other.

4.3 Ablation study

We used OULU's Protocol 1 to conduct an ablation study to verify the effectiveness of separate modules from the proposed network in four configurations, as shown in Table 1:

(i) Model A: The proposed model, but with SENet integration excluded. This model was designed to explicitly recognize the power of the channel-attention.

(ii) Model B: The independent AM was trained with a softmax loss. Accordingly, the feature maps that Model B extracted were merely able to accommodate to channel interaction, but they did not respond with region attention.

(ii) Model C: The AM was connected with the AAM, but the advantages of the AAS were removed.

(iv) Model D: This is the complete proposed model in which the attention is comprehensively and remarkably boosted in both channel and region spaces by means of SENets and the attended regions provided by the AAM through the adaptive translation offsets. Therefore, the performance has far more potential than that of the other three models.

Advantage of SENets. Similarly to Model A with the mere region attention, Model D, however offered with the extra understanding of channel interactions,

achieves the superiority (ACER of 0.6% to 1.4%) over its rival, Model A. This dominance demonstrates the impact of SENets on the model's perception in which the subnets greatly strengthen the power of the AM.
Advantage of the AAS. It can be seen clearly that Model D far outweighs Model C by an overwhelming margin of 0.2 percentage points due to the main effectiveness of the AAS.
Advantage of the AAM. In our approach, we used the AAS supervised by the regression loss J_D to alleviate difficulties in accommodating offsets to the optimal locations. Therefore, Model D with the channel and region attention, obtained 0.3% ACER lower than Model B with the deficiency of the region attention, indicating the impact of the AAM on the exploitation of the most relevant regions.

4.4 Intra-testing

Intra-testing was performed on the OULU-NPU and SiW datasets, which have a variety of practical protocols that we rigorously followed. The resulting comparison is shown in Tables 2 and 3.
SiW In Table 2, the results show that our method outperforms all of the state-of-the-art methods with a remarkable advantage, demonstrating considerable ameliorations in tackling or at least mitigating variations in media. To be more specific, the proposed model brings out a promising improvement at Protocol 3, which is the most demanding protocol for generalizability verification, with a reduction of nearly 42% compared with the next best method.
OULU-NPU The numerical results in Table 3 demonstrate that our approach significantly surpasses all of the existing methods with the 1^{st} position. Notably, we achieved state-of-the-art success over the former best methods at Protocols 1 and 4 by large margins, of around 60% and 33%, respectively. In spite of a high BPCER of 5.1% at Protocol 4, which is the most challenging protocol verifying the generalizability under variations of unknown sessions and capturing devices, the proposed model still dominate the others with the lowest ACER. The results partly suggest the robustness of our approach under PA variations in terms of pose, illumination, and capture device.

4.5 Cross-testing

We evaluated the effectiveness of our model on cross-datasets, in which the network alternately is trained on OULU-NPU and assessed on SiW and vice-versa, via the ACER metric. Our improvements of deeply mining AAS representations are more confident and transparent as some of the best results are shown. As shown in Table 4, our model achieved slightly worse results compared with the best model by approximately 41% and 22% at SiW with OULU's Protocol 2 and OULU with SiW's Protocol 1, respectively. We hypothesize that the reason is that the features extracted in the frequency domain by the above competitor have more correlations between both datasets. However, the rest still outweighs state-of-the-art results to a great extent, and especially the ACER of 1.9% for

Table 3. Intra-testing results on four protocols of OULU-NPU.

Protocol	Method	APCER (%)	BPCER (%)	ACER (%)
1	GRADIENT[34]	1.3	12.5	6.9
	BASN[10]	1.5	5.8	3.6
	STASN[9]	1.2	2.5	1.9
	Auxiliary[8]	1.6	1.6	1.6
	FaceDe-S[26]	1.2	1.7	1.5
	Ours	**1.1**	**0.1**	**0.6**
2	MixedFASNet[34]	9.7	2.5	6.1
	Auxiliary[8]	2.7	2.7	2.7
	BASN[10]	**2.4**	3.1	2.7
	GRADIANT	3.1	1.9	2.5
	STASN[9]	4.2	**0.3**	2.2
	Ours	2.7	1.0	**1.9**
3	GRADIENT	2.6 ± 3.9	5.0 ± 5.3	3.8 ± 2.4
	FaceDe-S[26]	4.0 ± 1.8	3.8 ± 1.2	3.6 ± 1.6
	Auxiliary[8]	2.7 ± 1.3	3.1 ± 1.7	2.9 ± 1.5
	STASN[9]	4.7 ± 3.9	$\mathbf{0.9 \pm 1.2}$	2.8 ± 1.6
	BASN[10]	$\mathbf{1.8 \pm 1.1}$	3.6 ± 3.5	2.7 ± 1.6
	Ours	1.9 ± 1.4	2.5 ± 2.1	$\mathbf{2.2 \pm 1.7}$
4	GRADIENT	5.0 ± 4.5	15.0 ± 7.1	10.0 ± 5.0
	Auxiliary[8]	9.3 ± 5.6	10.4 ± 6.0	9.5 ± 6.0
	STASN[9]	6.7 ± 10.6	8.3 ± 8.4	7.5 ± 4.7
	FaceDe-S[26]	$\mathbf{1.2 \pm 6.3}$	6.1 ± 5.1	5.6 ± 5.7
	BASN[10]	6.4 ± 8.6	$\mathbf{3.2 \pm 5.3}$	4.8 ± 6.4
	Ours	1.3 ± 1.1	5.1 ± 2.0	$\mathbf{3.2 \pm 0.9}$

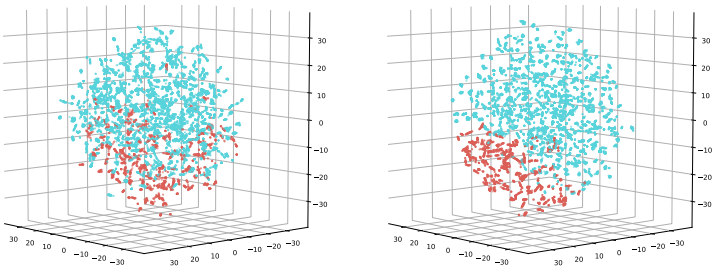
Table 4. Cross-testing results on SiW and OULU-NPU.

Training	Test	Method	ACER (%)
SiW	OULU 1	Auxiliary[8]	10.0
		DepthFreq[33]	9.3
		Ours	**3.8**
	OULU 2	Auxiliary[8]	14.0
		DepthFreq[33]	**7.8**
		Ours	11.0
	OULU 3	Auxiliary[8]	13.8 ± 5.7
		DepthFreq[33]	16.2 ± 5.0
		Ours	$\mathbf{9.7 \pm 1.6}$
	OULU 4	Auxiliary[8]	10.0 ± 8.8
		DepthFreq[33]	14.1 ± 8.3
		Ours	$\mathbf{1.9 \pm 0.6}$
OULU	SiW 1	DepthFreq	**7.28**
		Ours	8.90
	SiW 2	DepthFreq[33]	6.9 ± 1.1
		Ours	$\mathbf{6.7 \pm 0.9}$
	SiW 3	DepthFreq[33]	11.6 ± 4.7
		Ours	$\mathbf{6.1 \pm 0.02}$

SiW with OULU's Protocol 4 declines by five times compared with the best result. This verifies that our network is indeed competent at radically exploiting necessary information for PA detection based on both the AM and RPS with the multi-level fused features.

5 Discussion

Our approach provided notable improvement through outperforming the previous model by far[9], which determines the attended regions based only on binary supervision. Due to the simple constraints on supervision and the heavy reliance on the weak CNN-LSTM, the previous model has a great tendency to investigate



(a) Normal CNN-LSTM only

(b) Ours equipped with both channel and region attention mechanisms

Fig. 4. 3D visualization of feature distribution extracted from the last layer next to the classification block of (a) a vanilla CNN-LSTM and (b) our model comprising the AM, intensified with SENets for channel interaction, and the AAM, enhanced with region attention. Red points represent genuine cases and blue points are spoof cases. The figure is best viewed online in color.

inferior outer regions for the spoof cases, namely facial borders and backgrounds that may have no related clues for differentiation between genuine and spoof cases. As a result, we are aware of the issue that the effectiveness of the lower module, namely the AAM, relies heavily on the accuracy of the AM. Accordingly, we boosted the AM's performance to a great extent by adopting SENets that increase the interaction in channel. The offsets are also more likely to coincide with optimal locations due to the introduction of the AAS in a multi-task learning manner.

To better understand the effect of the AAS, the distribution of our multi-level fused feature at the channel-wise addition stage and that of a conventional CNN-LSTM on OULU-NPU's Protocol 1 is illustrated in Fig.4 through t-SNE[35]. It can be clearly observed that our model represents multi-level fused features with far better well-clustered properties than those of the conventional model. From the revealed information in the ablation study and the 3D visual distribution, our approach with the main purpose of fully resolving the issues of the excessive and inadequate exploitations in FAS partly yields effectiveness and obvious advancements in assuaging PA variations. Additional insight is presented in Fig. 5, where the attention to live images located by the AAM (orange boxes) has close ties to that of the AM (white boxes). However, the locations to discern spoof images in both modules are inconsistent with each other, proving that the initial advice of the AM is not always appropriate for the AAM to follow up.

Despite the benefits of our approach, it has some drawbacks. The model complexity means that training time is slightly longer than conventional ones as a result of the multi-step approach to guarantee the unified modules efficiently work together. Additionally, more memory must be allocated as well because of the module-driven scheme, but it is a worthwhile trade-off to improve classification performance.

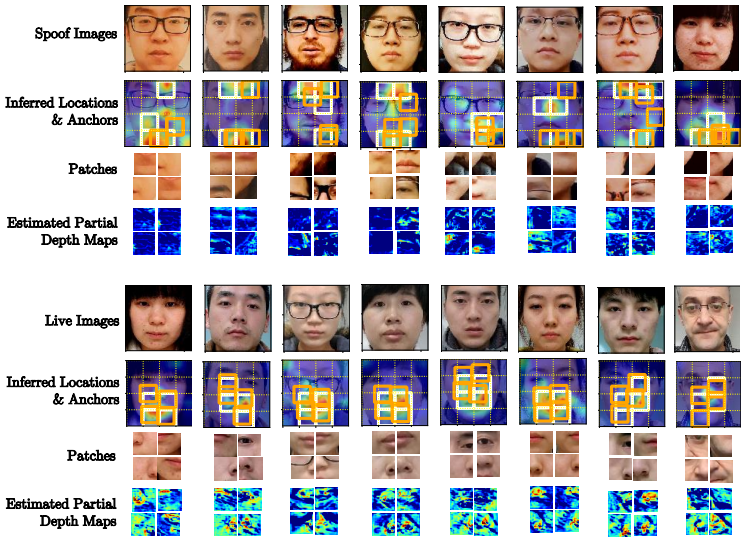


Fig. 5. Illustration of the model's perception of live and spoof cases in OULU-NPU Protocol 1. The first four columns in the first row are print attack cases, followed by four replay attack cases.

6 Conclusions

In this paper, we have described a novel initiative to overcome or at least mitigate the dilemma of PA variations in FAS expertise by leaving anxiety of the excessive and inadequate exploitations of the input behind. Specifically, introducing the extra channel interaction to the AM and the region attention to the AAM with the use of the AAS bears witness its feasibility in guiding the model to mine adequate distinct properties that it needs for PA detection. Our experimental outcomes also reflect a practical perspective that raising the model's consciousness to PA variations using attention mechanisms much more forceful than broadening the model's horizons in PA variations with some synthetic data.

References

1. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security **10** (2015) 746–761
2. 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), IEEE (2012) 1–7
3. Kim, W., Suh, S., Han, J.J.: Face liveness detection from a single image via diffusion speed model. IEEE transactions on Image processing **24** (2015) 2456–2465
4. Kollreider, K., Fronthaler, H., Faraj, M.I., Bigun, J.: Real-time face detection and motion analysis with application in liveness assessment. IEEE Transactions on Information Forensics and Security **2** (2007) 548–558
5. Pan, G., Sun, L., Wu, Z., Lao, S.: Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In: 2007 IEEE 11th International Conference on Computer Vision, IEEE (2007) 1–8
6. Luo, S., Kan, M., Wu, S., Chen, X., Shan, S.: Face anti-spoofing with multi-scale information. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE (2018) 3402–3407
7. 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), IEEE (2017) 319–328
8. 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. (2018) 389–398
9. Yang, X., Luo, W., Bao, L., Gao, Y., Gong, D., Zheng, S., Li, Z., Liu, W.: Face anti-spoofing: Model matters, so does data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3507–3516
10. Kim, T., Kim, Y., Kim, I., Kim, D.: Basn: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2019) 0–0
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
12. Määttä, J., Hadid, A., Pietikäinen, M.: Face spoofing detection from single images using micro-texture analysis. In: 2011 international joint conference on Biometrics (IJCB), IEEE (2011) 1–7
13. de Freitas Pereira, T., Anjos, A., De Martino, J.M., Marcel, S.: Lbp- top based countermeasure against face spoofing attacks. In: Asian Conference on Computer Vision, Springer (2012) 121–132
14. de Freitas Pereira, T., Anjos, A., De Martino, J.M., Marcel, S.: Can face anti-spoofing countermeasures work in a real world scenario? In: 2013 international conference on biometrics (ICB), IEEE (2013) 1–8
15. Yang, J., Lei, Z., Liao, S., Li, S.Z.: Face liveness detection with component dependent descriptor. In: 2013 International Conference on Biometrics (ICB), IEEE (2013) 1–6
16. Komulainen, J., Hadid, A., Pietikäinen, M.: Context based face anti-spoofing. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), IEEE (2013) 1–8
17. Patel, K., Han, H., Jain, A.K.: Secure face unlock: Spoof detection on smartphones. IEEE transactions on information forensics and security **11** (2016) 2268–2283

18. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: 2015 IEEE international conference on image processing (ICIP), IEEE (2015) 2636–2640
19. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. IEEE Transactions on Information Forensics and Security **11** (2016) 1818–1830
20. Bao, W., Li, H., Li, N., Jiang, W.: A liveness detection method for face recognition based on optical flow field. In: 2009 International Conference on Image Analysis and Signal Processing, IEEE (2009) 233–236
21. Siddiqui, T.A., Bharadwaj, S., Dhamecha, T.I., Agarwal, A., Vatsa, M., Singh, R., Ratha, N.: Face anti-spoofing with multifeature videolet aggregation. In: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE (2016) 1035–1040
22. Li, J., Wang, Y., Tan, T., Jain, A.K.: Live face detection based on the analysis of fourier spectra. In: Biometric Technology for Human Identification. Volume 5404., International Society for Optics and Photonics (2004) 296–303
23. Sun, L., Pan, G., Wu, Z., Lao, S.: Blinking-based live face detection using conditional random fields. In: International Conference on Biometrics, Springer (2007) 252–260
24. Li, L., Feng, X., Boulkenafet, Z., Xia, Z., Li, M., Hadid, A.: An original face anti-spoofing approach using partial convolutional neural network. In: 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE (2016) 1–6
25. Patel, K., Han, H., Jain, A.K.: Cross-database face antispoofing with robust feature representation. In: Chinese Conference on Biometric Recognition, Springer (2016) 611–619
26. Jourabloo, A., Liu, Y., Liu, X.: Face de-spoofing: Anti-spoofing via noise modeling. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 290–306
27. Liu, Y., Jourabloo, A., Ren, W., Liu, X.: Dense face alignment. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2017) 1619–1628
28. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. (2015) 2017–2025
29. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: ECCV. (2018)
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. (2017) 618–626
31. 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), IEEE (2017) 612–618
32. ISO/IEC JTC 1/SC 37 Biometrics.: information technology biometric presentation attack detection part 1: Framework. https://www.iso.org/obp/ui/iso (2006)
33. Huang, Y., Zhang, W., Wang, J.: Deep frequent spatial temporal learning for face anti-spoofing. arXiv preprint arXiv:2002.03723 (2020)
34. Boulkenafet, Z., Komulainen, J., Akhtar, Z., Benlamoudi, A., Samai, D., Bekhouche, S.E., Ouafi, A., Dornaika, F., Taleb-Ahmed, A., Qin, L., et al.: A competition on generalized software-based face presentation attack detection in

mobile scenarios. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE (2017) 688–696
35. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9** (2008) 2579–2605