

Look Locally Infer Globally: A Generalizable Face Anti-Spoofing Approach

Debayan Deb, *Student Member, IEEE*, and Anil K. Jain, *Life Fellow, IEEE*

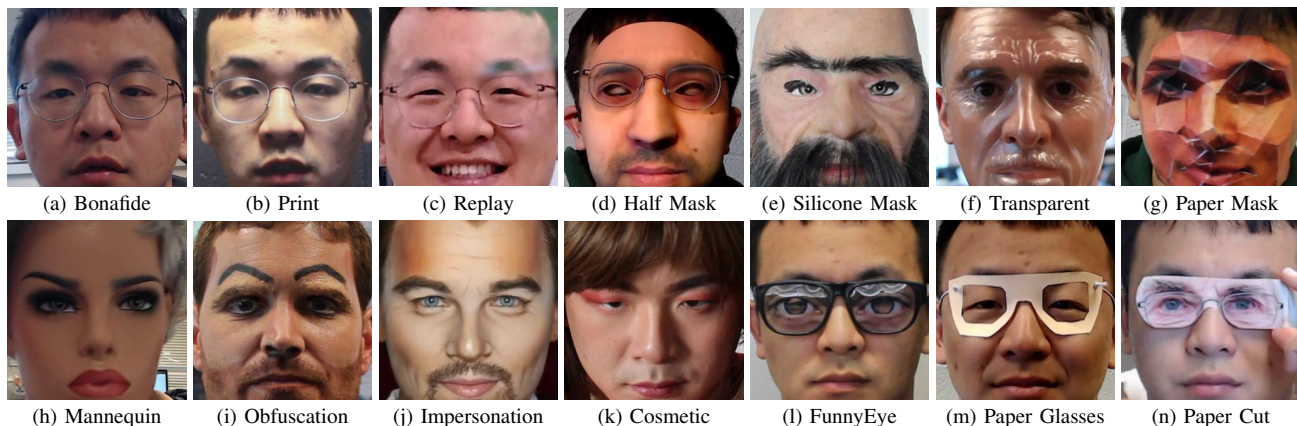


Fig. 1. Example presentation attack instruments: Simple attacks include (b) printed photograph, or (c) replaying the victim’s video. More advanced presentation attacks can also be leveraged such as (d-h) 3D masks, (i-k) make-up attacks, or (l-n) partial attacks [1]. A bonafide face is shown in (a) for comparison. Here, the presentation attacks in (b-c, k-n) belong to the same person in (a).

Abstract—State-of-the-art presentation attack detection approaches tend to overfit to the presentation attack instruments seen during training and fail to generalize to unknown presentation attack instruments. Given that face presentation attack detection is inherently a local task, we propose a face presentation attack detection framework, namely Self-Supervised Regional Fully Convolutional Network (*SSR-FCN*), that is trained to learn local discriminative cues from a face image in a self-supervised manner. The proposed framework (i) improves generalizability while maintaining the computational efficiency of holistic face presentation attack detection approaches (< 4 ms on a Nvidia GTX 1080Ti GPU), and (ii) is more interpretable since it localizes the parts of the face that are labeled as presentation attacks. Experimental results show that *SSR-FCN* can achieve $TDR = 65\%$ @ 2.0% FDR when evaluated on a dataset, SiW-M, comprising of 13 different presentation attack instruments under unknown attacks while achieving competitive performances under standard benchmark datasets (Oulu-NPU, CASIA-MFSD, and Replay-Attack).

Index Terms—Face anti-spoofing, presentation attack detection, regional supervision, fully convolutional neural network

I. INTRODUCTION

THE accuracy, usability, and touchless acquisition of state-of-the-art automated face recognition systems (AFR) have led to their ubiquitous adoption in a plethora of domains, including mobile phone unlock, access control systems, and payment services. The adoption of deep learning models over the past decade has led to prevailing AFR systems

with accuracies as high as 99% True Accept Rate at 10^{-6} False Accept Rate [2]. Despite this impressive recognition performance, current AFR systems remain vulnerable to the growing threat of *presentation attacks*¹.

Face attacks are “fake faces” which can be constructed with a variety of different instruments (presentation attack instruments), *e.g.*, 3D printed masks, printed paper, or digital devices (video replay attacks from a mobile phone) with a goal of enabling an attacker to impersonate a victim’s identity, or alternatively, obfuscate their own identity (see Figure 1). With the rapid proliferation of face images/videos on the Internet (especially on social media websites, such as Facebook, Twitter, or LinkedIn), replaying videos containing the victim’s face or presenting a printed photograph of the victim to the AFR system is a trivial task [6]. Even if a face presentation attack detection system could trivially detect printed photographs and replay video attacks (*e.g.*, with depth sensors), attackers can still attempt to launch more sophisticated attacks such as 3D masks [7], make-up, or even virtual reality [8].

The need for preventing face attacks is becoming increasingly urgent due to the user’s privacy concerns associated with spoofed systems. Failure to detect face attacks can be a major security threat due to the widespread adoption of automated face recognition systems for border control [9]. In 2011, a young individual from Hong Kong boarded a flight to Canada disguised as an old man with a flat hat

¹ISO standard IEC 30107-1:2016(E) defines presentation attacks as “*presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system*” [3]. Note that these presentation attacks are different from digital manipulation of face images, such as DeepFakes [4] and adversarial faces [5].

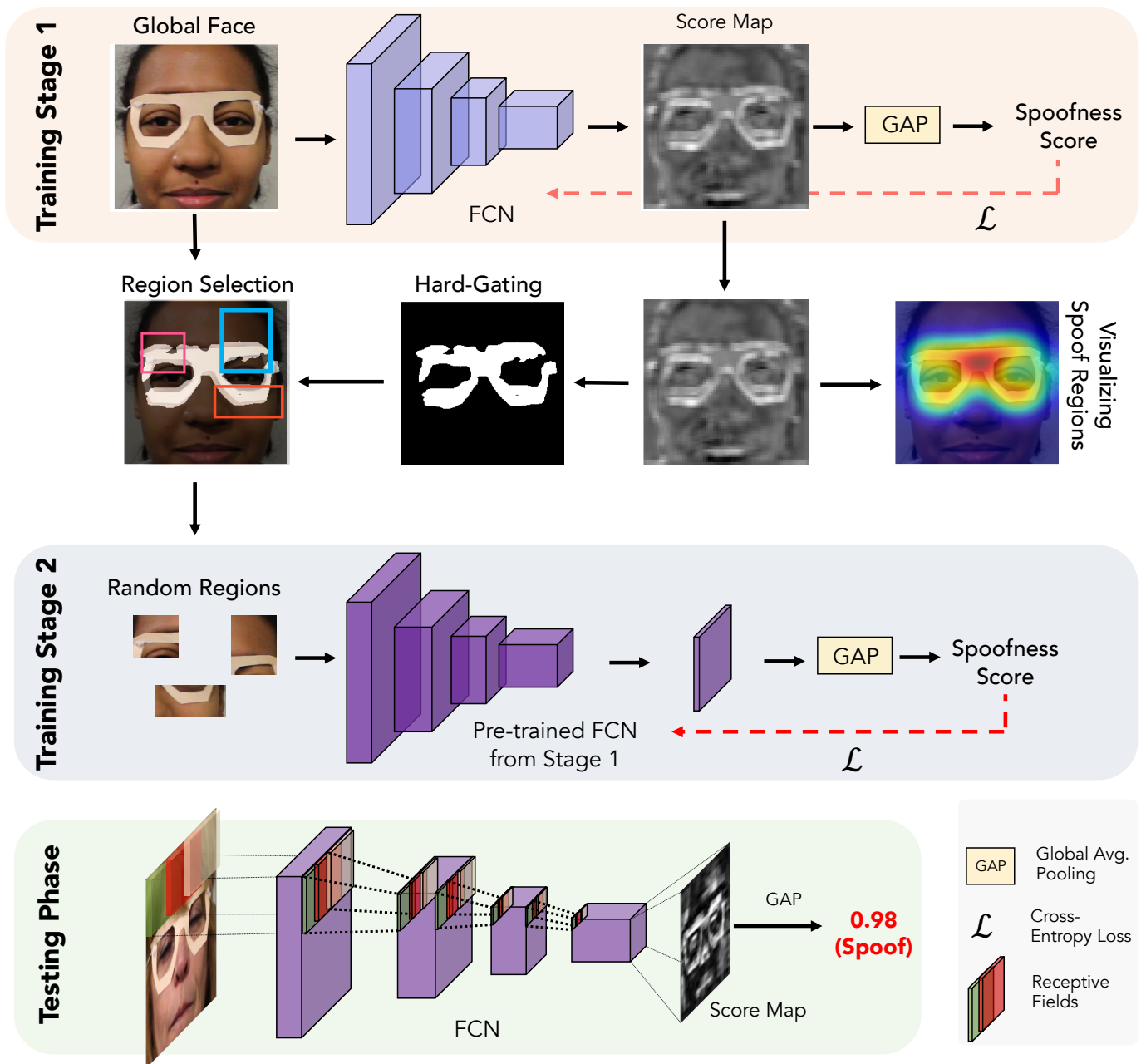


Fig. 2. An overview of the proposed Self-Supervised Regional Fully Convolution Network (SSR-FCN). We train in two stages: (1) Stage 1 learns global discriminative cues via training on the entire face image. The score map obtained from stage 1 is hard-gated to obtain presentation attack regions in the face image. We randomly crop arbitrary-size patches from the presentation attack regions and fine-tune our network in stage 2 to learn local discriminative cues. During test, we input the entire face image to obtain the classification score. The score map can also be used to visualize the presentation attack regions in the input image.

by wearing a silicone face mask to successfully fool the border control authorities [10]. Also consider that, with the advent of Apple’s iPhone X and Samsung’s Galaxy S8, all of us are carrying automated face recognition systems in our pockets embedded in our smartphones. Face recognition on our phones facilitates (i) unlocking the device, (ii) conducting financial transactions, and (iii) access to privileged content stored on the device. Failure to detect face presentation attacks on smartphones could compromise confidential information such as emails, banking records, social media content, and personal photos [11].

With numerous approaches proposed to detect face attacks, current face presentation attack detection methods have following shortcomings:

a) *Generalizability:* Since the exact presentation attack instrument may not be known beforehand, how to generalize well to “unknown”² attacks is of utmost importance. A ma-

²Unseen attacks are presentation attack instruments that are known to the developers whereby algorithms can be specifically tailored to detect them, but their data is never used for training. Unknown attacks are presentation attack instruments that are not known to the developers and neither seen during training.

TABLE I
A SUMMARY OF PUBLICLY AVAILABLE FACE PRESENTATION ATTACK DETECTION DATASETS.

Dataset	Year	Statistics		PIE Variations			# Presentation Attack Instruments					Total
		# Subj.	# Vids.	Pose	Expression Change	Illumination Change	Replay	Print	3D Mask	Makeup	Partial	
Replay-Attack [12]	2012	50	1,200	Frontal	No	Yes	2	1	0	0	0	3
CASIA-FASD [13]	2012	50	600	Frontal	No	No	1	1	0	0	0	2
3DMAD [14]	2013	17	255	Frontal	No	No	0	0	1	0	0	1
MSU-MFSD [15]	2015	35	440	Frontal	No	No	2	1	0	0	0	3
Replay-Mobile [16]	2016	40	1,030	Frontal	No	Yes	1	1	0	0	0	2
HKBU-MARs [17]	2016	35	1,009	Frontal	No	Yes	0	0	2	0	0	2
Oulu-NPU [18]	2017	55	4,950	Frontal	No	Yes	2	2	0	0	0	4
SiW [1]	2018	165	4,620	[−90°, 90°]	Yes	Yes	4	2	0	0	0	6
SiW-M [19]	2019	493	1,630	[−90°, 90°]	Yes	Yes	1	1	5	3	3	13

majority of the prevailing state-of-the-art face presentation attack detection techniques focus only on detecting 2D printed paper and video replay attacks, and are vulnerable to presentation attacks crafted from materials not seen during training of the detector. In fact, studies show a two-fold increase in error when presentation attack detection approaches encounter unknown presentation attack instruments [1]. In addition, current face presentation attack detection approaches rely on densely connected neural networks with a large number of learnable parameters (exceeding 2.7M), where the lack of generalization across unknown presentation attack instruments is even more pronounced.

b) Lack of Interpretability: Given a face image, face presentation attack detection approaches typically output a holistic face “attack score” which depicts the likelihood that the input image is bonafide or a presentation attack. Without an ability to visualize which regions of the face contribute to the overall decision made by the network, the global attack score alone may not be sufficient for a human operator to interpret the network’s decision.

In an effort to impart generalizability and interpretability to face presentation attack detection systems, we propose a face presentation attack detection framework specifically designed to detect unknown presentation attack instruments, namely, **Self-Supervised Regional Fully Convolutional Network (SSR-FCN)**. A Fully Convolutional Network (FCN) is first trained to learn global discriminative cues and automatically identify presentation attack regions in face images. The network is then fine-tuned to learn local representations via regional supervision. Once trained, the deployed model can automatically locate regions where attack occurs in the input image and provide a final attack score.

Our contributions are as follows:

- We show that features learned from local face regions have better generalization ability than those learned from the entire face image alone.
- We provide extensive experiments to show that the proposed approach, *SSR-FCN*, outperforms other local region extraction strategies and state-of-the-art face presentation attack detection methods on one of the largest publicly available dataset, namely, SiW-M, comprised of 13 different presentation attack instruments. The proposed method reduces the Equal Error Rate (EER) by (i) 14% relative to state-of-the-art [20] under the unknown attack setting, and (ii) 40% on known presentation attack instruments. In addition, *SSR-FCN* achieves competitive

performance on standard benchmarks on Oulu-NPU [18] dataset and outperforms prevailing methods on cross-dataset generalization (CASIA-FASD [13] and Replay-Attack [12]).

- The proposed *SSR-FCN* is also shown to be more interpretable since it can directly predict the parts of the faces that are considered as presentation attacks.

II. BACKGROUND

In order to mitigate the threats associated with presentation attacks, numerous face presentation attack detection techniques, based on both software and hardware solutions, have been proposed. Early software-based solutions utilized liveness cues, such as eye blinking, lip movement, and head motion, to detect print attacks [21–24]. However, these approaches fail when they encounter unknown attacks such as printed attacks with cut eye regions (see Figure 1n). In addition, these methods require active cooperation of user in providing specific types of images making them tedious to use.

Since then, researchers have moved on to *passive* face presentation attack detection approaches that rely on texture analysis for distinguishing bonafide and presentation attacks, rather than motion or liveness cues. The majority of face presentation attack detection methods only focus on detecting print and replay attacks, which can be detected using features such as color and texture [25–30]. Many prior studies employ handcrafted features such as 2D Fourier Spectrum [15, 31], Local Binary Patterns (LBP) [26, 32–34], Histogram of oriented gradients (HOG) [25, 35], Difference-of-Gaussians (DoG) [36], Scale-invariant feature transform (SIFT) [28], and Speeded up robust features (SURF) [37]. Some techniques utilize presentation attack detection beyond the RGB color spectrum, such as incorporating the luminance and chrominance channels [34]. Instead of a predetermined color spectrum, Li *et al.* [27] automatically learn a new color scheme that can best distinguish bonafide and presentation attacks. Another line of work extracts image quality features to detect presentation attacks [15, 29, 30]. Due to the assumption that presentation attack instruments are one of replay or print attacks, these methods severely suffer from generalization to unknown presentation attack instruments.

Hardware-based solutions in literature have incorporated 3D depth information [38–40], multi-spectral and infrared sensors [41], and even physiological sensors such as vein-flow information [42]. Presentation attack detection can be further enhanced by incorporating background audio signals [43].

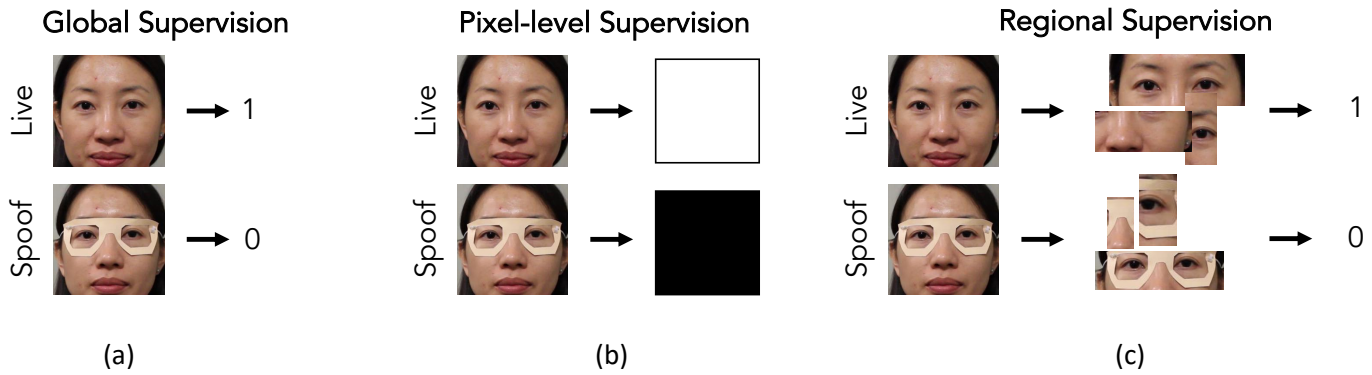


Fig. 3. Illustration of drawbacks of prior approaches. Top: example of a bonafide face; Bottom: example of a paper glasses presentation attack. In this case, the presentation attack artifact is only present in the eye-region of the face. (a) Classifier trained with global supervision overfits to the bonafide class since both images are mostly bonafide (the presentation attack instrument covers only a part of the face). (b) Pixel-level supervision assumes the entire image is either bonafide or presentation attack and constructs label maps accordingly. This is not a valid assumption in mask, makeup, and partial presentation attack instruments. Instead, (c) the proposed framework, trains on extracted regions from face images. These regions can be based on domain knowledge, such as eye, nose, mouth regions, or randomly cropped. The proposed *SSR-FCN* utilizes self-supervised region-selection.

However, with the inclusion of additional sensors along with a standard camera, the deployment costs can be exorbitant (*e.g.*, thermal sensors for iPhones cost over USD 400³).

State-of-the-art face presentation attack detection systems utilize Convolutional Neural Networks (CNN) so the feature set (representation) is learned that best differentiates bonafides from presentation attacks. Yang *et al.* were among the first to propose CNNs for face presentation attack detection and they showed about 70% decrease in Half Total Error Rate (HTER) compared to the baselines comprised of handcrafted features [44]. Further improvement in performance was achieved by directly modifying the network architecture [45–48]. Deep learning approaches also perform well for mask attack detection [7]. Incorporating auxiliary information (*e.g.* eye blinking) in deep networks can further improve the face presentation attack detection performance [1, 23].

Limited studies on generalizable face presentation attack detection focus on one-class classification approaches. These methods only model the distribution of bonafide face features using one-class classifiers such as one-class SVM [49], one-class GMM [50], or employ a distance metric loss [51]. However, these approaches have several drawbacks: (i) by only modeling the bonafide face feature distributions, the methods tend to overfit to bonafide class, and (ii) both one-class SVM and one-class GMM have been shown to perform poorly on public benchmark datasets (CASIA-FASD [13], Replay-Mobile [16], and MSU-MFSD [15]). A tree-based approach utilizing deep networks was proposed for generalizable face presentation attack detection [19]. In order to prevent face presentation attack detection methods from overfitting to the specific subject, environment, and presentation attack instrument, transfer learning has also been studied [52–54]. However, these methods share similar network architecture that are densely connected with thirteen convolutional layers exceeding 2.7M learnable parameters [1, 20, 52–56]. Due to this, a majority of the aforementioned presentation attack detection methods also suffer from poor generalization performance.

Table I outlines the publicly available face presentation attack detection datasets.

III. MOTIVATION

Our approach is motivated by following observations:

1) *Face Presentation Attack Detection is a Local Task*: It is now generally accepted that for print and replay attacks, “face presentation attack detection is usually a local task in which discriminative clues are ubiquitous and repetitive” [57]. However, in the case of masks, makeups, and partial attacks, the ubiquity and repetitiveness of presentation attack cues may not hold true. For instance, in Figure 3 (a-c), presentation attack artifact (the paper glasses) are only present in the eye regions of the face. Unlike face recognition, face presentation attack detection does not require the entire face image in order to predict whether the image is a presentation attack or bonafide. In fact, our experimental results and their analysis will confirm that the entire face image alone can adversely affect the convergence and generalization of networks.

2) *Global vs. Local Supervision*: Prior work can be partitioned into two groups: (i) *global supervision* where the input to the network is the entire face image and the CNN outputs a score indicating whether the image is bonafide or presentation attack [1, 20, 44–48, 55, 58], and (ii) *pixel-level supervision* where multiple classification losses are aggregated over each pixel in the feature map [56, 59]. These studies assume that all pixels in the face image is either bonafide or presentation attack (see Figure 3(b)). This assumption holds true for presentation attack instruments, such as replay and print attacks (which are the only presentation attack instruments considered by the studies), but not for mask, makeup, and partial attacks. Therefore, pixel-level supervision can not only suffer from poor generalization across a diverse range of presentation attack instruments, but also convergence of the network is severely affected due to noisy labels.

In summary, based on the 13 different presentation attack instruments shown in Figure 1, for which we have the data, we gain the following insights: (i) face presentation attack

³<https://amzn.to/2zJ6YW4>

TABLE II
ARCHITECTURE DETAILS OF THE PROPOSED FCN BACKBONE.

Layer	# of Activations	# of Parameters
Input	$H \times W \times 3$	0
Conv	$H/2 \times W/2 \times 64$	$3 \times 3 \times 3 \times 64 + 64$
Conv	$H/4 \times W/4 \times 128$	$3 \times 3 \times 64 \times 128 + 128$
Conv	$H/8 \times W/8 \times 256$	$3 \times 3 \times 128 \times 256 + 256$
Conv	$H/16 \times W/16 \times 512$	$3 \times 3 \times 256 \times 512 + 512$
Conv	$H/16 \times H/16 \times 1$	$3 \times 3 \times 512 \times 1 + 1$
GAP	1	0
Total		1.5M

Conv and GAP refer to convolutional and global average pooling operations.

detection is inherently a local task, and (ii) learning local representations can improve face presentation attack detection performance [56, 59]. Motivated by (i), we hypothesize that utilizing a Fully Convolutional Network (FCN) may be more appropriate for the face presentation attack detection task compared to a traditional CNN. The second insight suggests FCNs can be intrinsically regularized to learn local cues by enforcing the network to *look* at local spatial regions of the face. In order to ensure that these regions mostly comprise presentation attack patterns, we propose a *self-supervised* region extractor.

IV. PROPOSED APPROACH

In this section, we describe the proposed *Self-Supervised Regional Fully Convolutional Network (SSR-FCN)* for generalized face presentation attack detection. As shown in Figure 2, we train the network in two stages, (a) Stage I learns global discriminative cues and predicts score maps, and (b) Stage II extracts arbitrary-size regions from presentation attack areas and fine-tunes the network via regional supervision.

A. Network Architecture

In typical image classification tasks, networks are designed such that information present in the input image can be used for learning *global* discriminative features in the form of a feature vector without utilizing the spatial arrangement in the input. To this end, a fully-connected (FC) layer is generally introduced at the end of the last convolutional layer. This fully-connected layer outputs a D -dimension feature that aggregates decisions at various spatial regions to obtain a global description of the input image. However, this is not ideal for partial presentation attacks since the presentation attack artifact is not present in all spatial regions. Given the plethora of available presentation attack instruments, it is better to learn *local* representations and make decisions on local spatial inputs rather than global descriptors. Therefore, we employ a Fully Convolutional Network (FCN) by replacing the FC layer in a traditional CNN with a 1×1 convolutional layer followed by a global average pooling layer. The FCN leads to three major advantages over traditional CNNs:

- **Arbitrary-sized inputs:** By replacing the fully-connected layer with a global average pooling layer, the entire FCN can accept input images of any image size. This property can be exploited to learn discriminative features at local

spatial regions, regardless of the input size, rather than overfitting to a global representation of the entire face image.

- **Interpretability:** Since the proposed FCN is trained to provide decisions at a local level, the score map output by the network can be used to identify the presentation attack regions in the face.
- **Efficiency:** Via FCN, an entire face image can be inferred only once where local decisions are dynamically aggregated via the 1×1 convolution operator. Existing methods utilizing a traditional CNN which has a larger number of trainable parameters due to the fully connected layer at the end of the network. This necessitates a large training dataset which is limited in the face presentation attack detection literature (see Table I). FCN is more parameter-efficient and can be trained in an effective manner (to avoid overfitting).

B. Network Efficiency

A majority of prior work on CNN-based face presentation attack detection employs architectures that are densely connected with thirteen convolutional layers [1, 20, 55, 56, 60]. Even with the placement of skip connections, the number of learnable parameters exceed $2.7M$. As we see in Table I, only a limited amount of training data⁴ is generally available in face presentation attack detection datasets. Limited data coupled with the large number of trainable parameters causes current approaches to overfit, leading to poor generalization performance under unknown attack scenarios. Instead, we employ a shallower neural network comprising of only five convolutional layers with approximately $1.5M$ learnable parameters (see Table II).

C. Stage I: Training FCN Globally

We first train the FCN with global face images in order to learn global discriminative cues and identify presentation attack regions in the face image. Given an image, $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, we detect a face and crop the face region via 5 landmarks (two eyes, nose, and two mouth keypoints) in order to remove background information not pertinent to the task of face presentation attack detection. Here, H , W , and C refer to the height, width, and number of channels (3 in the case of RGB) of the input image. The face regions are then aligned and resized to a fixed-size (e.g., 256×256) in order to maintain consistent spatial information across all training data.

The proposed FCN consists of four downsampling convolutional blocks each coupled with batch normalization and ReLU activation. The feature map from the fourth convolutional layer passes through a 1×1 convolutional layer. The output of the 1×1 convolutional layer represents a score map $\mathbf{S} \in \mathbb{R}^{(H_S \times W_S \times 1)}$ where each pixel in \mathbf{S} represents a bonafide/presentation attack decision corresponding to its receptive field in the image. The height (H_S) and width (W_S) of the score map is determined by the input image size and

⁴The lack of large-scale publicly available face presentation attack detection datasets is due to the time and effort required along with privacy concerns associated in acquiring such datasets.



Fig. 4. Three presentation attack images and their corresponding binary masks extracted from predicted score maps. Black regions correspond to predicted bonafide regions, whereas, white regions indicate presentation attack.

the number of downsampling layers. For a $256 \times 256 \times 3$ input image, our proposed architecture outputs a $16 \times 16 \times 1$ score map.

The score map is then reduced to a single scalar value by globally average pooling. That is, the final classification score (s) for an input image is obtained from the $(H_S \times W_S \times 1)$ score map (\mathbf{S}) by,

$$s = \frac{1}{H_S \times W_S} \sum_i \sum_j \mathbf{S}_{i,j} \quad (1)$$

Using sigmoid activation on the final classification output (s), we obtain a scalar $p(c|\mathbf{x}) \in [0, 1]$ predicting the likelihood that the input image is a presentation attack, where $c = 0$ indicates bonafide and $c = 1$ indicates presentation attack.

We train the network by minimizing the Binary Cross Entropy (BCE) loss,

$$\mathcal{L} = -[y \times \log(p(c|\mathbf{x})) + (1 - y) \times \log(1 - p(c|\mathbf{x}))] \quad (2)$$

where y is the ground truth label of the input image.

D. Stage II: Training FCN on Self-Supervised Regions

In order to supervise training at a local level, we propose a regional supervision strategy. We train the network to learn local cues by only showing certain regions of the face where presentation attack patterns exist. In order to ensure that presentation attack artifacts/patterns indeed exist within the selected regions, the pre-trained FCN from Stage I (IV-C) can automatically guide the region selection process in presentation attack images. For bonafide faces, we can randomly crop a region from any part of the image.

Due to the absence of a fully connected layer, notice that FCN naturally encodes decisions at each pixel in feature map \mathbf{S} . In other words, higher intensity pixels within \mathbf{S} indicate a larger likelihood of a presentation attack pattern residing within the receptive field in the image. Therefore, discriminative regions (presentation attack areas) are automatically highlighted in the score map by training on entire face images (see Figure 2).

We can then craft a binary mask M indicating the bonafide/presentation attack regions in the input presentation attack images. First, we soft-gate the score map by min-max normalization such that we can obtain a score map $\mathbf{S}' \in [0, 1]$.

Let $f_{\mathbf{S}'}(i, j)$ represent the activation in the (i, j) -th spatial location in the scaled score map \mathbf{S}' . The binary mask M is designed by hard-gating,

$$M(i, j) = \begin{cases} 1, & \text{if } \mathbf{S}'(i, j) \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where, τ is a threshold that controls the size of the hard-gated region ($\tau = 0.5$ in our case). A larger τ leads to smaller regions and smaller τ can lead to spurious presentation attack regions. Examples of binary masks are shown in Figure 4. From the binary mask, we can then randomly extract a rectangular bounding box such that the center of the rectangle lies within the detected presentation attack regions. In this manner, we can crop rectangular regions of arbitrary sizes from the input image such that each region contains presentation attack artifacts according to our pre-trained global FCN. We constrain the width and height of the bounding boxes to be between MIN_{region} and MAX_{region} .

In this manner, we fine-tune our network to learn local discriminative cues.

E. Testing

Since FCNs can accept arbitrary input sizes and the fact that the proposed FCN has encountered entire faces in Stage I, we input the global face into the trained network and obtain the score map. The score map is then average pooled to extract the final classification output, which is then normalized by a sigmoid function in order to obtain an attack score within $[0, 1]$. That is, the final classification score is obtained by,

$$\frac{1}{1 + \exp(-s)}$$

In addition to the classification score, the score map (\mathbf{S}) can also be utilized for visualizing the presentation attack regions in the face by constructing a heatmap (see Figure 2).

V. EXPERIMENTAL SETUP

A. Datasets

The following four datasets are utilized in our study (Table I):

1) *Spoof-in-the-Wild with Multiple Attacks (SiW-M)* [19]: A dataset, collected in 2019, comprising 13 different presentation attack instruments, acquired specifically for evaluating generalization performance on unknown presentation attack instruments. Compared with other publicly available datasets (Table I), SiW-M is diverse in presentation attack instruments, environmental conditions, and face poses. We evaluate *SSR-FCN* under both *unknown* and *known* settings, and perform ablation studies on this dataset.

2) *Oulu-NPU* [18]: A dataset comprised of 4,950 high-resolution video clips of 55 subjects. Oulu-NPU defines four protocols each designed for evaluating generalization against variations in capturing conditions, attack devices, capturing devices and their combinations. We use this dataset for comparing our approach with the prevailing state-of-the-art face presentation attack detection methods on the four protocols.

3) *CASIA-FASD* [13] & *Replay-Attack* [12]: Both datasets, collected in 2012, are frequently employed in face presentation attack detection literature for testing *cross-dataset generalization* performance. These two datasets provide a comprehensive collection of attacks, including warped photo attacks, cut photo attacks, and video replay attacks. Low-quality, normal-quality, and high-quality videos are recorded under different lighting conditions.

All images shown in this paper are from SiW-M testing sets.

B. Data Preprocessing

For all datasets, we first extract all frames in a video. The frames are then passed through MTCNN face detector [61] to detect 5 facial landmarks (two eyes, nose and two mouth corners). Similarity transformation is used to align the face images based on the five landmarks. After transformation, the images are cropped to 256×256 . All face images shown in the paper are cropped and aligned. Before passing into network, we normalize the images by requiring each pixel to be within $[-1, 1]$ by subtracting 127.5 and dividing by 127.5.

C. Implementation Details

SSR-FCN is implemented in Tensorflow, and trained with a constant learning rate of $(1e - 3)$ with a mini-batch size of 128. The objective function, \mathcal{L} , is minimized using Adam optimizer [62]. It takes 20 epochs to converge. Following [1], we randomly initialize all the weights of the convolutional layers using a normal distribution of 0 mean and 0.02 standard deviation. We restrict the self-supervised regions to be at least $1/4$ of the entire image, that is, $MIN_{region} = 64$ and at most $MAX_{region} = 256$ which is the size of the global face image. Data augmentation during training involves random horizontal flips with a probability of 0.5. We train and test our proposed method on a single Nvidia GTX 1080Ti GPU. For evaluation, we compute the attack scores for all frames in a video and temporally average them to obtain the final classification score.

D. Evaluation Metrics

For all the experiments, we report the standard ISO/IEC 30107 [3] metrics:

- 1) Attack Presentation Classification Error Rate (APCER): the worst error rate among all the presentation attack instruments;
- 2) Bonafide Presentation Classification Error Rate (BPCER):
- 3) Average Classification Error Rate (ACER): the mean of APCER and BPCER.

TABLE III
GENERALIZATION ERROR ON LEARNING GLOBAL (CNN) VS. LOCAL (FCN) REPRESENTATIONS OF SiW-M [19].

Method	Metric (%)	Replay	Obfuscation	Paper Glasses	Overall
CNN	ACER	13.3	47.1	32.2	30.8 ± 17.0
	EER	12.8	44.6	23.6	27.0 ± 13.2
FCN (Stage I)	ACER	11.2	52.2	12.1	25.1 ± 23.4
	EER	11.2	37.6	12.4	20.4 ± 12.1

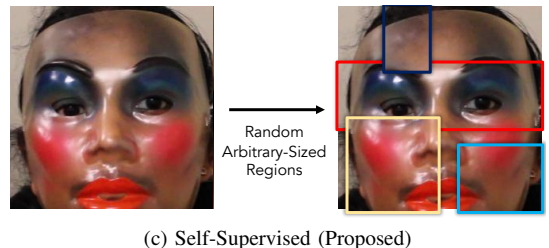
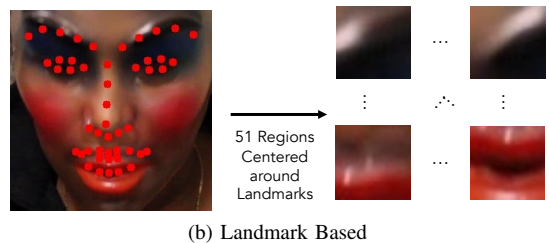
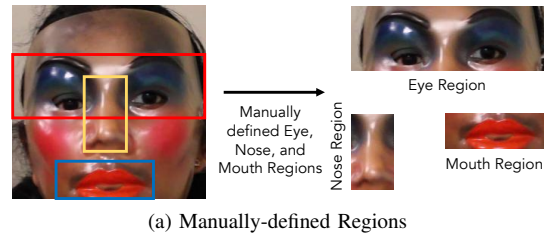


Fig. 5. Illustration of various region extraction strategies from training images. (a) and (b) are regions extracted via domain knowledge (manually defined) or landmark-based. (c) random regions extracted via proposed self-supervision scheme. Each color denotes a separate region.

In addition, we also report the Equal Error Rate (EER) and the True Detection Rate (TDR) at 2.0%⁵ False Detection Rate (FDR) for our evaluation. For a fair comparison with prior work, we report the Half Total Error Rate (HTER) for cross-dataset evaluation. *Except for EER and HTER, we employ a decision threshold of 0.5.*

VI. EXPERIMENTAL RESULTS














A. Evaluation of Global Descriptor vs. Local Representation

In order to analyze the impact of learning local embeddings as opposed to learning a global embedding, we conduct an ablation study on three presentation attack instruments in the SiW-M dataset [19], namely, Replay (Figure 1c), Obfuscation (Figure 1i), and Paper Glasses (Figure 1m). *The experiment is conducted under the unknown attack scenario (leave-one-instrument-out protocol).*

⁵Due to the small number of bonafide samples, thresholds at lower False Detection Rate (FDR) such as 0.2% (recommended under the IARPA ODIN program) cannot be computed.

TABLE IV

GENERALIZATION PERFORMANCE OF DIFFERENT REGION EXTRACTION STRATEGIES ON SIW-M DATASET. HERE, EACH COLUMN REPRESENTS AN UNKNOWN PRESENTATION ATTACK INSTRUMENT WHILE THE METHOD IS TRAINED ON THE REMAINING 12 PRESENTATION ATTACK INSTRUMENTS.

Method	Metric (%)	Replay	Print	Mask Attacks						Makeup Attacks			Partial Attacks			Mean \pm Std.
																
		99 vids.	118 vids.	72 vids.	27 vids.	88 vids.	17 vids.	40 vids.	23 vids.	61 vids.	50 vids.	160 vids.	127 vids.	86 vids.		
Global (Stage I)	ACER	11.2	15.5	12.8	21.5	35.4	6.1	10.7	52.2	50.0	20.5	26.2	12.1	9.6	22.6 \pm 15.3	
	EER	11.2	14.0	12.8	23.1	26.6	2.9	11.0	37.6	10.4	17.0	24.2	12.4	10.1	16.8 \pm 9.3	
Eye-Region	ACER	13.2	13.7	7.5	17.4	22.5	5.79	6.2	19.5	8.3	11.7	32.8	15.3	7.3	13.2 \pm 8.5	
	EER	12.4	11.4	7.3	15.2	21.5	2.9	6.5	20.2	7.8	11.2	27.2	14.7	7.5	12.3 \pm 6.2	
Nose-Region	ACER	17.4	10.5	8.2	13.8	30.3	5.3	8.4	37.4	5.1	18.0	35.5	31.4	7.1	17.6 \pm 12.0	
	EER	14.6	9.8	9.2	12.7	22.0	5.2	8.4	23.6	4.4	14.6	24.9	27.7	7.6	14.2 \pm 7.9	
Mouth-Region	ACER	20.5	20.7	22.9	26.3	30.6	15.6	17.1	44.2	18.1	24.0	38.0	47.2	8.5	25.7 \pm 11.4	
	EER	19.9	21.3	22.6	25.1	30.0	10.1	10.7	40.9	16.1	24.0	35.5	40.4	8.1	23.4 \pm 10.9	
Global + Eye + Nose	ACER	10.9	10.5	7.5	17.7	28.7	5.1	7.0	38.0	5.1	13.6	29.4	15.2	6.2	15 \pm 10.7	
	EER	10.2	10.0	7.7	15.8	21.3	1.8	6.7	21.0	3.0	12.3	22.5	12.3	6.5	11.6 \pm 6.8	
Landmark-Region	ACER	10.7	9.2	18.4	25.1	26.4	6.2	6.9	53.8	8.1	15.4	35.8	40.8	7.6	20.3 \pm 15.2	
	EER	8.0	10.1	12.2	23.1	18.8	8.9	4.1	40.1	9.9	15.6	17.7	25.6	4.9	15.3 \pm 10	
Global + Landmark	ACER	12.0	11.2	7.3	23.7	26.4	6.3	5.9	26.7	6.7	10.7	27.8	25.7	6.4	15.1 \pm 9.2	
	EER	11.5	10.1	7.2	19.0	4.9	6.6	4.6	25.6	6.7	10.9	23.5	18.5	4.7	11.8 \pm 7.4	
Random-Crop	ACER	9.2	6.7	7.3	19.9	30.9	9.1	6.9	44.0	6.5	13.8	31.8	28.6	5.9	17.0 \pm 12.8	
	EER	8.9	7.8	10.3	17.9	21.3	3.7	6.5	32.7	5.4	13.7	18.7	19.4	7.1	13.3 \pm 8.3	
Global + Random-Crop	ACER	12.3	10.7	6.5	18.2	22.9	6.2	6.1	18.6	4.9	11.6	32.7	16.1	7.2	13.4 \pm 8.1	
	EER	10.9	9.2	6.9	16.6	21.3	2.9	5.2	18.8	3.7	11.5	19.0	14.9	6.2	11.3 \pm 6.3	
SSR-FCN (Stage I \rightarrow Stage II)	ACER	7.4	19.5	3.2	7.7	33.3	5.2	3.3	22.5	5.9	11.7	21.7	14.1	6.4	12.4 \pm 9.2	
	EER	6.8	11.2	2.8	6.3	28.5	0.4	3.3	17.8	3.9	11.7	21.6	13.5	3.6	10.1 \pm 8.4	

In this experiment, a *traditional CNN* learning a global image descriptor is constructed by replacing the 1×1 convolutional layer with a fully connected layer. We compare the CNN to the proposed backbone *FCN* in Table II which learns local representations. For a fair comparison between CNN and FCN, we utilize the same meta-parameters and employ global supervision only (Stage I).

In Table III, we find that overall FCNs are more generalizable to unknown presentation attack instruments compared to global embeddings. For presentation attack instruments where the presentation attack affects the entire face, such as replay attacks, the differences between generalization performance of CNN and FCN are negligible. Here, presentation attack decisions at local spatial regions do not have any significant advantage over a single presentation attack decision over the entire image. Recall that CNNs employ a fully connected layer which strips away all spatial information. This explains why local decisions can significantly improve generalizability of FCN over CNN when presentation attack instruments are local in nature (*e.g.*, make-up attacks and partial attacks). Due to subtlety of obfuscation attacks and localized nature of paper glasses, FCN can exhibit a relative reduction in EER by 16% and 47%, respectively, relative to CNN.

B. Region Extraction Strategies

We considered 6 different region extraction strategies, namely, *Eye-Region*, *Nose-Region*, *Mouth-Region*, *Landmark-Region*, *Random-Crop*, and *Self-Supervised Regions (Proposed)* (see Figure 5). We also include results for a *Global* model which refers to training the FCN with the entire face image only (Stage I).

Since all face images are aligned and cropped, spatial information is consistent across all images in datasets. Therefore,

we can automatically extract facial regions that include eye, nose, and mouth regions (Figure 5a). We train the proposed FCN separately on each of the three regions to obtain three models: eye-region, nose-region, and mouth-region.

We also investigate extracting regions defined by face landmarks. For this, we utilize a state-of-the-art landmark extractor, namely DLIB [63], to obtain 68 landmark points. We exclude 17 landmarks defined around the jawline and define a subset of 51 landmark points around eyebrows (10 landmarks), eyes (12 landmarks), nose (9 landmarks), and mouth (20 landmarks). A total of 51 regions (with a fixed size 32×32) centered around each landmark are extracted and used to train a single FCN on all 51 regions.

Our findings are as follows: (i) almost all methods with regional supervision have lower overall error rates as compared to training with the entire face. Exception to this is when FCN is trained only on mouth regions. This is likely because a majority of presentation attack instruments may not contain presentation attack patterns across mouth regions (Figure 1). (ii) when both global and domain-knowledge strategies (specifically, eyes and nose) are fused, the generalization performance improves compared to the global model alone. Note that we do not fuse the mouth region since the performance is poor for mouth regions. Similarly, we find that regions cropped around landmarks when fused with the global classifier can achieve better generalization performance. (iii) sampling random local regions of the face also results in high error rates across the diverse set of presentation attack instruments. (iv) compared to all region extraction strategies, the proposed self-supervised region extraction strategy (Stage I \rightarrow Stage II) achieves the lowest generalization error rates across all presentation attack instruments with a 40% and 45% relative reduction in EER and ACER compared to the

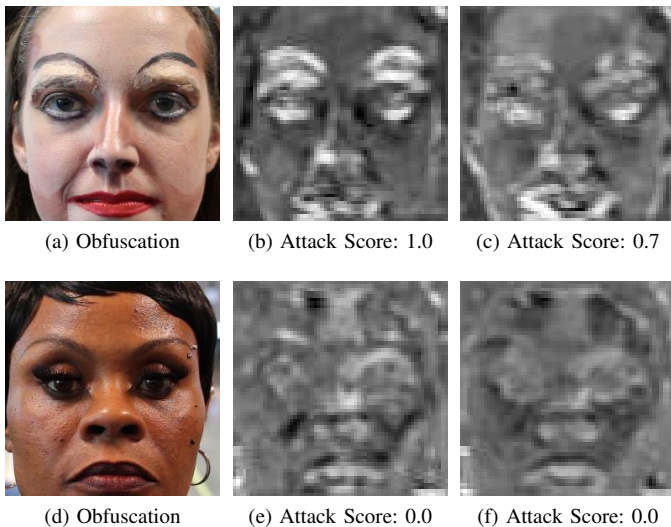


Fig. 6. (a) An example obfuscation presentation attack attempt where our network correctly predicts the input to be a presentation attack. (b, e) Score map output by our network trained via Self-Supervised Regions. (c, f) Score map output by FCN trained on entire face images. (d) An example obfuscation presentation attack attempt where our network *incorrectly* predicts the input to be a bonafide. Attack scores are given below the score maps. Decision threshold is 0.5.

Global model (Stage I). This supports our hypothesis that both Stage I and Stage II are required for enhanced generalization performance across unknown presentation attack instruments. A score-level fusion of the global FCN with self-supervised regions does not show any significant reduction in error rates. This is because we already trained the proposed FCN on global faces in Stage I.

The *Random-Crop* strategy can be viewed as a variant of training the proposed Stage II without any region-selection guidance from Stage I. In order to further investigate the benefit of the proposed self-supervision method, we train two models to distinguish between bonafide samples and *Paper Glasses* presentation attack samples: (i) *Random Crop* model is trained on patches randomly sampled from the input image such that the size of each region is between 64×64 and 256×256 , and (ii) *Self-Supervised Regions* model is trained on patches sampled from weakly labeled regions by the pre-trained model from Stage I. For a fair comparison, we do not employ the pre-trained model from Stage I to train the *Self-Supervised Regions* model. In Figure 7, we plot the training loss for both models over multiple training iterations. We can observe that the proposed self-supervised region method aids in network convergence. This is because random cropping can result in training with bonafide regions from presentation attack samples (see Figure 7), whereas, the proposed self-supervision method ensures that regions sampled from presentation attacks indeed contains presentation attack artifacts.

In Figure 6, we analyze the effect of training the FCN locally vs. globally on the prediction results. In the first row, where both models correctly predict the input to be a presentation attack, we see that FCN trained via random regions can correctly identify presentation attack regions such

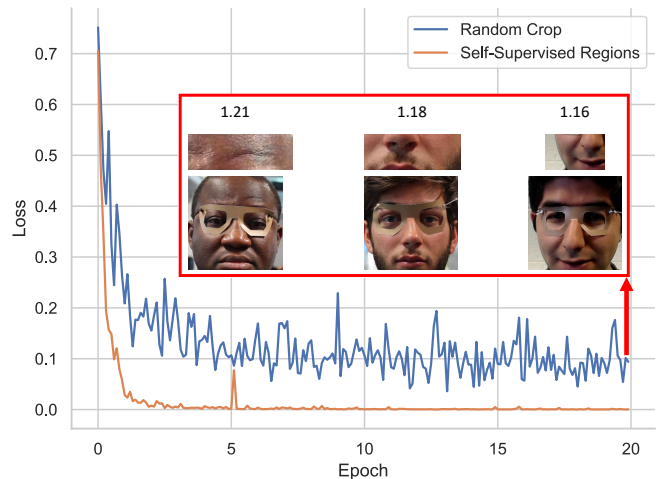


Fig. 7. Network convergence over a number of training iterations when a model trains on (a) randomly cropped patches (blue line), and (b) self-supervised regions extracted via pre-trained model from Stage I (orange line). Randomly cropping patches may result in noisy samples where bonafide samples from presentation attack samples may be used for training. Some example randomly cropped patches with high training loss are shown above the lines. Instead, we find that the proposed self-supervision aids in network convergence.

TABLE V
GENERALIZATION ERROR OF FCNs WITH RESPECT TO THE NUMBER OF TRAINABLE PARAMETERS.














Method	Metric (%)	Replay	Obfuscation	Paper Glasses	Mean \pm Std.
3-layers (76K)	ACER	13.9	57.6	12.3	27.9 ± 25.7
	EER	14.0	44.1	7.5	21.9 ± 19.5
5-layers (1.5M; proposed)	ACER	7.4	22.5	14.1	14.7 ± 7.6
	EER	6.8	17.8	13.5	12.7 ± 4.5
6-layers (3M)	ACER	11.2	32.9	19.8	21.3 ± 11.0
	EER	7.8	25.19	19.7	17.6 ± 7.2

as fake eyebrows. In contrast, the global FCN can barely locate the fake eyebrows. Since random regions increases the variability in the training set along with advantage of learning local features than global FCN, we find that proposed self-supervised regional supervision performs best.

C. Evaluation of Network Capacity

In Table V, we show the generalization performance of our model when we vary the capacity of the network. We consider three different variants of the proposed FCN: (a) 3-layer FCN (76K parameters), (b) 5-layer FCN (1.5M parameters; proposed), and (c) 6-layer FCN (3M parameters). This experiment is evaluated on three unknown presentation attack instruments, namely, *Replay*, *Obfuscation*, and *Paper Glasses*. We chose these presentation attack instruments due to their vastly diverse nature. Replay attacks consist of global presentation attack patterns, whereas obfuscation attacks are extremely subtle cosmetic changes. Paper Glasses are constrained only to eyes. While a large number of trainable parameters lead to poor generalization due to overfitting to the presentation attack instruments seen during training, whereas, too few parameters limits learning discriminative features. Based on this observation and experimental results, we utilize

TABLE VI
RESULTS ON SiW-M: UNKNOWN ATTACKS. HERE, EACH COLUMN REPRESENTS AN UNKNOWN PRESENTATION ATTACK INSTRUMENT WHILE THE METHOD IS TRAINED ON THE REMAINING 12 PRESENTATION ATTACK INSTRUMENTS.

Method	Metric	Replay	Print	Mask Attacks					Makeup Attacks			Partial Attacks			Mean \pm Std.
		 99 vids.	 118 vids.	 72 vids.	 27 vids.	 88 vids.	 17 vids.	 40 vids.	 23 vids.	 61 vids.	 50 vids.	 160 vids.	 127 vids.	 86 vids.	
SVM+LBP [18]	ACER	20.6	18.4	31.3	21.4	45.5	11.6	13.8	59.3	23.9	16.7	35.9	39.2	11.7	26.9 \pm 14.5
	EER	20.8	18.6	36.3	21.4	37.2	7.5	14.1	51.2	19.8	16.1	34.4	33.0	7.9	24.5 \pm 12.9
Auxiliary [1]	ACER	16.8	6.9	19.3	14.9	52.1	8.0	12.8	55.8	13.7	11.7	49.0	40.5	5.3	23.6 \pm 18.5
	EER	14.0	4.3	11.6	12.4	24.6	7.8	10.0	72.3	10.1	9.4	21.4	18.6	4.0	17.0 \pm 17.7
DTN [19]	ACER	9.8	6.0	15.0	18.7	36.0	4.5	7.7	48.1	11.4	14.2	19.3	19.8	8.5	16.8 \pm 11.1
	EER	10.0	2.1	14.4	18.6	26.5	5.7	9.6	50.2	10.1	13.2	19.8	20.5	8.8	16.1 \pm 12.2
CDC [20]	ACER	10.8	7.3	9.1	10.3	18.8	3.5	5.6	42.1	0.8	14.0	24.0	17.6	1.9	12.7 \pm 11.2
	EER	9.2	5.6	4.2	11.1	19.3	5.9	5.0	43.5	0.0	14.0	23.3	14.3	0.0	11.9 \pm 11.8
Proposed	ACER	7.4	19.5	3.2	7.7	33.3	5.2	3.3	22.5	5.9	11.7	21.7	14.1	6.4	12.4 \pm 9.2
	EER	6.8	11.2	2.8	6.3	28.5	0.4	3.3	17.8	3.9	11.7	21.6	13.5	3.6	10.1 \pm 8.4
	TDR*	72.0	51.0	96.0	55.9	39.0	100.0	95.0	31.0	90.0	44.0	33.0	42.9	94.7	65.0 \pm 25.9

*TDR evaluated at 2.0% FDR

TABLE VII
RESULTS ON SiW-M: KNOWN PRESENTATION ATTACK INSTRUMENTS.

Method	Metric (%)	Mask Attacks							Makeup Attacks			Partial Attacks			Mean \pm Std.
		Replay	Print	Half	Silicone	Trans.	Paper	Mann.	Obf.	Imp.	Cosm.	Funny Eye	Glasses	Paper Cut	
Auxiliary [1]	ACER	5.1	5.0	5.0	10.2	5.0	9.8	6.3	19.6	5.0	26.5	5.5	5.2	5.0	8.7 \pm 6.8
	EER	4.7	0.0	1.6	10.5	4.6	10.0	6.4	12.7	0.0	19.6	7.2	7.5	0.0	6.5 \pm 5.8
Proposed	ACER	3.5	3.1	1.9	5.7	2.1	1.9	4.2	7.2	2.5	22.5	1.9	2.2	1.9	4.7 \pm 5.6
	EER	3.5	3.1	0.1	9.9	1.4	0.0	4.3	6.4	2.0	15.4	0.5	1.6	1.7	3.9 \pm 4.4
	TDR*	55.5	92.3	69.5	100.0	90.4	100.0	85.1	92.5	78.7	99.1	95.6	95.7	76.0	87.0 \pm 13.0

*TDR evaluated at 2.0% FDR

the 5-layer FCN (see Table II) with approximately 1.5M parameters. A majority of prior studies employ 13 densely connected convolutional layers with trainable parameters exceeding 2.7M [1, 20, 56, 59].

D. Generalization across Unknown Attacks

The primary objective of this work is to enhance generalization performance across a multitude of unknown presentation attack instruments in order to effectively gauge the expected error rates in real-world scenarios. The evaluation protocol in SiW-M follows a leave-one-spoof-out testing protocol where the training split contains 12 different presentation attack instruments and the 13th presentation attack instrument is held out for testing. Among the bonafide videos, 80% are kept in the training set and the remaining 20% is used for testing bonafide. Note that there are no overlapping subjects between the training and testing sets. Also note that no data sample from the testing presentation attack instrument is used for validation since we evaluate our approach under unknown attacks. We report ACER and EER across the 13 splits. In addition to ACER and EER, we also report the TDR at 2.0% FDR.

In Table VI, we compare *SSR-FCN* with prior work. We find that our proposed method achieves significant improvement in comparison to the published results [20] (relative reduction of

14% on the average EER and 3% on the average ACER). Note that the standard deviation across all 13 presentation attack instruments is also reduced compared to prior approaches, even though some of them [1, 20] utilize auxiliary data such as depth and temporal information.

Specifically, we reduce the EERs of replay, half mask, transparent mask, silicone mask, paper mask, mannequin head, obfuscation, impersonation, and paper glasses relatively by 27%, 33%, 43%, 93%, 34%, 59%, and 6%, respectively. Among all the 13 presentation attack instruments, detecting obfuscation attacks is the most challenging. This is due to the fact that the makeup applied in these attacks are very subtle and majority of the faces are bonafide. Prior works were not successful in detecting these attacks and predict most of the obfuscation attacks as bonafide. By learning discriminative features locally, our proposed network improves the state-of-the-art obfuscation attack detection performance by 59% in terms of EER and 46% in terms of ACER.

E. SiW-M: Detecting Known Attacks

Here all the 13 presentation attack instruments in SiW-M are used for training as well as testing. We randomly split the SiW-M dataset into a 60%-40% training/testing split and report the results in Table VII. In comparison to a state-of-the-art face presentation attack detection method [1], our method

TABLE VIII
ERROR RATES (%) OF THE PROPOSED *SSR-FCN* AND AND COMPETING FACE PRESENTATION ATTACK DETECTORS UNDER THE FOUR STANDARD PROTOCOLS OF OULU-NPU [18].

Protocol	Method	APCER	BPCER	ACER
I	GRADIENT [64]	1.3	12.5	6.9
	Auxiliary [1]	1.6	1.6	1.6
	DeepPixBiS [56]	0.8	0.0	0.4
	TSCNN-ResNet [65]	5.1	6.7	5.9
	<i>SSR-FCN</i> (Proposed)	1.5	7.7	4.6
II	GRADIENT [64]	3.1	1.9	2.5
	Auxiliary [1]	2.7	2.7	2.7
	DeepPixBiS [56]	11.4	0.6	6.0
	TSCNN-ResNet [65]	7.6	2.2	4.9
	<i>SSR-FCN</i> (Proposed)	3.1	3.7	3.4
III	GRADIENT [64]	2.1 ± 3.9	5.0 ± 5.3	3.8 ± 2.4
	Auxiliary [1]	2.7 ± 1.3	3.1 ± 1.7	2.9 ± 1.5
	DeepPixBiS [56]	11.7 ± 19.6	10.6 ± 14.1	11.1 ± 9.4
	TSCNN-ResNet [65]	3.9 ± 2.8	7.3 ± 1.1	5.6 ± 1.6
	<i>SSR-FCN</i> (Proposed)	2.9 ± 2.1	2.7 ± 3.2	2.8 ± 2.2
IV	GRADIENT [64]	5.0 ± 4.5	15.0 ± 7.1	10.0 ± 5.0
	Auxiliary [1]	9.3 ± 5.6	10.4 ± 6.0	9.5 ± 6.0
	DeepPixBiS [56]	36.7 ± 29.7	13.3 ± 16.8	25.0 ± 12.7
	TSCNN-ResNet [65]	11.3 ± 3.9	9.7 ± 4.8	9.8 ± 4.2
	<i>SSR-FCN</i> (Proposed)	8.3 ± 6.8	13.3 ± 8.7	10.8 ± 5.1

outperforms for almost all of the individual presentation attack instruments as well as the overall performance across presentation attack instruments. *Auxiliary* [1] utilizes depth and temporal information for presentation attack detection which adds significant complexity to the network. We find that characterizing local spatial regions as bonafide/presentation attack in fact leads to better generalization on unknown attacks (see Table VI) and specialization on known attacks.

F. Evaluation on Oulu-NPU Dataset

We follow the four standard protocols defined in the OULU-NPU dataset [18] which cover the cross-background, cross-presentation-attack-instrument (cross-PAI), cross-capture-device, and cross-conditions evaluations:

- **Protocol I:** unseen subjects, illumination, and backgrounds;
- **Protocol II:** unseen subjects and attack devices;
- **Protocol III:** unseen subjects and cameras;
- **Protocol IV:** unseen subjects, illumination, backgrounds, attack devices, and cameras.

We compare the proposed *SSR-FCN* with the best performing method, namely GRADIENT [64], in IJCB Mobile Face Anti-Spoofing Competition [64] for each protocol. We also include some newer baseline methods, including Auxiliary [1], DeepPixBiS [56], and TSCNN [65]. We compare our proposed method with 10 baselines in total for each protocol. Additional baselines can be found in supplementary material.

In Table VIII, *SSR-FCN* achieves ACERs of 4.6%, 3.4%, 2.8%, and 10.8% in the four protocols, respectively. Among

TABLE IX
CROSS-DATASET HTER (%) OF THE PROPOSED *SSR-FCN* AND COMPETING FACE PRESENTATION ATTACK DETECTORS.

Method	CASIA → Replay	Replay → CASIA
CNN [44]	48.5	45.5
Color Texture [34]	47.0	49.6
FaceSpoofBuster [66]	43.3	53.0
Auxiliary [1]	27.6	28.4
De-Noising [57]	28.5	41.1
Damer & Dimitrov [67]	28.4	38.1
STASN [68]	31.5	30.9
SAPLC [59]	27.3	37.5
<i>SSR-FCN</i> (Proposed)	19.9	41.9

“CASIA → Replay” denotes training on CASIA and testing on Replay-Attack

the baselines, *SSR-FCN* even outperforms prevailing state-of-the-art methods in protocol III which corresponds to generalization performance for unseen subjects and cameras. The results are comparable to baseline methods in the other three protocols. Since Oulu-NPU comprises of only print and replay attacks, a majority of the baseline methods incorporate auxiliary information such as depth and motion. Indeed, incorporating auxiliary information could improve the results at the risk of overfitting and overhead cost and time.

G. Cross-Dataset Generalization

In order to evaluate the generalization performance of *SSR-FCN* when trained on one dataset and tested on another, following prior studies, we perform a cross-dataset experiment between CASIA-FASD [13] and Replay-Attack [12].

In Table IX, we find that, compared to 6 prevailing state-of-the-art methods, the proposed *SSR-FCN* achieves the lowest error (a 27% improvement in HTER) when trained on CASIA-FASD [13] and evaluated on Replay-Attack [12]. On the other hand, *SSR-FCN* achieves worse performance when trained on Replay-Attack and tested on CASIA-FASD. This can likely be attributed to higher resolution images in CASIA-FASD compared to Replay-Attack. This demonstrates that *SSR-FCN* trained with higher-resolution data can generalize better on poorer quality testing images, but the reverse may not hold true. We intend on addressing this limitation in future work.

Additional baselines can be found in supplementary material.

H. Failure Cases

Even though experiment results show enhanced generalization performance, our model still fails to correctly classify certain input images. In Figure 8, we show a few such examples.

Figure 8a shows incorrect prediction of bonafides as presentation attacks in the presence of inhomogeneous illumination. This is because the model predicts bonafide as being one of replay and print attacks which exhibit bright lighting patterns due to the recapturing media such as smartphones and laptops.

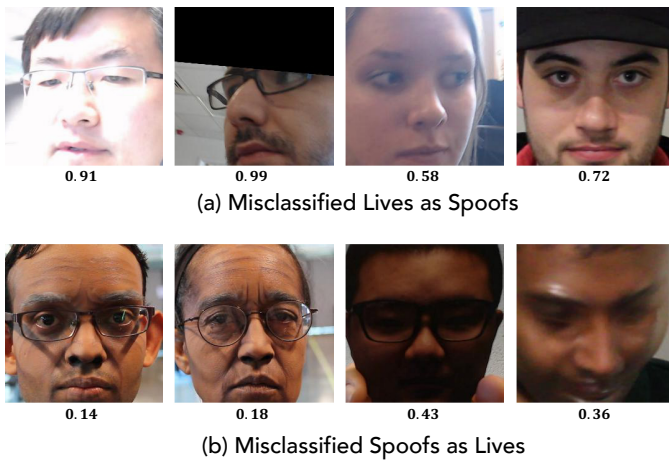


Fig. 8. Example cases where the proposed framework, *SSR-FCN*, fails to correctly classify bonafides and presentation attacks. (a) Bonafides are misclassified as presentation attacks likely due to bright lighting and occlusions in face regions. (b) Presentation attacks misclassified as bonafides due to the subtle nature of make-up attacks and transparent masks. Corresponding attack scores ($\in [0, 1]$) are provided below each image. Larger value of attack score indicates a higher likelihood that the input image is a presentation attack. Decision threshold is 0.5.

Since we fine-tune our network via regional supervision in Stage II, artifacts that obstruct parts of the faces can also adversely affect our model.

Figure 8b shows incorrect classification of presentation attack as bonafides. This is particularly true when presentation attack artifacts are very subtle, such as cosmetic and obfuscation make-up attacks. Transparent masks can also be problematic when the mask itself is barely visible.

I. Computational Requirement

Since face presentation attack detection modules are employed as a pre-processing step for automated face recognition systems, it is crucial that the presentation attack prediction time should be as low as possible. The proposed approach comprises of 1.5M trainable parameters compared to a traditional CNN [1] with 3M learnable parameters. The proposed *SSR-FCN* takes under 2 hours to train both Stage I and Stage II, and 4 milliseconds to predict a single (256×256) presentation attack/bonafide image on a Nvidia GTX 1080Ti GPU. In other words, *SSR-FCN* can process frames at 250 Frames Per Second (FPS) and the size of the model is only 11.8MB. Therefore, *SSR-FCN* is well suited for deployment where real-time decisions are required.

J. Visualizing Presentation Attack Regions

SSR-FCN can automatically locate the individual presentation attack regions in an input face image. In Figure 9, we show heatmaps from the score maps extracted for a randomly chosen image from all presentation attack instruments. Red regions indicate a higher likelihood of presentation attack.

For a bonafide input image, the presentation attack regions are sparse with low likelihoods. In the case of replay and print attacks, the predicted presentation attack regions are located

throughout the entire face image. This is because these presentation attacks contain global-level noise. For mask attacks, including half-mask, silicone mask, transparent mask, paper mask, and mannequin, the presentation attack patterns are identified near the eye and nose regions. Make-up attacks are harder to detect since they are very subtle in nature. Proposed *SSR-FCN* detects obfuscation and cosmetic attack attempts by learning local discriminative cues around the eyebrow regions. In contrast, impersonation make-up patterns exist throughout the entire face. We also find that *SSR-FCN* can precisely locate the presentation attack artifacts, such as funny eyeglasses, paper glasses, and paper cut, in partial attacks.

VII. DISCUSSION

We show that the proposed *SSR-FCN* achieves superior generalization performance on SiW-M dataset [19] compared to the prevailing state-of-the-art methods that tend to overfit on the seen presentation attack instruments. Our method also achieves comparable performance to the state-of-the-art in Oulu-NPU dataset [18] and outperforms all baselines for cross-dataset generalization performance (CASIA-FASD [13] \rightarrow Replay-Attack [12]).

In contrast to a number of prior studies [1, 20, 34, 55], the proposed approach does not utilize auxiliary cues for presentation attack detection, such as motion and depth information. While incorporating such cues may enhance performance on print and replay attack datasets such as Oulu-NPU, CASIA-MFSD, and Replay-Attack, it is at the risk of potentially overfitting to the two attacks and compute cost. A major benefit of *SSR-FCN* lies in its usability. A simple pre-processing step includes face detection and alignment. The cropped face is then passed to the network. With a *single* forward-pass through the FCN, we obtain both the score map and the final attack score.

Our proposed *SSR-FCN* outputs a bonafide/presentation attack decision at each pixel of intermediate feature maps. Due to the downsampling layers found after each convolutional operation (see Table II), the decisions are automatically aggregated. Therefore, the final feature map (referred to as *score map*) is of much smaller resolution (16×16 pixels) compared to the original image. Therefore, in the case of partial attacks such as paper eyeglasses, even though a small portion of the face image comprises of a presentation attack artifact, our final score map is an aggregated decision across multiple sliding windows (receptive field) of the original image. In Figure 10, we compute the average score map for all the paper eyeglasses presentation attacks. We find that, even though paper eyeglasses comprise of a small portion of the original image, majority of the pixels in the average *score map* comprise of high scores (indicating the presence of a presentation attack). In this paper, we obtain the final score via average pooling the score map. As future work, we intend on exploring other fusion mechanisms such as a weighted average via an attention mask.

Even though the proposed method is well-suited for generalizable face presentation attack detection, *SSR-FCN* is still limited by the amount and quality of available training data.

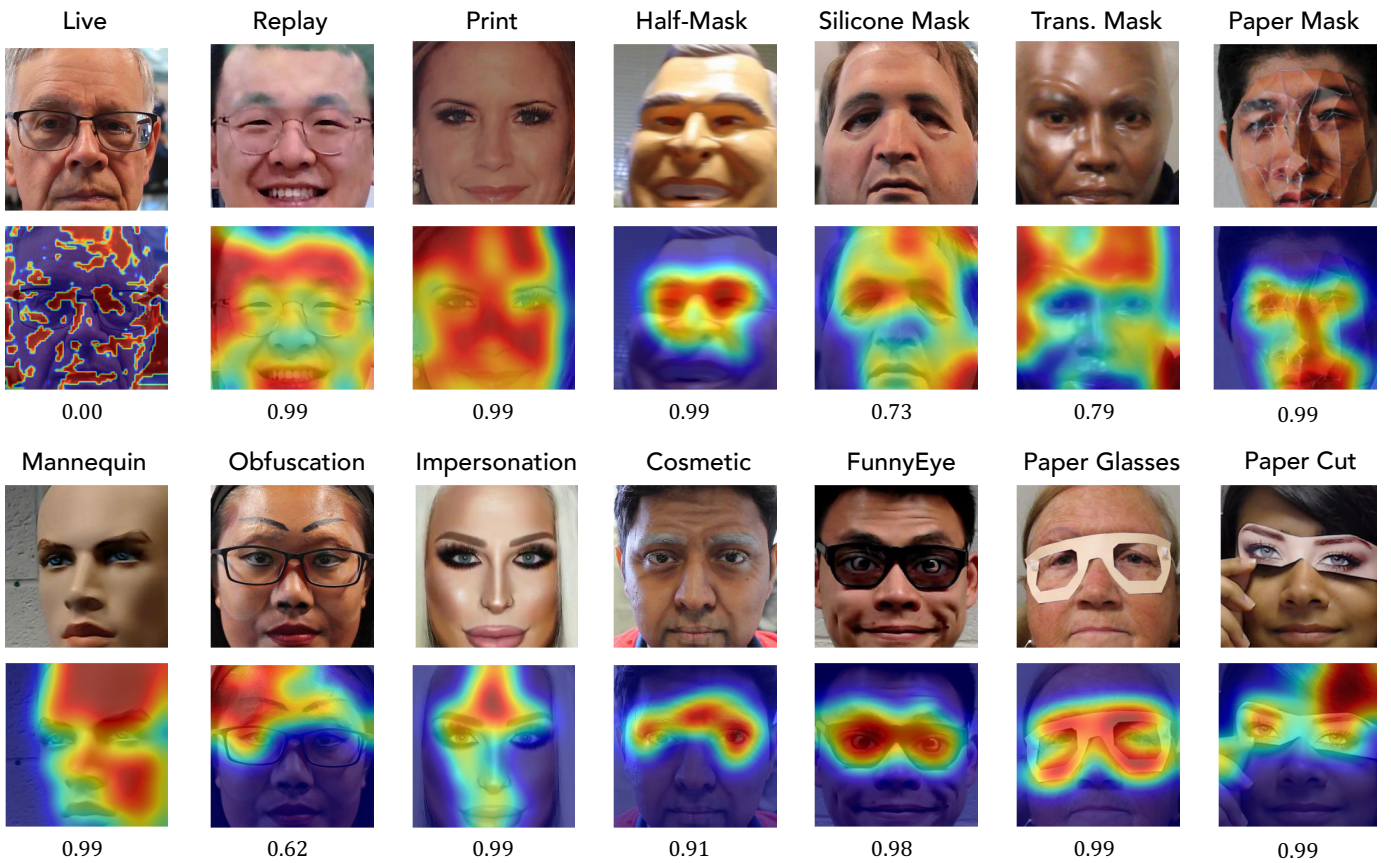


Fig. 9. Visualizing presentation attack regions via the proposed *SSR-FCN*. Red regions indicate higher likelihood of being a presentation attack region. Corresponding attack scores ($\in [0, 1]$) are provided below each image. Larger value of attack score indicates a higher likelihood that the input image is a presentation attack. Decision threshold is 0.5.

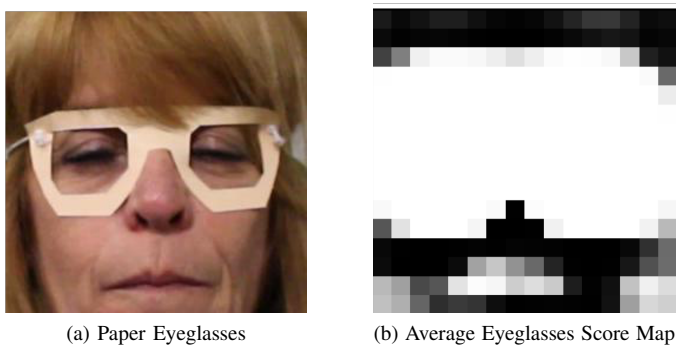


Fig. 10. A partial presentation attack artifact may be present in a small portion of the input 256×256 face image, such as (a) paper eyeglasses. However, since the proposed *SSR-FCN* dynamically aggregates decisions across multiple receptive fields in the image, a majority of the pixels in the final score map comprise of high scores (indicating the presence of a presentation attack). We visualize the average score map across all paper eyeglass attacks in (b).

For instance, when trained on a low-resolution dataset, namely Replay-Attack [12], cross-dataset generalization performance suffers.

VIII. CONCLUSION

Face presentation attack detection systems are crucial for secure operation of an automated face recognition system.

With the introduction of sophisticated presentation attacks, such as high resolution and tight fitting silicone 3D face masks, presentation attack detectors need to be robust and generalizable. We proposed a face presentation attack detection framework, namely *SSR-FCN*, that achieved state-of-the-art generalization performance against 13 different presentation attack instruments. *SSR-FCN* reduced the average error rate of competitive algorithms by 14% on one of the largest and most diverse face presentation attack detection dataset, SiW-M, comprised of 13 presentation attack instruments. It also generalizes well when training and testing datasets are from different sources. In addition, the proposed method is shown to be more interpretable compared to prior studies since it can directly predict the parts of the faces that are considered as presentation attacks. In the future, we intend on exploring whether incorporating domain knowledge in *SSR-FCN* can further improve generalization performance.

ACKNOWLEDGMENT

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2017 – 17020200004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA,

or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

- [1] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. CVPR*, Salt Lake City, UT, June 2018, pp. 389–398. 1, 3, 4, 5, 7, 10, 11, 12
- [2] P. J. Grother, M. Ngan, and K. Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT)," NIST, Intragency Report, 2020. [Online]. Available: https://pages.nist.gov/frvt/reports/1N/frvt_1N_report.pdf 1
- [3] International Standards Organization, "Information Technology Biometric Presentation Attack Detection Part 1: Framework," ISO/IEC 30107-1:2016, 2016. [Online]. Available: <https://www.iso.org/standard/53227.html> 1, 7
- [4] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time Face Capture and Reenactment of RGB Videos," in *Proc. CVPR*, 2016, pp. 2387–2395. 1
- [5] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces: Adversarial Face Synthesis," *arXiv:1908.05008*, 2019. [Online]. Available: <https://arxiv.org/abs/1908.05008> 1
- [6] S. Marcel, M. S. Nixon, and S. Z. Li, *Handbook of Biometric Anti-Spoofing*. Springer, 2014, vol. 1. 1
- [7] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 7, pp. 1713–1723, Mar. 2017. 1, 4
- [8] Y. Xu, T. Price, J.-M. Frahm, and F. Monrose, "Virtual U: Defeating Face Liveness Detection by Building Virtual Models from Your Public Photos," in *USENIX*, 2016, pp. 497–512. 1
- [9] FastPass, "FastPass - a harmonized, modular reference system for all European automated bordercrossing points." [Online]. Available: <https://www.fastpass-project.eu> 1
- [10] Daily Mail, "Police arrest passenger who boarded plane in Hong Kong as an old man in flat cap and arrived in Canada a young Asian refugee," 2011. [Online]. Available: <http://dailymail.com/2011/08/21/Police-arrest-passenger-who-boarded-plane-in-Hong-Kong-as-an-old-man-in-flat-cap-and-arrived-in-Canada-a-young-Asian-refugee/> 2
- [11] The Verge, "This \$150 mask beat Face ID on the iPhone X," 2017. [Online]. Available: <https://bit.ly/300bRoC> 2
- [12] I. Chingovska, A. Anjos, and S. Marcel, "On the Effectiveness of Local Binary Patterns in Face Anti-spoofing," in *Proc. BIOSIG*, 2012. 3, 7, 11, 12, 13
- [13] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face anti-spoofing database with diverse attacks," in *Proc. ICB*, 2012, pp. 26–31. 3, 4, 7, 11, 12
- [14] N. Erdogmus and S. Marcel, "Spoofing in 2D Face Recognition with 3D Masks and Anti-spoofing with Kinect," in *Proc. BTAS*, 2013. 3
- [15] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 746–761, Feb. 2015. 3, 4
- [16] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The REPLAY-MOBILE Face Presentation-Attack Database," in *Proc. BIOSIG*, 2016. 3, 4
- [17] S. Liu, B. Yang, P. C. Yuen, and G. Zhao, "A 3D Mask Face Anti-Spoofing Database with Real World Variations," in *Proc. CVPR*, 2016. 3
- [18] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. FG*, 2017, pp. 612–618. 3, 7, 10, 11, 12
- [19] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep Tree Learning for Zero-Shot Face Anti-Spoofing," in *Proc. CVPR*, 2019, pp. 4675–4684. 3, 4, 6, 7, 10, 12
- [20] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching Central Difference Convolutional Networks for Face Anti-Spoofing," in *Proc. CVPR*, 2020, pp. 5294–5304. 3, 4, 5, 10, 12
- [21] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun, "Real-time face detection and motion analysis with application in "liveness" assessment," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 548–558, Aug. 2007. 3
- [22] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcam," in *Proc. CVPR*, 2007, pp. 1–8.
- [23] K. Patel, H. Han, and A. K. Jain, "Cross-database face anti-spoofing with robust feature representation," in *Proc. CCBP*, 2016, pp. 611–619. 4
- [24] R. Shao, X. Lan, and P. C. Yuen, "Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3d mask face anti-spoofing," in *Proc. IJCB*, 2017, pp. 748–755. 3
- [25] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *Proc. BTAS*, 2013. 3
- [26] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "LBP-TOP based countermeasure against face spoofing attacks," in *Proc. ACCV*, 2012, pp. 121–132. 3
- [27] L. Li, Z. Xia, A. Hadid, X. Jiang, F. Roli, and X. Feng, "Face presentation attack detection in learned color-like space," *arXiv:1810.13170*, 2018. [Online]. Available: <https://arxiv.org/abs/1810.13170> 3
- [28] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 10, pp. 2268–2283, Jun. 2016. 3
- [29] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 710–724, Feb. 2014. 3
- [30] H. Li, S. Wang, and A. C. Kot, "Face spoofing detection with image quality regression," in *Proc. IPTA*, 2016, pp. 1–6. 3
- [31] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of fourier spectra," in *Proc. SPIE*, vol. 5404, 2004, pp. 296–303. 3
- [32] T. Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Can face anti-spoofing countermeasures work in a real world scenario?" in *Proc. ICB*, 2013, pp. 1–8. 3
- [33] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *Proc. IJCB*, 2011, pp. 1–7.
- [34] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 8, pp. 1818–1830, Aug. 2016. 3, 11, 12
- [35] J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection with component dependent descriptor," in *Proc. ICB*, 2013, pp. 1–6. 3
- [36] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Proc. ECCV*, 2010, pp. 504–517. 3
- [37] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing using speeded-up robust features and fisher vector encoding," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 141–145, Feb. 2017. 3
- [38] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection using 3d structure recovered from a single camera," in *Proc. ICB*, 2013. 3
- [39] Y. Wang, F. Nian, T. Li, Z. Meng, and K. Wang, "Robust face anti-spoofing with depth information," *J. Vis. Commun. Image Represent.*, vol. 49, pp. 332–337, Sep. 2017.
- [40] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "CASIA-SURF: A Dataset and Benchmark for Large-scale Multi-modal Face Anti-Spoofing," *arXiv:1812.00408*, 2019. [Online]. Available: <https://arxiv.org/abs/1812.00408> 3
- [41] V. Conotter, E. Bodnari, G. Boato, and H. Farid, "Physiologically-based detection of computer generated faces in video," *Proc. ICIP*, pp. 248–252, Jan. 2015. 3
- [42] Z. Zhang, D. Yi, Z. Lei, and S. Li, "Face liveness detection by learning multispectral reflectance distributions," in *Proc. FG*, 2011, pp. 436–441. 3
- [43] G. Chetty, "Biometric liveness checking using multimodal fuzzy fusion," in *Proc. WCCI*, 07 2010, pp. 1–8. 3
- [44] J. Yang, Z. Lei, and S. Z. Li, "Learn Convolutional Neural Network for Face Anti-Spoofing," *arXiv:1408.5601*, 2014. [Online]. Available: <https://arxiv.org/abs/1408.5601> 4, 11
- [45] N. N. Lakshminarayana, N. Narayan, N. Napp, S. Setlur, and V. Govindaraju, "A discriminative spatio-temporal mapping of face for liveness detection," in *Proc. ISBA*, 2017, pp. 1–7. 4
- [46] X. Tu and Y. Fang, "Ultra-deep neural network for face anti-spoofing," in *Proc. NIPS*, 2017, pp. 686–695.
- [47] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo, "Transfer learning using convolutional neural networks for face anti-spoofing," in *Proc. ICIAR*, 2017, pp. 27–34.
- [48] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *Proc. IPTA*, 2016, pp. 1–6. 4
- [49] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13 868–13 882, 2017. 4
- [50] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel, "On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing," in *Proc. ICB*, 2018, pp. 75–81. 4
- [51] D. Pérez-Cabo, D. Jiménez-Cabello, A. Costa-Pazo, and R. J. López-

- Sastre, "Deep anomaly detection for generalized face anti-spoofing," in *Proc. CVPR*, 2019. 4
- [52] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 7, pp. 1794–1809, 2018. 4
- [53] S. R. Arashloo, J. Kittler, and W. Christmas, "Face spoofing detection based on multiple descriptor fusion using multiscale dynamic binarized statistical image features," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 11, pp. 2396–2407, 2015.
- [54] J. Yang, Z. Lei, D. Yi, and S. Z. Li, "Person-specific face antispoofing with subject domain adaptation," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 797–809, 2015. 4
- [55] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in *Proc. IJCB*, 2017, pp. 319–328. 4, 5, 12
- [56] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," *arXiv:1907.04047*, 2019. [Online]. Available: <https://arxiv.org/abs/1907.04047> 4, 5, 10, 11
- [57] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Proc. ECCV*, 2018, pp. 290–306. 4, 11
- [58] C. Nagpal and S. R. Dubey, "A performance evaluation of convolutional neural networks for face anti spoofing," in *Proc. IJCNN*, 2019, pp. 1–8. 4
- [59] W. Sun, Y. Song, C. Chen, J. Huang, and A. C. Kot, "Face spoofing detection based on local ternary label supervision in fully convolutional networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3181–3196, 2020. 4, 5, 10, 11
- [60] W. Sun, Y. Song, H. Zhao, and Z. Jin, "A face spoofing detection method based on domain adaptation and lossless size adaptation," *IEEE Access*, vol. 8, pp. 66553–66563, 2020. 5
- [61] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with MTCNN," in *Proc. ICISCE*, 2017, pp. 424–427. 7
- [62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014. [Online]. Available: <https://arxiv.org/abs/1412.6980> 7
- [63] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009. 8
- [64] Z. Boulkenafet, J. Komulainen, Z. Akhtar, A. Benlamoudi, D. Samai, S. E. Bekhouche, A. Ouafi, F. Dornaika, A. Taleb-Ahmed, L. Qin *et al.*, "A competition on generalized software-based face presentation attack detection in mobile scenarios," in *Proc. IJCB*, 2017, pp. 688–696. 11
- [65] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, "Attention-Based Two-Stream Convolutional Networks for Face Spoofing Detection," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 578–593, Jun. 2019. 11
- [66] R. Bresan, A. Pinto, A. Rocha, C. Beluzo, and T. Carvalho, "Facespoofer: a presentation attack detector based on intrinsic image properties and deep learning," *arXiv:1902.02845*, 2019. [Online]. Available: <https://arxiv.org/abs/1902.02845> 11
- [67] N. Damer, K. Dimitrov, R. Wilson, E. Hancock, and W. Smith, "Practical view on face presentation attack detection," in *Proc. BMVC*, 2016. 11
- [68] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu, "Face anti-spoofing: Model matters, so does data," in *Proc. CVPR*, 2019, pp. 3507–3516. 11



Indian National Academy of Engineering and Chinese Academy of Sciences.

Anil K. Jain is a University distinguished professor in the Department of Computer Science and Engineering at Michigan State University. His research interests include pattern recognition and biometric authentication. He served as the editor-in-chief of the IEEE Transactions on Pattern Analysis and Machine Intelligence and was a member of the United States Defense Science Board. He has received Fulbright, Guggenheim, Alexander von Humboldt, and IAPR King Sun Fu awards. He is a member of the National Academy of Engineering and foreign fellow of the



Debayan Deb received his B.S. degree in computer science from Michigan State University, East Lansing, Michigan, in 2016. He is currently working towards a PhD degree in the Department of Computer Science and Engineering at Michigan State University, East Lansing, Michigan. His research interests include pattern recognition, computer vision, and machine learning with applications in biometrics.