This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2022.3169435, IEEE Transactions on Circuits and Systems II: Express Briefs

An Adaptive Spatio-temporal Global Sampling for Presentation Attack Detection

Usman Muhammad, Jiehua Zhang, Li Liu, and Mourad Oussalah *Senior, Member, IEEE*

***Abstract*—Without developing dedicated countermeasures, facial biometric systems can be spoofed with printed photos, replay attacks, silicone masks, or even a 3D mask of a targeted person. Thus, the threat of presentation attacks needs to be addressed to strengthen the security of the biometric systems. Since a 2D convolutional neural network (CNN) captures static features from video frames, the camera motion might hinders the performance of modern CNNs for video-based presentation attack detection (PAD). Inspired by the egomotion theory, we introduce an adaptive spatiotemporal global sampling (ASGS) technique to compensate the camera motion and use the resulting estimation to encode the appearance and dynamics of the video sequences into a single RGB image. This is achieved by adaptively splitting the video into small segments and capturing their global motion within each segment. The proposed global motion is estimated based on four key steps: dense sampling, FREAK feature extraction and matching, similarity transformation, and aggregation function. This allows using deep models pre-trained on images for video-based PAD detection. Moreover, the interpretation of ASGS reveals that the most important parts for supporting the decision on PAD are consistent with motion cues associated with the artifacts, i.e., hand movement, material reflection, and expression changes. Extensive experiments on four standard face PAD databases demonstrate its effectiveness and encourage further study in this domain.**

***Index Terms*—Dense sampling, Image warping, Face recognition, Global Motion, Presentation Attack Detection, Deep learning.**

I. Introduction

FACIAL recognition technology has been successfully applied in numerous real-world applications, such as mobile payments, automated teller machines (ATMs), automatic border control, and surveillance. However, there are various physical and digital attacks, such as face manipulation attacks (e.g., deepfake, face swap) [1], face morphing [2], face adversarial attacks [3], and face spoofing (i.e., presentation attacks) [4], that can be utilized to spoof the biometric systems. Thus, designing reliable approaches for presentation attack detection (PAD) is vital to enhance the security of face recognition systems.

The main issue for face PAD is to extract discriminative features to distinguish a bona fide face from an attack presentation. In the past few years, deep learning techniques [5], [6], have demonstrated significant improvements over traditional texture-based methods [7], [8] on several large-scale PAD databases. For instance, a novel framework based on central difference convolution (CDC) is introduced [9], which encapsulates intrinsic detailed patterns via accumulating both intensity and gradient information. Despite the success of deep learning methods, domain shift or domain adaptation (DA) is one of the main challenges that still need to be addressed. This refers to the degraded system performance when the PAD model is trained or tuned on the source domain and then tested on a completely unseen database (target domain) [10].

One way to address the problem of domain shift consists in using temporal feature learning. Existing works in this area can be roughly categorized into three streams: (i) extracting dynamic features through CNN network, *e.g.,* using optical flow, (ii) extracting Spatio-temporal features based on 3D CNN, and (iii) learning long-range (sequential) data *e.g.,* through Recurrent Neural Networks (RNN) [11]. Although there are other multimodal-based methods [12], [11], such as meta-teacher learning, self-supervised learning , or single-shot face anti-spoofing, these methods require a careful domain adaptation to transfer the knowledge from one domain to another. Moreover, these methods focus on enhancing the generalization ability of PAD methods from the perspective of domain generalization (DG), which intends to train a model by exploiting multiple available source domains without viewing any target data [13]. However, seeking a generalized feature space for the spoofed faces always remains challenging. Motion patterns in PAD videos contained in bona fide and attack videos are different. In bona fide face, the motion cues

A B

Fig. 1. Illustration of the black framing of the clip in CASIA dataset [23]. (A) We observe black framing issue at the boundary of the image in most of the estimated clips in [11]. (B) Our proposed method minimizes the black framing issue. This black framing provides artificial cues for PAD and should be removed [32].

Manuscript received April XX, XXXX; revised August XX, XXX. This work is conducted at the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland.

Usman Muhamamd, Jiehua Zhang, Mourad Oussalah are with Centre for Machine Vision and Signal Analysis (CMVS), Faculty of Information Technology and Electrical Engineering (ITEE), University of Oulu, Finalnd. Li Liu is with the National University of Defense Technology of China and also with the Center for Machine Vision and Signal analysis (CMVS), University of Oulu, Oulu, Finland. (e-mail: firstname.lastname@oulu.fi).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2022.3169435, IEEE Transactions on Circuits and Systems II: Express Briefs

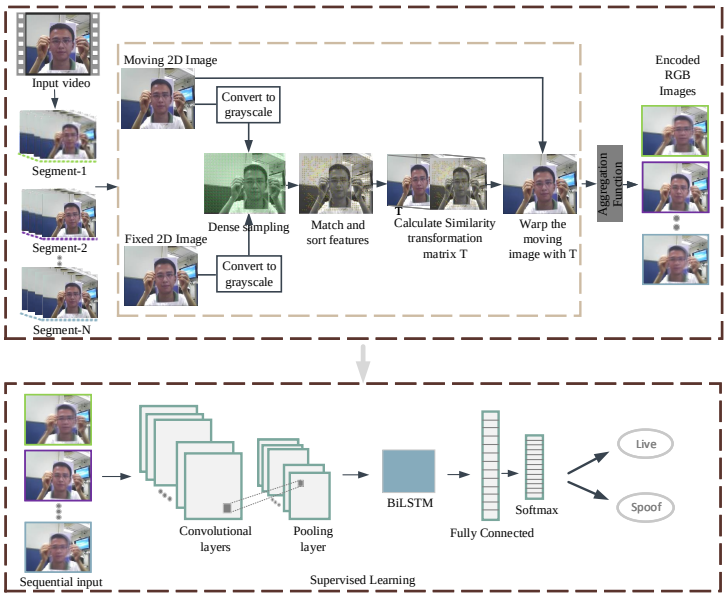

Fig. 2. Flow chart of our proposed method. Given a video of length S, we divide it into non-overlapping segments of smaller length s. ASGS estimates global motion by using dense keypoints matching and similarity transformation between consecutive frames. Using an aggregation function, we obtain a single RGB image from each segment of the video. Finally, off-the-shelf CNN features are provided to BiLSTM for final PAD detection.

depend on different factors such as eye blinking, head rotation, and mouth development. On the other hand, print attacks are mainly due to hand movement or material reflection, and artifacts caused by screen sloping are visible in a replay attack [6]. These motion cues are useful and important to analyze live and spoofed attacks where the relative motion between the background and the face region can be vital. However, extracting these motion cues remain difficult in many PAD databases [8], [14] due to noisy camera movements. Since the video data include not only camera movements but also motion information of the objects along the time axis, we need to determine whether the actual actions (e.g., eye blinking, hand trembling, head rotation) are caused by the objects or from the noisy (camera) motions. Thus, we argue that existing works do not explicitly focus on compensating this effect which might degrade important details to analyze live and spoofed attacks. Moreover, each PAD video might comprise hundreds to thousands of frames, not all of which are valuable. It makes not only the training much slower but also hard to extract meaningful information for the CNNs. Besides, the fixed-size spatio-temporal windows of analysis make the frame orders do not influence the performance of 2D CNN. Thus, effective handling of such spatiotemporal variations is pivotal to enhance the performance of PAD detection.

Inspired by the above discussion, our key idea is to cope with global (camera) motion while proposing a new video representation method that distills the motion information contained in video sequences into a single RGB image. To achieve this, the video is partitioned into a set of non-overlapping segments. Then, local features are extracted independently for each segment and their trajectory is evaluated to estimate inter-frame motion. Finally, an aggregation function is used to capture the gist of the dynamics and to encode the appearance information of the video sequences into a compact image. Intuitively, this idea has at least four main advantages. First, the encoded RGB image can be fed to any CNN architecture for a still image, where "still" captures the long-term dynamics in the video. Second, the encoded images reduce the analysis of video sequences to the analysis of a single RGB image and make the CNN model computationally attractive when accessing only a few images will be enough for subsequent analysis, instead of all frames during both training and test phases. Third, since the local motion vectors are calculated at consecutive frames using image registration, this aggregation contains motion cues without still spatial distribution information, which decreases the risk of over-fitting of human faces. Fourth, the representation provides a fast approximation in comparison to the expensive optical flow and allows the CNN model to enlarge the temporal receptive field with respect to a fixed-size temporal window.

In summary, our key contributions can be summarized in: (i) we propose a novel method called adaptive spatiotemporal global sampling (ASGS) for video representation, which encodes the appearance and dynamics of video sequences into a single RGB image. (ii) To model temporal correlation across multiple encoded images at different time steps, a unified CNN-BiLSTM is suggested to make full use of the motion cues across video frames for presentation attack detection. (iii) The effectiveness of the proposed approach is demonstrated on four PAD databases and the results show that our proposed method provides a state-of-the-art performance on three publicly available databases.

This work builds on our preliminary findings reported in [11]. More specifically, we extend [11] by (i) we propose to use dense representation and utilize similarity transformation that helps to minimize the black framing issue as shown in Fig.1 and provides good coverage of image features for improving the robustness of homography estimation; (ii) we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2022.3169435, IEEE Transactions on Circuits and Systems II: Express Briefs

TABLE I
PERFORMANCE EVALUATION USING MSU-MFSD (M), IDIAP (I), CASIA (C) AND OULU-NPU (O) DATABASES. COMPARISON RESULTS ARE OBTAINED FROM [12].

	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
Method	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
IDA [8]	66.67	27.86	55.17	39.05	28.35	78.25	54.20	44.59
Color Texture [7]	28.09	78.47	30.58	76.89	40.40	62.78	63.59	32.71
LBP-TOP [8]	36.90	70.80	42.60	61.05	49.45	49.54	53.15	44.09
Auxiliary [12]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
MADDG [25]	17.69	88.06	24.50	84.51	22.19	84.99	27.89	80.02
DAFL [10]	14.58	92.58	17.41	90.12	15.13	95.76	14.72	93.08
DR-MD [15]	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47
SSDG-R [13]	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54
MA-Net [16]	20.80	-	25.60	-	24.70	-	26.30	-
RFMetaFAS [5]	13.89	93.98	20.27	88.16	17.3	90.48	16.45	91.16
FAS-DR-BC(MT) [12]	11.67	93.09	18.44	89.67	11.93	94.95	16.23	91.18
ASGS (Ours)	**5.91**	**99.88**	**10.21**	**99.86**	45.84	76.09	**13.54**	**99.73**

provide extensive evaluations, especially from the perspective of the domain generalization (DG) with baselines for PAD, such as DAFL [10], DR-MD [15], SSDG-R [13], MA-Net [16], RFMetaFAS [5], and FAS-DR-BC(MT) [12]; (iii) we use off-the-shelf CNN features to avoid initializing pre-trained weights that cause increasing the computational resources; and (iv) a challenging dataset [24] is added for detecting silicone mask faces.

II. PROPOSED METHOD

The proposed approach is divided into two main components: (1) global motion estimation through an adaptive spatiotemporal global sampling, and (2) a joint CNN-BiLSTM network for PAD detection. We first explain the procedure of estimating the global motion and, then, the procedure of incorporating the CNN in BiLSTM model is described.

A. Spatiotemporal Global Sampling

As illustrated in Fig. 2, a video A is equally partitioned into s non-overlapping segments, $i.e., A = \{T_k\}_{k=1}^{s}$, where T_k is the k-th segment. The length of each segment is set to be $(l = 30)$ frames. In order to estimate the global motion, the trajectory of local motion vectors must be calculated between consecutive frames. For this, interest points play an important role as they determine the quality of the motion estimation. Moreover, the spatial distribution of the image features significantly impacts the performance of the calculated homography [17]. Taking this into consideration, we first define a region (grid) over the entire image, where the patch size 16×16 is used to detect dense features by selecting the sliding step equal to 1 pixel. Specifically, given an image $I, \Omega \longmapsto R^{Q}$ where $\Omega = \{0, 1, ..., g-1\} \times \{0, 1, ..., h-1\}$, g and h represent the number of rows and columns of image I. The sampling patch t is the number of sampled grids divided by the number of pixels in I; the objective is to determine a subset W of Ω for a given sampling patch t, such that:

$$W = \left\{ z | \, z \, \epsilon \, \Omega, \quad I(x) \; is \; informative, \quad \frac{\# \, Z}{g \times h} = t \right\} \quad (1)$$

where z denotes the local patches (i.e., grids) defined at the image pixel x, $I(x)$ is the response map at x and $\#Z$ represents the number of grids. In our work, the size of sampling patch is set to $t = 16 \times 16$. The content of W represents all of the patches (rectangular regions) on the image to be equal. Thus, by using the proposed sampling, Fast Retina Keypoint (FREAK) descriptor [18] is utilized to extract dense keypoints from each patch of the image I. The FREAK descriptor is famous due to its efficiency in smartphone deployment as compared to other local descriptors such as SURF [11]. Once these dense keypoints are extracted from the moved and fixed image, the second step is keypoints matching where Hamming distance (HD) is utilized in our work. The inter-frame motion parameters are estimated throughout the whole length of the segment by using the similarity transformation. A similarity transform is a special kind of geometric transformations that retains angles (shapes). More formally, let X denote the moving image keypoints and Y be the fixed image keypoints, we have

$$X = m * s * cos\theta - y * s * sin\theta + b, \quad (2)$$

$$Y = n * s * sin\theta - y * s * cos\theta + v, \quad (3)$$

where s, θ, and (b, v) are scaling, rotational, and translational differences between the images, respectively. These four parameters can be calculated based on the corresponding interest points in the images. The rotational difference is estimated from the angle of rotation between the lines corresponding to interest points in the images. The scaling is computed from the ratio of distances between the interest coordinates (points) in the images. To perform a scaling transformation, the translation variables (b, v) are obtained by replacing the coordinates of one of the correspondences into equations (1) and (2) and solving for b and v. Finally, the moving frame can be warped using these parameters to generate the final image.

To ensure optimal warping, we use the M-estimator SAmple Consensus (MSAC) [19] algorithm to detect outliers and remove false matching points before computing the final warped frame. MSAC is a variant of Random Simple Consensus (RANSAC) based on an improved cost function from the required functional relation. In order to estimate temporal information of all previous interframe motion vectors, we assume that warping is applied on segment k of video A to compute the homography $h_{a,k}$. Then the aggregation approach is used as follow:

$$h_a = \frac{1}{w_a} \sum_{k=1}^{w_a} h_a k, \quad (4)$$

TABLE II
COMPARISON RESULTS WITH LIMITED SOURCE DOMAINS ARE OBTAINED FROM [26].

	O&I to M		M&I to C		O&I to C		O&M to I		C&M to O	
Method	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
Supervised [26]	12.1	94.2	30.4	77.0	18.0	90.1	16.8	93.8	17.9	89.5
Mean-Teacher [27]	19.6	86.5	31.1	76.6	23.7	84.9	18.4	86.0	23.5	84.9
USDAN [28]	15.8	88.1	35.6	69.0	33.3	72.7	19.8	87.9	20.2	88.3
EPCR-labeled [26]	12.5	95.3	18.9	89.7	18.9	89.7	14.0	92.4	17.9	90.9
EPCR-unlabeled [26]	10.4	94.5	25.4	83.8	16.7	91.4	12.4	94.3	17.8	91.3
ASGS (Ours)	**8.2**	**97.3**	**23.4**	**91.8**	**13.9**	**96.8**	42.9	79.4	21.7	85.1

TABLE III
THE PERFORMANCE EVALUATION IN TERMS OF INTRA-DATASET. THE COMPARISON RESULTS ARE OBTAINED FROM THE ORIGINAL PAPER [24].

Method	SMFMVD Dataset			
	APCER(%)	BPCER(%)	ACER(%)	EER(%)
CDCN++ [9]	10.0	36.0	23.0	21.5
IDA [8]	35.0	4.00	19.5	14.1
Videolet [24]	4.0	73.0	38.5	22.1
CT [7]	9.0	15.0	12.0	12.3
Ref. [29]	8.0	51.0	29.5	34.1
MS-LBP [30]	6.0	9.0	7.5	8.1
VSFM [24]	4.0	2.0	3.0	1.1
ASGS (Ours)	**3.7**	**0.1**	**2.8**	1.6

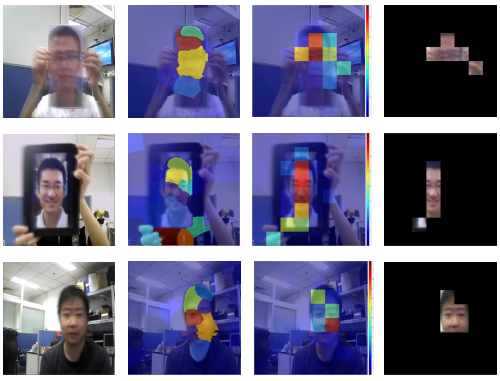

Fig. 3. Image explanation using LIME for ASGS encoded videos corresponding to a print attack (first row), video-replay attack (second row) and real face (third row).

where w_a is the number of selected frames for video A. By using the proposed aggregation function, we aggregate both temporal and spatial information of the frames to estimate the final RGB image. These enriched spatiotemporal encoded frames are the ones used for the end-to-end training of the joint CNN-BiLSTM model.

B. Recurrent Neural Network (RNN)

Mainstream CNN frameworks are connected to conventional statistical models, thus lacking the capacity to map sequence-to-sequence. To handle this issue, we first extract the high discriminative features of encoded video frames using the pooling layer of pretrained DenseNet-201 architecture [20]. Since the input of the CNN is video frames which comprise spatiotemporal patterns, the variations between video frames may accumulate complementary information for distinguishing live and spoofed faces. Thus, the Bidirectional Long Short-Term Memory Networks (BiLSTM) [21] is used to encode the temporal dynamic information across video frames. The BiLSTM computes long-range temporal relationships using the memory cell activation vector (M_c). It has an input gate (r_c), an output gate (e_c) and a forget gate (i_c). The three gates represent a fully connected layer, and its input is a vector and the output is a real number in $[0, 1]$.

III. EXPERIMENTS

To assess the generalization of the proposed face PAD approach, we considered four widely used publicly available databases consisting of bona fide and 2D face presentation attack videos, namely Idiap Replay-Attack database [22] (denoted as I), CASIA Face Anti-Spoofing database (denoted as C) [23], MSU Mobile Face Spoofing database [8] (denoted as M), and OULU-NPU database [14] (denoted as O). Three datasets are randomly selected for training and treated as source domains. The EER is computed on the source domain, and then Half Total Error Rate (HTER) is reported on the final testing set (target domain). To validate the performance of the proposed method on other types of attacks, a silicone mask face motion video dataset (SMFMVD) [24] is utilized that consists of 200 real and silicone masked facial motion videos.

A. Implementation details

All the images are resized to 224×224, and no data augmentation or face cropping is applied. For silicone mask dataset, the segment size was set to $l = 10$ frames. The off-the-shelf features from the last pooling layer of CNN (DenseNet-201) model are extracted and used as input to BiLSTM. For training, Adam optimizer is utilized by fixing a learning rate of 0.0001 with 500 hidden layer dimension. We do not use fixed epochs because an early stopping function is utilized that automatically stops the model when it starts over-fitting. *He* initializer is used to initialize recurrent weight of BiLSTM for multiple source domains while random initialization is performed in the case of silicone mask attacks and limited source domains evaluation. We argue that one must conduct a proper statistical analysis for initializing the network weights of BiLSTM.

B. Comparison against the state-of-the-art methods

In Table I, we evaluate our proposed method against state-of-the-art PAD methods. In the first row, O&C&I, O&M&I,

O&C&M and I&C&M represent the training sets while M, C, I and O are the testing sets, respectively. It can be observed that the model outperforms on three domain generalization test sets by a fair margin. Especially, in contrast to previous methods, [25], [15], [10], [25], our proposed method explicitly learns generalized features (e.g., eye blinking, hand movements, head rotation) across video frames and takes advantage of BiLSTM to capture the dynamic changes revealed by individuals. In Table II, we compare the domain generalization ability of our proposed method when limited source domain databases are accessible (i.e. only two source datasets).Furthermore, the proposed method is also evaluated on a silicone mask attacks. Based on the experimental results in Table III, we show that the proposed ASGS is not only effective for photo and replay attacks, but also outperform several state-of-the-art silicone-based PAD methods [9], [8], [7], [7], [29], [30], [24] under intra-dataset test environment. The main reason for the performance drops on O&C&M to I, O&M to I and C&M to O is due to the domain shift issue. For instance, two print attacks may be quite different in case of the same face if reprinted with different kinds of paper (e.g., glossy vs. rough paper). Thus, due to a large diversity of real-world environments, such as differences in (i) spoofing mediums (printing material, LCD screens), and (ii) the quality of video recording devices (different mobile phones, tablets), the proposed approach does not provide high accuracy on all cases.

Fig.3 illustrates the interpretability of the model using LIME (local interpretable model-agnostic explanations) [31] to understand the importance of the proposed ASGS and model's decision in a human-understandable way. The first row represents images for the print attack where the model focuses on face texture and hand movement cues are valuable for the prediction of the network. The second row represents images for a video-replay attack where one can see that the network gives importance to the material reflection and the tablet's edges provide salient information. The live class images are displayed in the third row where head motion and eye blinking contribute positively to distinguish live and spoofed faces. The masked images in the last column show the most important features used for final detection.

IV. Conclusion

This letter addresses the domain shift issue of 2D face PAD and presents a novel video representation method (ASGS) to compress the fine-grained motions (e.g., eye blinking, hand trembling, head rotation) into a single color image. Then a CNN-BiLSTM architecture is exploited to discriminate real and spoofed faces. Experimental results on four benchmark datasets demonstrate the proposed approach not only outperforms several state-of-the-art PAD methods for print or video attacks but also silicon-based masked attacks. However, a drawback of ASGS is that it falls behind real-time requirements and cannot be operated on those biometric systems that are operating on single facial images. Moreover, the estimation of feature trajectory can bring positioning errors in the case of significant motion blur. In future works, we will focus on building the model based on tracking the facial landmark-based trajectories for video-based PAD detection.

References

[1] Ibsen, Mathias, et al. Digital Face Manipulation in Biometric Systems. *In Handbook of Digital Face Manipulation and Detection*, pp. 27-43. Springer, Cham, 2022.

[2] Ramachandra, Raghavendra, et al. Towards making morphing attack detection robust using hybrid scale-space colour texture features. *In 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis* pp. 1-8.IEEE.2019.

[3] Xu, Y., Raja, K., et al. Adversarial Attacks on Face Recognition Systems. *In Handbook of Digital Face Manipulation and Detection* (pp. 139-161). Springer, Cham. 2022.

[4] Sánchez-Sánchez, M. Araceli, et al. Convolutional neural network approach for multispectral facial presentation attack detection in automated border control systems. *Entropy*, 22(11), 1296.2020.

[5] Shao, Rui, et al. "Regularized fine-grained meta face anti-spoofing."*IEEE Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 34. No. 07. 2020.

[6] Zhang, Zhuoyi, et al. "Two-stream Convolutional Networks for Multi-frame Face Anti-spoofing." *arXiv preprint*, 2108.04032, 2021.

[7] Boulkenafet, Zinelabidine, et al. "Face spoofing detection using colour texture analysis." *IEEE Transactions on Information Forensics and Security* 11, no. 8: 1818-1830, 2016.

[8] Wen, Di, et al. "Face spoof detection with image distortion analysis."*IEEE Transactions on Information Forensics and Security* 10.4: 746-761.2015.

[9] Yu, Zitong, et al. "Searching central difference convolutional networks for face anti-spoofing."*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2020.

[10] Saha, Suman, et al. "Domain agnostic feature learning for image and video based face anti-spoofing". *IEEE In Proceedings of Conference on Computer Vision and Pattern Recognition Workshops*, pp. 802-803, 2020.

[11] Muhammad, Usman, et al. "Self-supervised 2D face presentation attack detection via temporal sequence sampling." *Pattern Recognition Letters* 2022.

[12] Qin, Yunxiao, et al. "Meta-teacher for Face Anti-Spoofing". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021.

[13] Jia, Yunpei, et al. "Single-side domain generalization for face anti-spoofing." *IEEE In Proceedings of Conference on Computer Vision and Pattern Recognition*, pp. 8484-8493. 2020.

[14] Boulkenafet, Zinelabinde, et al. "Oulu-npu: A mobile face presentation attack database with real-world variations." *In 12th IEEE international conference on automatic face and gesture recognition*, pp. 612-618. IEEE, 2017.

[15] Wang, Guoqing, et al. "Cross-domain face presentation attack detection via multi-domain disentangled representation learning." *In Proceedings of Conference on Computer Vision and Pattern Recognition*, pp. 6678-6687. 2020.

[16] Liu, Ajian, et al. "Face Anti-Spoofing via Adversarial Cross-Modality Translation." *IEEE Transactions on Information Forensics and Security 16*, pp.2759-2772, 2021.

[17] Wang, H., Schmid, C. "Action recognition with improved trajectories." *In Proceedings of the IEEE international conference on computer vision*, pp. 3551-3558. 2013.

[18] Alahi, Alexandre, et al. "Freak: Fast retina keypoint." *2012 IEEE conference on computer vision and pattern recognition.* pp. 510-517, IEEE, 2012.

[19] Torr, Philip HS, and Andrew Zisserman. "Robust parameterization and computation of the trifocal tensor." *Image and vision Computing* 15.8 : 591-605, 1997.

[20] Huang, Gao, et al. "Densely connected convolutional networks." *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708. 2017.

[21] Schuster, et al. "Bidirectional recurrent neural networks." *IEEE transactions on Signal Processing* 45.11: 2673-2681, 1997.

[22] Chingovska, Ivana, et al. "On the effectiveness of local binary patterns in face anti-spoofing." *BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*. IEEE, (pp. 1-7), 2012.

[23] Zhang, Zhiwei, et al. "A face antispoofing database with diverse attacks." *In 2012 5th IAPR international conference on Biometrics (ICB)* (pp. 26-31). IEEE, 2012.

[24] Wang, Guangcheng, et al. "Silicone mask face anti-spoofing detection based on visual saliency and facial motion." *Neurocomputing* 458, pp.416-427.2021.

[25] Shao, Rui, et al. "Multi-adversarial discriminative deep domain generalization for face presentation attack detection". *In Proceedings of Conference on Computer Vision and Pattern Recognition*, pp. 10023-10031, 2019.

[26] Wang, Zezheng, et al. "Consistency Regularization for Deep Face Anti-Spoofing." *arXiv preprint arXiv* 2111.12320.2021.

[27] Tarvainen, Antti, and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results." *Advances in neural information processing systems* 30.2017.

[28] Jia, Yunpei, et al. "Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing." *Pattern Recognition* 115:107888.2021.

[29] Wang, Zezheng, et al. "Deep spatial gradient and temporal depth learning for face anti-spoofing." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.pp.5042-5051.2020.

[30] Erdogmus, Nesli, and Sebastien Marcel. "Spoofing face recognition with 3D masks." *IEEE transactions on information forensics and security* 9.7: 1084-1097.2014.

[31] Ribeiro, Marco, et al. " Why should i trust you? Explaining the predictions of any classifier." *Proceedings of international conference on knowledge discovery and data mining*, pp. 1135-1144, 2016.

[32] Wei, Donglai, et al. "Learning and using the arrow of time." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8052-8060, 2018.