Face anti-spoofing from the perspective of data sampling

Usman Muhammad[1,✉] and Mourad Oussalah[2]
[1]University of Oulu, Center for Machine Vision and Signal Analysis (CMVS), Oulu, Finland
[2]University of Oulu Faculty of Information Technology and Electrical Engineering, Center for Machine Vision and Signal Analysis (CMVS), Oulu, Finland
✉E-mail: muhammad.usman@oulu.fi

Without deploying face anti-spoofing countermeasures, face recognition systems can be spoofed by presenting a printed photo, a video, or a silicon mask of a genuine user. Thus, face presentation attack detection (PAD) plays a vital role in providing secure facial access to digital devices. Most existing video-based PAD countermeasures lack the ability to cope with long-range temporal variations in videos. Moreover, the key-frame *sampling* prior to the feature extraction step has not been widely studied in the face anti-spoofing domain. To mitigate these issues, this paper provides a data sampling approach by proposing a video processing scheme that models the long-range temporal variations based on Gaussian weighting function (GWF). Specifically, the proposed scheme encodes the consecutive t frames of video sequences into a single RGB image based on a Gaussian-weighted summation of the t frames. Using simply the data sampling scheme alone, it is demonstrated here that state-of-the-art performance can be achieved without any bells and whistles in both intra-database and inter-database testing scenarios for the three public benchmark datasets; namely, replay-Attack, MSU-MFSD, and CASIA-FASD. In particular, the proposed scheme provides a much lower error (from 15.2% to 7.6% on CASIA-FASD and 5.9% to 4.9% on replay-attack) compared to baselines in cross-database scenarios.

Introduction: During the past decade, we have witnessed great advancements in face anti-spoofing methods thanks to the emergence of deep learning models. However, with the growth of facial recognition technology, face identity threats are evenly increasing at the same rate and the problem is still unsolved due to the difficulty in the design of discriminative features. Therefore, how to effectively detect spoofing attacks remains a critical problem for both practitioners and the research community. Existing learning-based approaches can be classified based on static and dynamic information. Compared to static or image-based face PAD [1], video-based face anti-spoofing is more challenging because deep learning methods based on a 2D convolutional neural network (CNN) ignore the temporal dimension of the video and process each frame independently.

Although spatiotemporal feature learning based on optical flow [2], 3D CNN [3] or recurrent neural network (RNN) [4] have demonstrated their performance in the face anti-spoofing domain, there is a fundamental question that still needs to be addressed in the field—What are alternative approaches to improve the spatiotemporal representations for face anti-spoofing? Wang et al. [5] argued that temporal depth difference between live and spoof faces along with a contrastive depth loss can make impressive progress in improving the PAD performance. Zitong et al. [6] proposed a method based on central difference convolutional network (CDCN) via aggregating both intensity and gradient information. Moreover, state-of-the-art approaches address the domain generalization issue by aligning the feature distribution between source and target domain. For instance, adversarial learning [25], meta pattern or meta-teacher learning [8, 26], cross-adversarial learning [27], generative domain adaptation [28], hypothesis verification [29], or shuffled style assembly network [30] were introduced for improving the generalization ability of face anti-spoofing. However, most of them typically fit the models to a shared feature space which remains always challenging for a better generalization. Although previous works [4, 7] demonstrate that PAD performance can be improved by learning motion cues associated with the artifacts, that is, hand trembling or material reflection, such methods require complex feature engineering skills to generate the spatiotemporal representations.

We argue that most of the existing video-based countermeasures employ a fixed frame selection and overlook other important factors such as the importance of data understanding. Moreover, within a given video, not all the frames are of equal importance for PAD detection, and action-related frames may occur sparsely in a few frames. Thus, the key-frame *sampling* [9] is crucial before the feature extraction step and it remains an unsolved problem for video-based PAD. Moreover, different from existing domain generalization methods that focus on learning a common feature space as in [31], the proposed model aims to directly learn from the perspective of input data. To achieve this, we take the advantage of using temporal information and develop a simple strategy of selecting relevant frames according to their temporal variations. Motivated by the GWF [10], we introduce a simple data sampling mechanism, which aims to accumulate information from video sequences into a single RGB image and generate the most discriminative frames. Specifically, we demonstrate that appropriate data sampling provides an alternative solution in achieving a good generalization. Thus, our main objective is to provide the importance of data understanding, especially, (i) how differently does CNN or RNN behave concerning spatial-temporal modelling of video data? and (ii) conduct comprehensive experiments and analysis to demonstrate that our approach is extensible, which can be plugged into different deep learning-based face PAD models. In summary, the overall contributions of our work include:

1. We present a data-driven approach to capture the appearance and dynamics of video into a single RGB image.
2. Our analysis shows that the proposed temporal modelling can amplify important clues, for example, hand movements, and surface edges, to improve the detection accuracy.
3. We provide an interpretation of the decisions made by the employed model. The model revealed that the motion cues are the most important factors for distinguishing whether an input image is spoofed or not.
4. Experiments on four benchmark datasets, consisting of CASIA-MFSD, REPLAY-ATTACK, OULU-NPU, and MSU-MFSD databases, show that our proposed method provides competitive performance on three datasets in comparison to the state-of-the-art generalization methods used now.

Proposed method: The proposed data sampling is based on the assumption that adjacent frames in videos complement the temporal motion which is an important and universal signal. This also motivates us to adaptively select frames from videos. To achieve this, GWF is used to accumulate the video sequences into a fewer number of frames. Suppose that $\{S_n\}n \in N$ is an exhaustive non-overlapping sequence (pertaining to the full length of the video), which is given by

$$S_n = \{S_1, S_2, \ldots, S_k, \ldots\}, \tag{1}$$

Where $\{S_k\}$ represents the kth sub-sequence of $\{S_n\}$ and $k < n$. Specifically, GWF G, for a sub-sequence $\{S_k\}$, is denoted as follows:

$$G(S_k, M) = \sum_{q=1}^{Z} S_{k_q} * \frac{M_q}{\sum_{q=1}^{Z} M_q} \tag{2}$$

The function G takes a sub-sequence $\{S_k\}$, and Gaussian weight vector M as input, and accumulates the representation into a single RGB frame. In particular, M_q depicts the qth element of the Gaussian weight vector M. Z represents the size of the Gaussian weight vector. For instance, if the size of the Gaussian weight vector Z is 30 and the sub-sequence $\{S_k\}$ has thirty frames of the video. Then, the vector M is given by $M = [1, 2, 3, \ldots, 30]$. A single RGB image can be acquired based on the weighted summation of the thirty frames associated with the sub-sequence $\{S_k\}$ as illustrated in Equation (2). Thus, following Equation (2), different frame sizes can be selected to be accumulated into a single RGB image based on the GWF. Likewise, the same procedure is applied for the subsequent thirty frames belonging to the next sub-sequence and so on. The main steps are illustrated in Figure 1. Intuitively, this data sampling method has at least three main advantages. First, the encoded spatio-temporal information can be fed to any CNN

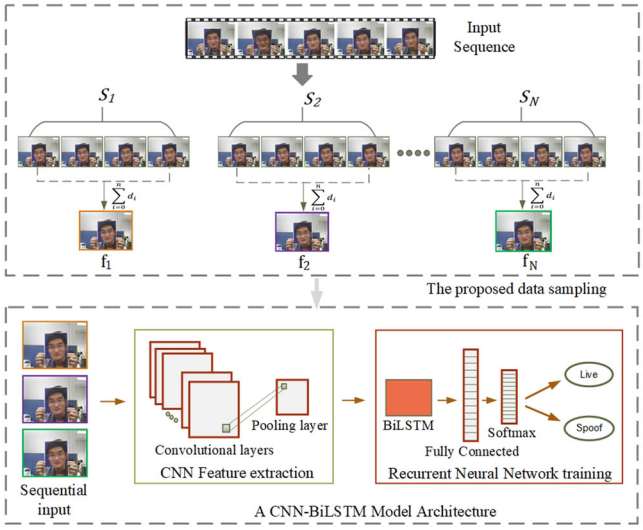

Fig. 1 *An illustration of the proposed method*

Table 1. Ablation study using cross-database evaluation

Method	Train CASIA-FASD Test replay-attack	Train replay-attack Test CASIA-FASD
Sub-sequence size (30)	3.1	15.3
Sub-sequence size (40)	4.9	7.6
Sub-sequence size (50)	3.8	12.4
Sub-sequence size (60)	4.6	10.1

architecture for a still image, where "still" captures long-range dynamic variations in videos. Second, the encoded images decrease the number of frames per video to be processed and force the learning-based classifiers to focus on discriminative clips. Third, our method accumulates raw video frames directly in comparison to the previous data preprocessing methods that uses complex processing such as estimation of global motion [7] or optical flow [2].

Implementation details: We conduct experiments on four major databases: CASIA face anti-spoofing [11] (denoted as C), Idiap replay-attack [12] (denoted as I), MSU mobile face spoofing [13] (denoted as M), and OULU-NPU [32] (denoted as O). In order to determine the performance, we report half total error rate (HTER), and equal error rate (EER) by following the existing testing protocols used in the current works [5, 21]. A pre-trained CNN (ResNet-101 [14]) is utilized to extract off-the-shelf CNN features using the pooling layer without data augmentation. All the images are resized to 224 × 224. The extracted features are then used as input to train the bidirectional long short-term memory networks (BiLSTM) [15] for final detection. To train the BiLSTM, the Adam optimizer is utilized by fixing a learning rate of 0.001, mini-batch size 32, validation frequency 30, and with 100 hidden layer dimension for all the intra-database protocols. An early stopping function [16] is utilized to reduce the risk of over-fitting since we do not set the fixed number of epochs. For the cross-database or inter-database scenario, the same experimental settings are used except the learning rate which was increased up to 0.0001 for the BiLSTM. Moreover, *He* initializer [17] is used for initializing recurrent weights that perform the best for all our experiments. In addition, experiments for multi-source domains (i.e., three datasets for training and the rest one for testing) were conducted with the hidden layer size of 500.

Ablation study: To investigate the influence of temporal variations in the videos, we thoroughly validate the performance of each sub-sequence size by performing an ablation study. The numerical results are reported in Table 1. The result shows that even capturing the small temporal variations leads to a competitive performance by using our proposed data sampling. It is worth mentioning that the model is trained on the training dataset of CASIA and uses the CASIA testing set as a validation set. Based on the EER of the CASIA testing set, we report HTER on a completely unseen replay-attack dataset. Similarly, the same experimental protocols are repeated for the replay-attack dataset. We note that the proposed data sampling provides consistent performance even with a sub-sequence size of 30. When the temporal length increases to 40, we achieved the best performance on the CASIA dataset and very competitive performance for the replay-attack dataset. Thus, we set the same sub-sequence size (40) for the intra-database evaluation.

Table 2. The results of cross-database testing.

Method	Train CASIA-FASD Test Replay-Attack	Train Replay-Attack Test CASIA-FASD
Color-LBP [18]	30.3	37.7
Auxiliary [19]	27.6	28.4
FaceDs [20]	28.5	41.1
STASN [21]	31.5	30.9
GDA [28]	15.1	29.7
DSGTD [5]	17.0	22.8
CDCN [6]	6.5	29.8
SSL learning [7]	5.9	15.2
VGG19-BiLSTM w/o DS	31.3	42.8
VGG19-BiLSTM w/DS	13.3	**11.4**
ResNet-BiLSTM w/o DS	27.4	33.2
ResNet-BiLSTM w/DS	**4.9**	**7.6**

Comparison against the state-of-the-art methods: We observed that the proposed method achieves 0.00, EER for the replay-attack, CASIA, and MSU datasets in the intra-database scenario. However, we focus on evaluating the performance on the most challenging scenario, that is, cross-dataset (single-source) and multi-source scenarios in terms of HTER. Table 2 shows the results when the model is trained on a single dataset and then evaluated on a completely unseen dataset. The proposed method improves the performance by up to 7% more than the previous state-of-the-art method [7] for the CASIA dataset. This is a remarkable improvement for the cross-database scenario. Furthermore, we also report the results of the ResNet-BiLSTM model without our proposed data sampling. One can see that the performance is improved up to 20% on both datasets. To demonstrate that the proposed data sampling can enhance the performance of other deep learning models, we also report the performance of VGG19 architecture [22] using a sub-sequence size (40). The results confirm the effectiveness of our approach. In Table 3, we compare our results with those state-of-the-art methods [25–31] which focus on domain generalization settings. Based on the experimental results, our method provides best performance on two benchmark databases (M and C) and a competitive performance on third dataset (I). Thus, we argue that sufficient data understanding is important in PAD detection. In addition, we provide heat map visualizations based on occlusion sensitivity maps [23] and local interpretable model-agnostic explanations (LIME) [24] to understand what patterns contribute to making a classification decision. The first column in Figure 2 shows the images generated by the proposed data sampling scheme. The second column represents heat maps based on occlusion sensitivity method. These maps show that the frame aggregation provides vital motion cues for the model prediction. Specifically, images of photo and replay-attack in the first and second row show that hand movement provides salient information. For further evaluation, LIME interpretation in third column of Figure 2 is provided to demonstrate that the proposed data sampling helps the models to focus on the paper's texture and hand rotation cues. The masked images in last column of Figure 2 represent the most important superpixels to view the model's decision in a human-understandable way.

Conclusions: In this paper, we addressed the face PAD issue by proposing a simple data sampling strategy that requires a GWF to aggregate the video sub-sequences into a single RGB image. In particular, the

Table 3. The results of cross-dataset testing on MSU-MFSD (M), CASIA-MFSD (C), replay-attack (I), and OULU-NPU (0).

	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(%)
ADL [25]	5.00	97.58	10.00	96.85	12.07	94.68	13.45	94.43
SSDG-R [31]	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54
HFN + MP [26]	5.24	97.28	9.11	96.09	15.35	90.67	12.40	94.26
Cross-ADD [27]	11.64	95.27	17.51	89.98	15.08	91.92	14.27	93.04
GDA [28]	9.2	98.0	12.2	93.00	10.00	96.00	14.40	92.60
SSAN-R [30]	6.67	98.75	10.00	96.67	08.88	96.79	13.72	93.63
FG +HV [29]	9.17	96.92	12.47	93.47	16.29	90.11	13.58	93.55
ResNet-BiLSTM w/DS	**4.12**	**99.93**	**7.04**	**99.87**	13.48	**97.42**	41.33	88.48

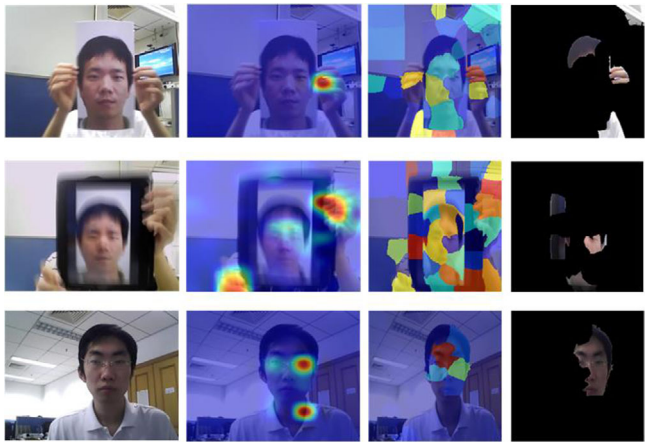

Fig. 2 *Image explanation using occlusion sensitivity maps and LIME corresponding to a print attack (first row), video-replay attack (second row) and real face (third row).*

proposed data sampling amplifies the motion cues which are naturally available in the format of video streams. We demonstrate that the generalization of the face anti-spoofing can be improved by learning from the data. This is different from existing methods where the domain generalization issue is mainly addressed by adversarial learning, generative domain adaptation and meta learning. Extensive experiments on four datasets demonstrate that our proposed method provides competitive performance on three datasets. Since the proposed data sampling is also required for encoding testing videos, the proposed method does not meet the real-time requirement. Thus, future work will focus on developing an end-to-end learning that can encode both appearance and temporal information effectively.

Author contributions: Usman Muhammad: Conceptualization, methodology, software, visualization, writing - original draft. Mourad Oussalah: Investigation, supervision, writing - review and editing.

Conflict of interest: The authors declare no conflict of interest.

Funding information: This work is supported by the Center for Machine Vision and Signal Analysis (CMVS) and the authors are grateful to the Academy of Finland Profi5 DigiHealth project.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: *22 August 2022* Accepted: *30 November 2022*
doi: 10.1049/ell2.12692

References

1 Nong, X., Zeng, Y., Hu, H.: Face anti-spoofing with refined triplet loss and multi-level attention constraint network. *Electron. Lett.* **57**(24), 912–914 (2021)
2 Li, L., Xia, Z., Wu, J., Yang, L., Han, H.: Face presentation attack detection based on optical flow and texture analysis. *J. King Saud Univ., Comput. Inf. Sci.* **34**(4), 1455–1467 (2022)
3 Gan, J., Li, S., Zhai, Y., Liu, C.: 3D convolutional neural network based on face anti-spoofing. In: 2017 2nd International Conference on Multimedia and Image Processing (ICMIP), pp. 1–5. IEEE, Piscataway, NJ (2017)
4 Muhammad, U., Zhang, J., Liu, L., Oussalah, M.: An adaptive spatio-temporal global sampling for presentation attack detection. *IEEE Trans. Circuits Syst. II Express Briefs* (2022). https://doi.org/10.1109/TCSII.2022.3169435
5 Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., Lei, Z.: Deep spatial gradient and temporal depth learning for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5042–5051. IEEE, Piscataway, NJ (2020)
6 Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5295–5305. IEEE, Piscataway, NJ (2020)
7 Muhammad, U., Yu, Z., Komulainen, J.: Self-supervised 2D face presentation attack detection via temporal sequence sampling. *Pattern Recognit. Lett.* **156**, 15–22 (2022)
8 Qin, Y., Yu, Z., Yan, L., Wang, Z., Zhao, C., Lei, Z.: Meta-teacher for face anti-spoofing. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(10), 6311–6326 (2021)
9 Zhi, Y., Tong, Z., Wang, L., Wu, G.: Mgsampler: An explainable sampling strategy for video action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1513–1522. IEEE, Piscataway, NJ (2021)
10 Basha, S.H., Pulabaigari, V., Mukherjee, S.: An information-rich sampling technique over spatio-temporal CNN for classification of human actions in videos. *Multimed. Tools Appl.* **81**(28), 40431–40449 (2022)
11 Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: 2012 5th IAPR International Conference on Biometrics (ICB), pp. 26–31. IEEE, Piscataway, NJ (2012)
12 Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG), pp. 1–7. IEEE, Piscataway, NJ (2012)
13 Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. *IEEE Trans. Inf. Forensics Secur.* **10**(4), 746–761 (2015)
14 He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE, Piscataway, NJ (2016)
15 Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**(11), 2673–2681 (1997)
16 Prechelt, L.: Early stopping-but when? In: *Neural Networks: Tricks of the trade*, Springer, Berlin, Heidelberg, pp. 55–69 (1998)

17 He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034. IEEE, Piscataway, NJ (2015)
18 Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: 2015 IEEE International Conference on ImageProcessing (ICIP), pp. 2636–2640. IEEE, Piscataway, NJ (2015)
19 Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 389–398. IEEE, Piscataway, NJ (2018)
20 Jourabloo, A., Liu, Y., Liu, X.: Face de-spoofing: Anti-spoofing via noise modeling. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 290–306. Springer, Cham (2018)
21 Yang, X., Luo, W., Bao, L., Gao, Y., Gong, D., Zheng, S., Liu, W.: Face anti-spoofing: Model matters, so does data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3507–3516. IEEE, Piscataway, NJ (2019)
22 Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv*:1409.1556 (2014)
23 Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision, pp. 818–833. Springer, Cham (2014)
24 Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. Association for Computing Machinery, New York, NY (2016)
25 Liu, M., Mu, J., Yu, Z., Ruan, K., Shu, B., Yang, J.: Adversarial learning and decomposition-based domain generalization for face anti-spoofing. *Pattern Recognit. Lett.* **155**, 171–177 (2022)
26 Cai, R., Li, Z., Wan, R., Li, H., Hu, Y., Kot, A.C.: Learning meta pattern for face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.* **17**, 1201–1213 (2022)
27 Huang, H., Xiang, Y., Yang, G., Lv, L., Li, X., Weng, Z., Fu, Y.: Generalized face anti-spoofing via cross-dversarial disentanglement with mixing augmentation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2939–2943. IEEE, Piscataway, NJ (2022)
28 Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Sheng, K., Ding, S., Ma, L.: Generative domain adaptation for face anti-spoofing. In: *European Conference on Computer Vision*, pp. 335–356. Springer, Cham (2022).
29 Liu, S., Lu, S., Xu, H., Yang, J., Ding, S., Ma, L.: Feature generation and hypothesis verification for reliable face anti-spoofing. *Proc. AAAI Conf. Artif. Intell.* **36**(2), 1782–1791 (2022)
30 Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., Wang, Z.: Domain generalization via shuffled style assembly for Face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4123–4133. IEEE, Piscataway, NJ (2022)
31 Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8484–8493. IEEE, Piscataway, NJ (2020)
32 Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: OULU-NPU: A mobile face presentation attack database with real-world variations. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 612–618. IEEE, Piscataway, NJ (2017)