Attention-Based Spatio-Temporal Feature Fusion for Efficient RGB Face Presentation Attack Detection

Tazeen Khan[1], M. Shahid Jabbar[2], M. Sohail Ibrahim[3], Aiman Helmi El-Maleh[1,3], Shujaat Khan[1,2]
[1]Department of Computer Engineering, College of Computing and Mathematics,
King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
[2]SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI),
King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
[3]Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS),
King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
Emails: {g202317010, muhammad.jabbar, muhammad.ibrahim, aimane, shujaat.khan}@kfupm.edu.sa

2026 International Wireless Communications and Mobile Computing (IWCMC) | 979-8-3315-5001-1/26/$31.00 ©2026 IEEE | DOI: 10.1109/IWCMC69287.2026.11579847

***Abstract*—Face Presentation Attack Detection (FacePAD) is a critical safeguard for modern face recognition systems, particularly in mobile and embedded environments where hardware and computational resources are constrained. Although RGB-based deep learning approaches have achieved strong performance, many existing methods rely on frame-wise processing or uniform temporal aggregation, limiting their ability to effectively exploit motion dynamics present in video sequences. This paper proposes an attention-based spatio-temporal FacePAD framework that adaptively aggregates frame-level deep features using an attention mechanism inspired by Bahdanau temporal attention. A convolutional backbone extracts spatial embeddings from sampled video frames, which are then fused through learnable attention weights to emphasize informative temporal cues while suppressing less discriminative frames. The resulting video-level representation is passed to a compact classifier for final bonafide/spoof classification. By avoiding computationally intensive recurrent or 3D convolutional architectures, the proposed method provides an effective yet structurally simple approach to temporal modeling in RGB-only FacePAD. Extensive experiments on the Replay-Attack, Replay-Mobile, ROSE-Youtu, and OULU-NPU datasets demonstrate the effectiveness of the proposed approach, achieving HTERs of 0% on Replay-Attack, 0% on Replay-Mobile, and 0.53% on ROSE-Youtu, and an ACER of 0.069% on OULU-NPU.**

***Index Terms*—Face presentation attack detection, face anti-spoofing, spatio-temporal attention, MobileNet, RGB video.**

I. Introduction

Face recognition technologies are widely deployed in applications such as mobile authentication, surveillance, and access control. However, these systems remain vulnerable to presentation attacks, including printed photographs, replayed videos, and 3D masks. Face presentation attack detection (FacePAD) addresses this vulnerability by distinguishing bonafide face presentations from spoofing attempts at the sensor level [1], [2]. Among available solutions, RGB-based FacePAD is particularly attractive due to its low cost and compatibility with existing camera hardware.

Corresponding author: Shujaat Khan

S. Khan would like to acknowledge the support received from King Fahd University of Petroleum and Minerals (KFUPM) under the Early Career Grant No. EC241027.

Early FacePAD methods relied on handcrafted texture and motion descriptors, such as local binary patterns and eye-blink detection, to reveal spoofing artifacts [3], [4]. Optical flow-based motion analysis was later introduced to capture dynamic inconsistencies between bonafide and fake presentations [5]. Although effective against simple attacks, these approaches struggled to generalize to high-quality replay and mask-based attacks.

The adoption of deep learning significantly improved performance by enabling end-to-end feature learning. CNN-based models extract discriminative spatial representations [6], [7], yet purely spatial approaches often overlook temporal cues essential for detecting video replay attacks. To address this limitation, spatio-temporal modeling techniques have been proposed. The work in [8] introduced a hybrid CNN-LSTM architecture to jointly learn spatial and temporal features, while Shao et al. [9] proposed deep dynamic texture learning for capturing discriminative motion patterns. More recent works explored temporal sequence sampling with self-supervised learning [10] and stacked LSTM architectures for enhanced temporal modeling [11]. Despite their effectiveness, recurrent networks and 3D convolutional methods are often computationally expensive, limiting real-time deployment.

To improve efficiency, lightweight CNN backbones, such as MobileNet [12], have been adopted for FacePAD. Building on this direction, the work in [13] demonstrated that integrating spatio-temporal feature extraction within a lightweight CNN can effectively capture both static and dynamic characteristics while maintaining real-time suitability for mobile and embedded platforms. In parallel, representation learning strategies such as fully-supervised contrastive learning coupled with temporal averaging have been explored to enhance feature separability between bonafide and spoof samples [14]. Although temporal information is crucial for detecting replay and video-based attacks, many existing approaches either treat frames uniformly during aggregation or rely on computationally intensive sequence models. Attention mechanisms may provide a principled way to adaptively weight informative

Authorized licensed use limited to: King Fahd University of Petroleum and Minerals. Downloaded on September 08,2026 at 18:13:40 UTC from IEEE Xplore. Restrictions apply.

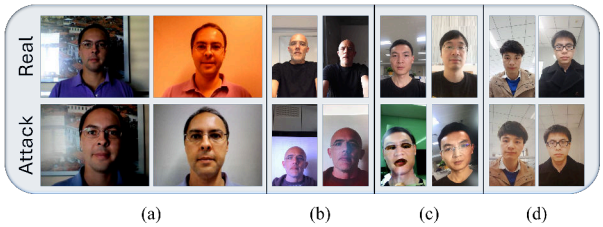

Fig. 1: Example of bonafide and attack video frames in the (a) Replay-Attack, (b) Replay-Mobile, (c) ROSE-Youtu, and (d) OULU-NPU datasets. The top row represents bonafide frames, while the bottom row presents attack frames using various attack instruments.

temporal cues rather than processing all frames equally.

To address the limitations of uniform temporal aggregation and computationally intensive sequence models, this paper proposes an attention-based spatio-temporal FacePAD framework built upon a lightweight CNN backbone. From each input video, a subset of frames is sampled and processed independently to extract frame-level spatial embeddings. These embeddings are then organized as a temporal sequence and passed to an attention module, which learns to assign adaptive attention weights across frames. By forming a weighted aggregation of temporal features, the proposed model emphasizes discriminative motion cues while suppressing redundant or less informative frames. The resulting video-level representation is finally fed into a compact classifier to perform the bonafide or spoof prediction. By avoiding recurrent architectures and 3D convolutional operators, the proposed design provides an effective and structurally simple approach to temporal modeling in RGB-only FacePAD.

II. Materials and Methods

A. Datasets

To evaluate the effectiveness of the proposed method, we utilize four widely-used face anti-spoofing datasets: Replay-Attack (RA) [15], Replay-Mobile (RM) [16], ROSE-Youtu (RY) [17], and OULU-NPU [18]. The details of each dataset are listed in Table I.

The Replay-Attack dataset [15] consists of 1,300 short video clips containing both bonafide presentations and spoofing attacks, including printed photos and video replays, from 50 subjects recorded under controlled and adverse lighting conditions, with predefined training, development, and test splits ensuring subject disjointness, example frames are shown in Fig. 1(a). The Replay-Mobile dataset [16] is tailored for mobile face recognition and FacePAD, comprising 1,190 video clips from 40 subjects captured under varying illumination using mobile devices, and includes both printed photo and digital replay attacks, as illustrated in Fig. 1(b). The ROSE-Youtu dataset [17] contains 3,350 video clips from 20 subjects recorded with multiple camera models under diverse lighting conditions, incorporating video replay, print photo, and paper mask attacks, and serves as a challenging benchmark for face anti-spoofing research (see Fig. 1(c)). Finally, OULU-NPU [18] is a high-resolution dataset comprising 4,950 video recordings from 55 subjects collected under varying illumination, background scenes, and device settings, example frames are shown in Fig. 1(d).

TABLE I: Dataset Description

Dataset	*Year*	*Subjects*	*Bonafide/Spoof Clips*
Replay-Attack [15]	*2012*	*50*	*300 / 1000*
Replay-Mobile [16]	*2016*	*40*	*550 / 640*
ROSE-Youtu [17]	*2018*	*20*	*1000 / 2350*
OULU-NPU [18]	*2017*	*55*	*990 / 3960*

B. Frame Sampling, Preprocessing, and Data Augmentation

Given a labeled video dataset $\mathcal{D} = \{(V_i, \ell_i)\}_{i=1}^{N}$, where V_i denotes a video and $\ell_i \in \{\text{bonafide}, \text{Spoof}\}$ its label, we construct fixed-length training clips consisting of τ frames.

Temporal Frame Sampling: For a video V containing N frames, we select a starting index s. During training, s is randomly sampled to increase temporal diversity:

$$s \sim \mathcal{U}\left(0,\, \max(0, N - \tau)\right). \tag{1}$$

A contiguous clip of length τ is then formed as

$$\{\mathbf{x}_t\}_{t=0}^{\tau-1}, \quad \text{where } \mathbf{x}_t = V[s + t]. \tag{2}$$

During validation and inference, s is fixed ($s = 0$) to ensure deterministic evaluation. After sampling, each video is represented as:

$$X = \{\mathbf{x}_t\}_{t=0}^{\tau-1} \in \mathbb{R}^{\tau \times 3 \times H \times W}. \tag{3}$$

Adaptive Center Cropping and Resizing: Each sampled frame $\mathbf{x}_t$ is spatially normalized in RGB space. To avoid geometric distortion during resizing, we apply adaptive center cropping. For a frame of spatial dimensions (H, W), we compute $\rho = \min(H, W)$ and extract a centered square region of size $\rho \times \rho$. The cropped region is then resized to a fixed spatial resolution of 224×224 using bilinear interpolation. After preprocessing, each clip is represented as $X \in \mathbb{R}^{\tau \times 3 \times 224 \times 224}$, which serves as input to the CNN backbone ϕ_Θ.

Data Augmentation: To improve robustness and reduce overfitting, spatial data augmentation is applied during training only. For each sampled clip, a single set of augmentation parameters is generated and applied identically to all frames $\{\mathbf{x}_t\}_{t=0}^{\tau-1}$, preserving temporal consistency. The augmentation strategy includes:

- *Random rotation:* $\theta \sim \mathcal{U}(-180°, 180°)$ with probability 0.5; otherwise $\theta = 0$.
- *Random scaling:* $s \sim \mathcal{U}(0.7, 1.3)$ with probability 0.5; otherwise $s = 1$.

The same transformation parameters (θ, s) are applied to all frames in the clip to maintain temporal coherence. No data augmentation is applied during validation or inference.

Authorized licensed use limited to: King Fahd University of Petroleum and Minerals. Downloaded on September 08,2026 at 18:13:40 UTC from IEEE Xplore. Restrictions apply.

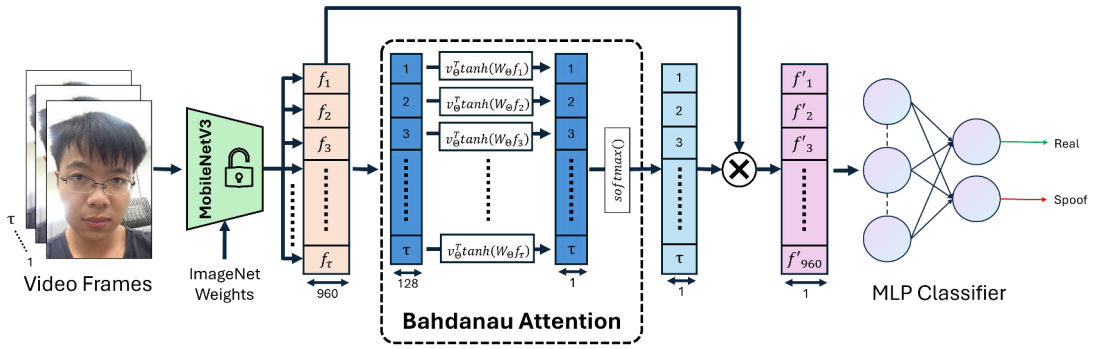

Fig. 2: Architecture of the proposed attention-based spatio-temporal FacePAD framework. Sampled video frames are processed by a pre-trained MobileNetV3 backbone to extract frame-level features, which are adaptively aggregated using Bahdanau temporal attention and fed into a classifier for bonafide/spoof prediction.

C. Proposed Method and Network Architecture

The proposed FacePAD framework operates on RGB video clips $X = \{\mathbf{x}_t\}_{t=0}^{\tau-1}$, as mentioned in Sec. II-B. The overall framework, presented in Fig. 2, is summarized in Algorithm 1, consisting of three main model components of spatial feature extraction, Bahdanau temporal attention-based fusion, and bonafide/spoof classification.

Each frame $\mathbf{x}_t \in \mathbb{R}^{3\times224\times224}$ is processed independently by a CNN backbone ϕ_Θ. In this work, we employ MobileNetV3-Large as the feature extractor due to its favorable performance–efficiency trade-off, which is critical for real-time FacePAD deployment on edge and embedded platforms [13], [19], [20]. MobileNetV3-Large CNN backbone provides compact yet discriminative spatial representations while maintaining low computational complexity. For each time step t, $\mathbf{f}_t = \phi_\Theta(\mathbf{x}_t) \in \mathbb{R}^D$, where $D = 960$ denotes the feature dimension of the backbone. The sequence of frame-level embeddings form

$$F = [\mathbf{f}_0, \ldots, \mathbf{f}_{\tau-1}] \in \mathbb{R}^{\tau\times D}. \tag{4}$$

We adopt an additive attention mechanism inspired by [21] to model temporal dependencies across frames. Instead of treating all frames equally, this attention module learns to assign relatively higher weights to temporally informative frames to better discriminate between *bonafide* and *Spoof* classes.

For each frame-level feature $\mathbf{f}_t$, an attention score is computed as

$$e_t = v_\Theta^\top \tanh(W_\Theta \mathbf{f}_t), \tag{5}$$

where W_Θ and v_Θ are learnable parameters. The scores are normalized across time using softmax:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=0}^{\tau-1}\exp(e_k)}, \quad t = 0, \ldots, \tau-1. \tag{6}$$

The final spatio-temporal representation is obtained via a weighted aggregation:

$$\mathbf{h} = \sum_{t=0}^{\tau-1} \alpha_t \mathbf{f}_t \in \mathbb{R}^D. \tag{7}$$

This adaptive fusion mechanism emphasizes frames containing discriminative motion cues, such as subtle facial dynamics while suppressing redundant or less informative observations. Importantly, temporal modeling is achieved without introducing recurrent units or computationally expensive 3D convolutions, thereby preserving architectural efficiency.

The aggregated representation $\mathbf{h}$ is passed to a fully connected classifier

$$\mathbf{z} = g_\Theta(\mathbf{h}) = [z_{\text{Bonafide}}, z_{\text{Spoof}}] \in \mathbb{R}^2. \tag{8}$$

The softmax function produces class probabilities, and the predicted label is obtained as

$$\hat{y} = \arg\max\left(z_{\text{Bonafide}}, z_{\text{Spoof}}\right). \tag{9}$$

The proposed architecture integrates lightweight spatial encoding with adaptive temporal attention to capture both appearance-based and motion-based cues relevant to presentation attack detection. By combining MobileNetV3-Large with Bahdanau temporal attention, the framework effectively models spatio-temporal inconsistencies in spoofing attempts while maintaining computational efficiency.

D. Implementation Details

The proposed model is implemented using the PyTorch deep learning framework and trained on an NVIDIA GeForce RTX A4500 GPU. All experiments are conducted using RGB video clips with 224×224 spatial resolution and τ sampled frames. The network parameters Θ, including the MobileNetV3-Large backbone, Bahdanau attention module, and classifier head, are optimized using the Adam optimizer with a learning rate of 10^{-4}. Binary cross-entropy loss is employed for bonafide/Spoof classification.

The training is conducted for a maximum of 100 epochs with a batch size of 16. Early stopping is applied with a patience of 10 epochs based on the validation loss to prevent overfitting. The model checkpoint achieving the lowest validation loss is selected for final evaluation. No data augmentation is applied during validation or testing, ensuring deterministic evaluation.

Authorized licensed use limited to: King Fahd University of Petroleum and Minerals. Downloaded on September 08,2026 at 18:13:40 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Proposed Attention-based Spatio-Temporal FacePAD Framework.

Input: Θ (trained parameters), V (input video), τ (number of frames)
Output: $\hat{y} \in \{\text{Bonafide}, \text{Spoof}\}$
Model: CNN backbone $\phi_\Theta : \mathbb{R}^{3\times H\times W} \to \mathbb{R}^D$ $(D = 960)$;
Bahdanau attention (W_Θ, v_Θ); classifier $g_\Theta : \mathbb{R}^D \to \mathbb{R}^2$.
1) Frame Sampling: Sample $\{\mathbf{x}_t\}_{t=0}^{\tau-1}$ from V, $\mathbf{x}_t \in \mathbb{R}^{3\times H\times W}$; apply preprocessing.
2) Feature Extraction: for $t = 0, \ldots, \tau - 1$ **do**
  $\mathbf{f}_t = \phi_\Theta(\mathbf{x}_t) \in \mathbb{R}^D$
end
$F = [\mathbf{f}_0, \ldots, \mathbf{f}_{\tau-1}] \in \mathbb{R}^{\tau\times D}$.
3) Temporal Attention: for $t = 0, \ldots, \tau - 1$ **do**
  $e_t = v_\Theta^\top \tanh(W_\Theta \mathbf{f}_t)$
end
$\alpha_t = \frac{\exp(e_t)}{\sum_{k=0}^{\tau-1}\exp(e_k)}$; Score normalization
$\mathbf{h} = \sum_{t=0}^{\tau-1} \alpha_t \mathbf{f}_t \in \mathbb{R}^D$.
4) Classification: $\mathbf{z} = g_\Theta(\mathbf{h})$; apply softmax;
$\hat{y} = \arg\max(\mathbf{z})$.
return $\hat{y}$

E. Performance Evaluation Metrics

The proposed framework is evaluated using standard FacePAD metrics for performance. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Accuracy is computed as $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$. Sensitivity (TPR) and specificity (TNR) are defined as $\text{Sensitivity} = \frac{TP}{TP+FN}$ and $\text{Specificity} = \frac{TN}{TN+FP}$, respectively.

The false acceptance rate (FAR) and false rejection rate (FRR) are defined as $\text{FAR} = \frac{FP}{FP+TN}$ and $\text{FRR} = \frac{FN}{FN+TP}$. The half total error rate (HTER) is computed as $\text{HTER} = \frac{\text{FAR}+\text{FRR}}{2}$. The equal error rate (EER) corresponds to the operating point where FAR = FRR.

The area under the ROC curve (AUC-ROC) provides a threshold-independent evaluation by integrating the receiver operating characteristic curve across varying decision thresholds. Youden's Index is defined as $\text{Sensitivity}+\text{Specificity}-1$, indicating the optimal balance between true positive and true negative rates.

For the OULU-NPU dataset, we additionally report APCER, BPCER, and ACER, computed as $\text{ACER} = \frac{\text{APCER}+\text{BPCER}}{2}$, where APCER and BPCER are defined as attack presentation classification error rate and bonafide presentation classification error rate, respectively.

In all experiments, the temporal length τ used during training is kept identical during testing to ensure consistent temporal modeling between optimization and inference. These metrics collectively provide a comprehensive evaluation of the proposed FacePAD system in detecting presentation attacks.

III. Experimental Results

A. Ablation Study

An ablation study is conducted to analyze the impact of the temporal sampling parameter τ on the performance of the proposed framework using the ROSE-Youtu dataset. This dataset is specifically selected due to its diverse and challenging attack scenarios, ranging from printed photo attacks to complex paper mask presentations, thereby providing a rigorous evaluation setting.

TABLE II: Ablation study on the RY dataset for different temporal lengths τ.

Metric \ τ	$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 10$	$\tau = 20$
Accuracy (%)	98.06	99.03	98.79	98.74	99.43
AUC-ROC	0.999	0.9994	0.9993	0.9994	0.9996
EER (%)	1.39	1.39	1.39	0.92	0.38
HTER (%)	1.47	1.29	1.12	0.80	0.53
FAR (%)	1.39	1.69	1.00	0.92	0.38
FRR (%)	1.56	0.89	0.46	0.67	0.67
Youden's Index	0.971	0.974	0.978	0.984	0.989

Table II presents the ablation analysis of the temporal parameter τ on the ROSE-Youtu dataset. The results indicate a clear relationship between the number of sampled frames and the discriminative capability of the proposed method. While even a single-frame configuration ($\tau = 1$) achieves strong performance, incorporating additional temporal context consistently improves performance. As τ increases from 1 to 20, both HTER and EER show a substantial reduction, with EER decreasing from 1.39% to 0.38% and HTER from 1.47% to 0.53%. This trend highlights the importance of temporal attention mapping in capturing motion-based inconsistencies present in replay and mask attacks. The corresponding computational cost at inference for different τ values is reported in Table III.

Although the overall accuracy and AUC-ROC values remain consistently high across all configurations, the more discriminative metrics such as EER, HTER, FAR, and FRR reveal meaningful improvements with longer temporal windows. In particular, the steady increase in Youden's Index from 0.971 to 0.989 reflects enhanced separability between bonafide and spoof classes as more temporal information is incorporated. The reduction in FAR with larger τ values further suggests improved suppression of spoof samples, while FRR remains relatively stable, indicating that the model maintains reliable bonafide detection.

Overall, the ablation study confirms that the proposed adaptive spatio-temporal aggregation benefits from richer temporal context, with performance progressively improving as τ increases. At the same time, promising results even at smaller τ values demonstrate that the proposed method effectively identifies informative frames without requiring excessively long sequences.

Authorized licensed use limited to: King Fahd University of Petroleum and Minerals. Downloaded on September 08,2026 at 18:13:40 UTC from IEEE Xplore. Restrictions apply.

TABLE III: Impact of number of frames on computational cost at inference.

Number of frames	Parameters	GFLOPs	Latency (ms)	Throughput
$\tau = 1$	3.589M	0.47	4.68±0.28	214.36 ±13.92
$\tau = 3$	3.589M	1.40	4.33±0.39	233.00 ±20.35
$\tau = 5$	3.589M	2.33	4.77±0.14	209.84 ±7.43
$\tau = 10$	3.589M	4.65	4.30±0.42	234.71 ±20.57
$\tau = 20$	3.589M	9.30	5.44±0.02	183.76 ±0.53

Results are reported as mean ± standard deviation over 1000 repeated forward passes on NVIDIA GeForce RTX A4500 GPU, with batch size 1.

B. Comparison with State-of-the-Art Face PAD Methods

Table IV presents a consolidated comparison of the proposed method with recent state-of-the-art approaches across the Replay-Attack, Replay-Mobile, and ROSE-Youtu datasets. On the RA and RM datasets, the proposed method consistently achieves an HTER of 0.0% across all tested temporal configurations, matching or surpassing the best-performing contemporary methods, including ResNet50V2 [14], depth-augmented teacher–student frameworks [19], and other recent spatio-temporal approaches [13]. Compared to earlier deep architectures such as 3D ConvNet [22] and EfficientNet-B0 [23], which report significantly higher error rates, the proposed framework demonstrates clear improvements in robustness while maintaining a structurally simpler temporal modeling strategy.

The RY dataset presents a more challenging evaluation scenario due to its diverse attack types and recording conditions. In this setting, the proposed method exhibits a clear performance trend as the temporal parameter τ increases. While smaller temporal windows ($\tau = 3$ and $\tau = 5$) already provide competitive results, increasing the number of sampled frames progressively reduces both HTER and EER. Notably, with $\tau = 20$, the proposed method achieves one of the lowest reported error rates among RGB-only approaches, outperforming transformer-based models such as ViViT [24] and several recent depth-augmented and adversarially enhanced methods [19], [20], while maintaining competitive performance compared to the Spatio-Temporal method [13] with $\tau = 50$ frames. This performance gain highlights the effectiveness of the proposed attention-based method in capturing discriminative motion patterns that are particularly relevant for complex spoofing scenarios.

Overall, the comparative results demonstrate that the proposed attention-based spatio-temporal fusion strategy generalizes well across datasets of varying difficulty. The consistent zero-error performance on RA and RM, combined with progressively improved results on the more challenging RY dataset, confirms the importance of leveraging attention-based temporal weighting to enhance robustness in RGB-based FacePAD systems.

Table V presents the comparative performance evaluation on the OULU-NPU dataset. The proposed method achieves the lowest ACER of 0.069%, outperforming all listed state-of-the-art approaches. In particular, it significantly improves upon recent Spatio-Temporal [13] and Depth-augmented student models [19], which report ACER values above 0.5%. The proposed framework also achieves an APCER of 0.139% and a BPCER of 0.00%, indicating strong attack detection capability while maintaining perfect bonafide classification. Compared to earlier handcrafted texture-based methods and more complex 3D CNN architectures, the results demonstrate that the proposed attention-based temporal aggregation strategy provides superior discrimination performance on challenging and high-quality attack scenarios.

Overall, the comparative evaluations across Replay-Attack, Replay-Mobile, ROSE-Youtu, and OULU-NPU demonstrate the consistent effectiveness of the proposed attention-based spatio-temporal framework. The proposed method achieves state-of-the-art or highly competitive performance across all benchmarks, including perfect HTER on Replay-Attack and Replay-Mobile, progressively improved results on the challenging ROSE-Youtu dataset with increasing temporal context, and the lowest reported ACER on OULU-NPU. These

TABLE IV: Comparison with state-of-the-art methods on Replay-Attack, Replay-Mobile, and ROSE-Youtu datasets.

Method	Year	RA	RM	RY
3D ConvNet [22]	2023	11.7	8.7	21.3/-
Image Diffusion [25]	2023	0.09	1.14	4.92/4.95
ResNet50V2 [14]	2023	0.03	0.0	2.53/2.64
ViViT [24]	2023	–	–	13.28/2.46
EfficientNet-B0 [23]	2024	36.88	4.62	9.54/-
Spatio-Temporal [13], [19]	2025	0.0	0.0	1.47/0.85
Deformable Conv. [26]	2025	0.0	0.0	1.26/0.80
AdvSpoofGuard [20]	2025	0.0	0.0	1.97/1.08
Depth-aug. Teacher [19]	2025	0.0	0.0	1.02/1.15
Depth-aug. Student [19]	2025	0.0	0.0	1.01/1.39
Proposed Method ($\tau = 3$)	2026	0.0	0.0	1.29/1.39
Proposed Method ($\tau = 5$)	2026	0.0	0.0	1.12/1.39
Proposed Method ($\tau = 10$)	2026	0.0	0.0	0.80/0.92
Proposed Method ($\tau = 20$)	2026	0.0	0.0	0.53/0.38

Results are HTER (%) for RA & RM, HTER/EER (%) for RY.

TABLE V: Performance comparison of the proposed method with the state-of-the-art methods on OULU-NPU dataset.

Method	Year	APCER (%)	BPCER (%)	ACER (%)
CS-LBP [27]	2020	14.9	10.5	12.7
CS-VGG [27]	2020	3.3	9.4	6.35
ED-LBP (VAR) [28]	2021	11.3	8.4	9.9
Texture (VAR) [29]	2021	14.5	15	14.8
Fake-Net (VAR) [30]	2021	5.4	6.9	6.2
OFT (VAR) [31]	2022	5.7	2.7	4.2
3D-LCN [32]	2024	1.5	0.5	1.0
Spatio-Temporal [13] ‡	2025	1.94	0.55	1.25
Deformable Conv. [26]	2025	0.13	0.27	0.21
Depth-aug. Student (MB-V3) [19]	2025	1.39	0.83	1.11
Depth-aug. Student (MB-V2) [19]	2025	0.28	0.83	0.56
Proposed Method ($\tau = 3$)	**2025**	**0.139**	**0.00**	**0.069**

CS: Compact Space, VAR: Video Attack Return, OFT: Optical Flow + Texture, MB: MobileNet. ‡ means the method is retested on this dataset.

Authorized licensed use limited to: King Fahd University of Petroleum and Minerals. Downloaded on September 08,2026 at 18:13:40 UTC from IEEE Xplore. Restrictions apply.

results collectively validate the effectiveness of the proposed adaptive temporal aggregation in enhancing discriminative capability for RGB-based FacePAD, confirming that the proposed framework generalizes well across diverse datasets and attack scenarios.

IV. Conclusion

This paper introduced an attention-based spatio-temporal framework for RGB-only Face Presentation Attack Detection that adaptively aggregates frame-level features using Bahdanau temporal attention. By learning to emphasize discriminative motion cues while suppressing less informative frames, the proposed method overcomes the limitations of uniform temporal aggregation without relying on recurrent or 3D CNN architectures. Experimental results demonstrate strong and consistent performance across multiple benchmark datasets, achieving HTERs of 0% on Replay-Attack and Replay-Mobile, 0.53% on ROSE-Youtu, and an ACER of 0.069% on OULU-NPU. These results confirm that adaptive temporal weighting significantly enhances discrimination capability in RGB-based FacePAD and provides a robust solution across diverse and challenging attack scenarios.

References

[1] J. Galbally *et al.*, "Biometric anti-spoofing methods: A survey in face recognition systems," *IEEE Security and Privacy*, vol. 12, no. 5, pp. 30–37, 2014.

[2] P. Grother, P. Grother, M. Ngan, and K. Hanaoka, "Face recognition vendor test (frvt) part 2: Identification," 2019.

[3] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, 2011, pp. 1–7.

[4] G. Pan, L. Sun, Z. Wu, and Y. Wang, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, 2007.

[5] W. Bao, H. Li, N. Li, and W. Jiang, "Liveness detection for face recognition based on optical flow field," *2009 International Conference on Image Analysis and Signal Processing*, pp. 233–236, 2009.

[6] 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 Transactions on Information Forensics and Security*, vol. 15, pp. 578–593, 2019.

[7] 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 Transactions on Information Forensics and Security*, vol. 15, pp. 3181–3196, 2020.

[8] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 389–398.

[9] R. Shao, X. Lan, and P. C. Yuen, "Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 4, pp. 923–938, 2018.

[10] U. Muhammad, Z. Yu, and J. Komulainen, "Self-supervised 2d face presentation attack detection via temporal sequence sampling," *Pattern Recognition Letters*, 2022.

[11] C. Dhiman, A. Antil, A. Anand, and S. Gakhar, "A deep face spoof detection framework using multi-level elbps and stacked lstms," *Signal, Image and Video Processing*, pp. 1–14, 2024.

[12] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.

[13] S. Khan, T. H. M. Siddique, M. S. Ibrahim, A. J. Siddiqui, and K. Huang, "Spatio-temporal deep learning for improved face presentation attack detection," *Knowledge-Based Systems*, p. 113059, 2025.

[14] M. O. Alassafi, M. S. Ibrahim, I. Naseem, R. AlGhamdi, R. Alotaibi, F. A. Kateb, H. M. Oqaibi, A. A. Alshdadi, and S. A. Yusuf, "Fully supervised contrastive learning in latent space for face presentation attack detection," *Applied Intelligence*, vol. 53, no. 19, pp. 21 770–21 787, 2023.

[15] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*. IEEE, 2012, pp. 1–7.

[16] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2016, pp. 1–7.

[17] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 1794–1809, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4624345

[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 *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 612–618.

[19] M. S. Jabbar, T. H. M. Siddique, K. Huang, and S. Khan, "Knowledge distillation with predicted depth for robust and lightweight face presentation attack detection," *Knowledge-Based Systems*, p. 114325, 2025.

[20] T. H. M. Siddique, S. Khan, Z. Wang, and K. Huang, "Advspoofguard: Optimal transport driven robust face presentation attack detection system," *Knowledge-Based Systems*, p. 113759, 2025.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[22] S. Giurato, A. Ortis, and S. Battiato, "Real-time multiclass face spoofing recognition through spatiotemporal convolutional 3d features," in *International Conference on Image Analysis and Processing*. Springer, 2023, pp. 356–367.

[23] V. D. Huszár and V. K. Adhikarla, "Securing phygital gameplay: Strategies for video-replay spoofing detection," *IEEE Access*, 2024.

[24] M. Marais, D. Brown, J. Connan, and A. Boby, "Facial liveness and anti-spoofing detection using vision transformers," in *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, vol. 8, 2023, p. 2023.

[25] M. O. Alassafi, M. S. Ibrahim, I. Naseem, R. AlGhamdi, R. Alotaibi, F. A. Kateb, H. M. Oqaibi, A. A. Alshdadi, and S. A. Yusuf, "A novel deep learning architecture with image diffusion for robust face presentation attack detection," *IEEE Access*, vol. 11, pp. 59 204–59 216, 2023.

[26] S. M. Ibrahim, M. S. Ibrahim, S. Khan, Y.-W. Ko, and J.-G. Lee, "Improving face presentation attack detection through deformable convolution and transfer learning," *IEEE Access*, 2025.

[27] L. Li, Z. Xia, X. Jiang, F. Roli, and X. Feng, "Compactnet: learning a compact space for face presentation attack detection," *Neurocomputing*, vol. 409, pp. 191–207, 2020.

[28] X. Shu, H. Tang, and S. Huang, "Face spoofing detection based on chromatic ed-lbp texture feature," *Multimedia Systems*, vol. 27, no. 2, pp. 161–176, 2021.

[29] N. Daniel and A. Anitha, "Texture and quality analysis for face spoofing detection," *Computers & Electrical Engineering*, vol. 94, p. 107293, 2021.

[30] M. Alshaikhli, O. Elharrouss, S. Al-Maadeed, and A. Bouridane, "Face-fake-net: The deep learning method for image face anti-spoofing detection: Paper id 45," in *2021 9th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2021, pp. 1–6.

[31] Z. Li *et al.*, "Facepad: a critical review of face presentation attack detection systems," *Pattern Recognition*, vol. 113, p. 107749, 2022.

[32] Z. Ning, W. Zhang, and J. Yang, "Face anti-spoofing based on 3d learnable convolutional operators," in *2024 36th Chinese Control and Decision Conference (CCDC)*. IEEE, 2024, pp. 4034–4040.

Authorized licensed use limited to: King Fahd University of Petroleum and Minerals. Downloaded on September 08,2026 at 18:13:40 UTC from IEEE Xplore. Restrictions apply.