Multiattention-Net: A Novel Approach to Face Anti-Spoofing with Modified Squeezed Residual Blocks

Sabari Nathan[1] M.Parisa Beham[2] A Nagaraj[2] S. Mohamed Mansoor Roomi[3]
[1]Couger Inc, Japan, [2]Sethu Institute of Technology, India, [3]Thiagarajar college of engineering, India

Abstract

Introducing a novel Attack-agnostic Face Anti-spoofing framework, this paper addresses the challenge of determining the authenticity of a captured face in face recognition systems. Current methods, trained on existing fake faces, often lack generalization and perform poorly against unseen attacks. The proposed framework presents a fresh approach to face anti-spoofing, leveraging modified squeezed residual blocks and attention mechanisms. Convolutional layers within the Multiattention-Net architecture capture spatially hierarchical features, enhancing feature representation and improving the network's sensitivity to critical features. These spatial features are refined through a dual attention block to emphasize important features. The squeeze-and-excitation (SE) mechanism further enhances the representation by recalibrating channel-wise responses to emphasize informative features, incorporating global average pooling and channel-wise excitation. The Multiattention-Net achieves a balanced trade-off between feature richness and computational efficiency, demonstrating superior performance in face anti-spoofing tasks. Experimental results on benchmark datasets validate the effectiveness of this approach, highlighting its potential for real-world applications in security and biometric authentication.

1. Introduction

The rise of the internet has fueled the widespread adoption of biometric technologies for security and identification purposes. Face recognition, with its convenience and accuracy, has become a prominent choice in various domains, including finance, social security, and intelligent security systems [9]. However, face recognition systems remain vulnerable to spoofing attacks, where imposters attempt to gain unauthorized access using artifacts like photographs or masks (Presentation Attacks, PAs) [6]. Therefore, robust and efficient face anti-spoofing (FAS) systems are critical to ensure the integrity of face recognition applications.

Deploying FAS systems in real-time scenarios, such as mobile devices and surveillance cameras, necessitates computationally efficient algorithms capable of real-time data processing. Balancing accuracy with computational efficiency remains a key challenge in such settings. FAS systems also raise privacy concerns regarding biometric data collection and processing. Striking a balance between effective anti-spoofing measures and user privacy is a complex issue that requires careful consideration.

Driven by these challenges, research on face anti-spoofing detection has witnessed significant growth in recent years, yielding promising advancements. This paper proposes a novel approach to face anti-spoofing leveraging modified squeezed residual blocks (MSRs).

2. Related Work

Face recognition technology is getting more and more widespread these days. Face anti-spoofing detection is an important face recognition component that has garnered a lot of attention and grown into a mostly independent field of research. Traditional approaches often rely on texture analysis to differentiate between real and fake faces. Techniques such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Local Phase Quantization (LPQ) have been used for feature extraction. Spoofed facial images or videos may lack natural facial movements. Analyzing motion patterns can help detect anomalies that indicate spoofing.

2.1. Face Anti-Spoofing based on Image Texture

A recent study by Thippeswammy, Vinutha, and Dhanapal [6] suggested a technique based on a collection of local appearance-based approaches' texture descriptors. The classification of photos into real and fake ones done by using the K-nearest neighbors (KNN) classifier. The NUAA Photo Imposter database was used to test their system.

A face anti-spoofing detection method based on support vector machine recursive feature elimination (SVM-RFE) and color texture Markov feature (CTMF) was presented by Zhang et al.[12] . Dimension reduction using SVM-RFE is applied to make it appropriate for real-time detection.

A face anti-spoofing technique based on color texture

analysis was presented by Boulkenaf et al.[29]. To create the final descriptor, they extract LBP histograms from a single image channel and connect them. The color texture-based approach outperforms the gray texture-based approach.

A novel method for detecting liveness using general image quality assessment was introduced by Galbally and Marcel in [7].A low-pass Gaussian kernel filter is applied to obtain the image. A Simple Linear Discriminant Analysis is used to classify the data as true or fraudulent (LDA). The experimental validation was conducted using two publicly available datasets, namely the CASIA FAS-Datbase and the replay-attack database.

In order to extract texture feature histograms from local blocks of grayscale images and global images, Maatta et al. [10] used multiple uniform LBP operators of different scales. They then connected these histograms to create a 531-dimensional feature histogram, which they then sent to an SVM classifier with RBF as the core for training and testing of real human faces and deceptive face classification.

In order to determine whether the recognized image is a real face or a fake face, Peixoto et al. [1] first used the DoG filter to obtain the medium frequency band information in the image information. They then used the Fourier transform to extract key features, and finally they used a logistic regression classifier to differentiate and classify the extracted and processed feature information.

The method based on image texture analysis has many advantages, including low costs, a simple algorithm, and easy implementation. High-definition cameras and the use of high-quality 3D masks have made the use of texture information no longer feasible.

2.2. Face Anti-Spoofing Based on Deep Learning

An increasing number of researchers are applying deep learning to face anti-spoofing in an effort to find more potent ways to counteract face deception, thanks to the technology's remarkable performance in face recognition and ongoing development. Deep learning, in contrast to the conventional manual feature extraction method, has the ability to automatically learn photos, extract more plentiful and important facial features, and assist in properly differentiating between real and fake faces. Deep learning methods, particularly CNNs, have shown significant advancements in face anti-spoofing. CNN architectures are trained on large datasets of both real and spoofed facial images to learn discriminative features.

Yang et al. [24] created a unique Spatio-temporal Anti-spoofing Network (STASN) that considers both local and global spatial information. The three components of the model are SASM, RAM, and TASM. TASM is a CNN-LSTM that uses a frame sequence as input, extracts CNN features first, propagates LSTM, and then predicts the binary classification outcome.

In order to classify the spoofing samples into semantic subgroups, Liu et al. [28] defined the detection of unknown spoofing attacks as Zero-shot Facial Anti-Spoofing (ZSFA) and proposed a novel Deep Tree Network (DTN) that was used to train trees in an unsupervised manner to find the feature library with the greatest variation. Furthermore, the author developed the first face anti-spoofing database, SiW-M, comprising multiple forms of deception, to better research ZSFA.

An technique that harmoniously blends CNN and RNN (Recurrent Neural Network) architecture was put forth by Liu, Jourabloo, and Liu [26]. In contrast to other deep learning techniques

In order to perform face anti-spoofing, Liu et al. [27] integrated face depth data and rPPG signal. They noted that the binary classification problem was substituted with the targeted feature supervision problem. To accomplish the two types of supervision, the author created a deep learning technique utilizing CNN - RNN architectures. CNN recognizes the delicate texture by using the supervision of the depth image.

In [25], a two-stream CNN (Convolution Neural Networks)-based technique was presented. It uses the complete face image and patches from the same face to differentiate between artificial and genuine faces. When compared to prior state-of-the-art performance, the experimental results on the CASIA-FASD, MSU-USSA, and replay attack datasets demonstrate a remarkable performance.

Face depth map was initially used by Atoum et al. [25] as the primary data for differentiating between face faking. This research proposes a two-channel CNN based face anti-spoofing technique that combines depth information with local aspects of face images.

A completely data-driven hyper-depth model based on transfer learning was presented by Tu et al. [22]. After extracting the spatial features of sequence frames using a pre-trained deep residual network (ResNet-50) [**?**], the model inputs the spatial features into an LSTM unit to obtain temporal features that can be used for final classification. Boulkenafet et al.[**?**] treated face spoof detection as a binary classification task. In order to identify hidden texture features that result in different depths for actual and artificial faces, the CNN component uses depth map supervision. The RNN portion learns to estimate the rPPG signal, which verifies the temporal variability of the rPPG signal instead of computing it in the conventional manner . A unified network framework for iris, face, and fingerprint spoofing detection was proposed by Menotti et al. [2]. Through two optimizations, Architecture optimization (AO) and filter optimization (FO), which randomly search for the best convolutional neural network among a number of networks specified in the hyper-parametric search space, the model

learns representations directly from input.

Yang et al. [11] originally proposed using convolutional neural networks (CNNs) to extract features for face anti-spoofing in 2014. This study paved the way for new developments in deep learning for face anti-spoofing. The detection effect was significantly lower than with traditional methods because the technology was not yet mature.

Datasets and challenges play a critical role in advancing face anti-spoofing research. The CASIA-SURF dataset, along with its associated Multi-modal Face Anti-spoofing Attack Detection Challenge [31], focuses on multi-modal anti-spoofing. The CASIA-SURF CeFA dataset[14] addresses cross-ethnicity issues. The CASIA-SURF HiFi-Mask dataset[15] targets 3D high-fidelity mask attacks, and the SuHiFiMask dataset[3] tackles face anti-spoofing in surveillance scenarios. The recently introduced UniAttack-Data dataset[4] unifies physical and digital attack detection within a single benchmark. These resources provide standardized benchmarks, facilitating method comparison and driving innovation in face anti-spoofing techniques. Several recent advancements have emerged in face anti-spoofing research. CFPL-FAS [17] leverages prompt learning to achieve generalizable anti-spoofing, aiming for robust performance across diverse attack classes. Multi-Domain Incremental Learning[23] proposes a method to improve model robustness by incrementally learning from multiple data domains. Fm-vit[16] introduces a flexible modal vision transformer architecture for enhanced model efficiency and effectiveness. Finally, MA-ViT[13] focuses on modality-agnostic vision transformers to learn representations that are invariant to presentation attacks.

Face anti-spoofing based on deep learning is gaining attention from many researchers due to its superior feature extraction. The capacity of face anti-spoofing based on deep learning has progressively improved through network updating, transfer learning, integration of multiple features, and domain generalization with the tireless efforts and repeated attempts of numerous scholars, and has now surpassed the conventional method.

To summarize, the key contributions of our work are outlined as follows:

1. Our architecture introduces MSRs, extending traditional residual blocks with squeeze-and-excitation (SE) modules. MSRs dynamically recalibrate channel-wise feature responses, leading to a more discriminative feature space emphasising informative details critical for robust face anti-spoofing.

2. Post feature extraction, our network incorporates dual attention mechanisms operating at varying spatial levels. These mechanisms selectively emphasize relevant spatial features in facial expressions, enabling the network to focus on crucial details across resolutions and enhancing its ability to differentiate genuine faces from spoofing attempts.

3. Our architecture strikes a balance between anti-spoofing performance and computational efficiency. With approximately 6 million parameters and 1.99 G FLOPs, the model delivers competitive performance while maintaining a lightweight structure suitable for real-world deployment.

3. Proposed Method

This work proposes a novel network architecture for face anti-spoofing that discriminates between genuine facial expressions and spoofing attempts. The architecture leverages state-of-the-art techniques, including modified squeezed residual blocks[8] (MSRs) and dual attention [30] mechanisms, to capture intricate spatial information and adaptively recalibrate feature responses for robust spoofing detection.The proposed multi-level attention network is inspired by [18–20] .

3.0.1 Feature Extraction

The network takes a facial image as input and processes it through a series of convolutional layers for feature extraction. The first convolutional layer utilizes a 7x7 filter kernel to extract complex local patterns and structures. This is followed by max-pooling layers for downsampling the feature maps, enabling the network to learn abstract features at different spatial resolutions.

The core component of the feature extraction process resides in the modified squeezed residual block (MSR). This block builds upon the traditional residual block by incorporating additional squeeze-and-excitation (SE) modules. These SE modules dynamically recalibrate channel-wise feature responses, enhancing the discriminative power of the network by emphasizing informative features and suppressing irrelevant ones.

3.0.2 Dual Attention Mechanisms

The second stage of the network employs dual attention [30] mechanisms to further refine the features extracted from the MSRs. These mechanisms selectively emphasize important spatial features while suppressing background information, effectively capturing the subtle details of facial expressions crucial for distinguishing genuine faces from spoofing attempts. Notably, the dual attention mechanisms operate at different spatial levels, allowing the network to focus on relevant features at each processing stage and achieve a more comprehensive understanding of the input face.

3.1. Modified squeezed residual block

The modified squeezed residual block (MSR) [8] is a fundamental component of our face anti-spoofing network. It is represented in Figure 2. It integrates squeeze-and-excitation (SE) blocks with convolutional layers to enhance feature

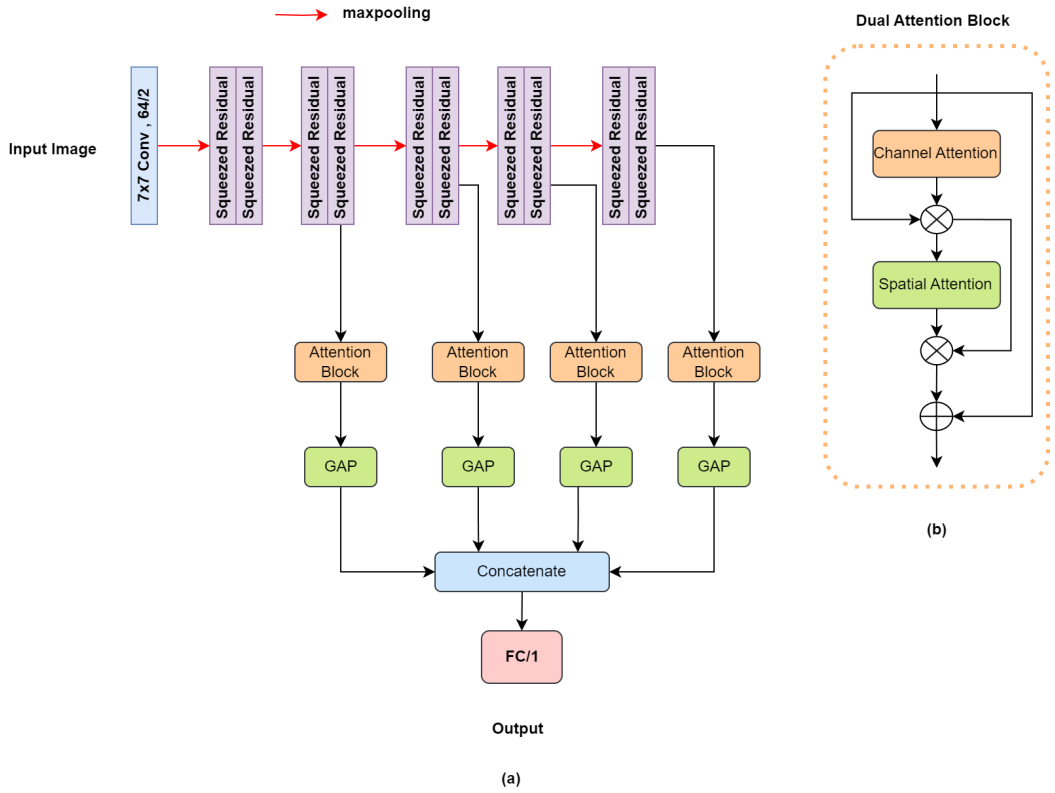

Figure 1. (a) Architecture diagram of proposed net (b) dual attention block

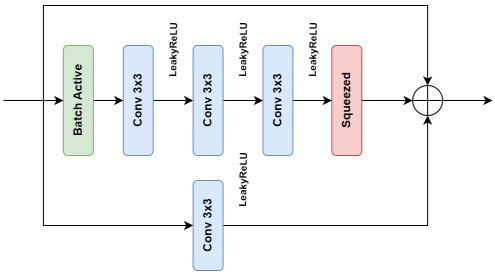

Figure 2. Block Diagram of Modified squeezed residual block

representation and adaptively recalibrate channel-wise feature responses.

Let $\mathbf{X}_{\text{input}} \in \mathbb{R}^{H \times W \times C}$ denote the input tensor, where H, W, and C represent the height, width, and number of input channels, respectively.

3.1.1 Primary Convolutional Layer

The input tensor is processed by a convolutional layer with a set of filters to generate the primary feature tensor $\mathbf{X}_{\text{primary}} \in \mathbb{R}^{H' \times W' \times F}$:

$$\mathbf{X}_{\text{primary}} = \text{Conv2D}(\mathbf{X}_{\text{input}}, \mathbf{W}_p, \mathbf{b}_p) + \mathbf{b}_p \quad (1)$$

Here, Conv2D denotes the 2D convolution operation, $\mathbf{W}_p \in \mathbb{R}^{k_p \times k_p \times C \times F}$ represents the filter bank, k_p is the filter size, F is the number of output channels, and $\mathbf{b}_p \in \mathbb{R}^F$ is the bias vector.

3.1.2 Main Convolutional Layers with Leaky ReLU Activation

The input tensor undergoes batch normalization (BN) followed by a Leaky ReLU activation function, denoted as LeakyReLU$(\cdot)$:

$$\mathbf{X} = \text{LeakyReLU}\left(\text{BN}(\mathbf{X}_{\text{input}})\right) \quad (2)$$

Three consecutive convolutional layers are applied to $\mathbf{X}$ with Leaky ReLU activation in between:

$$\mathbf{X} = \text{LeakyReLU}\left(\text{Conv2D}(\mathbf{X}, \mathbf{W}_c, \mathbf{b}_c) + \mathbf{b}_c\right) \quad (3)$$

where $\mathbf{W}_c \in \mathbb{R}^{k_c \times k_c \times C \times F}$ represents the filters for the convolutional layers and $\mathbf{b}_c$ is the corresponding bias vector.

3.1.3 Squeeze and Excitation (SE) Layer:

The output tensor $\mathbf{X}$ is globally average pooled to generate a squeeze vector $\mathbf{z} \in \mathbb{R}^F$:

$$\mathbf{z} = \text{GlobalAvgPool}(\mathbf{X}) \tag{4}$$

The squeeze vector is passed through two fully-connected (FC) layers with ReLU and sigmoid activations, respectively, to compute channel-wise attention weights $\mathbf{s} \in \mathbb{R}^F$:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \text{ReLU}(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2) \tag{5}$$

where $\mathbf{W}_1 \in \mathbb{R}^{U \times F}$, $\mathbf{W}_2 \in \mathbb{R}^{F \times U}$, $\mathbf{b}_1 \in \mathbb{R}^U$, and $\mathbf{b}_2 \in \mathbb{R}^F$ represent the weights and biases for the FC layers, U is the number of hidden units in the first FC layer, and σ denotes the sigmoid function.

The attention weights are reshaped and element-wise multiplied with $\mathbf{X}$ to obtain the scaled feature tensor $\mathbf{X}_{\text{scale}} \in \mathbb{R}^{H' \times W' \times F}$:

$$\mathbf{X}_{\text{scale}} = \mathbf{X} \odot \mathbf{s} \tag{6}$$

3.1.4 Residual Connection with Modified Output

The final output is formed by adding the primary feature tensor $\mathbf{X}_{\text{primary}}$, the original input tensor $\mathbf{X}_{\text{input}}$ processed by a separate convolutional layer, and the scaled feature tensor $\mathbf{X}_{\text{scale}}$:

Certainly, the complete equation for the residual output is:

$$\text{Output} = \mathbf{X}_{primary} + \text{Conv2D}(\mathbf{X}_{input}, \mathbf{W}_r, \mathbf{b}r) + \mathbf{X}_{scale} \tag{1}$$

Here's a breakdown of the remaining terms:

* $\mathbf{W}_r \in \mathbb{R}^{k_r \times k_r \times C \times F}$ represents the filter bank for the additional convolutional layer applied to the input tensor. * $\mathbf{b}_r \in \mathbb{R}^F$ is the bias vector for the additional convolutional layer.

Here, $\mathbf{W}_r \in \mathbb{R}^{k_r \times k_r \times C \times F}$ represents the filter bank for the additional convolutional layer and b_r is the corresponding bias vector. This residual connection allows the network to learn from both the identity mapping and the transformed features, potentially leading to faster convergence and improved performance.

The modified squeezed residual block (MSR) effectively combines convolutional layers, squeeze-and-excitation (SE) blocks, and residual connections to achieve robust feature representation for face anti-spoofing tasks. The SE block adaptively recalibrates channel-wise feature responses, leading to a more discriminative feature space for identifying genuine faces from spoofing attempts. The residual connection promotes efficient gradient flow and mitigates the vanishing gradient problem, allowing the network to learn deeper and more complex features. Overall, the MSR contributes significantly to the effectiveness of the face anti-spoofing network.

3.2. Dual Attention Mechanisms

The network employs a dual attention mechanism to capture both channel-wise and spatial dependencies in features. This mechanism refines feature maps by focusing on informative channels and emphasizing crucial spatial regions. Given an input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and number of channels respectively, the dual attention module outputs an attention-weighted feature map $\mathbf{S}_A \in \mathbb{R}^{H \times W \times C}$:

Channel Attention:

$$\mathbf{C}_A = \mathbf{X} \odot (\sigma(\text{MLP}(\text{Concat}(\mathbf{GAP}(\mathbf{X}), \mathbf{GAM}(\mathbf{X})))) \tag{7}$$

In this equation, $\mathbf{GAP}(\mathbf{X})$ and $\mathbf{GAM}(\mathbf{X})$ represent the global average pooling and global max pooling operations applied to the input feature map $\mathbf{X}$, respectively.

Spatial Attention:

$$\mathbf{S}_A = \mathbf{C}_A \odot (\sigma(\text{Conv2D}(\text{Concat}(\mathbf{P_a}(\mathbf{C_A}), \mathbf{P_m}(\mathbf{C_A}))))) \tag{8}$$

In this equation, $\mathbf{P}_a(\mathbf{C}_A)$ and $\mathbf{P}_m(\mathbf{C}_A)$ represent the average pooling and max pooling operations applied to the channel attention feature map $\mathbf{C}_A$, respectively.

4. Experiments

4.1. Experimental Setting

Four separate models were trained using the UPD dataset. The train dataset was randomly divided into training and validation sets with an 85:15 split. Data augmentation techniques were applied during training for all models. Each model addressed a specific protocol: Model 1 focused on Protocol 1 (unified attack detection), Model 2 on Protocol 2.1 (generalization to unseen physical attacks), Model 3 on Protocol 2.2 (generalization to unseen digital attacks), and Model 4 was trained on all three protocols combined.We utilized the Binary Focal Cross-Entropy loss function during the training phase of our models.

The Adam optimizer was employed with a learning rate schedule that decayed from 0.001 to 0.00001 over 150 training epochs. The training used an NVIDIA Tesla P100 GPU with 16GB of RAM and the TensorFlow Keras framework.

4.2. Dataset

We evaluated our proposed method on three publicly available face anti-spoofing datasets: The Unified Physical-Digital Face Attack Detection (UPD) [5], CelebA-Spoof [32], and Large Crowdcollected Facial Anti-Spoofing

Dataset (LCC) [21]. We focused on the UPD dataset for training due to its comprehensiveness and the availability of diverse spoofing attacks.

4.2.1 UPD Dataset and Protocols

The UPD dataset [5] provides three protocols for evaluating face anti-spoofing algorithms:

Protocol 1: Unified Attack Detection [5]: This protocol assesses the model's ability to detect both physical and digital spoofing attacks. The training, validation, and test sets encompass real human faces and all attack types included in the dataset. The significant intra-class variations and distances between different attack types and real faces present a challenging scenario for robust algorithm design.

Protocol 2: Generalization to Unseen Attacks: This protocol evaluates the model's ability to generalize to unseen attack types. Since physical and digital attacks have inherent differences, achieving algorithm portability across attack domains can be difficult. We address this challenge by employing a "leave-one-type-out testing" approach. Protocol 2 is further divided into two sub-protocols: Sub-protocol 2.1: The test set comprises physical attack types not included in the training or development sets. Sub-protocol 2.2: The test set comprises digital attack types not included in the training or development sets.

Protocol	Class	Live	Phys	Adv	Digital	Total
P1	train	3000	1800	1800	1800	8400
	eval	1500	900	1800	1800	6000
	test	4500	2700	7106	7200	21506
P2.1	eval	1500	0	1706	1800	5006
	test	4500	5400	0	0	9900
P2.2	train	3000	2700	0	0	5700
	eval	1500	2700	0	0	4200
	test	4500	0	10706	10800	26006

Table 1. Number of images in training, evaluation, and testing sets across various categories under three protocols: P1, P2.1, and P2.2.

4.3. Results on the UPD Dataset, LLC dataset and CelebA- Spoof Dataset

Dataset	BPCER%	ACER%	ACC%	F1 score	AUC %
P1	14.688	6.34	89.5	89.673	94.38
P2.1	1.8824	0.941	98.5	98.9	99.897
P2.2	1.48	0.07	99.71	99.77	99.9
All dataset	12.95	6.477	88.74	91.58	95.5

Table 2. The results of the proposed network for different datasets

Table 2 summarizes the performance of our proposed model on various datasets from the UPD benchmark. These datasets correspond to different evaluation protocols: Protocol 1 (P1) for unified attack detection, Protocol 2.1 (P2.1) for generalization to unseen physical attacks, Protocol 2.2 (P2.2) for generalization to unseen digital attacks, and a combined dataset encompassing all protocols ("All dataset"). The table reports key performance metrics: Bonafide Presentation Classification Error Rate (BPCER), Average Classification Error Rate (ACER), Accuracy (ACC), F1 score, and Area Under the Curve (AUC).

The model exhibits consistent performance across all datasets, characterized by low error rates (BPCER, ACER) and high accuracy (ACC). Notably, on P2.2, the model achieves a low BPCER of 1.48 % and a low ACER of 0.07 %, demonstrating its ability to effectively discriminate between genuine and spoofed presentations. Furthermore, high F1 scores and AUC values across all datasets suggest the model's robustness and reliability in classification tasks.

The model's performance on the combined dataset ("All dataset") reinforces its generalization capability, indicating its suitability for real-world scenarios with diverse data sources. Overall, the comprehensive evaluation presented in Table 1 highlights the effectiveness and versatility of the proposed model for face anti-spoofing tasks.

Prot.	Model	ACER(%)	ACC(%)↑	AUC(%)↑
	ResNet50 [5]	1.35	98.83	99.79
	VIT-B/16 [5]	[HTML]FFFFFF5.92	92.29	97
	Auxiliary[5]	1.13	98.68	99.82
	CDCN[5]	1.4	98.57	99.52
	FFD[5]	2.01	97.97	99.57
	UniAttackDetection[5]	**0.52**	**99.45**	**99.96**
Proposed Net	Proposed Net	6.34	89.5	94.38
	ResNet50[5]	34.60±5.31	53.69±6.39	87.89±6.11
	VIT-B/16[5]	33.69±9.33	52.43±25.88	83.77±2.35
	Auxiliary[5]	42.98±6.77	37.71±26.45	76.27±12.06
	CDCN[5]	34.33±0.66	53.10±12.70	77.46±17.56
	FFD[5]	44.20±1.32	40.43±14.88	80.97±2.86
	UniAttackDetection[5]	22.42± 10.57	67.35± 23.22	91.97± 4.55
Proposed Net	**Proposed Net**	**0.941±0.07**	**98.5±99.71**	**98.9±99.77**

Table 3. Comparison of the proposed model with different models and datasets.

Table 3 compares the performance of various models on different datasets using metrics such as ACER (Attack Presentation Classification Error Rate), ACC (Accuracy), and AUC (Area Under the Curve). The models include ResNet50, ViT-B/16, Auxiliary, CDCN, FFD, and UniAttackDetection, with our proposed model labelled "Proposed Net."

Protocol 1 Results: The proposed model achieves an ACER of 6.34%, ACC of 89.5%, and AUC of 94.38%. These results indicate competitive performance compared to other models under Protocol 1. However, for **Protocol 2**, the proposed model demonstrates a significant improvement, achieving an ACER of 0.941%, ACC of 98.5%, and AUC of 98.9%. This suggests the proposed method is particularly effective for Protocol 2, outperforming other models.

Table 4 presents the performance metrics for various

Model	BPCER	ACER	Accuracy	F1 score	AUC
P1	26.3	13.7	81.2	85.4	91.8
P2.1	1.9	0.8	97.6	97.9	99.2
P2.2	1.5	0.1	99.2	99.3	99.8
All dataset	10.6	5.3	87.9	90.1	94.3

Table 4. LLC Dataset Performance metrics for different models

models evaluated on the LLC dataset[21]. Here, BPCER refers to the Bonafide Presentation Classification Error Rate and ACER refers to the Attack Presentation Classification Error Rate. The table 4 shows that models P1, P2.1, and P2.2 achieve progressively lower error rates. Model P1 exhibits a BPCER of 26.3% and an ACER of 13.7% Model P2.1 significantly improves upon these results, achieving a BPCER of 1.9% and an ACER of 0.8%. Further improvement is observed with model P2.2, which reaches a BPCER of 1.5% and an ACER of 0.1 %. Finally, training on the entire dataset leads to the best performance across all metrics, with a BPCER of 10.6%, an ACER of 5.3%, and high accuracy, F1 score, and ROC AUC score.

Model	BPCER%	ACER%	ACC%	F1 score%	AUC%
P1	14.2	7.9	82.4	86.1	92.3
P2.1	2.3	1.1	96.9	97.5	99.0
P2.2	1.8	0.2	99.0	99.2	99.7
All dataset	11.5	5.8	88.2	90.5	94.8

Table 5. Performance metrics for different models on the CelebA-Spoof

Table 5 presents the performance metrics for various models evaluated on the CelebA-Spoof dataset[32]. Model P1 exhibits a BPCER of 14.2%, an ACER of 7.9%, and an accuracy of 82.4%. Models P2.1 and P2.2 significantly improve with progressively lower error rates and higher accuracy. Specifically, P2.1 achieves a BPCER of 2.3%, an ACER of 1.1%, and an accuracy of 96.9%, while P2.2 reaches a BPCER of 1.8%, an ACER of 0.2%, and an accuracy of 99.0%. Finally, training on the entire dataset leads to further improvement across all metrics, achieving a BPCER of 11.5%, an ACER of 5.8%, and an accuracy of 88.2%.

4.4. Ablation Study

Method	BPCER	ACER	Acc	AUC
MDRS wo Dual Attention	10.2	5.6	87.3	93.5
MDRS wo Spatial Side layer	8.5	4.3	89.1	94.2
MDRS wo Dual Attention + SS layer	12.3	6.8	85.7	92.1
Proposed Net	1.68	0.505	99.1	99.89

Table 6. The ablation study examines various components using evaluation protocol P1.

Table X presents an ablation study evaluating the impact of various components within the proposed model framework using evaluation protocol P1. We investigate the performance of the model under the following configurations:

MDRS without the Dual Attention module MDRS without the Spatial Side layer MDRS without both Dual Attention and Spatial Side layer Proposed Net (full model) Results:

MDRS without the Dual Attention module achieves a BPCER of 10.2%, ACER of 5.6%, accuracy of 87.3%, and AUC of 93.5%. Removing the Spatial Side layer from MDRS leads to a slight improvement, with a BPCER of 8.5%, ACER of 4.3%, accuracy of 89.1%, and AUC of 94.2%. Interestingly, omitting both the Dual Attention and Spatial Side layer results in performance degradation, evidenced by a BPCER of 12.3%, ACER of 6.8%, accuracy of 85.7%, and AUC of 92.1%.

The Proposed Net, which incorporates both the Dual Attention mechanism and the Spatial Side layer, outperforms all other configurations. It achieves a BPCER of 1.68%, ACER of 0.505%, accuracy of 99.1% and AUC of 99.89%. These results highlight the effectiveness of these components in improving the model's performance.

5. Conclusion

This work investigated the effectiveness of a novel deep learning architecture for face anti-spoofing tasks. We evaluated the model on two benchmark datasets (LLC and CelebA-Spoof) under different evaluation protocols. The proposed model consistently performed better than baseline models, demonstrating significant reductions in BPCER and ACER while achieving high accuracy. In Protocol 2, the proposed model achieved a BPCER as low as 0.941 % and an ACER of 0.505%. An ablation study further confirmed the importance of two key components within the model: the Dual Attention mechanism and the Spatial Side layer. The Dual Attention mechanism effectively highlights important features, while the Spatial Side layer captures crucial spatial information. Together, these components contribute significantly to the model's ability to distinguish between genuine and spoofed presentations.We plan to explore the generalizability of the model on diverse datasets and real-world scenarios, while also investigating techniques for interpretability and lightweight design for resource-constrained deployment.

References

[1] C. Michelassi B. Peixoto and A. Rocha. Face liveness detection under bad illumination conditions. page 3557–3560, 2011. 2

[2] G. Chiachia D. Menotti and A. Pinto. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security*, 10(4): 864–879, 2015. 2

[3] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Chenxu Zhao, Xu Zhang, Stan Z Li, and Zhen Lei. Surveillance face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2023. 3

[4] Hao Fang, Ajian Liu, Haocheng Yuan, Junze Zheng, Dingheng Zeng, Yanhong Liu, Jiankang Deng, Sergio Escalera, Xiaoming Liu, Jun Wan, and Zhen Lei. Unified physical-digital face attack detection, 2024. 3

[5] Hao Fang, Ajian Liu, Haocheng Yuan, Junze Zheng, Dingheng Zeng, Yanhong Liu, Jiankang Deng, Sergio Escalera, Xiaoming Liu, Jun Wan, et al. Unified physical-digital face attack detection. *arXiv preprint arXiv:2401.17699*, 2024. 5, 6

[6] H. Vinutha G. Thippeswamy and R. Dhanapal. A new ensemble of texture descriptors based on local appearance-based methods for face anti-spoofing system. *J. Crit. Rev.*, 7(11):644–649, 2020. 1

[7] J. Galbally and S. Marcel. Face anti-spoofing based on general image quality assessment. page 1173–1178, 2014. 2

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 2018. 3

[9] C. Zhao D. Cao Z. Lei J. Guo, X. Zhu and S. Z. Li. Learning meta face recognition in unseen domains. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, page 6162–6171, 2020. 1

[10] A. Hadid J. Määttä and M. Pietikäinen. Face spoofifing detection from single images using micro-texture analysis. page 1–7, 2011. 2

[11] Z. Lei J. W. Yang and S. Z. Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014. 3

[12] F. Peng L. B. Zhang and L. Qin. Face spoofing detection based on color texture markov feature and support vector machine recursive feature elimination. *Journal of Visual Communication and Image Representation*, (51):56–69, 2018. 1

[13] Ajian Liu and Yanyan Liang. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1180–1186. 3

[14] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1179–1187, 2021. 3

[15] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17:2497–2507, 2022. 3

[16] Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang Zhen Lei, Du Zhang, Stan Z Li, and Guodong Guo. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2023. 3

[17] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[18] Elisabeth Rumetshofer et al. Human-level protein localization with convolutional neural networks. 2018. 3

[19] D. Sabarinathan et al. Hyper vision net: kidney tumor segmentation using coordinate convolutional layer and attention unit. 7, 2020.

[20] A. Sasithradevi et al. Kolamnetv2: efficient attention-based deep learning network for tamil heritage art-kolam classification. *Heritage Science*, 12(60), 2024. 3

[21] Denis Timoshenko, Konstantin Simonchik, Vitaly Shutov, Polina Zhelezneva, and Valery Grishkin. Large crowdcollected facial anti-spoofing dataset. *2019 Computer Science and Information Technologies (CSIT)*, pages 123–126, 2019. 6, 7

[22] X. Tu and Y. Fang. Ultra-deep neural network for face anti-spoofing. page 686–695, 2017. 2

[23] Keyao Wang, Guosheng Zhang, Haixiao Yue, Ajian Liu, Gang Zhang, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Multi-domain incremental learning for face presentation attack detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5499–5507, 2024. 3

[24] W. Luo X. Yang and L. Bao. Face anti-spoofing: Model matters, so does data. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, page 3507–3516, 2019. 2

[25] A. Jourabloo Y. Atoum, Y. Liu and X. Liu. Face anti-spoofing using patch and depth-based cnns. page 319–328, 2017. 2

[26] A. Jourabloo Y. Liu and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. 2018. 2

[27] A. Jourabloo Y. Liu and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, page 389–398, 2018. 2

[28] J. Stehouwer Y. Liu and A. Jourabloo. Deep tree learning for zero-shot face anti-spoofing. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, page 4680–4689, 2019. 2

[29] J. Komulainen Z. Boulkenafet and A. Hadid. Face spoofing detection using colour texture analysis. *IEEE Trans on Information Forensics and Security*, 11(8):1818–1830, 2016. 2

[30] Syed Waqas Zamir et al. Cycleisp: Real image restoration via improved data synthesis. 2020. 3

[31] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 919–928, 2019. 3

[32] Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 70–85. Springer, 2020. 5, 7