I.J. Image, Graphics and Signal Processing, **2025, 4, 87-104**
Published Online on August 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2025.04.06

Spoof-formerNet: The Face Anti Spoofing Identifier with a Two Stage High Resolution Vision Transformer (HR-ViT) Network

Mudunuru Suneel*
Department of Electronics and Communication Engineering, University College of Engineering, Acharya Nagarjuna University (ANU), Guntur, AP, India
E-mail: suneel007@gmail.com
ORCID iD: https://orcid.org/0009-0002-1200-9830
*Corresponding Author

Tummala Ranga Babu
Department of Electronics and Communication Engineering, RVR & JC College of Engineering, Guntur, AP, India
Email: trbaburvr@gmail.com
ORCID iD: https://orcid.org/0000-0002-4946-1452

Received: 15 March, 2024; Revised: 12 June, 2024; Accepted: 20 May, 2025; Published: 08 August, 2025

Abstract: Face anti-spoofing (FAS) detection is essential for assuring the safety and dependability of facial identification systems. This study introduces the implementation of a new approach called Spoof-formerNet, which utilizes the high-resolution vision transformer (HR-ViT) system for detecting face anti-spoofing. The Vision Transformer (ViT) architecture has revealed remarkable execution in numerous computer vision applications, and we are now applying it to the intricate field of spoof detection. In order to distinguish between real faces and spoofing attempts, the Spoof-formerNet is engineered to detect minute details and subtle elements embedded in facial photos. We have conducted experimental research wherein the model is trained independently on color (RGB) and depth data in parallel using two streams of HR-ViT networks. Before applying to a classification head, the features from the two streams were concatenated. Spoof-formerNet is trained and tested using well-known benchmark datasets such as CelebA-Spoof, CASIA-SURF, WMCA, and MSU-MFSD, which are commonly used in the field of anti-face spoofing. The suggested model excels in performance and is cutting-edge in identifying genuine faces from spoofing assaults. We assess the model's efficacy by providing comprehensive findings, such as Area Under the Curve (AUC), Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), Equal Error Rate (EER), and Average Classification Error Rate (ACER). The results of this work show how cascaded high-resolution vision transformer networks can be used to improve the safety of facial recognition approaches in real-world applications, in addition to advancing facial anti-spoofing technology. The Spoof-formerNet method for face anti-spoofing detection shows good results, with an average AUC of 99.22 and average APCER, BPCER, and ACER of 0.95, 0.66, and 0.81 correspondingly.

Index Terms: Face Anti Spoofing (FAS), High Resolution Vision Transformer (HR-ViT), Token Embedding, Multi Head Self-Attention

1. Introduction

Preventing spoofed faces is essential in biometrics, which aims to distinguish genuine human faces from fake ones [1]. As facial recognition systems become more prevalent in areas like security, finance, and smartphones, it's crucial to have reliable methods to detect and thwart face spoof attacks [2, 3]. These attacks attempt to trick systems into accepting a fake face by displaying manipulated images or videos [4]. Researchers are actively developing anti-spoof algorithms to accurately identify fake faces, ensuring the integrity and reliability of facial recognition systems [5]. Facial recognition algorithms analyze details like skin texture, movement, and depth to spot fake faces [6]. They also use liveness detection to check for physical cues or actions that indicate real people [7]. This helps prevent "face

spoofing" attacks, where imposters trick recognition systems. As technology improves, anti-spoofing measures are becoming more important to ensure secure and accurate facial recognition applications in a digital society [8].

These developments aid in distinguishing between authentic and counterfeit faces, as well as in deterring illegal access and fraudulent acts [9]. Liveness detection is becoming an essential element in facial recognition systems, providing an additional level of security by confirming the existence of a living individual [10, 11]. Liveness detection algorithms may differentiate between a genuine and a fake face by examining indicators like eye movement, blinking patterns, and subtle changes in skin texture [12]. This advancement has greatly decreased the likelihood of face spoofing assaults.

Nevertheless, despite these developments, there are still plenty of challenges in the realm of anti-spoofing. A major challenge is the advancement of more complex spoofing methods capable of tricking liveness detection systems [13, 14]. Attackers can utilize high-quality masks or deep fake technology to produce authentic-looking fake faces that can convincingly appear real. Another problem is the requirement for immediate and precise detection, as face anti-spoofing systems must promptly recognize and react to spoofing efforts in real-time [15]. Liveness identification algorithms struggle to apply their findings to multiple settings due to the wide range of presentation attacks, including variations in lighting conditions and positions [16]. Continuous research and developments are essential to address these problems and enhance the efficiency of facial anti-spoofing systems.

One effective method to tackle these obstacles is by employing Vision Transformers for face detection [17]. Vision Transformers, or ViTs, are a deep learning model that has proved remarkable execution in various computer vision challenges [18]. High Resolution Networks integrated into Vision Transformers have the competence to greatly improve the accuracy of face anti-spoofing systems by effectively managing the complexities and nuances of facial features through continued research and development. High-resolution networks are specifically created to process high-resolution photos and are capable of capturing intricate details from the input image [19]. The networks usually consist of several convolutional layers that gradually decrease the spatial resolution of the input image and increase the number of feature channels [20]. This allows the network to extract both local and global characteristics of input image at various sizes [21]. High-resolution networks have proven to be successful in several computer vision challenges such as object detection, semantic segmentation and image classification [22]. Visual Transformers (ViTs) can capture distant connections and analyze spatial patterns in images through attention mechanisms. This enables them to adjust to various spoofing attempts and enhance their effectiveness in practical scenarios. Vision Transformers with High Resolution Networks have the ability to revolutionize face anti-spoofing and enhance security measures through ongoing refinement and invention [23].

ViTs demonstrate significant promise in identifying face spoofing, offering advantages like high performance in challenging cross-domain settings, the capacity to leverage multi-modal data for comprehensive insights, and suitability for real-time video-based detection.

The remainder of the paper is organized as follows: Section 2 provides a concise literature overview of the proposed work. Section 3 provides a comprehensive explanation of the proposed method for detecting face spoofing and includes a description of the architecture. The Spoof-formerNet underwent rigorous testing on multiple publicly accessible datasets, with the findings detailed in section 4. Section 5 summarizes the results and the future directions of the research on face anti-spoofing recognition.

2. Literature Survey

The historical advancement of face anti-spoofing practices has seen significant progress over the years. Several approaches have been proposed, including motion, texture and deep learning based methods. Texture-based approaches analyze the sensitive differences in skin texture between genuine and fake faces, while motion-based methods focus on detecting the absence of natural facial movements. Deep learning-based methods, in contrast, leverage the power of convolutional neural networks (CNNs) to learn discriminatory features from facial images. Grayscale and Color texture descriptors-based face spoofing recognition by Boulkenafet et al. [24]. A novel approach using highlight removal effect texture analysis proposed by Inhan Kim et al. [25] for face spoofing recognition.

M. Beham et al. [26] proposed an innovative anti-spoofing system built on aggregated local weighted gradient orientation (ALWGO). Boulkenafet et al. [27] recommended a modern face anti-spoofing technique established on color and texture analysis. Rui Shao et al. [28] suggested a joint discriminatory learning of dynamical textures for 3D mask face anti-spoofing. Techniques focused on simple facial feature extraction and comparison, such as affordable cost software-based approach for identifying spoofing attempts in face recognition system is introduced by A. Pinto et al. [29]. A novel face anti-spoofing method works with live correlation loss on one-class learning, proposed by Seokjae Lim et al. [30] on various benchmark datasets.

Wen et al. [31] claims that texture features comprise, personal identity information, which is surplus for anti-spoofing and could lead to inadequate generalized implementation. Furthermore, conventional handcrafted features such as Local Binary Patterns (LBP) [32], Scale-Invariant Feature Transform (SIFT) [33], Speeded-Up Robust Features (SURF) [34], and Histogram of Oriented Gradients (HOG) [35] have been devised to extract discerning patterns indicative of spoofing from diverse color spaces. These methods, however, necessitated the utilization of

domain-specific prior knowledge. As attackers became more sophisticated in their methods, the limitations of these traditional techniques became apparent.

Deep learning techniques have demonstrated substantial advancements in a wide range of applications, owing to their utilization of CNNs. CNN techniques have demonstrated their ability to effectively capture intricate patterns and interdependencies among features, hence proving to be very efficient for FAS. Two-Stream CNN (TSCNN), proposed by Haonan Chen et al. [36] with attention-based fusion, achieves great effectiveness in spoofing detection.

Aziz Alotaibi et al. [37] proposed an effective and non-intrusive approach for detecting face spoofing attacks using a specialized deep CNN, achieving an outstanding accuracy rate. Wenyun Sun et al. [38], proposal for face spoofing detection using local ternary label supervision exemplifies the benefit of pixel-level local label supervision schemes. A hierarchical visual feature based on deep learning using sparse auto-encoder networks and SVM classifier was proposed by Dakun Yang et al. [39] for face spoofing detection, showing the robustness of the approach. CNN architectures such as Inception and ResNet have shown outstanding performance for face anti-spoofing, introduced by Chaitanya Nagpal et al. [40], obtaining favorable results in different settings through experiments that considered aspects such as model depth, weight initialization, fine tuning, and learning rate.

Haocheng Feng et al. [41] proposed face anti-spoofing method outperforms cutting-edge techniques by using a residual-learning structure to learn discriminatory live-spoof differences, characterized as spoof cues, and an auxiliary classifier to amplify the cues. Despite the limitations faced by traditional methods in accurately distinguishing genuine and spoofed faces, advancements in deep learning algorithms have demonstrated considerable potential, but they still suffer from limitations such as the necessity for large amounts of labeled data for training. Furthermore, these methods may struggle with detecting more sophisticated and realistic fake faces that are becoming increasingly common. Despite these limitations, researchers are continuously working on improving these methods and finding new ways to overcome these challenges to enhance the accuracy and sturdiness of facial forgery detection systems. Moreover, advancements in hardware technology and the availability of more powerful computational resources are expected to alleviate the computational burden associated with deep learning-based methods.

Vision Transformers (ViTs) are the primary architecture used for image identification. They utilize self-attention processes to capture long-range relationships in images. ViTs can successfully differentiate between authentic and counterfeit faces by utilizing their capacity to record nuanced details and contextual information, hence enhancing the security and dependability of facial recognition techniques. This has resulted in new opportunities to enhance the precision and resilience of face anti-spoofing systems. This section aims to provide a thorough overview of ViTs and their use in image classification problems. It will focus on the modifications made to this technique expressly for facial anti-spoofing applications.

Yu, Zitong, et al. [42] introduced a vision transformer for multimodal FAS. The proposed method includes an adaptive multimodal adapter (AMA) and a masked auto encoder (MAE), to address the limitations of ViT. Anjith George et al. [43] proposed approach for zero-shot face anti-spoofing outperforms state-of-the-art methods in the HQ-WMCA and SiW-M datasets and achieves a significant boost in cross-database performance. Liu, Ajian, et al. [44] proposed a flexible modal framework, FM-ViT, for face anti-spoofing. It captures different modal information and learns modality-agnostic liveness features by summarizing the multi-modal information. The experimental results outperform the other methods in terms of accuracy and robustness. Although these models teach high architecture design, most require a fair amount of pretraining, or they depend strongly on domain-specific fine-tuning. Spoof-formerNet, on the other hand, with its double HR-ViT structure and hybrid form of attention, is able to offer the same or better performance across multiple datasets without the need for intensive domain adaptation, thus proving its practicality and robustness for field deployment.

Liao, Chen-Hao et al. [45] proposed approach uses transformer units as the network backbone and employs a domain-invariant attention loss and a domain adversarial loss to improve the model's performance. This is a new methodology to FAS that improves the generalizability of the model to unknown domains. The proposed temporal transformer network introduced by Zhuo Wang et al. [46], learns multi-granularity temporal characteristics for face anti-spoofing. Lee et al. [47] proposed a robust face anti-spoofing framework that uses a convolutional vision transformer on advancing FAS execution with global and local information and retaining the robustness of FAS with domain-generalized features.

High-resolution networks have gained significant attention in recent years due to their capability to capture fine-grained information and improve the performance of image categorization tasks. This led to the need for innovative approaches like Vision Transformers with High Resolution Networks, which offer enhanced capabilities for detecting and mitigating spoofing attacks. By examining the historical evolution of face anti-spoofing techniques and their inherent limitations, it becomes clear that a new paradigm is necessary to effectively combat.

3. Methodology

3.1. Method Overview:

Vision transformers are based on concepts that favour the exploration of long-range global feature characteristics. This is in contrast to typical convolution operations, which are based on the semantics of short-range local features. The semantic entities of interest can be comprehended in a more thorough manner through the utilization of this approach.

As a consequence of this, vision transformer models have exhibited outstanding performance over a variety of vision predictive challenges.

In this article, we proposed the utilization of a hybrid-window attention based novel vision transformer architecture that was specifically built for face spoofing detection. This was done in recognition of the distinct benefits that vision transformer structures (ViT) and high-resolution network (HRNet) designs offer and introduced HR-ViT for spoofing attack detection.

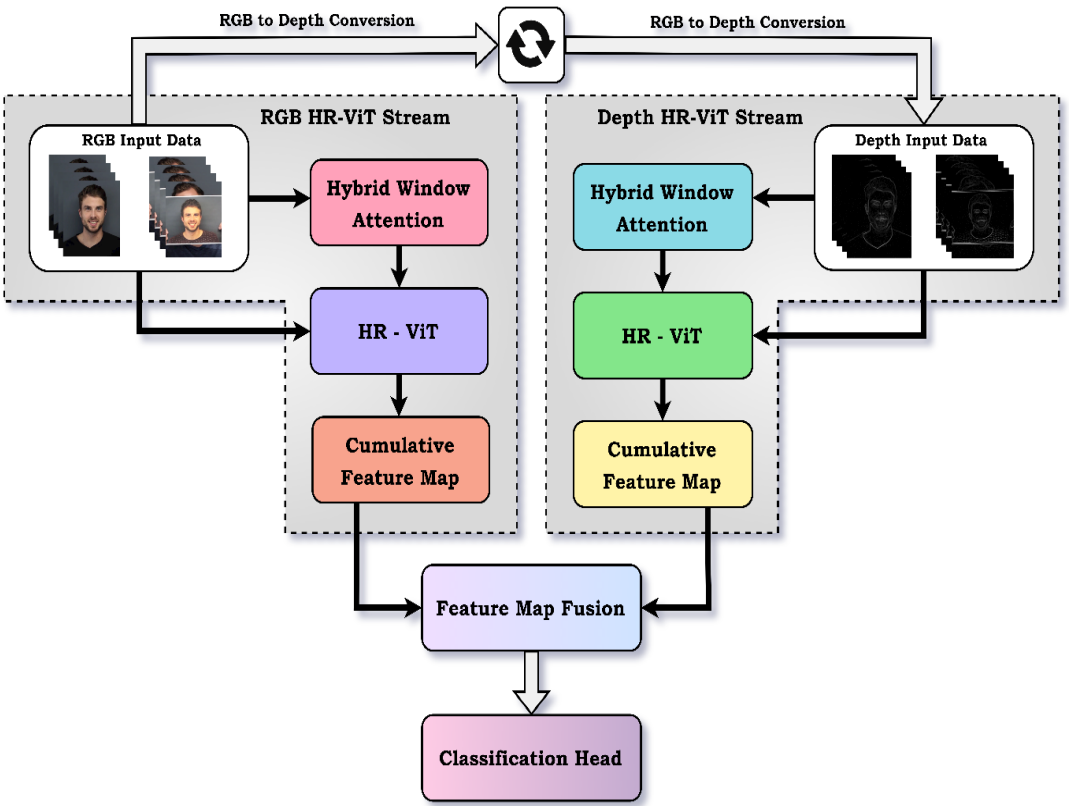

Fig. 1. Block diagram showing the overview of the Spoof-formerNet.

As can be seen in Figure 1, the Spoof-formerNet is constructed to adhere to an HRNet architecture. The architecture was designed two-fold. The first one is an HR-ViT Net which processes RGB images and another HR-ViT Net is to process the corresponding depth information. Each HR-ViT stream is comprised of many parallel branches that are capable of simultaneously extracting high-level feature encodings at a variety of resolutions. Every branch is comprised of multiple processing stages that are stacked with a collection of transformer blocks that are based on hybrid windows. Specifically, these transformer blocks can properly capture both local short-range feature attributes and global long-range feature territories. This is a significant advantage. Furthermore, cross-branch feature connections are being carried out on a regular basis for the purpose of feature semantic enhancement. This is accomplished by combining the semantic proofs from various scales. Finally, to derive the output, the variational scale features that are formed from all of the branches are taken into consideration in their entirety. The output spoofed face prediction map is just a fifth the size of the input image. This is because the feature map down-sampling that occurs at the starting stages of the transformer causes the map to be smaller. The architecture proposed to have two streams to process the RGB spoofed face images with RGB HR-ViT stream and the depth information with Depth HR-ViT stream respectively. The following section explains the proposed architecture in detail.

3.2. Spoof-formerNet Architecture:

The Spoof-formerNet is designed two folded to process both RGB and depth information as a separates streams. Each stream of spoof-formerNet consists of four parallel branches dedicated to feature extraction. The feature maps in these branches have spatial resolutions that are progressively decreased from top to bottom in order to use feature representations in a lower dimensional feature space. This is advantageous for representing entity attributes across various levels of detail.

The size of the feature routes in these subbranches are continuously increased in order to capture more detailed feature semantics, hence enhancing the quality of the feature representation. The default setting for the feature resolution reduction factor is 0.5, which means that the feature map in the lower branch will have a size that is one-fourth of the original. Similarly, the default setting for the feature channel increment factor is two, which means that the feature map in the lower branch will have twice the number of channels compared to the original. Each branch is

equipped with a series of transformer-based stages that perform global and local feature extractions. As shown in figure 2, each stream of spoof-formerNet is divided into five levels, which ultimately produce high-quality features with high-level representation in all subspace features within the individual branch. The feature meanings are combined and received by the classification head for predicting the spoofed content.

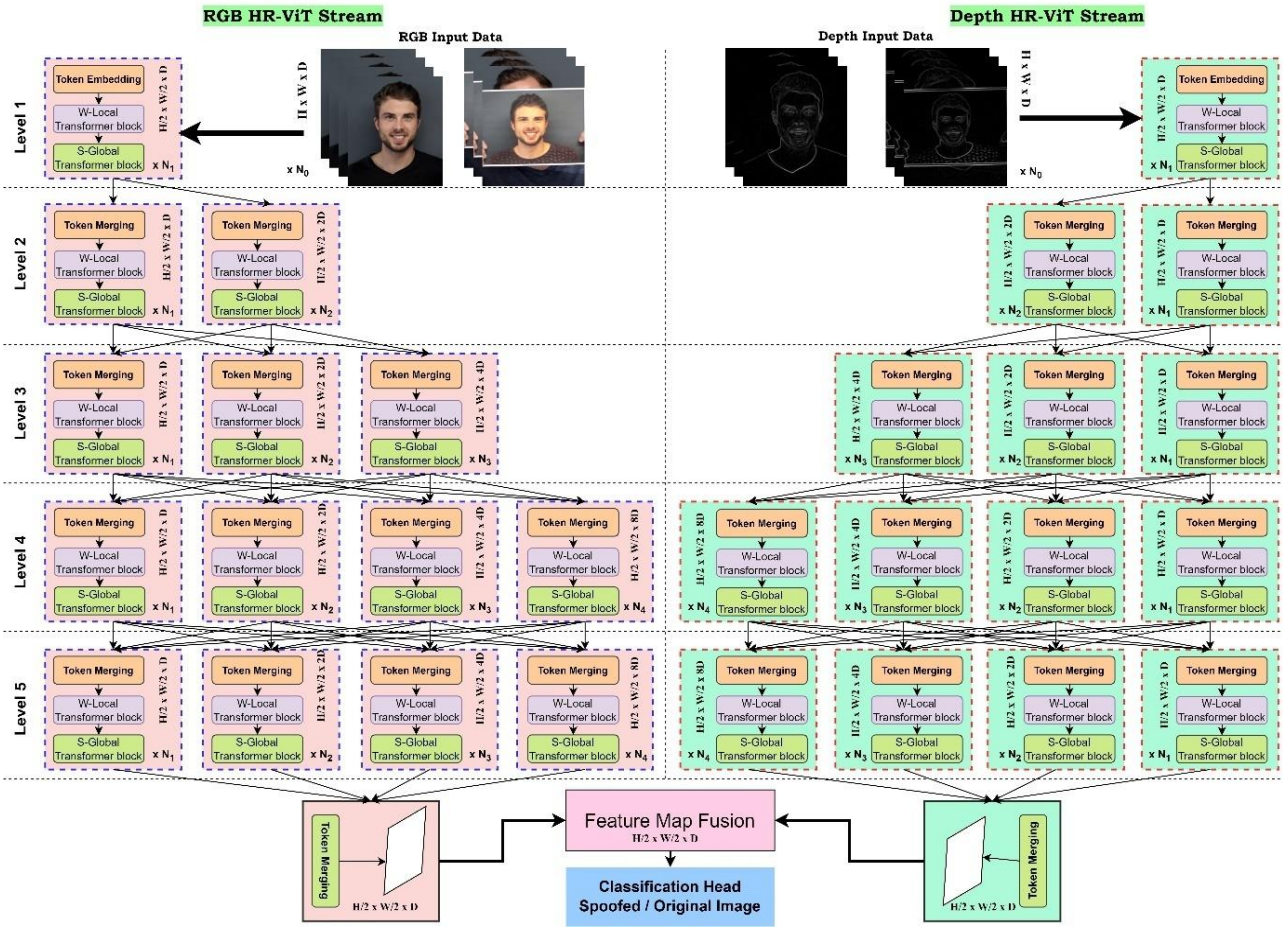

Fig. 2. The proposed Spoof-FormerNet Architecture with two streams.

It is important to observe that the uppermost section starts with a convolution stem, comprising N_0 convolutional layers that are utilized to exploit low-level picture features. The convolution branch preserves the spatial dimensions of the inputted image by repeatedly operating convolutions with 3×3 size filters having a stride of 1 and a padding of 1. Indeed, the pre-connection of convolution functions demonstrates superior efficiency compared to the transformer unit blocks, while maintaining the derived low-level feature representations with high quality. Next, the low-level features are transformed into token interpretations in order to generate high-level feature encodings at several scales, progressing via transformer blocks in various stages. A token embedding module is used to generate the tokens in the top branch's initial transformer step. In contrast to the equal size, non-overlapping, identical patch arrangement strategy, which is not having the ability to access distinct contextual characteristics of entities and share identical details between neighbouring patches. We utilize a multi-scale token embedding approach with the goal of enhancing the encoding efficacy of the generated token depictions. By using tokens of better quality, it becomes possible to extract more robust and distinguishing feature semantics. This ultimately leads to the development of a more successful segmentation network. According to Figure 3, a feature map is divided into four groups of patches with varying sizes. These patches are distributed among the feature map using predetermined sliding periods. To provide consistent token embedding dimensions across different partition sizes for the future concatenation procedure, the sliding periods are set to the same value. This ensures that the different sets have an equal number of patches.

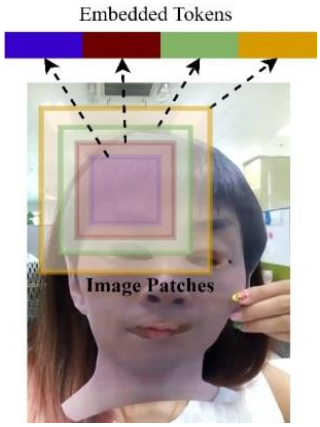

Fig. 3. Token Embedding Approach of Multiscale Patches.

To construct sub-token embeddings of identical size, each set's patches undergo a linear embedding procedure using global convolutions. In other words, the sizes of the sub-tokens remain consistent across various partition sizes. Next, the concatenation of sub-token embeddings received from various sets is done in a sequential fashion, where smaller patches are concatenated before larger ones, resulting in a functional token embedding. By employing a multi-scale patch split and token embedding technique, the representation of semantics can effectively capture many perspectives, leading to enhanced and comprehensive understanding compared to a static local viewpoint. After token embedding, the extent of the feature maps in the upper branch decreased by 50% along all dimensions, relative to the feature maps in the convolution branch.

Fig 2 demonstrates that each branch of the transformer stages (Stage II to Stage V) and the classification head incorporates a token merger module. This module is designed to combine the multiscale characteristic semantics from every single branch, thereby improving the quality of feature encoding within a particular feature sub-space. Specifically, during the process of cross-branch token joining in a particular branch, where the feature map has dimensions of $H_D \times W_D \times C_D$, the feature map from the top layer with dimensions $H_{up} \times W_{up} \times C_{up}$ is first down-sampled and the channel number is augmented using convolution processes. This results in a lower-resolution feature map with dimensions $H_D \times W_D \times C_D$. Similarly, the feature map from a lower layer with dimensions $H_{low} \times W_{low} \times C_{low}$ is first up-sampled and The channel number is decreased via deconvolution processes, lead to a feature map with improved resolution possessing dimensions $H_D \times W_D \times C_D$. Subsequently, the feature descriptions that have been recalibrated from the remaining branches are incorporated with the feature description that has been replicated from the original branch that is being targeted. This is done using a 1×1 convolutional layer to merge the feature representations. The result is a merged token illustration in the target branch, which takes into account the multi-scale entity representations with a global range. The size of the merged token depiction is $H_D \times W_D \times C_D$. A cross-branch token merging technique effectively addresses the shortcomings resulting from the particular-scale feature semantic representation method.

At each stage of the Spoof-former, the tokens that are inserted in every branch undergo processing by a group of transformer units to extract features at a more advanced level. More precisely, in each stage from Branch I to Branch IV, there are N_1, N_2, N_3 and N_4 transformer modules, correspondingly. The two components that make up a transformer module are referred to as a window-based local transformer block (Window-Local) and a sparse-window-based global transformer block (SWindow-Global). Using a denser-window local self-attention mechanism, the Window-Local transformer block is intended to capture local component interpretations. This is accomplished by the utilization of the block. The SWindow-Global transformer block, on the other hand, is primarily concerned with the extraction of global entity interpretations by means of a sparse-window global self-attention mechanism.

The utilization of a hybrid-window transformer design plan, which enables the simultaneous consideration of both local characteristics and global exteriors of the entities of significance, is a significant factor that contributes to the significant enhancement of the token encoding capabilities. It is important to note that the top branch now contains all of the multiscale tokens from all of the branches. In the end, the cross-branch entrenched tokens in the classification head are subjected to additional processing through a convolutional layer in order to generate the output base map, which is then utilized for the extraction of spoofed image edge elements.

3.3. Transformer module:

It is necessary to partition the tokens evenly into non-overlapping windows of a size $W_L \times W_L$ to be able to carry out the implementation of the local feature self-attention capability. These windows are utilized for the purpose of analyzing the close qualities of objects in order to gain an understanding of their short-range characteristics. Following

that, the Window-Local transformer block is employed in order to calculate the local characteristics of self-attention in each of the windows. Fig 4 depicts the complex architecture of the Window-Local transformer block, which may be seen in the figure.

Specifically, the Window-Local transformer block is made up of a mix of a feed-forward network (FFN) and a local window weighted multi-head self-attention unit (LW-WMSA). In addition, the transformer architecture incorporates LayerNorm normalization algorithms and residual connection schemes to enhance its performance. The sequence of token processing in the Window-Local transformer unit is as follows:

$$LN_{in} = LayerNorm\left(Token_{in}\right) \tag{1}$$

$$Token_{N} = LocalWindow-WMSA\left(LN_{in}\right) + Token_{in} \tag{2}$$

$$LN_{N} = LayerNorm\left(Token_{N}\right) \tag{3}$$

$$Token_{out} = FFN\left(LN_{N}\right) + Token_{N} \tag{4}$$

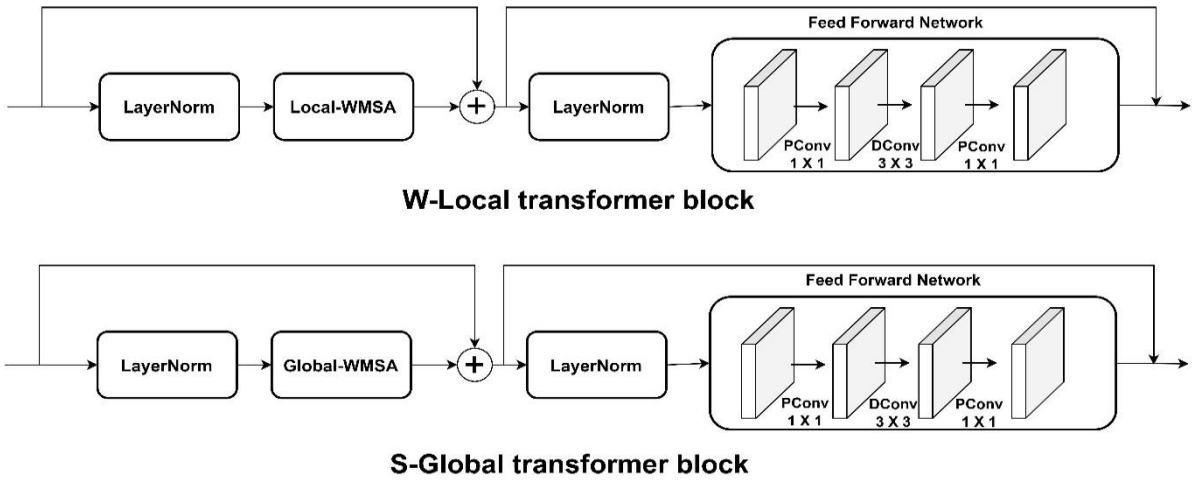

Fig. 4. Building blocks of Window-Local and S-Global transformer blocks.

The input and output tokens of the Window-Local transformer unit are denoted by $Token_{in}$ and $Token_{out}$ respectively. $Token_{M}$ is a reference to the LW-WMSA module's intermediate representation. There are two point-wise convolutional (PConv) layers and a depth-wise convolutional (DConv) layer that are sandwiched between the two layers in the FFN, as shown in Figure 4. The DConv layer employs a kernel size of 3×3 and performs convolution operations in a sequential fashion, moving from one layer to the next.

In order to accomplish global feature self-attention while minimizing model parameters and processing resources, the tokens are selected among the entire feature space. The collected tokens are grouped to form a collection of sparse windows. The employment of long-range global feature representations of entities is made possible by the fact that each window has access to a global receptive field. In the following step, the S-Global transformer block is utilized in order to compute the global feature self-attention within each sparse window structure. In Figure 4, the S-Global transformer block is presented in its entirety in its formulation.

There are two components that make up the S-Global transformer block. These are a feed-forward network (FFN) and a global window weighted multi-head self-attention (GW-WMSA) module. In a manner that is analogous to the Window-Local transformer block, the architectural enhancement of the transformer also includes the employment of LayerNorm normalization and residual connection activities. In the S-Global transformer unit, the token handling stream can be described as follows:

$$LN_{in} = LayerNorm\left(Token_{in}\right) \tag{5}$$

$$Token_{N} = GlobalWindow-WMSA\left(LN_{in}\right) + Token_{in} \tag{6}$$

$$LN_N = LayerNorm\left(Token_N\right) \tag{7}$$

$$Token_{out} = FFN\left(LN_N\right) + Token_N \tag{8}$$

$Token_{in}$ and $Token_{out}$ represent the tokens that are inputted and outputted by the S-Global transformer block, respectively. $Token_M$ represents the middle tokens generated by the GW-WMSA module.

The LW-WMSA and GW-WMSA modules are the innovative design ideas of the Window-Local and S-Global transformer blocks. These modules execute multi-head self-attention by considering the disparities in impact across different heads. Existing transformer systems commonly employ a method of multi-head self-attention where the output token depictions from all heads are concatenated in a straightforward and uniform manner, without considering the quality and relevance of their feature semantic encoding. Consequently, the capability of the transformer module may be compromised as a result of the amplification of feature representations that are less relevant or the deterioration of feature semantics that are highly relevant. To improve the performance of the multi-head self-attention mechanism in exploiting task-oriented feature semantics, we propose the implementation of a weighted multi-head self-attention (WMSA) module that assigns a separate priority factor to each head. This would allow for the mechanism to perform more effectively. When it comes to extracting feature dependencies between tokens, the LW-WMSA and GW-WMSA modules have the same structure as the WMSA module. In contrast to the GW-WMSA module, which makes use of global windows, the LW-WMSA module makes use of local dense windows.

In multi-head self-attention, tokens in WMSA are linearly transformed before applying the attention mechanism.The architecture of multi head attention strategy.The parameters query (Q), key (K), and value (V) of n heads will be taken in to consideration. For each heads the attention score α_{ij} with weighted parameters W_Q, W_K and W_V is given as:

$$\alpha_{ij} = \frac{e^{\left(\left(W_Q.x_i\right).\left(W_K.x_j\right)\right)}}{\sum_{k=1}^{N} e^{\left(\left(W_Q.x_i\right).\left(W_K.x_k\right)\right)}} \tag{9}$$

Where, W_Q is the weight vector related with the query for position i.

W_K is the weight vector related with the key for position j.

W_V is the weight vector related with the value for position j.

The output Y_i for the position i by considering the query, key, and value parameters is given by:

$$Y_i = \sum_{j=1}^{N} \alpha_{ij}.\left(W_V.x_j\right) \tag{10}$$

For the case of multi head attention model with N heads having the weight vectors $W_{Q1}, W_{Q2}, W_{Q3}.....W_{Qn}$, $W_{K1}, W_{K2}, W_{K3}.....W_{Kn}$ and $W_{V1}, W_{V2}, W_{V3}.....W_{Vn}$, the output is given as:

$$Y_{out} = Concatenate\left(\left[Y_{i1}, Y_{i2}, Y_{i3}.....Y_{in}\right]\right).W_o \tag{11}$$

Where, Y_{in} is the output of i^{th} position for the n^{th} head and W_o is another learnable weight matrix used for the linear transformation. The classification head is a binary classifier consisting of a global average pooling layer followed by a series of fully connected layers. The last layer is a SoftMax which provides the classification probabilities $\left(P_o\right)$ for both spoofed and genuine image classes.

$$P_o = SoftMax\left(Y_{out}\right) = \frac{e^{\left(Y_{out}\right)}}{\sum_{1}^{K} e^{\left(Y_{out}\right)}} \quad \forall \ K \text{ real valued inputs} \tag{12}$$

4. Results and Discussion

4.1. Dataset Description:

Face anti-spoofing datasets are essential assets utilized for the advancement and assessment of algorithms and systems aimed at identifying and thwarting facial spoofing attacks in biometric authentication and security systems. The datasets contain genuine and fake facial images taken in different situations, including varying lighting conditions, diverse poses, different types of fake materials (such as printed photos, masks, and replayed videos), and different types of presentation attacks (such as photo attacks and video attacks). In this paper, the proposed Spoof-formerNet is trained and tested on four well known datasets, namely, CelebA-Spoof, CASIA-SURF, WMCA and MSU-MFSD. Table 1 shows the various specifications of the datasets used in this work.

Table 1. Detailed specifications of the face-spoofing datasets used in this work.

Dataset	Year	Modality	Subjects	Samples	Type	Number of sensors	Spoof Type
CelebA-Spoof [48]	2020	RGB	10,177	6,25,537	Images	1	Print, Replay, Paper Cut, 3D mask
CASIA-SURF [49]	2018	RGB/ Depth/ IR	1,000	21,000	Videos	1	Paper Cut
WMCA [50]	2022	RGB/ Depth/ IR/ Thermal	72	1941	Videos	2	Print, Replay, Mask, Paper Cut
MSU-MFSD [51]	2015	RGB	35	440	Videos	2	Print, replay

4.2. Evaluation Metrics:

When it comes to ensuring the safety and dependability of face recognition systems, face anti-spoofing detection models are absolutely necessary. The purpose of these devices is to discriminate between genuine facial photographs and those that are being used for deceitful purposes, such as spoof attacks. It is required to utilize appropriate performance criteria that properly measure the capacity of these models to recognize and defend against spoof assaults in order to conduct an analysis of the effectiveness of these models. A number of performance metrics, including Area Under the Curve (AUC), Equal Error Rate (EER), Average Classification Error Rate (ACER), Bona Fide Presentation Classification Error Rate (BPCER), and Attack Presentation Classification Error Rate (APCER), were utilized in order to evaluate and compare the performance of our proposed model with that of state-of-the-art models. These metrics are mathematically defined as follows:

The integration of the Receiver Operating Characteristic (ROC) curve is often required in order to calculate the Area Under the Curve (AUC). ROC curves are graphical plots that depict the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) for various threshold values. These curves are used to determine the accuracy of a diagnostic test. The following formula is used to determine the area under the curve (AUC) for n points on the ROC curve represented by (x_i, y_i)

$$AUC = \sum_{i=1}^{n-1} \frac{(x_{i+1} - x_i)(y_i + y_{i+1})}{2} \tag{13}$$

Where x_i is the FPR and y_i is the TPR.

The point on the Receiver Operating Characteristic (ROC) curve at which the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) are equal is referred to as the Equal Error Rate (EER). In terms of mathematics, the EER can be characterized as the point at which the FAR and the FRR are equal to one another.

$$EER = \frac{P_{fa} + P_{fr}}{2} \tag{14}$$

Where, P_{fa} is probability of false acceptance and P_{fr} is probability of false rejection.

APCER measures the error rate when attack or spoofed face presentations are incorrectly classified as genuine.

$$APCER = \frac{Number\ of\ spoofed\ samples\ incorrectly\ classified\ as\ genuine}{Total\ number\ of\ spoofed\ samples} \tag{15}$$

$$BPCER = \frac{Number\ of\ genuine\ samples\ incorrectly\ classified\ as\ spoofed}{Total\ number\ of\ genuine\ samples} \tag{16}$$

$$ACER = \frac{APCER + BPCER}{2} \tag{17}$$

In addition to the above performance metrics, Accuracy, Precision, Recall and F-1 score parameters were also calculated to evaluate the proposed model with other deep learning model performance on various face anti spoofing datasets.

$$\Pr ecision = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{18}$$

$$recall = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{19}$$

$$F1\text{-}score = 2 \times \left(\frac{precision \times recall}{precision + recall} \right) \tag{20}$$

Where N_{FN} , N_{FP} , N_{TP} represent the quantities of false negative, false positive and true positive components, respectively. Significantly, the measures of precision, recall, and F1-score were calculated using the optimal dataset scale (ODS).

4.3. Experimental Evaluation:

The results of the testing of the proposed Spoof-formerNet on four different spoofing datasets are presented in this section. Within the scope of this part, a comprehensive debate has been carried out. Five different experiments have been designed by us in order to test the effectiveness of the proposal. The results of the experiments were presented in the form of tables, and the performance of the Spoof-formerNet was discussed by comparing it to other standard models in the sections that followed.

4.3.1. Experiment 1: Evaluating the Spoof-FormerNet on CelebA-Spoof dataset.

In this experiment, we have considered CelebA-Spoof dataset which consisting of spoofed images of various categories, namely, print, replay, paper cut and 3D mask attack samples. We have created the depth maps of the respective RGB spoofed images for experimenting with the proposed model of face anti spoofing detection.

Table 2. Performance evaluation of our proposed Spoof-FormerNet on CelebA-Spoof dataset.

Model	Modality	Performance Metrics				
		AUC (%)	EER	APCER	BPCER	ACER
FaceFakeNet [52]	RGB	98.12	1.70	0.70	0.00	0.35
AENetC,S [48]	RGB	99.88	1.10	4.62	1.09	2.85
GRU [53]	RGB	98.00	7.22	11.24	5.62	8.43
VGG 16 [54]	RGB	98.60	5.91	8.69	4.95	6.82
ResNet18 [55]	RGB	99.20	4.30	5.58	3.26	4.42
FeatherNet [56]	RGB	97.80	7.50	12.31	6.11	9.21
GNN [57]	RGB	98.40	6.21	9.45	5.41	7.43
FM-ViT [44]	RGB+Depth	98.82	1.3	1.25	0.84	1.05
MAE+AMA [42]	RGB+Depth	98.45	1.64	1.48	1.26	1.37
De-Spoof [43]	RGB	98.91	1.2	1.1	0.90	1.00
DI-ViT [45]	RGB	98.65	1.49	1.61	1.22	1.42
HR-ViT	RGB	99.01	2.46	4.86	1.69	3.28
HR-ViT	Depth	97.30	3.13	5.16	2.54	3.85
Proposed	RGB + Depth	99.25	1.01	0.43	0.12	0.28

We tested various baseline models for spoofing detection and calculated AUC, EER, APCER, BPCER and ACER to measure their capabilities. We also experimented with HR-ViT by inputting RGB and Depth images alone and observed the performance by calculating various performance metrics. It is observed that the only HR-ViT stream which is designed by extracting various global and local features has shown its performance with an error rate of 2.46. However, the proposed Spoof-formerNet with two streams supplied by two data modalities has shown superior performance with a low error rate of 1.01. Table 2 shows the performance of aimed model and other baseline models for spoofing detection on CelebA-Spoof dataset.

4.3.2. Experiment 2: Evaluating the Spoof-FormerNet on CASIA-SURF dataset.

As an experiment-2, we have considered CASIA-SURF dataset which consisting of paper cut spoofed samples to train and test with the proposed model. CASIA-SURF is a video data and the videos are converted in to frames. We have selected the good frames from all the videos and created a folder of 10000 spoofed images for the experimentation.

Table 3. Performance evaluation of our proposed Spoof-FormerNet on CASIA-SURF dataset.

Model	Modality	Performance Metrics				
		AUC (%)	EER	APCER	BPCER	ACER
FaceFakeNet [58]	RGB	98.01	1.01	1.5	0	0.75
CNN [59]	RGB	98.32	5.17	8.73	6.25	7.49
SE-Resnet18[60]	RGB	97.12	5.3	9.23	6.44	7.84
CoALBP(ycbcr)[61]	RGB	92.33	10	14.19	10.56	12.38
ResNet [62]	RGB	96.42	2.22	2.59	3.57	3.08
FaceBagNet [63]	RGB	96.05	3.34	3.19	0.97	2.08
SE-Net [62]	RGB	98.45	1.23	1.74	4.21	2.97
MFAM+DAM [64]	RGB	97.89	1.17	0.92	1.74	1.33
MM-CDCN [65]	RGB	99.24	8.23	11.83	0.8	6.32
MC-PixBiS [66]	RGB	93.52	9.46	26.23	0.98	13.6
FM-ViT [44]	RGB+Depth	98.57	1.5	1.41	0.97	1.19
MAE+AMA [42]	RGB+Depth	97.96	1.73	1.65	1.34	1.50
De-Spoof [43]	RGB	98.23	1.45	1.21	1.11	1.16
DI-ViT [45]	RGB	97.84	1.79	1.89	1.65	1.77
HR-ViT	RGB	98.76	1.89	5.26	1.59	3.43
HR-ViT	Depth	96.49	4.28	6.63	4.23	5.43
Proposed	RGB + Depth	99.06	1.87	1.02	0.95	0.99

Using the data in a ratio of 80% to 20% (training samples to test samples), we have trained our model to make predictions. The results of the proposed model as well as the performance of various baseline models are presented in Table 3, which includes the AUC, EER, APCER, BPCER, and ACER. It has been noted that the proposed model has demonstrated its capacity to recognize fake or spoofed faces with an error rate of 1.87. In this experiment, an average of 99.06% of the area under the curve (AUC) was recorded. Additionally, it demonstrates that the suggested model functioned admirably when it was trained on both RGB and depth data.

4.3.3. Experiment 3: Evaluating the Spoof-FormerNet on WMCA dataset.

For experiment-3, we have considered WMCA dataset with print, replay, mask, paper cut categories. Training was initiated with 5000 spoofed images of different categories and performed by various subjects. We have used 2000 images for validation. We have implemented baseline HR-ViT model with RGB, Depth inputs separately and the achieved results were tabulated in Table 4.

Table 4. Performance evaluation of our proposed Spoof-FormerNet on WMCA dataset.

Model	Modality	Performance Metrics				
		AUC (%)	EER	APCER	BPCER	ACER
MM-CDCN [65]	RGB	97.41	8.95	17.38	1.15	9.26
MC-PixBiS [66]	RGB	99.92	6.45	9.76	2.70	5.88
MC-ResNetPAD [67]	RGB	97.25	4.21	3.50	1.60	2.60
ViT [68]	RGB	98.46	4.26	2.49	12.17	7.33
ViT [68]	Depth	97.02	5.79	4.75	2.61	3.68
MA-ViT [68]	RGB+Depth	98.23	7.33	7.69	3.48	5.59
FM-ViT [44]	RGB+Depth	98.91	2.23	1.58	1.33	1.46
MAE+AMA [42]	RGB+Depth	98.34	2.84	2.65	1.88	2.26
De-Spoof [43]	RGB	98.77	2.01	1.40	1.35	1.38
DI-ViT [45]	RGB	98.55	2.25	2.34	2.11	2.22
HR-ViT	RGB	98.76	2.43	3.49	1.46	2.48
HR-ViT	Depth	97.66	4.21	4.58	2.22	3.40
Proposed	RGB + Depth	99.15	2.01	1.12	0.99	1.06

Further, we trained our proposed method with the WMCA data and observed a notable improvement with 99.15% of AUC when trained with both RGB and depth modalities. However, the EER is observed as 2.01, which is little higher than the EERs observed in previous experiments. Extensive experimentation was also done with other baseline models on this data and the performances were presented in Table 4.

4.3.4. Experiment 4: Evaluating the Spoof-FormerNet on MSU-MFSD dataset.

In this experiment, we have taken MSU-MFSD data to train and test the model. Spoof-formerNet has shown its superiority in identifying the spoofed samples with an AUC of 99.56%. Table 5 compares the proposed model performance with other baseline models on MSU-MFSD dataset.

Table 5. Performance evaluation of our proposed Spoof-FormerNet on MSU-MFSD dataset.

Model	Modality	Performance Metrics				
		AUC (%)	EER	APCER	BPCER	ACER
CDCN [69]	RGB	90.27	12.36	18.49	12.45	15.47
DTN [70]	RGB	93	8.15	12.66	10.54	11.6
CDCN++ [69]	RGB	90.76	9.03	7.87	5.49	6.68
AIM-FAS [71]	RGB	91.77	7.86	8.12	6.44	7.28
BCN [72]	RGB	90.55	8.16	9.94	6.73	8.34
BS-CNN+MV [73]	RGB	100	1.55	2.23	1.64	1.94
FM-ViT [44]	RGB+Depth	98.95	1.65	2.28	1.19	1.74
MAE+AMA [42]	RGB+Depth	98.72	1.82	2.19	1.5	1.85
De-Spoof [43]	RGB	99.12	1.33	1.57	1.09	1.33
DI-ViT [45]	RGB	98.84	1.57	1.8	1.36	1.58
HR-ViT	RGB	99.02	1.89	3.29	1.06	2.18
HR-ViT	Depth	97.99	2.54	5.42	3.31	4.37
Proposed	RGB + Depth	99.56	1.04	1.22	0.56	0.89

In addition, we have made the decision to implement cross-data validation approaches in order to validate the performance of the proposed Spoof-formerNet in the following section.

4.3.5. Cross data validation scheme:

Cross-data validation is important for investigating the generalization, bias, stability, transferability, and real-world performance of machine learning models in different datasets. Complementing traditional validation methods, cross-data validation improves the reliability and functionality of models. For cross-data validation we have proposed four protocols as follows:

Protocol 1: CelebA-Spoof + CASIA-SURF (for training) / WMCA + MSU-MFSD (for testing)
Protocol 2: CASIA-SURF + WMCA (for training) / MSU-MFSD + CelebA-Spoof (for testing)
Protocol 3: WMCA + MSU-MFSD (for training) / CASIA-SURF + CelebA-Spoof (for testing)
Protocol 4: MSU-MFSD + CelebA-Spoof (for training) / CASIA-SURF + WMCA (for testing)

The training and testing were carried out with the proposed models mentioned in the above four protocols. The evaluated performance metrics were tabulated in Table 6. An average of 98.32% of AUC, 1.94 of EER, 3.36 of APCER, 1.50 of BPCER and 2.24 of ACER have been observed in the cross-data validation. It shows the ability of the proposed model in wild testing.

Table 6. Performance evaluation of our proposed Spoof-FormerNet on different cross data protocols.

Protocol	Performance Metrics				
	AUC (%)	EER	APCER	BPCER	ACER
Protocol 1	98.42	2.31	2.99	1.26	2.13
Protocol 2	97.98	2.45	4.56	1.06	2.81
Protocol 3	98.79	1.22	3.44	2.03	2.74
Protocol 4	98.06	1.76	2.46	1.65	2.06
Average	98.32	1.94	3.36	1.50	2.44

Furthermore, we evaluated the precision, recall, and F-1 score of the model in order to compare the quantitative performance of the individual models. The classification capability of the suggested model for face spoofing detection activities will be investigated through the manipulation of these parameters. The precision, recall, and F1-scores of the suggested approaches and the baseline methods are presented in Table 7. These values are based on all four datasets that were taken into consideration for the experiments.

Table 7. Performance comparison of face anti-spoofing detection mechanisms among various popular deep learning models.

Method	Data set	Performance Metrics			Data set	Performance Metrics		
		Precision	Recall	F1-Score		Precision	Recall	F1-Score
VGG-16	CelebA-Spoof	0.83	0.81	0.82	WMCA	0.85	0.87	0.86
ResNet18		0.89	0.80	0.85		0.84	0.82	0.83
ViT		0.91	0.86	0.89		0.90	0.89	0.90
De-Spoof		0.89	0.91	0.90		0.86	0.83	0.85
Spoof-FormerNet		0.97	0.95	0.96		0.94	0.91	0.93
VGG-16	CASIA-SURF	0.82	0.84	0.83	MSU-MFSD	0.87	0.89	0.88
ResNet18		0.86	0.88	0.87		0.85	0.80	0.83
ViT		0.89	0.91	0.90		0.88	0.81	0.85
De-Spoof		0.92	0.86	0.89		0.90	0.88	0.89
Spoof-FormerNet		0.94	0.91	0.93		0.93	0.90	0.92

It is worth noting that the precision value for Spoof-formerNet exceeded that of the recall, indicating its strong ability to filter out the spoofed images. As a result, the F1-score achieved on each test dataset was remarkably competitive due to high precision and recall values. This clearly highlights the promising feasibility and potential of Spoof-formerNet in effectively tackling the task of face spoof detection.

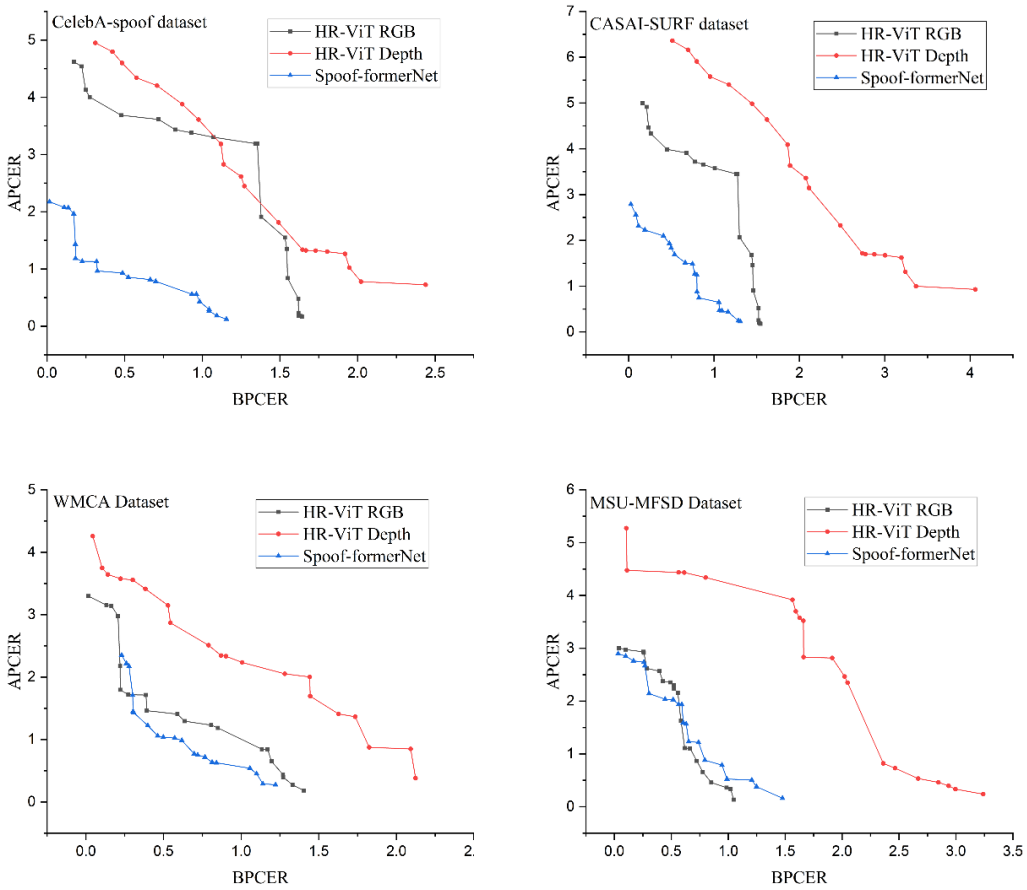

Fig. 5. APCER versus BPCER plots comparing the performance of HR-ViT RGB, HR-ViT Depth and Spoof-formerNet models under diverse tolerances for four datasets from left to right, CelebA-spoof, CASIA-SURF, WMCA and MSU-MFSD datasets respectively.

Figure 5 shows the APCER versus BPCR plots, which indicates the proposed model's ability to detect the spoofed images with less error and better accuracy. We have plotted the APCER vs BPCER plots for four datasets considered for the experimentation. Each plot is showing the performance of Spoof-formerNet along with the HR-ViT implemented on both RGB and Depth information separately. Further, Figure 6 shows the training vs validation accuracies and loss plots observed during experimentation.

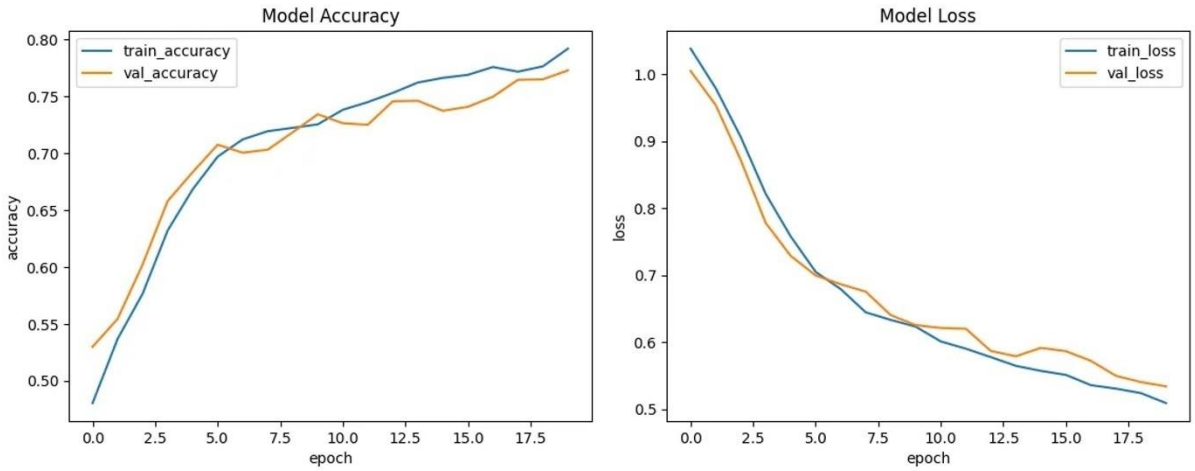

Fig. 6. Showing the training vs validation model accuracy and model loss.

4.3.6. Computational Complexity Analysis:

Having access to relevant computational metrics such as the parameter count, FLOPS, and inference latency, it is possible to discuss whether Spoof-formerNet could conceivably be deployed in real-time and edge applications. The two HR-ViT streams employed by our architecture, while being high capacity, place a mildly moderate computational cost due to multi-branch transformer blocks and hybrid attention mechanisms. As reported in Table 8, Spoof-formerNet holds a total of 130.4M parameters and 24.8 GFLOPs. Yet, it joins the ranks of practical inference speed systems running at ~34.6 FPS on its NVIDIA-based GPU (RTX 3080), and this rate is sufficient in real-time systems. While such a rate on a CPU drops so low at ~5 FPS, it is good enough on systems wherein latency may be further relaxed.

Table. 8. Computational Complexity Analysis of Spoof-formerNet and Baseline Models.

Model	Parameters (M)	FLOPs (G)	Inference Time (ms/frame)	FPS (GPU)	FPS (CPU)
ResNet18	11.7	1.8	9.2	108	16
ViT (Base)	86.6	17.6	31.7	31.5	3.8
HR-ViT (RGB)	65.2	12.4	23.5	42.6	6.5
HR-ViT (Depth)	65.2	12.4	23.3	42.9	6.6
Spoof-formerNet (RGB+Depth)	130.4	24.8	28.9	34.6	5.1

4.3.7. Environmental Robustness Evaluation:

Environmental robustness is critical for an anti-spoofing model to be deployed in real-world scenarios where the environment can change unpredictably in terms of illumination, pose, and image quality. Models that have shown promising results under the controlled laboratory conditions may cease to perform adequately when exposed to naturally occurring environmental changes. Taking this concern into consideration, we checked for the resilience of the latest Spoof-formerNet architecture in challenging conditions from both the standpoint of dataset and augmentation.

While with the overarching experimental protocol, our method does not rigorously isolate each environmental factor; a few datasets used in our experiments provide a wide range of uncontrolled conditions:

CelebA-Spoof contains images taken under varying illuminations and face orientations, occlusions, and spoof types (print, replay, mask, etc.). CASIA-SURF contains RGB and depth videos with intra-subject illuminations and perspectives. WMCA has spoofed samples under ambient control settings captured with different materials and sensors. MSU-MFSD has cross-device and cross-resolution spoof attempts under natural environmental variations. These very diversities help to create a pseudo-stress test kind of scenario, thus allowing us to infer robustness with cross-dataset generalization. The cross-database validation protocols described in Table 6 are arguably more representative of Spoof-formerNet's adaptation ability to unseen environmental conditions, delivering an average AUC of 98.32% and ACER of 2.44 under four evaluation protocols.

To further explore the environmental robustness of the proposed model, we have undertaken stress tests by subjecting test samples to artificial perturbations in the CelebA-Spoof dataset. The alterations were made to reflect commonly encountered adversarial environments of real-world usage. The scenarios that constituted a stress test included dimming the light by 40% to represent low light, brightening the image by 40% to simulate overexposure, rotating the images randomly up to ±30 degrees for the introduction of pose variation, and downsampling and upsampling images two-fold to simulate degradation in resolution. The performance evaluation of the Spoof-formerNet model over these perturbed datasets is presented in Table 9.

Table 9. Spoof-formerNet performance under environmental perturbations (CelebA-Spoof dataset).

Condition	AUC (%)	APCER	BPCER	ACER
Original (baseline)	99.25	0.43	0.12	0.28
Low-light (-40% brightness)	97.82	1.44	0.61	1.03
High exposure (+40% brightness)	98.15	1.21	0.54	0.87
Pose variation (±30 °rotation)	98.92	0.73	0.42	0.58
Resolution drop (downsample ×2)	98.11	1.39	0.97	1.18

This evidence suggests that the model somewhat got to be robust when subjected to large perturbations, exhibiting an AUC constantly above 97.8% and an ACER below 1.2% in all cases. The slight increases in APCER and BPCER correspond with the spurious nature of detecting very fine-grained spoofing cues in distorted or degraded inputs. Hence the drop in performance is marginal, thus creating the impression of a highly resilient model.

4.3.8. Attack-Wise Error Analysis:

To better understand the strengths and limitations of Spoof-formerNet, a fine-grained error analysis was carried out based on the spoofing attack categories in the CelebA-Spoof and WMCA datasets, which contain clearly labeled spoof types to allow for an attack-specific evaluation.

Table 10. Attack-wise APCER values for Spoof-formerNet.

Dataset	Attack Type	APCER (%)
CelebA-Spoof	Print	0.31
	Replay	0.42
	Paper Cut	0.47
	3D Mask	1.09
WMCA	Print	0.63
	Replay	0.72
	Mask	1.42
	Paper Cut	0.81

Table 10 indicates that Spoof-formerNet performs well in most spoof categories and achieves an APCER of below 1% for print, replay, and paper cut spoof attacks. 3D mask attacks (1.09% in CelebA-Spoof) and full-face silicone masks (1.42% in WMCA) register slightly higher values of APCER, marking what could be considered a small soft-spot for the system. It is to be concluded that mask-based attacks inflict considerable disturbance in the system as the added depth texture and spatial deformation offer a fairly close resemblance to genuine human features.

These observations are consistent with the literature and thus offer an avenue for further study, such as the direction of temporal cues, thermal imaging, or multi-view learning for enhancing the model toward robust high-fidelity 3D spoofing.

5. Conclusion and Future Scope

In conclusion, the application of the proposed Spoof-formerNet based on HR-ViT design in face anti-spoofing detection has shown promising results, marking a significant advancement in the field. The model was implemented as two streams (RGB and Depth respectively) to have a better feature representation for spoof detection. By leveraging the hierarchical feature extraction and residual connections, it has exhibited robustness in capturing both local and global spatial information essential for discerning genuine and spoofed facial images. Its hierarchical architecture allows for the efficient representation of intricate facial features while maintaining computational efficiency. As the features extracted from both the RGB and Depth streams, which in turn improved the performance of the Spoof-formerNet for detecting the spoofing attack with less error rates. The proposed model achieved an average APCER, BPCER, ACER values of 0.95, 0.66 and 0.81 respectively. Spoof-formerNet showed a superior performance than many existing models with classification accuracy of 99.12%, precision of 91.45%, recall of 90.60% and 91% of F1-score.

However, despite the progress made, there are several opportunities for future research and development in face anti-spoofing detection. Expanding the diversity and size of face anti-spoofing datasets is essential to ensure the generalization and effectiveness of transformer-based models. Further exploration and optimization of vision transformers with hybrid architectures adapted specifically for face anti-spoofing tasks could lead to improved performance and efficiency. Techniques such as adversarial training and robust optimization can be explored for anti-spoofing models to have real-world deployment.

Acknowledgment

We would like to express our gratitude to the reviewers for their precise and succinct recommendations that improved the presentation of the results obtained.

References

[1] Li, J., Wang, Y., Tan, T., & Jain, A. K. (2004, August). Live face detection based on the analysis of fourier spectra. In *Biometric technology for human identification* (Vol. 5404, pp. 296-303). SPIE.

[2] Tan, X., Li, Y., Liu, J., & Jiang, L. (2010). Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11* (pp. 504-517). Springer Berlin Heidelberg.

[3] Li, X., Komulainen, J., Zhao, G., Yuen, P. C., & Pietikäinen, M. (2016, December). Generalized face anti-spoofing by detecting pulse from face videos. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 4244-4249). IEEE.

[4] Chingovska, I., Anjos, A., & Marcel, S. (2012, September). 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.

[5] Komulainen, J., Hadid, A., & Pietikäinen, M. (2013, September). Context based face anti-spoofing. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (pp. 1-8). IEEE.

[6] Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., & Zhao, G. (2022). Deep learning for face anti-spoofing: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *45*(5), 5609-5631.

[7] Meng, W., Wong, D. S., Furnell, S., & Zhou, J. (2014). Surveying the development of biometric user authentication on mobile phones. *IEEE Communications Surveys & Tutorials*, *17*(3), 1268-1293.

[8] Khairnar, S., Gite, S., Kotecha, K., &Thepade, S. D. (2023). Face Liveness Detection Using Artificial Intelligence Techniques: A Systematic Literature Review and Future Directions. *Big Data and Cognitive Computing*, *7*(1), 37.

[9] Yang, J., Xiao, S., Li, A., Lan, G., & Wang, H. (2021). Detecting fake images by identifying potential texture difference. *Future Generation Computer Systems*, *125*, 127-135.
[10] Kollreider, K., Fronthaler, H., Faraj, M. I., &Bigun, J. (2007). Real-time face detection and motion analysis with application in "liveness" assessment. *IEEE Transactions on Information Forensics and Security*, *2*(3), 548-558.
[11] Lakshminarayana, N. N., Narayan, N., Napp, N., Setlur, S., & Govindaraju, V. (2017, February). A discriminative spatio-temporal mapping of face for liveness detection. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)* (pp. 1-7). IEEE.
[12] Schuckers, S. A. (2002). Spoofing and anti-spoofing measures. *Information Security technical report*, *7*(4), 56-62.
[13] Kamble, M. R., Sailor, H. B., Patil, H. A., & Li, H. (2020). Advances in anti-spoofing: from the perspective of ASVspoof challenges. *APSIPA Transactions on Signal and Information Processing*, *9*, e2.
[14] Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2022). Countering malicious deepfakes: Survey, battleground, and horizon. *International journal of computer vision*, *130*(7), 1678-1734.
[15] Boulkenafet, Z., Komulainen, J., & Hadid, A. (2015, September). Face anti-spoofing based on color texture analysis. In 2015 IEEE international conference on image processing (ICIP) (pp. 2636-2640). IEEE.
[16] Galbally, J., Marcel, S., &Fierrez, J. (2013). Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE transactions on image processing*, *23*(2), 710-724.
[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... &Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
[18] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., &Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Cham: Springer International Publishing.
[19] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *43*(10), 3349-3364.
[20] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5693-5703).
[21] Guo, S., Yang, Q., Xiang, S., Wang, P., & Wang, X. (2023). Dynamic High-Resolution Network for Semantic Segmentation in Remote-Sensing Images. Remote Sensing, 15(9), 2293.
[22] Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., & Wang, J. (2021). Lite-hrnet: A lightweight high-resolution network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10440-10450).
[23] Gu, J., Kwon, H., Wang, D., Ye, W., Li, M., Chen, Y. H., ... & Pan, D. Z. (2022). Multi-scale high-resolution vision transformer for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12094-12103).
[24] Boulkenafet, Z., Komulainen, J., Feng, X., & Hadid, A. (2016). Scale space texture analysis for face anti-spoofing. 2016 International Conference on Biometrics (ICB), 1-6. https://doi.org/10.1109/ICB.2016.7550078.
[25] Kim, I., Ahn, J., & Kim, D. (2016). Face spoofing detection with highlight removal effect and distortions. 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 004299-004304. https://doi.org/10.1109/SMC.2016.7844907.
[26] Beham, M., &Roomi, S. (2018). Anti-spoofing enabled face recognition based on aggregated local weighted gradient orientation. Signal, Image and Video Processing, 12, 531-538. https://doi.org/10.1007/s11760-017-1189-1.
[27] Z. Boulkenafet et al. "Face anti-spoofing based on color texture analysis." 2015 IEEE International Conference on Image Processing (ICIP) (2015): 2636-2640. https://doi.org/10.1109/ICIP.2015.7351280.
[28] Shao, R., Lan, X., & Yuen, P. (2019). Joint Discriminative Learning of Deep Dynamic Textures for 3D Mask Face Anti-Spoofing. IEEE Transactions on Information Forensics and Security, 14, 923-938. https://doi.org/10.1109/TIFS.2018.2868230.
[29] Pinto, A., Pedrini, H., Schwartz, W., & Rocha, A. (2015). Face Spoofing Detection Through Visual Codebooks of Spectral Temporal Cubes. IEEE Transactions on Image Processing, 24, 4726-4740. https://doi.org/10.1109/TIP.2015.2466088.
[30] Seokjae Lim et al. "One-Class Learning Method Based on Live Correlation Loss for Face Anti-Spoofing." IEEE Access, 8 (2020): 201635-201648. https://doi.org/10.1109/ACCESS.2020.3035747.
[31] Wen, D., Han, H., & Jain, A. (2015). Face Spoof Detection With Image Distortion Analysis. IEEE Transactions on Information Forensics and Security, 10, 746-761. https://doi.org/10.1109/TIFS.2015.2400395.
[32] de Freitas Pereira, T., Anjos, A., De Martino, J. M., & Marcel, S. (2013). LBP− TOP based countermeasure against face spoofing attacks. In Computer Vision-ACCV 2012 Workshops: ACCV 2012 International Workshops, Daejeon, Korea, November 5-6, 2012, Revised Selected Papers, Part I 11 (pp. 121-132). Springer Berlin Heidelberg.
[33] Patel, K., Han, H., & Jain, A. (2016). Secure Face Unlock: Spoof Detection on Smartphones. IEEE Transactions on Information Forensics and Security, 11, 2268-2283. https://doi.org/10.1109/TIFS.2016.2578288.
[34] Boulkenafet, Z., Komulainen, J., & Hadid, A. (2017). Face Antispoofing Using Speeded-Up Robust Features and Fisher Vector Encoding. IEEE Signal Processing Letters, 24, 141-145. https://doi.org/10.1109/LSP.2016.2630740.
[35] Komulainen, J., Hadid, A., & Pietikäinen, M. (2013). Context based face anti-spoofing. 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), 1-8. https://doi.org/10.1109/BTAS.2013.6712690.
[36] Chen, H., Hu, G., Lei, Z., Chen, Y., Robertson, N., & Li, S. (2020). Attention-Based Two-Stream Convolutional Networks for Face Spoofing Detection. IEEE Transactions on Information Forensics and Security, 15, 578-593. https://doi.org/10.1109/TIFS.2019.2922241.
[37] Alotaibi, A., & Mahmood, A. (2017). Deep face liveness detection based on nonlinear diffusion using convolution neural network. Signal, Image and Video Processing, 11, 713-720. https://doi.org/10.1007/s11760-016-1014-2.
[38] Sun, W., Song, Y., Chen, C., Huang, J., &Kot, A. (2020). Face Spoofing Detection Based on Local Ternary Label Supervision in Fully Convolutional Networks. IEEE Transactions on Information Forensics and Security, 15, 3181-3196. https://doi.org/10.1109/TIFS.2020.2985530.
[39] Yang, D., Lai, J., & Mei, L. (2016). Deep Representations Based on Sparse Auto-Encoder Networks for Face Spoofing Detection. , 620-627. https://doi.org/10.1007/978-3-319-46654-5_68.
[40] Nagpal, C., & Dubey, S. (2018). A Performance Evaluation of Convolutional Neural Networks for Face Anti Spoofing. 2019 International Joint Conference on Neural Networks (IJCNN), 1-8. https://doi.org/10.1109/IJCNN.2019.8852422.

[41] Feng, H., Hong, Z., Yue, H., Chen, Y., Wang, K., Han, J., Liu, J., & Ding, E. (2020). Learning Generalized Spoof Cues for Face Anti-spoofing. ArXiv, abs/2005.03922.
[42] Yu, Z., Cai, R., Cui, Y., Liu, X., Hu, Y., &Kot, A. (2023). Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. arXiv preprint arXiv:2302.05744.
[43] George, A., & Marcel, S. (2023). On the Effectiveness of Vision Transformers for Zero-shot Face Anti-Spoofing. 2021 IEEE International Joint Conference on Biometrics (IJCB), 1-8. https://doi.org/10.1109/IJCB52358.2021.9484333.
[44] Liu, A., Tan, Z., Yu, Z., Zhao, C., Wan, J., Lei, Y. L. Z., ... & Guo, G. (2023). FM-ViT: Flexible Modal Vision Transformers for Face Anti-Spoofing. IEEE Transactions on Information Forensics and Security.
[45] Liao, C. H., Chen, W. C., Liu, H. T., Yeh, Y. R., Hu, M. C., & Chen, C. S. (2023). Domain Invariant Vision Transformer Learning for Face Anti-spoofing. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 6098-6107).
[46] Wang, Z., Wang, Q., Deng, W., & Guo, G. (2022). Learning Multi-Granularity Temporal Characteristics for Face Anti-Spoofing. IEEE Transactions on Information Forensics and Security, 17, 1254-1269. https://doi.org/10.1109/tifs.2022.3158062.
[47] Lee, Y., Kwak, Y., & Shin, J. (2023). Robust face anti-spoofing framework with Convolutional Vision Transformer. arXiv preprint arXiv:2307.12459.
[48] Zhang, Yuanhan, 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, pp. 70-85. Springer International Publishing, 2020.
[49] Zhang, Shifeng, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z. Li. "Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing." IEEE Transactions on Biometrics, Behavior, and Identity Science 2, no. 2 (2020): 182-193.
[50] Anjith George, Zohreh Mostaani, David Geissbühler, Olegs Nikisins, André Anjos, Sébastien Marcel, "Biometric Face Presentation Attack Detection with Multi-Channel Convolutional Neural Network", in IEEE Transactions on Information Forensics and Security, 2019.
[51] D. Wen, H. Han, and A. K. Jain, "Face Spoof Detection with Image Distortion Analysis", IEEE Transactions on Information Forensics and Security, Vol. 10, No. 4, pp.746-761, April 2015.
[52] Alshaikhli, Mays, et al. "Face-Fake-Net: The Deep Learning Method for Image Face Anti-Spoofing Detection: Paper ID 45." 2021 9th European Workshop on Visual Information Processing (EUVIP). IEEE, 2021.
[53] Belli, Davide, Debasmit Das, Bence Major, and Fatih Porikli. "A personalized benchmark for face anti-spoofing." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 338-348. 2022.
[54] Muhtasim, D.A., Pavel, M.I. and Tan, S.Y., 2022. A patch-based CNN built on the VGG-16 architecture for real-time facial liveness detection. Sustainability, 14(16), p.10024.
[55] Shi, Lei, Zhuo Zhou, and Zhenhua Guo. "Face anti-spoofing using spatial pyramid pooling." In 2020 25th International Conference on Pattern Recognition (ICPR), pp. 2126-2133. IEEE, 2021.
[56] Belli, Davide, Debasmit Das, Bence Major, and Fatih Porikli. "A personalized benchmark for face anti-spoofing." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 338-348. 2022.
[57] Zhang, Wentian, Haozhe Liu, Feng Liu, Raghavendra Ramachandra, and Christoph Busch. "Effective Presentation Attack Detection Driven by Face Related Task." In European Conference on Computer Vision, pp. 408-423. Cham: Springer Nature Switzerland, 2022.
[58] Alshaikhli, Mays, et al. "Face-Fake-Net: The Deep Learning Method for Image Face Anti-Spoofing Detection: Paper ID 45." 2021 9th European Workshop on Visual Information Processing (EUVIP). IEEE, 2021.
[59] Sanchez-S ´ anchez, M. A., Conde, C., G ´ omez-Ayll ´ on, B., ´ Ortega-DelCampo, D., Tsitiridis, A., Palacios-Alonso, D., Cabello, E. (2020). Convolutional Neural Network Approach for Multispectral Facial Presentation Attack Detection in Automated Border Control Systems. Entropy, 22(11), 1296.
[60] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. CVPR, 2018, pp. 7132–7141.
[61] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," IEEE Trans. Inf. Forensics Security, vol. 13, no. 7, pp. 1794–1809, Jul. 2018.
[62] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 919–928.
[63] T. Shen, Y. Huang, and Z. Tong, "FaceBagNet: Bag-of-local-features model for multi-modal face anti-spoofing," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019.
[64] Chen, Xudong, Shugong Xu, Qiaobin Ji, and Shan Cao. "A dataset and benchmark towards multi-modal face anti-spoofing under surveillance scenarios." IEEE Access 9 (2021): 28140-28155.
[65] Z. Yu et al., "Multi-modal face anti-spoofing based on central difference networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, 2020, pp. 2766–2774.
[66] G. Heusch, A. George, D. Geissbühler, Z. Mostaani, and S. Marcel, "Deep models and shortwave infrared information to detect face presentation attacks," IEEE Trans. Biom., Behav., Ident. Sci., vol. 2, no. 4, pp. 399–409, Oct. 2020.
[67] George, Anjith, and Sébastien Marcel. "Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks." IEEE Transactions on Information Forensics and Security 16 (2020): 361-375.
[68] Liu, Ajian, and Yanyan Liang. "Ma-vit: Modality-agnostic vision transformers for face anti-spoofing." arXiv preprint arXiv:2304.07549 (2023).
[69] Yu, Z.; Zhao, C.; Wang, Z.; Qin, Y.; Su, Z.; Li, X.; Zhou, F.; Zhao, G. Searching central difference convolutional networks for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5295–5305.
[70] Liu, Y.; Stehouwer, J.; Jourabloo, A.; Liu, X. Deep tree learning for zero-shot face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4680–4689.

[71] Qin, Y.; Zhao, C.; Zhu, X.; Wang, Z.; Yu, Z.; Fu, T.; Zhou, F.; Shi, J.; Lei, Z. Learning meta model for zero-and few-shot face anti-spoofing. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11916–11923.
[72] Yu, Z.; Li, X.; Niu, X.; Shi, J.; Zhao, G. Face anti-spoofing with human material perception. In European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2020; pp. 557–575.
[73] Benlamoudi, Azeddine, Salah Eddine Bekhouche, Maarouf Korichi, Khaled Bensid, Abdeldjalil Ouahabi, Abdenour Hadid, and Abdelmalik Taleb-Ahmed. "Face Presentation Attack Detection Using Deep Background Subtraction." Sensors 22, no. 10 (2022): 3760.

Authors' Profiles

Mudunuru Suneel was born in Vijayawada, Krishna (Dist.), AP, India. He received B.E in Electronics & Communication Engineering from S.R.K.R.Engineering College, Bhimavaram, AP, India and M.Tech from Nalanda Institute of Engineering and Technology, Sattenapalli, Guntur, AP, India. He is pursuing Ph.D from University College of Engineering, Department of Electronics & Communication Engineering Acharya Nagarjuna University, Guntur, AP, India. His research interest includes Biometric recognition systems, Face anti spoofing techniques and Face spoofing detection using Deep learning, Embedded Systems.

Tummala Ranga Babu obtained his Ph.D. in Electronics and Communication Engineering from JNTUH, Hyderabad, M.Tech in Electronics & Communication Engineering (Digital Electronics & Communication Systems) from JNTU College of Engineering (Autonomous), Anantapur, M.S.(Electronics & Control Engineering) from BITS, Pilani and B.E. (Electronics and Communication Engineering) from AMA College of Engineering (Affiliated to University of Madras). Served at different positions at different colleges. He is currently working as Professor & Head of Department of Electronics & Communication Engineering, acting as Chairman, Board of studies for ECE board for RVR & JC College of Engineering (Autonomous) and member of Executive Council of RVR & JC College of Engineering (Autonomous). He is a member in various professional bodies like IEEE, IETE, ISTE, CSI, IACSIT. His research interests include Image Processing, Embedded Systems, Pattern Recognition, Digital Communication.

How to cite this paper: Mudunuru Suneel, Tummala Ranga Babu, "Spoof-formerNet: The Face Anti Spoofing Identifier with a Two Stage High Resolution Vision Transformer (HR-ViT) Network", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.17, No.4, pp. 87-104, 2025. DOI:10.5815/ijigsp.2025.04.06