TransConvNet: A Unified Deep Learning Framework for Multimodal Biometric Spoofing Verification

Abstract

Multimodal biometric systems are commonly used in different sectors, including banks, finance, and healthcare. These systems use various methods, such as face recognition, voice verification, and fingerprint matching, to prove the identity of users. Although these systems have many practical applications, they are still vulnerable to presentation attacks such as printed photos, 3D masks, replayed audio, and molded fingerprints. Anti-spoofing solutions are designed to guard biometric systems from these types of attacks. Existing solutions often rely on separate models for each modality to prevent spoofing attacks, making them impractical for embedded devices. To address this issue, we proposed TransConvNet, a unified multimodal deep learning framework that leverages the strengths of ConvNext and Vision Transformer (ViT) to detect spoofing across three modalities: face, voice, and fingerprint. ConvNext extracts local texture details, while ViT captures global dependencies using bottleneck adapter modules to enable cross-domain generalization. Experiments on diverse datasets (e.g., CelebA-Spoof, ASV-Spoof 2019, LivDet) show outstanding performance, with TransConvNet outperforming individual ConvNext and ViT models in AUC (up to 1.4% higher) and EER (as low as 1.5%). Cross-domain Leave-One-Out (LOO) testing results confirm strong generalization, showing that TransConvNet is suitable for multimodal biometric security in real-world applications.

1. Introduction

Multimodal biometric systems are crucial in enhancing security across various industries. These systems allow users to prove their identity using face recognition, voice verification, or fingerprint matching. Although these systems have remarkable capabilities, they are still sensitive to presentation attacks. For example, facial recognition systems (FRS) can be deceived by printed photos, 3D masks, and replayed images. Audio verification systems face logical access (LA) attacks, such as replay attacks, and physical access (PA) attacks, which include voice conversion (VC) and text-to-speech (TTS) attacks generated synthetically. In addition, fingerprint authentication is vulnerable to spoofing attempts involving fake fingerprint types.

People commonly use multimodal biometric systems in everyday life applications. For example, banks, financial institutions, consumer device makers, and healthcare providers widely use facial recognition technology. In contrast, automatic speaker verification (ASV) is increasingly used in banking, teleconferencing, call centers, and voice authentication. Similarly, fingerprint authentication has applications in law enforcement, time-attendance tracking, mobile device security, and ATMs.

Many authentication systems use multiple modalities to verify users, such as cell phones that authenticate users based on either fingerprint or face recognition. Similarly, voice and fingerprint biometrics are used in banking and healthcare. These multimodal biometric systems utilize separate models to validate spoofing attempts across multiple modalities, which may require more resources like memory and computation. Due to these limitations, they are not practical for embedded systems and smart devices. So, it is essential to develop a robust and unified antispoofing solution that can identify multiple spoofing attacks using a single model.

Researchers proposed various security-enhancing techniques to address challenges faced by multimodal biometric systems. These methods protect facial recognition systems from presentation attacks, secure speaker verification systems against logical and physical intrusions, and defend fingerprint biometrics from fake fingerprint molds. Figure. 1 provides an abstract view of our proposed unified multi-model anti-spoofing framework.

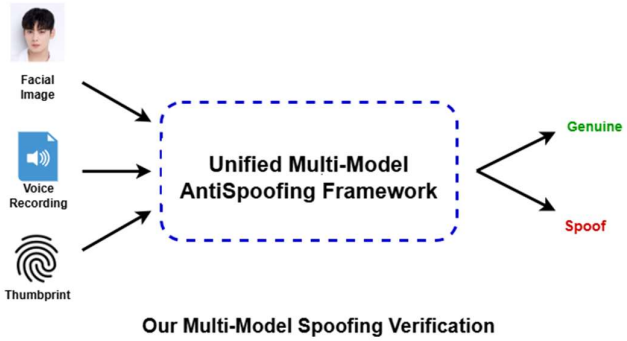

Figure 1: Overview of the proposed Unified Multi-Model Anti-Spoofing Framework, integrating facial, voice, and fingerprint modalities for robust spoof detection.

In facial recognition, traditional antispoofing methods have utilized handcrafted features [4,5,6,7] and temporal characteristics [1,2,3,8] of facial images. However, due to the limitations of these conventional techniques, deep-learning-based techniques have also been proposed [9,10,11,12]. Convolutional Neural Networks (CNNs) are now commonly used [15,16] to extract low-level features, while Vision Transformers have been used [13,14] to capture global features. Regardless of these improvements, current face antispoofing solutions perform well in intra-domain scenarios but often perform poorly in cross-domain scenarios due to significant differences in data distributions. To improve generalizability approaches such as data augmentation [24,25], transfer learning [22], transfer learning [22], ensemble learning [8,23], and pixel-wise supervision [18,19] have been examined.

In audio verification, methods exploiting temporal [26,27,28] and spectral features have been developed to address vulnerabilities in Automatic Speaker Verification systems. The ASV community has examined a variety of public

datasets and benchmarks for both logical access (LA) and physical access (PA) attacks on audio biometrics [29,30,31,32]. Earlier work focused on spectral features such as Constant Q Cepstral Coefficients (CQCC) [33,34], Linear Frequency Cepstral Coefficients (LFCC) [35,36], Mel Frequency Cepstral Coefficients (MFCC) [37], Short-Time Fourier Transform (STFT) [38], Constant-Q Transform (CQT) [39], and Mel Spectrogram [40]. These spectral features have limited performance against complex audio spoofing attacks. Recently, researchers have processed raw audio directly with deep learning architectures [26,28].

Fingerprint antispoofing solutions are usually divided into hardware-based and software-based methods [41,42,43]. Hardware-based methods use additional sensors to detect features for liveness verification, like blood pressure, skin distortion, and odor. On the other hand, software-based methods [44,45,46] rely only on fingerprint images and require no additional sensors.

Although existing solutions can efficiently identify unknown attacks and manage samples from diverse datasets, there is a continuous demand for a single unified solution that works for face, voice, and fingerprint spoofing verification and should be more effective, robust, and generalizable across multiple modalities.

In this work, we proposed a unified deep learning architecture called TransConvNet, which combines the features of ConvNext and Vision Transformer to improve overall performance. The ability of TransConvNet to detect spoofing attempts across multiple modalities makes it suitable for embedded devices with lower power consumption. ConvNext is appropriate for extracting low-level features through convolutional mechanisms, while Vision Transformer utilizes an attention-based mechanism to capture global features. We use a bottleneck adaptive Transformer within the Vision Transformer (ViT) by freezing the original layers and training only the adapter modules. Unlike Huang et al. [5], who use ensemble adapters within the adaptive Transformer, we demonstrate that a single native adapter is enough for effective performance. Our unified framework supports three modalities —face, voice, and fingerprint— and provides an adaptable solution for multimodal systems.

The main contributions of this research are listed below:

- We propose **TransConvNet,** a unified antispoofing framework that effectively detects various facial, voice, and fingerprint spoofing attacks.
- We introduce an adaptive Vision Transformer with bottleneck adapters to enable few-shot cross-domain generalization.
- We propose an innovative feature computation technique that effectively integrates ConvNext and Vision Transformer to capture local and global features.
- A detailed experimental study, along with cross-corpora analysis, was conducted on six diverse datasets to evaluate the effectiveness of our unified antispoofing framework.

2. Literature Review

This section provides a detailed overview of existing research methodologies on face, voice, and fingerprint antispoofing techniques. The subsequent sections discuss the progression in each modality from traditional approaches to state-of-the-art methods.

2.1. Face AntiSpoofing

Facial recognition systems are vulnerable to spoofing/presentation attacks such as printed photos, paper cuts, replayed images or videos, and 3D Masks. Various generalized and flexible solutions from traditional and deep learning categories have been explored to guard FRS from these threats.

Traditional methods have been extensively studied in face antispoofing research. These methods usually rely on handcrafted features such as Local Binary Patterns (LBP) [47,48], Histogram of Oriented Gradients (HOG) [49], and Difference of Gaussians (DoG) [50,51] to capture spoofing-oriented features from the images. These features are then passed to traditional machine learning (ML) classifiers like Support Vector Machines (SVMs) and Tree-Based Classifiers (TBC) to classify images as spoof or genuine. Motion-based signals such as head movements,

eye blinks, and image quality [52,53] are also researched to identify attacks like printed photos or paper cuts. While these methods can detect specific spoofing attacks with considerable success, they perform worse against modern spoofing techniques, such as replay attacks with high-quality videos. Also, these methods generally give saturated performance in varying conditions, such as lighting, pose, or resolution changes.

Due to the limited performance of methods based on handcrafted features, deep learning-based techniques have been studied. At the start, Convolutional Neural Networks have been extensively used to extract hierarchical features, including edges, contours, textures, and gradients [5,16]. CNN-based methods have performed well on challenging spoofing datasets by learning discriminative features directly from the data. Vision Transformers [54] have also been explored as a robust choice for facial antispoofing problems [13,14]. It splits images into small patches and then processes them using self-attention mechanisms. CNNs are limited to extracting local relationships across the image pixels, while ViT can capture global relationships and show significant performance in detecting challenging spoofing attempts. This ability of Vision Transformers demonstrated that they are practical for addressing presentation attacks.

Although deep learning-based methods perform well on samples from the same domain, they are worse against instances resembling common characteristics like illumination, background, or camera settings due to shifts in data distribution. Therefore, researchers have focused on cross-domain adaptation and generalization techniques to further improve the performance of deep learning-based methods. These methods perform well on entirely unseen domains by learning domain-invariant features. In [5], adapter modules are used inside Vision Transformer blocks to enable few-shot cross-domain generalization. This technique effectively captures global features from the given input but fails to capture low-level hierarchical features, such as edges and textures, which are critical for spoofing verification. In contrast, our method can extract local and global features by exploiting combined features from ConvNext and Vision Transformer.

2.2. Voice AntiSpoofing

Automatic Speaker Verification systems are vulnerable to attacks like replay, text-to-speech synthesis, voice conversion, and synthetic deep audio manipulations. Research in this area has evolved from using handcrafted features to developing sophisticated end-to-end systems.

At the start, handcrafted frontend methods were widely used in voice antispoofing systems. Traditional techniques utilize features from frequency and temporal characteristics. These include Mel-Frequency Cepstral Coefficients (MFCC) [37], Linear Frequency Cepstral Coefficients (LFCC) [35,36], Constant-Q Cepstral Coefficients (CQCC) [33,34], and Mel Spectrograms [40]. In [70], a detailed study is performed to measure the effectiveness of different frontends for spoofing verification. Traditional classifiers such as SVMs and Gaussian Mixture Models (GMMs) are then used to verify spoofs by employing these frontends. Over time, deep learning models like CNNs became popular to learn features from spectrograms. Along with that, Recurrent Neural Networks (RNNs) [55] and Long Short-Term Memory (LSTM) [56] networks have been used to capture temporal dependencies in audio signals. The method proposed in [71] processes raw audio using an RNN and leverages temporal information to separate genuine and fake audio samples. However, it lacks the low-level spatial information necessary for spoofing-related tasks.

As manual feature extraction is computationally expensive and time-consuming, end-to-end systems that can directly process raw audio have gained popularity. Therefore, models such as the RawNet family [57], Conformer [59], Wav2Vec [58], WaveNet [60], and LEAF [61] have been proposed. A series of ASVspoof challenges [29,30,31,32] have become standard benchmarks for evaluating these systems and testing them against various attacks, such as physical access (PA), logical access (LA), and deepfakes (DF). Metrics like Equal Error Rate (EER) and Tandem Detection Cost Function (t-DCF) are widely recognized for evaluating the effectiveness of such systems.

In short, existing antispoofing solutions process 2D audio signals using either CNN or Vision Transformer. However, we utilize both to take advantage of local and global features.

2.3. Fingerprint AntiSpoofing

Various materials like silicone, gelatin, and latex mimic human skin's properties, as they are flexible enough to create fingerprint molds. Antispoofing solutions offer an extra layer of security to prevent the fingerprint biometric system from such attacks. Fingerprint antispoofing solutions are categorized into hardware and software-based methods.

Additional sensors have been used in hardware-based methods to confirm fingerprint authenticity. For example, optical sensors have been used to detect skin moisture and micro-sweat pores, while capacitive sensors are used to measure electrical conductivity to detect spoofing attempts. Furthermore, temperature sensors distinguish between bonafide and fake samples by comparing the heat emitted from the fingerprints with the normal body temperature[63,64]. In [62], a comprehensive study is performed to analyze the performance of different sensors and the suitability of their applications.

In contrast, software-based methods depend on visual information extracted from fingerprint scans to separate genuine samples from fake ones. For this purpose, handcrafted techniques such as texture analysis, ridge distortion detection [66], and Local Binary Patterns (LBP) [65] have been utilized in traditional software-based methods. However, with the elevation of deep learning, models that use CNNs [67,68] and ViT [69] as backbone architectures have been widely accepted. In [69], a ViT-based unified architecture performs fingerprint recognition and spoofing authentication simultaneously. [66] Focus on ridge analysis and noise patterns present in the valleys between ridges.

In conclusion, existing antispoofing solutions target one specific modality. Due to this reason, multimodal biometric systems require separate models for each modality, which is not ideal for embedded devices. To address these limitations, we proposed TransConvNet, a unified framework capable of handling spoofing attacks launched on different modalities at once.

3. Proposed Method

This section presents **TransConvNet,** an end-to-end deep learning-based unified framework to verify spoofing from multiple modalities. This framework is designed to process three distinct types of biometric signals: 1) facial images, 2) voice recordings, and 3) fingerprint images by processing combined features from ConvNext and Vision Transformer. TransConvNet takes advantage of the unique capabilities of both architectures to provide an efficient and robust antispoofing solution. The convolution mechanism of ConvNext allows our model to capture local texture details such as edges, gradients, and repeating structures. In contrast, the Vision Transformer utilizes attention layers that help to extract global dependencies and long-term relationships across the spoofing samples.

Initially, TransConvNet standalone extracts features from both architectures and then combines them using a non-adaptive fusion strategy, which we discussed below in detail. This mechanism utilizes local and global information to classify genuine and spoofed biometric samples. At last, fused features are passed to a multi-layer perceptron (MLP) with two layers to get the final binary classification output, which indicates whether it is fake or genuine.

3.1. Pre-Processing

Before feeding the input signal into **TransConvNet**, we transform it into a (224×224) matrix with three channels to ensure consistency across all modalities. For face spoofing datasets that include video clips, we extract five equidistant frames from each video to ensure fair sampling of temporal information. We then employ the Multi-Task Cascaded Convolution Neural Network (MTCNN) to crop only the face portion from the entire image, as it contains the desired information for spoof detection. However, cropping performed using bounding box values returned by MTCNN sometimes cuts off the ears, top hairs, and chin, resulting in the loss of important information that may degrade the verification performance. To mitigate this issue, we add ten pixels of padding along all four sides to expand the area of the cropping frame. This cropping and padding process is visually

depicted in Figure 2. After this adjustment, we transformed the images into a $(224 \times 224 \times 3)$ format to make them compatible with the proposed architecture.

Figure 2: Comparison of face cropping methods for TransConvNet: Original Image, Tightly Crop (MTCNN without padding), and Square Crop (MTCNN with 10-pixel padding.

Raw audio recordings cannot be directly processed since our architecture is designed for 2-D signals. We convert the audio signals into a 2-D format to overcome this limitation by generating mel-spectrograms. A mel-spectrogram is a two-dimensional time-frequency audio representation where the temporal information is expressed on the x-axis. At the same time, the y-axis represents frequency on the mel scale. In contrast, the color of the mel-spectrogram represents the signal's energy at a given time-frequency point. The mathematical form of the mel-spectrogram transformation is elaborated in Eq. 1.

$$S_m(t,f) = -\log\left(\sum_{k=0}^{N-1} |X(k)|^2 \cdot M(f,k)\right) \tag{1}$$

Here, $S_m(t,f)$ is the mel-spectrogram at time t and mel-frequency $f, X(k)$is the Fourier transform of the audio signal, and $M(f,k)$ is the mel-filterbank matrix. Compared to a raw spectrogram, the mel-spectrogram comprises high-frequency details and emphasizes perceptually essential features. Mel-spectrograms are highly effective in voice antispoofing because they preserve critical speaker-specific characteristics such as the spectral envelope shaped by the vocal tract, harmonic structures, pitch patterns, speaking style, and energy variations. This transformation of raw audio into a mel-spectrogram allows the model to analyze audio data effectively in the spatial domain. Finally, we resize the mel-spectrograms to (224×224) resolution.

Similarly, fingerprint images also need pre-processing to meet the required size and to ensure a clear depiction of the fingerprint portion. However, currently, no pre-trained recognition model specifically detects the fingerprint portion in an image. We finetune the YOLO-9 model on fingerprint images to detect and isolate the fingerprint area before conducting spoofing verification. After isolating the fingerprint, we resize the images into a shape of (224×224) pixels with three color channels to maintain compatibility.

3.2. TransConvNet Architecture

As previously stated, TransConvNet integrates embeddings generated from ConvNext and Vision Transformer. Initially, it gets the embedding of dimension 1024 from ConvNext and then extracts embeddings of the same dimension from the Vision Transformer. We then use an averaging approach to combine these embeddings, resulting in a final dimension of 1024. This simple but effective fusion strategy permits TransConvNet to leverage the complementary strengths of convolutional and transformer-based representations while maintaining computational efficiency. After that, we pass them to a multi-layer perceptron (MLP) that includes two layers. The first layer consists of 1024 units, while the second layer has 512 hidden units, both layers followed by the ReLU activation function. Finally, we use two output units to classify input samples as spoofed or genuine. This lightweight classifier design allows TransConvNet to efficiently transform fused feature representations into binary classification outputs without adding unnecessary complexity. Figure 3 illustrates the complete architecture of our proposed TransConvNet, describing the integration of ConvNeXt and Vision Transformer embeddings followed by the MLP classifier.

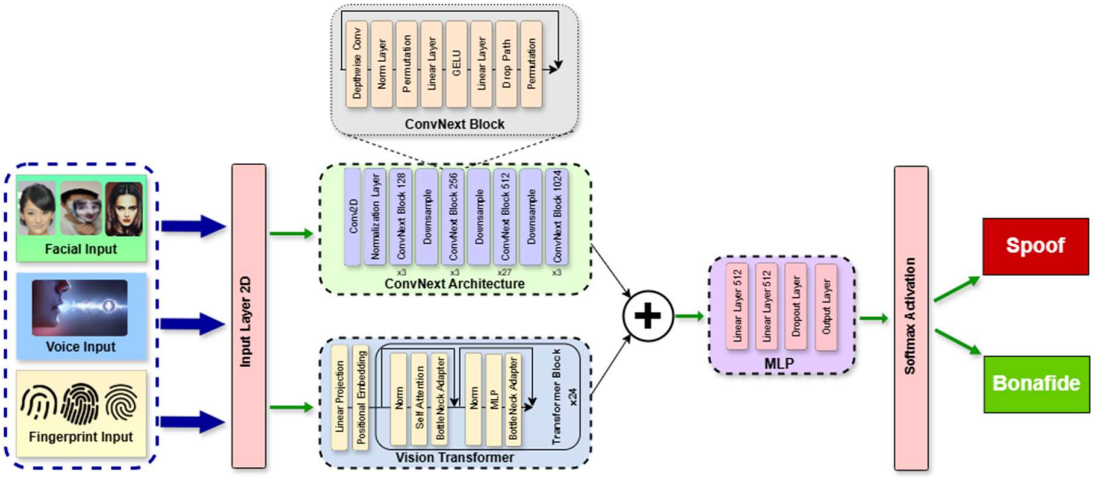

Figure 3: The proposed TransConvNet's architecture integrates ConvNeXt and Vision Transformer modules for processing facial, voice, and fingerprint inputs. Then, using an MLP and a softmax layer, the inputs are classified into spoof or bona fide.

To better understand the components of TransConvNet, we provide an overview of ConvNext and Vision Transformer. ConvNext is a modern convolutional neural network that bridges the gap between convolutional networks and vision transformers in terms of performance and has been widely adopted for many vision-related tasks. It utilizes a depthwise convolution layer to improve computational efficiency by splitting convolutions into pointwise operations. In our work, we use the pre-trained weights of the ConvNext base model from Torchvision. These weights were initially trained on ImageNet-21k and later finetuned on ImageNet-1k. We get the embeddings after the last convolutional block, which has a dimension of 1024. Our approach to extracting embeddings after the final convolutional stage ensures that low-level textures and high-level semantic features contribute to the final spoofing decision. The feature extraction from the ConvNext model can be represented as follows:

$$F_c = \text{ConvNext}(I; \theta_c) \in R^{1024} \quad (2)$$

In Eq. 2, F_c is the feature embedding from ConvNext, I is the input image $(224 \times 224 \times 3)$, $and\ \theta_c$ are the pre-trained ConvNext parameters.The Vision Transformer is the transformer-based architecture designed explicitly for image-related tasks. It processes images by splitting them into small patches, typically 16x16 pixels, treating each patch as a token in the same way as words are treated in natural language processing (NLP). ViT introduces

positional embeddings to encode the relative positions of these image patches. These positional embeddings are combined with patch embeddings before being passed into the transformer block to preserve spatial information within the image. By incorporating ViT, our architecture benefits from the model's strength in capturing global contextual relationships, which are crucial for identifying slight variations typical in spoofed samples. The following equation represents the feature extraction process using the Vision Transformer (ViT):

$$F_v = \mathrm{ViT}\big(P(I) + E_{\mathrm{pos}}; \theta_v\big) \in R^{1024} \tag{3}$$

where F_v is the ViT feature embedding, $P(I)$ is the patch embedding of input image I, E_{pos} is the positional embedding, and θ_v are the ViT parameters. To integrate ViT into TransConvNet, we use pre-trained weights from a large Vision Transformer model (ViT-L/16) that was trained on the ImageNet-1K dataset using the SWAG (Semi-Weakly Supervised Pretraining) linear finetuning approach. Importantly, we introduce two bottleneck adapter modules in each transformer block to facilitate few-shot learning and enhance cross-domain generalization, which is another significant contribution to our work. One adapter module is placed after the attention layer, while the other is placed after the MLP inside each transformer block. Each adapter module consists of two linear layers: the first layer reduces the embedding dimension to a lower bottleneck dimension (set to 64 in this work). In contrast, the second layer returns the output to match the adapter's input dimension. A GELU activation function is applied between these linear layers to introduce non-linearity. Additionally, we have a residual connection within the adapter module to maintain the flow of information. The adapter module operation can be mathematically formulated as:

$$A(x) = W_2 \cdot \mathrm{GELU}(W_1 \cdot x) + x \tag{4}$$

Here, $A(x)$ is the adapter output, x is the input embedding, $W_1 \in R^{64 \times d}$ reduces to bottleneck dimension 64, $W_2 \in R^{64 \times d}$ restores the dimension, and d is the input dimension. By training only these lightweight adapter modules and freezing all original layers, we achieve better domain adaptability with fewer training resources, making the model suitable for scenarios with limited labeled spoofing data.

We experiment with adaptive and non-adaptive techniques to combine feature embeddings from ConvNext and ViT. In the non-adaptive approach, we use two methods: averaging and concatenation. In averaging, we compute the element-wise average of the embeddings extracted from both models, while in concatenation, we directly combine the embeddings across the feature dimension. In contrast, adaptive techniques leverage the dynamic mechanism to learn the contribution of each embedding vector. Specifically, we employ an adaptive attention mechanism by passing the combined embeddings to the attention layer. This layer computes dynamic attention scores for the embeddings captured from both models. These scores represent the relative importance of each embedding for the specific task. After that, these scores are normalized using the softmax function to weight the embeddings adaptively. Finally, the weighted embeddings are summed up to produce the fused representation of the same dimensionality as the original embeddings. However, non-parametric averaging techniques outperform concatenation and adaptive attention mechanisms in our task setting. Therefore, we adopt an averaging approach in our proposed TransConvNet, prioritizing simplicity, stability, and better generalization over additional complexity. The final fused representation is obtained by averaging the ConvNext and ViT embeddings, as shown below:

$$F_{\mathrm{fused}} = \frac{F_c + F_v}{2} \in R^{1024} \tag{5}$$

In Eq. 5, F_{fused} is the fused feature embedding, and F_c, F_v are the ConvNext and ViT embeddings, respectively. After feature fusion, we utilize an MLP that consists of two fully connected layers to process fused embeddings. The first layer consists of 512 hidden neurons and aims to reduce the dimension of the fused features. The ReLU activation function and dropout layer are applied after the first layer to introduce non-linearity and prevent overfitting. The final output layer consists of two neurons. This layer maps the classification outcomes into the final binary decision as spoofed or genuine. These two layers of MLP facilitate TransConvNet in performing

robust classification by effectively managing non-linear interactions in the fused feature space. The final classification decision is obtained by passing the fused features through a multi-layer perceptron (MLP), as defined below:

$$y = \text{softmax}(W_3 \cdot \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 \cdot F_{\text{fused}} + b_1) + b_2) + b_3) \quad (6)$$

Here, $y \in R^{2}$ is the classification output (spoof/genuine), $W_1 \in R^{1024\times1024}, W_2 \in R^{1024\times512}, W_3 \in R^{512\times2}$ and $b1, b2, b3$ are biases.

3.3. Training Details

As we mentioned before, TransConvNet is a unified framework that provides an antispoofing solution for three distinct modalities. We train it using facial images, voice recordings, and fingerprint images simultaneously. To ensure fairness and prevent the model from being biased toward samples from a specific modality, we include an equal number of samples from each modality in every training batch. In our training setup, we select eight samples from each modality, resulting in a total batch size of 24. Moreover, to avoid the model from being biased toward a specific class, we take four spoofed and four genuine samples from a particular modality.

We use the PyTorch framework to implement TransConvNet and the AdamW optimizer to update the model weights with an initial learning rate of $1 \times e^{-4}$ and a weight decay rate of $1 \times e^{-6}$. We employ the cosine learning rate scheduler to adjust the learning rate during training. The cross-entropy loss function is used to compute the loss, and we train the model for a total of 1000 iterations. The mathematical details of the cross-entropy loss are mentioned in Eq. 7.

$$\mathcal{L} = -\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}[\mathrm{y_i}\log(\hat{y}_\mathrm{i}) + (1-\mathrm{y_i})\log(1-\hat{y}_\mathrm{i})] \quad (7)$$

Here, $\mathcal{L}$ is the cross-entropy loss, N Is the batch size, y_i is the actual label, and $\hat{y}_\mathrm{i}$ is the predicted probability for the sample i.

4. Experiments & Results

In this section we provides a comprehensive overview of the datasets used for each modality, the evaluation metrics employed to assess the model's performance, and the results achieved. Additionally, we discuss comparative analysis with existing models and explore TransConvNet's cross-domain generalization capabilities.

4.1. Datasets

Experiments are performed on multiple diverse datasets to evaluate the effectiveness of our proposed method. Every dataset contains several types of spoofing attacks as well as bonafide samples. We discuss the details of these datasets in the following subsections.

4.1.1. Face AntiSpoofing

We utilized five distinct datasets for face antispoofing verification: CelebA-Spoof, Oulu-NPU, CASIA-FASD, MSU-MFSD, and Idiap Replay Attack. CelebA-Spoof contains 561,575 images from 10,177 subjects and includes 10 spoof-type annotations. The primary attack categories are printed paper, paper cut, replayed, and 3D mask. The remaining four datasets include video clips. We capture five evenly spaced frames from each video to ensure diversity. Oulu-NPU is one of the largest and most diverse datasets, comprising 5,940 real-world mobile device recordings under varying conditions. CASIA-FASD consists of replayed, warped photos and cut photo attacks, with each attack photo in four different resolutions across 50 subjects, resulting in 600 samples (50 subjects × 3 attack types × 4 resolutions). MSU-MFSD is a small dataset of 280 videos with high-quality print and replay attacks recorded using tablets and smartphones. Lastly, the Idiap Replay Attack is a challenging dataset of 1300

samples with 50 subjects recorded under controlled conditions and using various spoof mediums. A summary of the datasets used in our experiments, including their modality, number of subjects and data samples, sensors, and spoof types, is provided in Table 1.

Table 1: The datasets of Face Anti-Spoofing under different illumination and environment conditions (V means video and I means Image)

Dataset	Year	Modality	#Subjects	#Data(V/I)	#Sensor	Spoof Type
Replay-Attack	2012	RGB	50	1200(V)	2	1 Print, 2 Replay
CASIA-MFSD	2012	RGB	50	600(V)	3	1 Print, 2 Replay
MSU-MFSD	2015	RGB	35	440(V)	2	1 Print, 2 Replay
Oulu-NPU	2017	RGB	55	5990(V)	6	2 Print, 2 Replay
CelebA-Spoof	2020	RGB	10,177	561,575(I)	>10	3 Print, 3 Replay,1 3D, 3 Paper Cut

For the Face Antispoofing assessment, we followed two distinct evaluation protocols:

1. Evaluation of the CelebA Dataset: We conducted multi-class classification using the CelebA dataset in our initial protocol. We allocated 494,405 images for training and reserved 67,170 images for testing. Our final model includes five output classes: four for different types of attacks and one for predicting genuine samples.

2. Leave-One-Out Testing (LOUT): This protocol involved cross-domain generalization evaluation across four datasets: MSU-MFSD (M), OULU-NPU (O), Replay-Attack (R), and CASIA (C). Each dataset varies concerning illumination, background complexity, and image resolution. The training process employs three datasets, leaving one for testing. This procedure is meticulously replicated four times, ensuring that each dataset serves as the test set precisely once throughout the evaluation.

4.1.2. Voice AntiSpoofing

For voice antispoofing, we utilized the ASV-Spoof 2019 dataset, which comprises two primary categories of samples: (1) Logical Access (LA), including Voice Conversion (VC) and Text-to-Speech (TTS) samples, and (2) Physical Access (PA), which includes replay attack samples.

Both categories are further divided into three parts: training, development, and evaluation. The LA section contains six known attack IDs (A1 to A6), while the evaluation dataset includes 13 unknown attack IDs (A7 to A19). The total samples of the LA part are 122,157: 25,380 for train, 24,844 for dev, and 71,933 for eval. In contrast, the PA section features 27 known attacks, incorporating various playback devices, recording devices, and environmental acoustics, with the evaluation set again comprising 13 unknown attacks and configurations. The training set contains 54,000 samples, the dev set contains 33,534, and the eval set contains 153,522 samples from the PA collection.

4.1.3. Fingerprint AntiSpoofing

This section provides an overview of the LivDet 2013 dataset to evaluate our model's performance with fingerprint images. This dataset was designed to benchmark fingerprint spoof detection algorithms across various spoofing materials and sensors. LivDet 2013 consists of 16,000 images, distributed equally, 50% for training and 50% for testing. This dataset includes authentic fingerprint images captured by sensors such as Biometrika, Italdata, Crossmatch, and Swipe. However, spoofing samples were generated using materials like Gelatin, Wood glue, Latex, Ecoflex, and Modasil.

4.2. Evaluation Metrics

We employed several classification metrics to assess the performance of our proposed method, including Precision, Recall, and Accuracy, in addition to AUC (Area Under the Curve) and EER (Equal Error Rate) for Face and Fingerprint evaluations. We utilized the minimum tandem Detection Cost Function (min-tDCF) alongside EER for voice antispoofing. Lower values of EER and t-DCF indicate better model performance, with EER representing the point where the False Acceptance Rate (FAR)—the proportion of non-spoof samples incorrectly

accepted as legitimate—equals the False Rejection Rate (FRR)—the proportion of spoof samples incorrectly rejected. Mathematically, EER is defined as:

$$\mathrm{EER} = \mathrm{FAR}(\theta^*) = \mathrm{FRR}(\theta^*), \quad \theta^* = \arg\min_{\theta} |\mathrm{FAR}(\theta) - \mathrm{FRR}(\theta)| \tag{8}$$

In Eq. 8, FAR is the False Acceptance Rate, FRR is the False Rejection Rate, and θ^* is the optimal threshold.

4.3. Results Analysis

Table 2 summarizes TransConvNet's results, indicating Precision, Recall, F1-Score, EER, and AUC scores across three biometric modalities: face, voice, and fingerprint. As mentioned in Section 3, TransConvNet leverages features extracted from both the ConvNext and ViT backbones. This experimentation aims to compare the results of our proposed model with individual ConvNext and ViT backbones. The reported results demonstrate that TransConvNet outperforms both the ConvNext and ViT models with a significant margin.

We conducted multi-class classification by treating each spoof category as a separate class, along with the genuine samples class. To compute the AUC score in this multi-class scenario, we adopted a One-vs-Rest (OvR) approach, treating each class as positive while considering all others as negative. We determined the final AUC score through a macro-averaging method. TransConvNet outperforms both ConvNext and ViT by 0.16% and 1.03%, respectively, in terms of AUC score.

Table 2: Performance comparison of ConvNeXt, Vision Transformer, and the proposed TransConvNet across face, voice, and fingerprint modalities using key metrics such as Accuracy, AUC, F1 Score, EER, and min-tDCF.

Modality	Dataset	Model	Accuracy%	AUC%	F1 Score%	EER	min-tDCF
Face	CelebA	CovNext	83.17	99.79	86.41	0.022	-
		Vision Transformer	91.50	98.93	93.63	0.053	-
		TransConvNet (ours)	**98.12**	**99.95**	**98.64**	**0.009**	**-**
Voice	ASV-Spoof	CovNext	89.68	81.0	90.73	8.54	0.12
		Vision Transformer	89	95.31	92.94	7.32	0.063
		TransConvNet (ours)	**90.84**	**97.01**	**94.67**	**6.59**	**0.041**
Fingerprint	LivDet	CovNext	92.10	97.65	93.86	0.09	-
		Vision Transformer	94.61	98.96	95.25	0.043	-
		TransConvNet (Multi)	96.60	99.72	96.58	0.025	-
		TransConvNet (Binary)	**98.07**	**99.75**	**98.02**	**0.012**	**-**

To process samples belonging to voice modality, we employ numerous 2D signal representations of audio recordings before inputting them into our model. These representations include Mel Spectrogram, Linear Frequency Cepstral Coefficients (LFCC), Mel-Frequency Cepstral Coefficients (MFCC), Short-Time Fourier Transform (STFT), and Constant Q Transform (CQT). Our experiments indicate that we achieve superior results with the Mel Spectrogram and CQT representations. Consequently, incorporating all modalities, we have utilized the Mel Spectrogram for the final combined batch training. For the ASV-Spoof benchmark dataset, the results are as follows: ConvNext yields a min-tDCF of 0.002 and an EER of 3.4, while ViT results in a min-tDCF of 0.04 and an EER of 5.4. By incorporating both ConvNext and ViT, we achieved a min-tDCF score of 0.0034 and an EER of 1.5.

We conducted binary and multi-class classification experiments using the Livdet 2013 dataset for fingerprint antispoofing. This study treated images captured by the Biometrika, Italdata, Crossmatch, and Swipe sensors as separate categories. Our findings reveal that binary classification yields better results than multi-class classification because it focuses only on distinguishing between live and spoof samples without needing to model inter-sensor variability. The Italdata class is significantly challenging to differentiate between legitimate samples.

TransConvNet achieves an AUC score of 1.2% higher than ConvNext and 1.4% higher than ViT, demonstrating its robustness and effectiveness.

In short, TransConvNet beats both ConvNext and ViT across all three biometric modalities. It can handle sensor variation and complex spoof types, and these results confirm that it is practical for real-world multimodal biometric security systems.

4.4. Comparison with SOTA methods and Cross-Domain Generalizability

To compare our model with existing state-of-the-art methods and to assess our approach's cross-domain generalization capabilities, we perform Leave-One-Out (LOO) testing on the widely adopted MICO protocol, which consists of four datasets: MSU MFSD (M), IDIAP REPLAY ATTACK (I), CASIA-FASD (C), and OULU-NPU (O).In LOO testing, we train the model using three source domain datasets and then evaluate its performance on an unknown target domain dataset. We follow the standard convention to present the results in Table 3. For example, "O&C&M to I" indicates that OULU-NPU (O), CASIA-FASD (C), and MSU MFSD (M) are used as source domain datasets, while IDIAP REPLAY ATTACK (I) is the target domain.

Table 3: Cross-domain generalization evaluation on the standard MICO Leave-One-Out (LOO) testing protocol. The proposed TransConvNet demonstrates superior performance in three out of four settings, achieving the best HTER and AUC across domain shifts.

Methods	**I &C&M to O**		**O &C&M to I**		**O &C&I to M**		**O &M&I to C**	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
LBPTOP[14]	53.15	44.09	49.45	49.54	36.90	70.80	42.60	61.05
MSLBP[33]	50.29	49.31	50.30	51.64	29.76	78.50	54.28	44.98
ColorTexture[4]	63.59	32.71	40.40	62.78	28.09	78.47	30.58	76.89
BinaryCNN[46]	29.61	77.54	34.47	65.88	29.25	82.87	34.88	71.94
MMD-AAE[24]	40.98	63.08	31.58	75.18	27.08	83.19	44.59	58.29
MADDG[39]	27.98	80.02	22.19	84.99	17.69	88.06	24.50	84.51
RFM[40]	16.45	91.16	17.30	90.48	13.89	93.98	20.27	88.16
SSDG-M[21]	25.17	81.83	18.21	94.61	16.67	90.47	23.11	85.45
D2AM[7]	15.27	90.87	15.43	91.22	12.70	95.66	20.98	85.58
DRDG[30]	15.63	91.75	15.56	91.79	12.43	95.81	19.05	88.79
ANRL[29]	15.67	91.90	16.03	91.04	10.83	96.75	17.85	89.26
SSAN[44]	19.51	88.17	14.00	94.58	10.42	94.76	16.47	90.81
AMEL[61]	11.31	93.96	18.60	88.79	10.23	96.62	11.88	94.39
EBDG[11]	15.66	92.02	18.69	92.28	9.56	97.17	18.34	90.01
IADG	8.86	97.14	10.62	94.50	**5.41**	98.19	**8.70**	**96.44**
TransConvNet(ours)	**7.23**	**98.59**	**6.78**	**99.59**	5.77	**98.81**	10.02	95.12

As shown in Table 3, TransConvNet achieves the highest performance across three out of four cross-domain settings in the MICO protocol, clearly outperforming all other state-of-the-art methods. It is worth noticing that TransConvNet achieves an AUC of 98.59 (I&C&M to O), 99.59 (O&C&M to I), and 98.81 (O&C&I to M), demonstrating strong generalization to unseen domains. This outstanding performance is attributed to TransConvNet's dual-backbone architecture, which captures both local texture and global structure cues. However, in the O&M&I to C setting, the model shows relatively lower performance (AUC 95.12%) compared to its performance on other domains. This may be due to CASIA-FASD's high variability in acquisition conditions and limited intra-class diversity, making it more challenging for the model to generalize without seeing similar samples during training.

4.5. Ablation Study

We conducted an ablation study to analyze the effectiveness of various activation functions. We also evaluate the significance of different audio frontends, such as MFCC and STFT, to transform voice recordings into 2D signal representations.

4.5.1. Effectiveness of Different Activation Functions

We examine the importance of the activation function by conducting several experiments using different activation functions on various datasets. Specifically, we utilize three activation functions: ReLU, GELU, and Leaky ReLU. Table 4 presents the results of all activation functions in terms of the AUC score. Since evaluating a single dataset is unreliable, we assess five datasets: OULU-NPU, Replay Attack, CASIA-FASD, MSU-MFSD, and CelebA-Spoof. Figure 4 illustrates the results in a bar plot for clarity. It is evident that ReLU consistently performs better than both GELU and Leaky ReLU. This is likely because ReLU is simple and helps the model learn faster by avoiding problems like vanishing gradients.

Table 4: Ablation study analyzing the effectiveness of different activation functions across datasets. ReLU consistently performs better in most settings, demonstrating its suitability for our architecture.

Datasets	Activation Functions		
	GELU	RELU	Leaky ReLU
CelebA	99.76	99.63	99.83
I &C&M to O	98.23	98.59	97.0
O &C&M to I	95.46	99.59	93.68
O &C&I to M	96.0	98.81	92.07
O &M&I to C	96.01	95.51	91.66

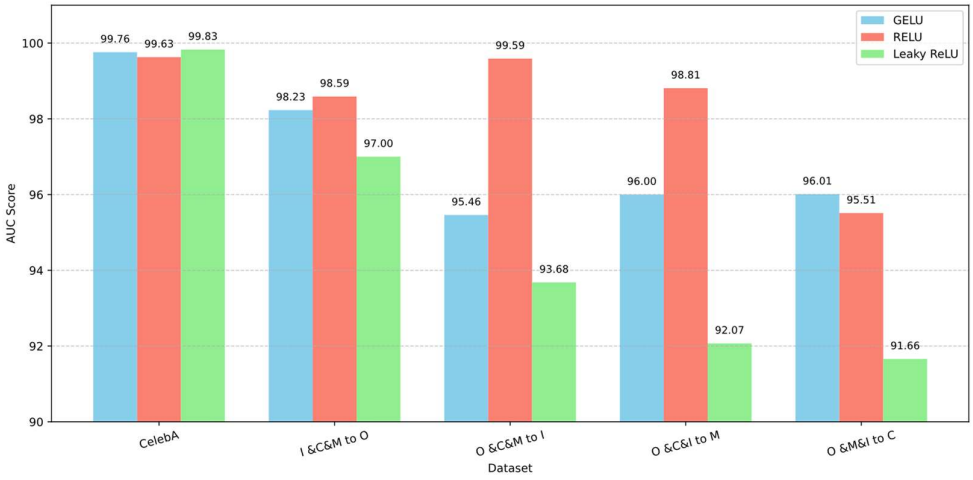

Figure 4: This figure shows the impact of GELU, ReLU, and Leaky ReLU on AUC scores for five dataset settings. GELU and ReLU generally outperform Leaky ReLU, with ReLU achieving the best overall performance.

4.5.2. Analysis of Audio Frontends

We conducted experiments using five popular audio frontends to assess the impact of different audio representations: Mel Spectrogram, LFCC, MFCC, STFT, and CQT. Each method converts raw audio into a 2D time-frequency representation, which serves as input for our model. A comparative analysis of different audio frontends is reported in Table 5.

Table 5: Evaluation of various audio feature extraction methods for audio signal processing. Mel-Spectrogram outperforms other frontends in terms of overall AUC, and EER, indicating its effectiveness in capturing spoof-related cues.

Audio Frontend	Accuracy%	Precision%	Recall%	F1 score%	AUC%	min-tDCF	EER
Mel-Spectrogram	90.84	98.94	90.74	94.67	97.01	0.09245	8.59
LFCC	76.95	99.60	74.60	85.31	93.35	0.25402	21.94
MFCC	86.38	98.73	85.92	91.88	93.32	0.14081	11.95
STFT	86.77	99.70	85.50	92.06	96.01	0.14495	13.58
CQT	89.04	99.11	88.57	93.54	96.18	0.11427	10.48

The Mel Spectrogram achieved the best overall performance among all the frontends, with the highest AUC score of 97.01 and the lowest Min t-DCF (0.09245) and EER (8.59). These results demonstrate that Mel Spectrograms are a discriminative representation for distinguishing between bonafide and spoofed audio samples.

In comparison, LFCC performed the worst across all the metrics. This may be because it does not effectively capture cues essential for spoofing verification. In contrast, MFCC, STFT, and CQT also gave good results, but they didn't perform as well as the Mel Spectrogram in most metrics. Among them, CQT achieved the second-highest AUC (96.18) and a low EER (10.48), showing that it could be a strong alternative.

These results indicate that selecting an audio frontend affects the model's performance. Mel Spectrograms are particularly well-suited for our model architecture, which utilizes both local patterns (via CNNs) and global context (via transformers).

5. Conclusion

Our proposed method is the first attempt to verify spoofing samples from three distinct modalities: facial images, voice recordings, and fingerprint molds. We introduced TransConvNet, a unified multimodal antispoofing framework that combines features from ConvNext and Vision Transformer using an averaging method. ConvNext allows our model to extract local features from the input sample, which is crucial for spoofing verification. In contrast, ViT effectively captures long-term global dependencies to make our model more robust. Inside each transformer block, we introduced bottleneck adapter modules, and during training, we froze all original layers of ViT and trained only the newly added adapters. We evaluated the performance of our method using state-of-the-art benchmark datasets. Our results demonstrate the cross-domain generalization and few-shot learning capabilities of our approach. TransConvNet is suitable for verifying spoofing attempts in multimodal recognition systems, such as smartphones, that employ facial and fingerprint authentication.

References

[1] K. H. Fronthaler and J. Bigun, "Kollreider, K., Fronthaler, H., & Bigun, J. (2009). Non-intrusive liveness detection by face images. Image and Vision Computing, 27(3), 233-244.," *Image and Vision Computing,* vol. 27(3), no. 4, pp. 233-244., 2009.

[2] B. J. Komulainen and K. , "Boulkenafet, Z., Komulainen, J., & Hadid, A. (2016). Face spoofing detection using colour texture analysis. IEEE Transactions on Information Forensics and Security, 11(8), 1818-1830.," *Information Forensics and Security,* no. 5, pp. 1818-1830, 2016.

[3] C. A. Anjos and S. Marcel, "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.," *biometrics special interest group,* no. 6, pp. 1-7, 2012.

[4] M. A. Hadid and . M. Pietikäinen, "Määttä, J., Hadid, A., & Pietikäinen, M. (2011, October). Face spoofing detection from single images using micro-texture analysis. In 2011 international joint conference on Biometrics (IJCB) (pp. 1-7). IEEE.," *international joint conference on Biometrics,* no. 7, pp. 1-7, 2011.

[5] H. D. Sun and Y. Liu, "Huang, H. P., Sun, D., Liu, Y., Chu, W. S., Xiao, T., Yuan, J., ... & Yang, M. H. (2022, October). Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In European conference on computer vision (pp. 37-54). Cham: Springer Nature Swit," *computer vision,* no. 9, pp. 37-54, 2022.

[6] R. C. Z. Yu and C. Kong, "Cai, R., Yu, Z., Kong, C., Li, H., Chen, C., Hu, Y., & Kot, A. C. (2024). S-adapter: Generalizing vision transformer for face anti-spoofing with statistical tokens. IEEE Transactions on Information Forensics and Security.," *Information Forensics and Security,* no. 10, 2024.

[7] Zitong, A. Liu and K. H. Cheng, "Yu, Z., Liu, A., Zhao, C., Cheng, K. H., Cheng, X., & Zhao, G. (2023). Flexible-modal face anti-spoofing: A benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6346-6351).," *Computer Vision and Pattern Recognition,* no. 11, pp. 6346-6351, 2023.

[8] Liu, Stehouwer and J. Stehouwer,, "Liu, Y., Stehouwer, J., Jourabloo, A., & Liu, X. (2019). Deep tree learning for zero-shot face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4680-4689).," *Computer Vision and Pattern Recognition,* no. 12, pp. 4680-4689, 2019.

[9] A. J. and Y. , "Atoum, Y., Liu, Y., Jourabloo, A., & Liu, X. (2017, October). Face anti-spoofing using patch and depth-based CNNs. In 2017 IEEE international joint conference on biometrics (IJCB) (pp. 319-328). IEEE.," *biometrics,* no. 15, pp. 319-328, 2017.

[10] Y. J. and Z. Lei, "Yang, J., Lei, Z., & Li, S. Z. (2014). Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601.," *face anti-spoofing,* no. 16, p. 1408.5601, 2014.

[11] George, A. and S. Marcel, "George, A., & Marcel, S. (2021, August). On the effectiveness of vision transformers for zero-shot face anti-spoofing. In 2021 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-8). IEEE.," *Biometrics,* no. 13, pp. 1-8, 2021.

[12] L. and Chen-Hao, "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).," *Applications of Computer Vision,* no. 14, pp. 6098-6107, 2023.

[13] N. G. and S. Marcel, "Nikisins, O., George, A., & Marcel, S. (2019, June). Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing. In 2019 International Conference on Biometrics (ICB) (pp. 1-8). IEEE.," *Biometrics,* no. 17, pp. 1-8, 2019.

[14] Z. X. Li and Y. , "Yu, Zitong, Xiaobai Li, Jingang Shi, Zhaoqiang Xia, and Guoying Zhao. "Revisiting pixel-wise supervision for face anti-spoofing." IEEE Transactions on Biometrics, Behavior, and Identity Science 3, no. 3 (2021): 285-295.," *Biometrics, Behavior, and Identity Science,* vol. 3, no. 18, pp. 285-295, 2021.

[15] G. and S. Marcel, "George, A., & Marcel, S. (2019, June). Deep pixel-wise binary supervision for face presentation attack detection. In 2019 International Conference on Biometrics (ICB) (pp. 1-8). IEEE.," *Biometrics,* no. 19, pp. 1-8, 2019.

[16] C. and R. , "Cai, Rizhao, Zhi Li, Renjie Wan, Haoliang Li, Yongjian Hu, and Alex C. Kot. "Learning meta pattern for face anti-spoofing." IEEE Transactions on Information Forensics and Security 17 (2022): 1201-1213.," *Information Forensics and Security,* vol. 17, no. 20, pp. 1201-1213, 2022.

[17] Q. and Y. , "Qin, Y., Yu, Z., Yan, L., Wang, Z., Zhao, C., & Lei, Z. (2021). Meta-teacher for face anti-spoofing. IEEE transactions on pattern analysis and machine intelligence, 44(10), 6311-6326.," *analysis and machine intelligence,* vol. 44, no. 21, pp. 6311-6326, 2021.

[18] V. S. and G. Gadelha, "Verissimo, S., Gadelha, G., Batista, L., Janduy, J., & Falcão, F. (2023). Transfer learning for face anti-spoofing detection. IEEE Latin America Transactions, 21(4), 530-536.," *face anti-spoofing detection,* vol. 21, no. 22, pp. 530-536., 2023.

[19] U. and M. , "Muhammad, U., Laaksonen, J., Romaissa Beddiar, D., & Oussalah, M. (2024). Domain generalization via ensemble stacking for face presentation attack detection. International Journal of Computer Vision, 1-24.," *Computer Vision,* no. 23, pp. 1-24, 2024.

[20] W. W. and . P. Liu, "Wang, W., Liu, P., Zheng, H., Ying, R., & Wen, F. (2023). Domain generalization for face anti-spoofing via negative data augmentation. IEEE Transactions on Information Forensics and Security, 18, 2333-2344.," *Information Forensics and Security,* vol. 18, no. 24, pp. 2333-2344., 2023.

[21] O. and D. v. d. Haar, "Orfao, J., & van der Haar, D. (2023, February). Analysis of Generative Data Augmentation for Face Antispoofing. In International Conference on Pattern Recognition Applications and Methods (pp. 69-94). Cham: Springer Nature Switzerland.," *Pattern Recognition Applications and Methods,* no. 25, pp. 69-94, 2023.

[22] H. and A. B. J. Teoh, "Hua, Guang, Andrew Beng Jin Teoh, and Haijian Zhang. "Towards end-to-end synthetic speech detection." IEEE Signal Processing Letters 28 (2021): 1265-1269.," *Signal Processing Letters,* vol. 28, no. 26, pp. 1265-1269, 2021.

[23] G. . T. and Wanying, "Ge, Wanying, Jose Patino, Massimiliano Todisco, and Nicholas Evans. "Raw differentiable architecture search for speech deepfake and spoofing detection." arXiv preprint arXiv:2107.12212 (2021).," *speech deepfake and spoofing detection,* no. 27, p. 2107.12212, 2021.

[24] Tak and H. , "Tak, Hemlata, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. "End-to-end anti-spoofing with rawnet2." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6369," *Acoustics, Speech and Signal Processing (ICASSP),* no. 28, p. 6369, 2021.

[25] K. and S. , "Kinnunen, T., Sahidullah, M., Héctor Delgado, E. U. R. E. C. O. M., Massimiliano Todisco, E. U. R. E. C. O. M., Nicholas Evans, E. U. R. E. C. O. M., Yamagishi, J., & Lee, K. A. (2018). The 2nd Automatic Speaker Verification Spoofing and Countermeasures C," *Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) Database,* vol. 2, no. 29, 2018.

[26] S. K. and D. , "Kinnunen, Tomi, et al. "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection." Interspeech 2017. International Speech Communication Association, 2017.," *International Speech Communication Association,* no. 30, pp. 2-6, 2017.

[27] M. and T. , "Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., ... & Lee, K. A. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. arXiv preprint arXiv:1904.05441.," *Future horizons in spoofed and fake audio detection,* no. 31, p. 1904.05441., 2019.

[28] Y. and Y. , "Yamagishi, Junichi, et al. "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection." ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge. 2021.," *Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge,* no. 32, 2021.

[29] T. and M. , "Todisco, Massimiliano, et al. "ASVspoof 2019: Future horizons in spoofed and fake audio detection." arXiv preprint arXiv:1904.05441 (2019).," *Future horizons in spoofed and fake audio detection,* no. 33, p. 1904.0544, 2019.

[30] T. and D. , "Todisco, Massimiliano, Héctor Delgado, and Nicholas Evans. "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification." Computer Speech & Language 45 (2017): 516-535.," *Computer Speech & Language 45,* no. 34, pp. 516-535., 2017.

[31] L. and E. Li, "Luo, Anwei, et al. "A capsule network based approach for detection of audio spoofing attacks." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.," *Acoustics, Speech and Signal Processing,* no. 35, pp. 6359-6363, 2021.

[32] Zhang and F. Jiang,, "Zhang, Y., Jiang, F., & Duan, Z. (2021). One-class learning towards synthetic voice spoofing detection. IEEE Signal Processing Letters, 28, 937-941.," *IEEE Signal Processing Letters,* vol. 28, no. 36, pp. 937-941, 2021.

[33] A. and M. , "Alzantot, M., Wang, Z., & Srivastava, M. B. (2019). Deep residual neural networks for audio spoofing detection. arXiv preprint arXiv:1907.00501.," *Deep residual neural networks for audio spoofing detection,* no. 37, p. 1907.00501, 2019.

[34] F. and Z. Teng,, "Fu, Quchen, et al. "Fastaudio: A learnable audio front-end for spoof speech detection." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.," *Acoustics, Speech and Signal Processing,* no. 38, pp. 3693-3697, 2022.

[35] Y. and J. , "Yang, Jichen, et al. "Modified magnitude-phase spectrum information for spoofing detection." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 1065-1078.," *Audio, Speech, and Language Processing,* vol. 29, no. 39, pp. 1065-1078., 2021.

[36] C. N. and M. Dua, "Chakravarty, N., & Dua, M. (2024). An improved feature extraction for Hindi language audio impersonation attack detection. Multimedia Tools and Applications, 1-26.," *Multimedia Tools and Applications,* no. 40, pp. 1-26, 2024.

[37] M. M. and N. , "Memon, S., Manivannan, N., Noor, A., Balachadran, W., & Boulgouris, N. V. (2012). Fingerprint sensors: Liveness detection issue and hardware based solutions. Sensors & Transducers, 136(1), 35.," *Sensors & Transducers,* vol. 35, no. 41, p. 136(1), 2012.

[38] E. J. and K. Cao, "Engelsma, J. J., Cao, K., & Jain, A. K. (2017). RaspiReader: An open source fingerprint reader facilitating spoof detection. arXiv preprint arXiv:1708.07887.," *An open source fingerprint reader facilitating spoof detection,* no. 42, p. 1708.07887., 2017.

[39] W. and . X. , "Wang, Xuerui, et al. "Attacks and defenses in user authentication systems: A survey." Journal of Network and Computer Applications 188 (2021): 103080.," *Network and Computer Applications,* vol. 188, no. 43, p. 103080, 2021.

[40] Sharma and . R. Prakash, "Sharma, R. P., & Dey, S. (2019). Fingerprint liveness detection using local quality features. The Visual Computer, 35(10), 1393-1410.," *he Visual Computer,* vol. 35(10, no. 44, pp. 1393-1410., 2019.

[41] Rajaram,, Kanchana, and . B. A. NG, "Rajaram, Kanchana, Bhuvaneswari Amma NG, and Ashwin S. Guptha. "CLNet: a contactless fingerprint spoof detection using deep neural networks with a transfer learning approach." Multimedia Tools and Applications 83.9 (2024): 27703-27722.," *Multimedia Tools and Applications,* vol. 83.9, no. 45, pp. 27703-27722, 2024.

[42] Adami and , Banafsheh, "Adami, Banafsheh, et al. "A universal anti-spoofing approach for contactless fingerprint biometric systems." 2023 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2023.," *International Joint Conference on Biometrics,* no. 46, pp. 1-8, 2023.

[43] Chingovska, , Ivana, and A. Anjos,, "Chingovska, Ivana, André Anjos, and Sébastien Marcel. "On the effectiveness of local binary patterns in face anti-spoofing." 2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG). IEEE, 2012.," *international conference of biometrics special interest group (BIOSIG),* no. 47, pp. 1-7, 2012.

[44] Li, , L.,, Feng, and X, "Li, Lei, et al. "Face spoofing detection with local binary pattern network." Journal of visual communication and image representation 54 (2018): 182-192.," *Journal of visual communication and image representation,* vol. 54, no. 48, pp. 182-192., 2018.

[45] Ganjoo, , Romit, and . A. Purohit, "Ganjoo, Romit, and Anjali Purohit. "Anti-spoofing door lock using face recognition and blink detection." 2021 6th international conference on inventive computation technologies (ICICT). IEEE, 2021.," *nternational conference on inventive computation technologies,* no. 49, pp. 1090-1096, 2021.

[46] Tan, , X., , Li,, Y., and Liu, "Tan, Xiaoyang, et al. "Face liveness detection from a single image with sparse low rank bilinear discriminative model." Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings," *Conference on Computer Vision,* no. 51, pp. 504-517, 2010.

[47] Jingade, and R. Raghavendra, "Jingade, R. R., & Kunte, R. S. (2022). DOG-ADTCP: A new feature descriptor for protection of face identification system. Expert Systems with Applications, 201, 117207.," *Expert Systems with Applications,* vol. 201, no. 51, p. 117207., 2022.

[48] Gang, and Pan, "Pan, Gang, et al. "Eyeblink-based anti-spoofing in face recognition from a generic webcamera." 2007 IEEE 11th international conference on computer vision. IEEE, 2007.," *international conference on computer vision,* no. 52, pp. 1-8, 2007.

[49] Hao and Ge, "Ge, Hao, et al. "Face anti-spoofing by the enhancement of temporal motion." 2020 2nd International Conference on Advances in Computer Technology, Information Science and Communications (CTISC). IEEE, 2020.," *Advances in Computer Technology, Information Science and Communications,* no. 53, pp. 106-111, 2020.

[50] Alexey and D. , "Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).," *Transformers for image recognition,* no. 54, p. 11929, 2010.

[51] Williams, , Ronald J. and . D. Zipser, "Williams, Ronald J., and David Zipser. "A learning algorithm for continually running fully recurrent neural networks." Neural computation 1.2 (1989): 270-280.," *Neural computation,* no. 55, pp. 270-280., 1989.

[52] Graves and , Alex, "Graves, Alex, and Alex Graves. "Long short-term memory." Supervised sequence labelling with recurrent neural networks (2012): 37-45.," *Supervised sequence labelling with recurrent neural networks,* no. 56, pp. 37-45., 2012.

[53] Jung, and Jee-weon, "Jung, Jee-weon, et al. "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification." arXiv preprint arXiv:1904.08104 (2019).," *text-independent speaker verification,* no. 57, p. 1904.08104, 2019.

[54] Gulati, and Anmol, "Gulati, Anmol, et al. "Conformer: Convolution-augmented transformer for speech recognition." arXiv preprint arXiv:2005.08100 (2020).," *Convolution-augmented transformer for speech recognition,* no. 59, 2020.

[55] Steffen and Schneider, , "Schneider, Steffen, et al. "wav2vec: Unsupervised pre-training for speech recognition." arXiv preprint arXiv:1904.05862 (2019).," *Unsupervised pre-training for speech recognition,* no. 59, p. 1904.05862, 2019.

[56] V. D. O. Aaron, "Van Den Oord, Aaron, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 12 (2016).," *A generative model for raw audio,* vol. 12, no. 60, p. 1609.03499, 2016.

[57] Zeghidour and , Neil, "Zeghidour, Neil, et al. "LEAF: A learnable frontend for audio classification." arXiv preprint arXiv:2101.08596 (2021).," *A learnable frontend for audio classification,* no. 61, p. 2101.08596, 2021.

[58] Yu and Yirong, "Yu, Yirong, et al. "A review of fingerprint sensors: Mechanism, characteristics, and applications." Micromachines 14.6 (2023): 1253.," *Mechanism, characteristics, and applications. Micromachines,* vol. 14.6, no. 62, p. 1253, 2023.

[59] Zhang, and Yilong, "Zhang, Yilong, et al. "3D CNN-based fingerprint anti-spoofing through optical coherence tomography." Heliyon 9.9 (2023).," *fingerprint anti-spoofing through optical coherence tomography,* vol. 9.9, no. 63, 2023.

[60] Sadasivuni, and K. Kumar, "Sadasivuni, Kishor Kumar, et al. "Anti-spoofing device for biometric fingerprint scanners." 2017 IEEE International Conference on Mechatronics and Automation (ICMA). IEEE, 2017.," *IEEE International Conference on Mechatronics and Automation,* no. 64, pp. 683-687, 2017.

[61] . S. Schuckers and Bozhao,, "Tan, Bozhao, and Stephanie Schuckers. "Spoofing protection for fingerprint scanner by fusing ridge signal and valley noise." Pattern Recognition 43.8 (2010): 2845-2857.," *Pattern Recognition,* vol. 43.8, no. 66, pp. 2845-2857, 2010.

[62] Mohan,, Lekshmy S. and J. James., "Mohan, Lekshmy S., and Joby James. "Fingerprint spoofing detection using HOG and local binary pattern." International Journal of Advanced Research in Computer and Communication Engineering 6.4 (2017): 586-593.," *Advanced Research in Computer and Communication Engineering,* vol. 6.4, no. 65, pp. 586-593., 2017.

[63] Sajjad, and Muhammad, "Sajjad, Muhammad, et al. "CNN-based anti-spoofing two-tier multi-factor authentication system." Pattern Recognition Letters 126 (2019): 123-131.," *Pattern Recognition Letters,* vol. 126, no. 67, pp. 123-131, 2019.

[64] Zhang,, Y., and Yu,, "Zhang, Yilong, et al. "3D CNN-based fingerprint anti-spoofing through optical coherence tomography." Heliyon 9.9 (2023).," *fingerprint anti-spoofing through optical coherence tomography,* vol. 9.9, no. 68, 2023.

[65] Goyal, Tushar and Khandelwal, "Goyal, Tushar, et al. "Fingerprint Anti-spoofing Analysis: From Minutiae to Transformers." International Conference on Computer Vision and Image Processing. Cham: Springer Nature Switzerland, 2023.," *Conference on Computer Vision and Image Processing,* no. 69, pp. 394-409, 2023.

[66] M. U. and M. Oussalah, "Muhammad, U., & Oussalah, M. (2023, January). Self-supervised face presentation attack detection with dynamic grayscale snippets. In 2023 IEEE 17th International Conference

on Automatic Face and Gesture Recognition (FG) (pp. 1-6). IEEE.," *Automatic Face and Gesture Recognition,* no. 3, pp. 1-6, 2023.

[67] S. R, Y. PC and X. Lan, "Shao, R., Lan, X., & Yuen, P. C. (2020, April). Regularized fine-grained meta face anti-spoofing. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 11974-11981).," *artificial intelligence,* vol. 34, no. 1, pp. 11974-11981, 2020.

[68] Z. Z. X. Zhong and Y. Zhang, "Zhang, Z., Jiang, C., Zhong, X., Song, C., & Zhang, Y. (2021). Two-stream convolutional networks for multi-frame face anti-spoofing. arXiv preprint arXiv:2108.04032.," *arXiv preprint arXiv,* no. 1, p. 2108.04032, 2021.

[69] M. . U. and . M. Z. Hoque, "Muhammad, U., Hoque, M. Z., Oussalah, M., & Laaksonen, J. (2023, October). Deep ensemble learning with frame skipping for face anti-spoofing. In 2023 Twelfth International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-6). IEE," *Image Processing Theory, Tools and Applications (IPTA),* no. 8, pp. 1-6, 2023.