# Beyond Intra-Dataset Accuracy: Cross-Dataset Generalization of Scratch-Trained and Pretrained CNNs for Face Presentation Attack Detection

Nawal Alsufyani
Department of Computer Science, College of Computers and Information Technology,
Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia

***Abstract*—Face presentation attack detection (PAD) systems often achieve strong performance under intra-dataset evaluation but experience substantial degradation when exposed to unseen acquisition conditions. This study presents an empirical investigation of the discrepancy between intra-dataset performance and cross-dataset generalization by comparing a lightweight convolutional neural network trained from scratch (SimpleCNN) with an ImageNet-pretrained ResNet18. Both models were trained and evaluated using OULU-NPU, followed by strict zero-shot evaluation on CASIA-FASD without target-domain fine-tuning, adaptation, or threshold recalibration. On OULU-NPU, ResNet18 achieved an Average Classification Error Rate (ACER) of 10.97%, an Equal Error Rate (EER) of 10.22%, and an Area Under the ROC Curve (AUC) of 0.967, outperforming SimpleCNN, which achieved an ACER of 42.67%, an EER of 17.85%, and an AUC of 0.894. However, both models experienced substantial degradation on CASIA-FASD, with ACER values of 50.80% and 50.43% for SimpleCNN and ResNet18, respectively. The corresponding AUC values decreased to 0.403 and 0.305, accompanied by extremely high bona fide presentation classification error rates. These findings reveal substantial limitations in the transferability of the evaluated source-trained classifiers under the investigated cross-dataset protocol. Since the compared configurations differ in both architecture and initialization, the results do not isolate the independent contribution of ImageNet pretraining. Furthermore, the evaluation is limited to one transfer direction and frame-level predictions. The study highlights the importance of rigorous cross-dataset assessment, careful interpretation of class-specific errors, and broader experimental validation when evaluating the robustness of face PAD systems.**

***Keywords*—Face presentation attack detection; face anti-spoofing; cross-dataset evaluation; domain shift; transfer learning; deep learning; biometrics**

## I. Introduction

Face recognition has become one of the most widely adopted biometric technologies because of its convenience, non-contact operation, and increasingly high recognition accuracy. It is now extensively employed in applications such as mobile authentication, access control, surveillance, banking, and border management. Despite these advances, face recognition systems remain vulnerable to presentation attacks, in which an adversary attempts to impersonate an authorized user using artifacts such as printed photographs, replayed videos, or three-dimensional masks. Consequently, Face Presentation Attack Detection (PAD), also commonly referred to as face anti-spoofing, has become an essential component of secure face recognition systems [1]–[4]. Recent surveys emphasize that the increasing realism and diversity of presentation attacks continue to pose significant challenges to reliable biometric authentication [5], [6].

Early face PAD approaches primarily relied on handcrafted visual descriptors designed to capture texture, color, image-quality, and motion differences between bona fide and spoof presentations. Local Binary Patterns (LBP), micro-texture analysis, dynamic texture descriptors, color-texture representations, and image-distortion analysis demonstrated that presentation artifacts could provide useful discriminatory information [7]–[12]. However, these manually designed representations are sensitive to variations in illumination, camera characteristics, display devices, printing quality, and attack media. Their limited representation capacity consequently restricts their ability to generalize to complex and previously unseen attack conditions [5], [6].

The emergence of deep learning substantially changed the direction of face PAD research. Convolutional neural networks (CNNs) enable discriminative representations to be learned directly from facial images rather than relying exclusively on manually engineered features. Several studies have consequently explored CNN architectures and specialized convolutional operators for detecting spoof-specific texture and structural cues [13]–[16]. Transfer learning further facilitates this process by exploiting representations learned from large-scale image datasets. Residual networks such as ResNet are particularly attractive because residual connections enable the training of deeper models while mitigating optimization difficulties [17]. Nevertheless, strong within-dataset classification performance does not necessarily imply robust PAD performance under unseen acquisition conditions [5], [6].

More recent studies have therefore moved beyond conventional binary classification. Auxiliary-supervision approaches employ additional cues such as facial depth or physiological signals to encourage models to learn information related to the physical properties of bona fide faces [13]. Pixel-wise supervision and fine-grained local learning have similarly been investigated to reduce dependence on dataset-specific global appearance patterns [18], [19]. In parallel, transformer-based architectures have been introduced to face PAD because self-attention can model long-range spatial relationships and complementary global information [20], [21].

Despite these developments, cross-dataset generalization remains one of the most difficult problems in face PAD [5], [6]. A model trained on one benchmark may encounter substantially different illumination, sensors, compression charac-

teristics, subjects, backgrounds, and spoofing instruments when deployed on another dataset. Such distribution differences can cause a classifier to rely on dataset-specific correlations rather than intrinsic liveness information. Consequently, models reporting excellent intra-dataset performance may exhibit substantial degradation when transferred to an unseen domain. This problem has motivated domain-generalization approaches designed specifically to learn representations that remain discriminative across previously unseen domains [22]–[25]. For example, Single-Side Domain Generalization (SSDG) explicitly addresses the distribution discrepancy among domains and reports improved performance on unseen face PAD datasets.

The importance of realistic generalization assessment is also reflected in the design of major PAD benchmarks. OULU-NPU was specifically developed to evaluate face PAD under real-world variations involving different mobile devices, illumination conditions, backgrounds, and attack scenarios [26]. The database contains 4,950 bona fide and attack videos captured using six mobile devices under three sessions, making it an important benchmark for assessing robustness to realistic acquisition variability. CASIA-FASD provides another widely used benchmark containing diverse spoofing conditions and is frequently employed together with OULU-NPU and other datasets in cross-dataset face PAD evaluation [27].

Although increasingly sophisticated architectures have been proposed, an important practical question remains: to what extent can conventional transfer learning itself improve generalization compared with a lightweight CNN trained from scratch? Many existing studies combine specialized architectures, auxiliary supervision, domain adaptation, domain-generalization objectives, or multi-domain training [13], [22]–[25]. While these approaches achieve competitive performance, the contribution of the basic learning strategy can become difficult to isolate. A controlled comparison between training from scratch and transfer learning therefore provides useful insight into how pretrained representations affect both within-domain discrimination and cross-dataset robustness.

To investigate this issue, this study presents a controlled comparison between two representative learning strategies: a lightweight CNN trained from scratch (SimpleCNN) and an ImageNet-pretrained ResNet18 model [17], [28]. Both models are evaluated under the same preprocessing pipeline, training protocol, and evaluation criteria. The primary experiments are conducted on OULU-NPU [26], while CASIA-FASD [27] is used as an unseen target domain for strict zero-shot cross-dataset evaluation. No target-domain fine-tuning, adaptation, or threshold recalibration is performed. This design enables the effect of transfer learning on generalization to be examined independently of more complex domain-adaptation mechanisms.

The main contributions of this study are summarized as follows:

- An empirical comparison of two conventional face PAD configurations, SimpleCNN trained from scratch and ImageNet-pretrained ResNet18, under a common experimental and evaluation protocol.
- An investigation of the performance gap between intra-dataset evaluation on OULU-NPU and strict zero-shot cross-dataset evaluation on CASIA-FASD, without target-domain adaptation or threshold recalibration.
- A detailed examination of class-specific classification errors using APCER, BPCER, ACER, EER, confusion matrices, and ROC analysis to characterize the behavior of both classifiers under domain shift.

The study is positioned as an empirical analysis of cross-dataset robustness rather than a proposal of a new PAD architecture or a controlled isolation of pretraining effects.

The remainder of this paper is organized as follows. Section II reviews relevant developments in face presentation attack detection and cross-domain generalization. Section III describes the datasets, model configurations, and experimental procedures. Section IV presents the evaluation results and discusses the observed generalization limitations. Section V summarizes the main findings and identifies directions for further investigation.

## II. Related Work

### A. Evolution of Face Presentation Attack Detection

Face Presentation Attack Detection (PAD) has evolved considerably as presentation attacks have become more diverse and realistic. Early research primarily focused on distinguishing bona fide faces from simple print and replay attacks using handcrafted visual descriptors. Texture-based approaches were particularly influential, as spoofing media often introduce detectable differences in local facial appearance. Micro-texture analysis based on Local Binary Patterns (LBP) demonstrated that these differences could be exploited for spoof detection [7], [8]. Dynamic-texture analysis subsequently incorporated temporal information from video sequences, while color-texture approaches exploited chromatic distortions introduced by printing and display devices [9]–[11]. Image-quality and distortion characteristics were also investigated as indicators of recaptured facial content [12].

Although handcrafted methods provided effective baselines, their performance was strongly influenced by acquisition conditions, illumination, camera characteristics, and the quality of the presentation medium. These limitations encouraged the transition toward deep learning approaches capable of automatically learning discriminative representations from training data [5], [6]. CNN-based PAD methods subsequently demonstrated improved representation capability by learning hierarchical spatial features associated with spoofing artifacts. Auxiliary supervision further extended this direction by incorporating additional information such as facial depth and physiological signals rather than treating PAD solely as a binary classification problem [13].

The development of specialized architectures marked another important stage in face PAD research. Central Difference Convolutional Networks (CDCN), for example, were designed to capture fine-grained intensity-gradient information that is particularly relevant for distinguishing bona fide facial surfaces from presentation artifacts [14]. Other approaches investigated architecture design, local regions, pixel-level supervision, and fine-grained representations to improve the discrimination of subtle spoofing cues [15], [18], [19]. These developments significantly strengthened within-dataset PAD performance but

did not completely resolve the problem of generalization to unseen domains.

Consequently, research increasingly shifted toward domain generalization and cross-dataset PAD. Multi-domain representation learning and single-side domain generalization were introduced to reduce dependence on dataset-specific characteristics and encourage the extraction of more transferable representations [22]–[25]. This shift is particularly important because PAD datasets differ substantially in sensors, illumination, backgrounds, subjects, compression, and attack instruments. Therefore, strong performance on a single benchmark may not accurately represent a model's robustness under real deployment conditions.

More recently, transformer-based approaches have expanded the PAD landscape by exploiting self-attention to capture global and long-range dependencies. Vision Transformer-based PAD methods complement the predominantly local representations learned by CNNs and provide an alternative mechanism for modeling spatial relationships associated with presentation attacks [20], [21]. Nevertheless, recent surveys indicate that robust generalization to unseen attacks and acquisition domains remains an important unresolved challenge [5], [6].

Overall, the evolution of face PAD can be viewed as a progression from handcrafted texture and image-quality analysis, through CNN-based representation learning and auxiliary supervision, toward domain-generalized and transformer-based architectures. This progression has substantially improved PAD performance; however, the continuing gap between within-dataset and cross-dataset performance motivates controlled investigations of how different learning strategies affect generalization. This issue forms the central motivation for the comparative evaluation presented in this study.

### *B. CNN-Based Face Presentation Attack Detection*

The introduction of convolutional neural networks (CNNs) significantly advanced face presentation attack detection by replacing manually designed descriptors with automatically learned hierarchical representations. Unlike handcrafted approaches, CNN-based methods can learn discriminative spatial patterns directly from facial images, including texture irregularities, printing artifacts, display characteristics, and other subtle cues associated with spoof presentations [13]–[17]. Transfer learning further strengthened this paradigm by allowing models pretrained on large-scale image datasets to be adapted to face PAD with comparatively limited biometric training data [17].

Early CNN-based PAD studies primarily formulated the problem as binary classification between bona fide and attack presentations. Although these models generally improved intra-dataset performance, subsequent research demonstrated that conventional CNNs may still learn dataset-specific characteristics rather than intrinsic liveness cues. Yang et al. [29], for example, demonstrated that both network architecture and training-data diversity substantially influence face anti-spoofing performance and proposed a spatio-temporal framework designed to exploit subtle spoofing evidence from multiple facial regions.

To improve the representation of fine-grained spoofing characteristics, Yu et al. [14] introduced Central Difference Convolutional Networks (CDCN). Central Difference Convolution combines conventional intensity information with local gradient information, enabling the network to emphasize detailed structural patterns associated with presentation artifacts. The resulting CDCN and CDCN++ architectures achieved strong performance across several face PAD benchmarks and demonstrated that specialized convolutional operations can provide more discriminative representations than standard convolutions.

CNN-based PAD subsequently evolved toward more localized and fine-grained representation learning. Wang et al. [30] proposed PatchNet, which reformulated face anti-spoofing as a fine-grained patch-recognition problem. Instead of relying predominantly on global facial appearance, PatchNet learns local capture characteristics associated with acquisition devices and presentation materials. This strategy improved robustness to previously unseen spoofing conditions and demonstrated the importance of local information for cross-domain PAD.

Other CNN-based studies have incorporated auxiliary supervision, pixel-level prediction, depth estimation, temporal information, and multi-modal cues to encourage networks to learn more physically meaningful spoofing characteristics [13], [18], [19]. Multi-modal CNN architectures have additionally combined RGB, depth, and infrared information to capture complementary presentation-attack evidence [31]. Such approaches generally provide richer representations than conventional RGB binary classifiers, although they may require additional sensors or supervision during training.

Despite substantial improvements in intra-dataset performance, CNN-based face PAD remains vulnerable to domain shift. Variations in cameras, illumination, compression, backgrounds, subjects, and attack media can alter the visual statistics learned during training. Consequently, high-capacity CNNs may exploit dataset-specific shortcuts rather than representations that consistently characterize bona fide and spoof presentations across domains [5], [6]. This limitation is particularly relevant to the present study because it investigates whether ImageNet-pretrained ResNet18 provides better generalization than a lightweight CNN trained from scratch when both models are evaluated under identical conditions.

Overall, CNN-based approaches established the foundation of modern face PAD and demonstrated the effectiveness of automatically learned representations. However, the continuing discrepancy between within-dataset and cross-dataset performance motivated the development of auxiliary-supervision and domain-generalization strategies, which are discussed in the following sections.

### *C. Auxiliary Supervision and Physiological Cue-Based Methods*

Conventional CNN-based face PAD methods typically formulate the task as binary classification between bona fide and spoof presentations. Although this formulation can achieve strong intra-dataset performance, binary supervision may encourage networks to exploit dataset-specific visual correlations rather than intrinsic liveness characteristics. To address this limitation, auxiliary-supervision methods introduce additional

learning objectives that guide the network toward physically meaningful spoofing cues. A representative approach was proposed by Liu et al. [13], who argued that binary supervision alone may provide insufficient guidance for learning generalizable PAD representations. Their framework jointly estimates facial depth using pixel-wise supervision and remote photoplethysmography (rPPG) signals using temporal supervision. Because genuine faces exhibit three-dimensional facial structure and physiological pulse information that are generally absent or distorted in conventional print and replay attacks, these auxiliary signals encourage the network to learn more meaningful liveness characteristics. Their experiments demonstrated improvements in both intra-dataset and cross-dataset evaluation.

Depth-based supervision has also been explored by combining local appearance information with estimated facial geometry. Atoum et al. [32] proposed patch- and depth-based CNNs in which local image patches capture spoof-specific texture information while depth estimation provides complementary geometric supervision. This strategy reflects an important shift from learning only a final live/spoof label toward learning intermediate representations associated with the physical structure of genuine facial presentations.

Another direction is pixel-wise supervision. George and Marcel [33] proposed a CNN framework employing deep pixel-wise binary supervision rather than relying exclusively on a single image-level classification label. Their approach demonstrated competitive intra- and cross-dataset performance while using frame-level RGB information, showing that dense supervision can provide stronger learning guidance without necessarily requiring additional sensing modalities.

Beyond depth and pixel-level supervision, researchers have investigated material-related characteristics of presentation attacks. Yu et al. [34] formulated face anti-spoofing from the perspective of human material perception and proposed Bilateral Convolutional Networks to capture macro- and micro-level material information. Multi-level feature refinement and multi-head supervision were employed to encourage the extraction of robust representations associated with differences among genuine skin and spoofing materials such as paper, displays, and masks. The approach demonstrated competitive results under both intra- and cross-dataset evaluation.

Overall, auxiliary supervision represents an important development in face PAD because it constrains the learning process using information beyond a simple binary label. Depth, physiological, pixel-level, and material-related supervision can encourage models to capture cues that are more closely associated with the physical properties of bona fide and spoof presentations. Nevertheless, these approaches may require additional annotations, pseudo-label generation, temporal sequences, or specialized training objectives. Furthermore, improved auxiliary supervision does not completely eliminate performance degradation when acquisition conditions differ substantially between training and testing domains. These limitations have motivated increasing attention toward explicit domain-generalization strategies, which are discussed in the following subsection.

### *D. Domain Generalization and Cross-Dataset Face PAD*

Despite substantial improvements in face presentation attack detection, cross-dataset generalization remains a major challenge. Models trained and evaluated on the same dataset often achieve strong performance because the training and testing samples share similar acquisition characteristics. However, when a trained model is evaluated on an unseen dataset, differences in camera sensors, illumination, image resolution, compression, backgrounds, subject populations, and attack instruments can cause substantial performance degradation [5], [6]. Cross-dataset evaluation is therefore particularly important because it provides a more realistic assessment of whether a PAD model has learned transferable liveness cues rather than dataset-specific characteristics.

Domain generalization (DG) aims to address this problem by learning representations from one or more source domains that remain effective on previously unseen target domains. Unlike domain adaptation, DG assumes that target-domain data are unavailable during training. This setting is especially relevant to practical face PAD deployment, where the characteristics of future cameras, environments, and presentation attacks cannot be known in advance.

Shao et al. [22] introduced a multi-adversarial discriminative domain-generalization framework for face PAD. Their approach learns a generalized feature space shared across multiple source domains through adversarial learning and a dual-force triplet-mining constraint. Auxiliary facial-depth supervision was additionally incorporated to improve the discrimination between bona fide and spoof presentations. The study demonstrated that explicitly reducing source-domain discrepancies can improve performance on previously unseen PAD datasets.

Wang et al. [23] subsequently proposed multi-domain disentangled representation learning for cross-domain face PAD. Their framework separates PAD-relevant information from subject-discriminative characteristics and combines the resulting representations with multi-domain learning. By reducing the influence of subject-specific information and emphasizing domain-independent PAD features, the method achieved improved cross-dataset performance across several public benchmarks.

Jia et al. [24] proposed Single-Side Domain Generalization (SSDG), which challenged the assumption that bona fide and spoof distributions should be aligned in the same manner. SSDG encourages bona fide samples from different source domains to form a compact distribution while preserving the diversity of spoof samples across domains. An asymmetric triplet loss and single-side adversarial learning are used to construct a more generalizable decision boundary. This approach demonstrated that explicitly considering the asymmetric characteristics of bona fide and attack distributions can improve generalization to unseen domains.

Meta-learning has also been investigated as a mechanism for simulating domain shifts during training. Shao et al. [25] proposed regularized fine-grained meta-learning, in which the model is repeatedly exposed to simulated domain-shift scenarios. Domain knowledge is incorporated as a regularization mechanism, while fine-grained meta-learning enables the model to learn from multiple domain-shift configurations.

This strategy encourages optimization directions that are less dependent on a particular source domain and therefore more transferable to unseen attacks. More recently, Sun et al. [35] reconsidered the conventional objective of forcing all source domains into a completely domain-invariant representation. Their Separability and Alignment face anti-spoofing framework instead preserves domain separability while aligning the transition between bona fide and spoof samples across domains. The study showed that feature shifts may persist in unseen domains even when source-domain representations appear well aligned, highlighting the complexity of achieving reliable cross-domain PAD.

Collectively, these studies demonstrate that cross-dataset robustness generally requires explicit mechanisms for controlling domain-dependent information, such as adversarial learning, representation disentanglement, meta-learning, or domain-aware alignment [22]–[35]. However, these approaches typically employ multiple source datasets or specialized domain-generalization objectives. In contrast, the present study deliberately does not incorporate domain adaptation or domain-generalization training. Instead, SimpleCNN and ImageNet-pretrained ResNet18 are trained under the same conventional binary-supervision protocol on OULU-NPU and subsequently evaluated directly on CASIA-FASD without fine-tuning or threshold recalibration. This strict zero-shot setting is used to quantify how much generalization can be obtained from conventional CNN training and transfer learning alone.

Accordingly, the objective of the present work is not to outperform specialized domain-generalization methods. Rather, it establishes a controlled baseline that exposes the magnitude and nature of the generalization gap when explicit domain-generalization mechanisms are absent. The resulting comparison provides a useful reference for interpreting the additional improvements reported by more sophisticated cross-domain PAD approaches.

### *E. Vision Transformer and Hybrid Architectures for Face PAD*

The emergence of Vision Transformers (ViTs) has introduced a new direction in face presentation attack detection by providing an alternative to the predominantly local feature extraction of convolutional neural networks. Unlike CNNs, which progressively aggregate information through local receptive fields, transformers employ self-attention to model relationships among image regions over a broader spatial context [36]. This capability is potentially valuable for face PAD because presentation attacks may produce both localized artifacts and global inconsistencies related to facial structure, display boundaries, illumination, and recapture characteristics.

Dosovitskiy et al. [36] demonstrated that images can be represented as sequences of patches and processed using a standard transformer architecture. Although ViT was originally developed for general image classification rather than PAD, it established the foundation for subsequent transformer-based anti-spoofing approaches. Hierarchical transformer architectures such as Swin Transformer further introduced shifted-window attention to combine local modeling with progressively aggregated global representations [37]. These developments motivated researchers to investigate whether attention-based representations could improve the robustness of face PAD.

Ming et al. [21] proposed ViTransPAD, a video transformer architecture specifically designed for face presentation attack detection. The method combines convolutional processing with self-attention to exploit spatial and temporal information from facial video sequences. By incorporating convolutional inductive biases together with transformer-based global modeling, ViTransPAD illustrates the potential benefit of hybrid CNN–Transformer designs for capturing complementary spoofing characteristics.

Transformer architectures have also been investigated specifically from the perspective of cross-domain robustness. George and Marcel [20] examined Vision Transformers for zero-shot face anti-spoofing and showed that transformer-based representations can provide promising generalization behavior when evaluated on unseen presentation-attack conditions. Their findings suggest that the global relationships captured through self-attention may complement the local texture information traditionally emphasized by CNN-based PAD systems.

Liao et al. [38] subsequently introduced the Domain Invariant Vision Transformer (DiVT) for face anti-spoofing. Rather than relying solely on the representational capacity of a transformer backbone, DiVT explicitly targets domain generalization by learning representations intended to suppress domain-specific information while retaining spoof-discriminative characteristics. This work demonstrates that transformer architectures can be combined with domain-generalization objectives to address the distribution shifts commonly encountered in cross-dataset PAD evaluation. More recently, vision-language pretraining has extended the transformer-based PAD paradigm beyond conventional image classification. Srivatsan et al. [39] proposed FLIP, a cross-domain face anti-spoofing framework that exploits language guidance and large-scale pretrained visual-language representations. This direction suggests that representations learned from broader semantic supervision may reduce dependence on the relatively limited visual distributions available in individual PAD datasets.

Overall, transformer-based and hybrid CNN–Transformer methods provide promising mechanisms for combining local spoofing evidence with global contextual relationships. However, their improved representational capacity does not inherently eliminate the domain-generalization problem. Robust cross-dataset performance still depends on the learning objective, supervision strategy, training-domain diversity, and ability to suppress dataset-specific shortcuts [5], [6], [22]–[25], [35], [38], [39], [43]. Therefore, increasingly sophisticated architectures should not be interpreted as a substitute for rigorous cross-dataset evaluation.

In the context of the present study, transformer-based methods are reviewed to illustrate the progression of modern PAD architectures rather than as direct experimental competitors. The objective of this work remains to establish a controlled baseline using a lightweight CNN and an ImageNet-pretrained ResNet18. Quantifying their behavior under strict OULU-NPU-to-CASIA-FASD transfer provides a transparent reference point against which more sophisticated transformer and domain-generalization strategies can be interpreted.

### F. Research Gap and Positioning of this Study

Despite substantial progress in face presentation attack detection, a persistent gap remains between strong intra-dataset performance and reliable cross-dataset generalization. Existing studies have addressed this problem using auxiliary supervision, domain-generalization strategies, and transformer-based architectures [13], [22]–[25], [35], [38], [39], [42], [43]. However, these approaches often introduce specialized learning objectives, multiple training domains, or additional supervision.

This study therefore investigates the following focused question: How do the intra-dataset and cross-dataset behaviors of a scratch-trained lightweight CNN and an ImageNet-pretrained ResNet18 differ under a common face PAD experimental protocol? To address this question, both models are developed using OULU-NPU and subsequently evaluated on CASIA-FASD without target-domain fine-tuning, domain adaptation, or threshold recalibration. The objective is not to claim state-of-the-art performance, but to establish a transparent baseline illustrating the extent to which strong source-domain discrimination persists under an unseen domain shift.

## III. Methodology

### A. Experimental Framework

This study investigates the generalization capability of two representative face presentation attack detection configurations under a common experimental setting: (1) a lightweight convolutional neural network trained entirely from scratch, referred to as SimpleCNN, and (2) an ImageNet-pretrained ResNet18 fine-tuned for binary face PAD. Common preprocessing, training, and evaluation procedures are used to support a consistent comparison between the two configurations.

The experimental framework (as shown in Fig. 1) consists of two evaluation stages. First, both models are trained and evaluated using OULU-NPU Protocol 1 to measure intra-dataset performance. Second, the OULU-NPU-trained models are directly evaluated on CASIA-FASD without fine-tuning, domain adaptation, or threshold recalibration. This strict zero-shot cross-dataset setting is used to quantify the ability of each learning strategy to generalize to previously unseen acquisition conditions.

The complete experimental pipeline therefore consists of video frame extraction, face localization, preprocessing and augmentation, model training, intra-dataset evaluation on OULU-NPU, and zero-shot cross-dataset evaluation on CASIA-FASD. The experimental design maintains consistent preprocessing, optimization settings, and evaluation criteria across the two configurations. However, the architectures differ in model capacity and initialization. Accordingly, the observed performance differences reflect the combined influence of these factors and should not be interpreted as an isolated measurement of the effect of ImageNet pretraining.

### B. Datasets and Evaluation Protocol

Two publicly available face PAD benchmarks were employed: OULU-NPU and CASIA-FASD. OULU-NPU was used for model development and intra-dataset evaluation, whereas CASIA-FASD was reserved exclusively for zero-shot cross-dataset testing. Table I summarize characteristics of the datasets used in this study.

*1) OULU-NPU Dataset:* It is a widely used benchmark designed to evaluate face PAD under realistic mobile acquisition conditions [26]. It contains 4,950 video sequences from 55 subjects recorded using six smartphone cameras across three acquisition sessions. The dataset includes bona fide presentations as well as print and replay attacks captured under variations in illumination and background conditions.

This study follows Protocol 1, which evaluates generalization to previously unseen environmental conditions while maintaining consistent acquisition devices and attack instruments. The experimental subset used in this work contains 1,158 training samples, 1,118 development samples, and 2,076 testing samples. The frame-level OULU-NPU test set comprised 2,076 frames obtained from 150 video sequences, including 449 bona fide frames and 1,627 presentation-attack frames. No temporal aggregation was applied; therefore, each frame was evaluated independently. These values are those reported in the current experimental implementation.

*2) CASIA-FASD Dataset:* It was employed exclusively as the unseen target domain for cross-dataset evaluation [27]. The benchmark contains multiple presentation-attack configurations, including warped-photo, cut-photo, and replay attacks, and differs from OULU-NPU in acquisition conditions, camera characteristics, and attack configurations. CASIA-FASD was not used during training, validation, hyperparameter selection, fine-tuning, domain adaptation, or threshold recalibration. The cross-dataset evaluation set comprised 65,786 frames extracted from 240 video sequences, including 10,128 bona fide frames and 55,658 presentation-attack frames. Predictions were generated independently at the frame level without temporal aggregation. This strict separation between the source and target domains provides a zero-shot assessment of cross-dataset generalization.

The evaluation subsets contain unequal numbers of bona fide and presentation-attack frames. In the OULU-NPU test subset, 449 of the 2,076 evaluated frames are bona fide, whereas 1,627 are attacks. In the CASIA-FASD target subset, 10,128 of the 65,786 frames are bona fide, whereas 55,658 are attacks. These class distributions are explicitly reported because they can influence aggregate performance measures and the interpretation of classification errors. APCER and BPCER are therefore presented separately, alongside ACER, to avoid relying exclusively on overall classification performance.

### C. Face Detection and Preprocessing

A common preprocessing pipeline was applied to both datasets. For OULU-NPU, up to 15 frames were sampled at approximately uniform temporal intervals from each video to reduce temporal redundancy while retaining representative facial information. Face localization was performed during preprocessing using the Viola–Jones Haar Cascade detector [40]. For each sampled frame, the largest detected facial region was selected, and the corresponding bounding box was expanded by 30% to preserve contextual information around the facial boundary. The resulting region was cropped and resized to 224×224 pixels. When no face was detected,

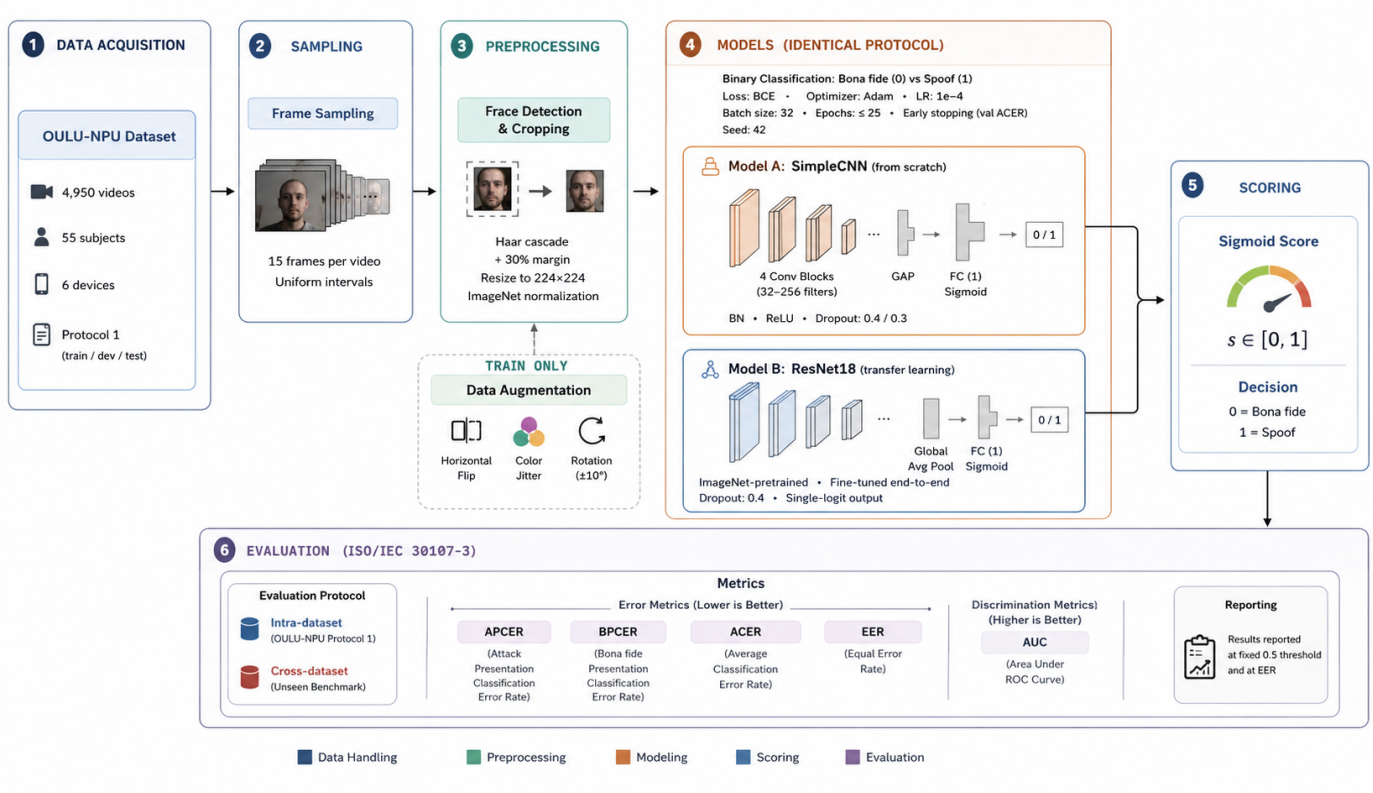


Fig. 1. The overall experimental framework.

TABLE I. CHARACTERISTICS OF THE DATASETS USED IN THIS STUDY

| Characteristic | OULU-NPU | CASIA-FASD |
|---|---|---|
| Role | Source domain | Unseen target domain |
| Video sequences evaluated | 150 | 240 |
| Evaluated frames | 2,076 | 65,786 |
| Bona fide frames | 449 | 10,128 |
| Attack frames | 1,627 | 55,658 |
| Evaluation level | Frame-level | Frame-level |
| Temporal aggregation | None | None |
| Training use | Yes | No |
| Validation use | Yes | No |
| Target-domain adaptation | — | None |

TABLE II. IMAGE PREPROCESSING PIPELINE

| Step | Description |
|---|---|
| Frame extraction | Up to 15 frames sampled at approximately uniform temporal intervals |
| Face detection | Viola–Jones Haar Cascade |
| Face selection | Largest detected face |
| Bounding box | Expanded by 30% |
| Image size | $224 \times 224$ pixels |
| Training augmentation | Horizontal flip, brightness, contrast, saturation, and rotation |
| Validation/Test | Resize only |
| Normalization | ImageNet mean and standard deviation |
| Detection failure | Full image retained as fallback |

the full image was retained as a fallback and resized to the same input dimensions. Data augmentation was applied only during training and consisted of random horizontal flipping with probability 0.5, brightness, contrast, and saturation adjustments of ±20%, and random rotation within ±10°. No augmentation was applied during validation or testing. Finally, all images were normalized using the ImageNet channel-wise mean (0.485,0.456,0.406) and standard deviation (0.229,0.224,0.225). The same image preparation and normalization settings were used for both architectures to maintain consistent input conditions (see Table II).

### D. Model Architectures

*1) SimpleCNN:* serves as the lightweight from-scratch baseline. The architecture consists of four convolutional blocks containing 32, 64, 128, and 256 filters, respectively. Each block consists of convolution, batch normalization, ReLU activation, and max pooling. The convolutional feature extractor is followed by adaptive global average pooling, dropout with a probability of 0.4, a fully connected layer mapping 256 features to 64 dimensions, ReLU activation, dropout with a probability of 0.3, and a final linear layer producing a single binary-classification logit. The network is randomly initialized and contains approximately 0.406 million trainable parameters.

*2) ResNet18:* The second model is ResNet18, representing the transfer-learning strategy. ResNet architectures employ residual connections that facilitate optimization of deep convolutional networks [17]. In this study, ResNet18 is initialized using ImageNet-pretrained weights. The original classification

layer is replaced with a dropout layer with probability 0.4, followed by a linear transformation from 512 features to a single output. All network layers are fine-tuned end-to-end on the OULU-NPU training data. For both architectures, the final logit is transformed using the sigmoid function:

$$p = \sigma(z) = \frac{1}{1 + e^{-z}}, \tag{1}$$

where $z$ denotes the model output and $p$ represents the predicted probability of the presentation-attack class. A probability approaching 0 represents a bona fide presentation, whereas a probability approaching 1 represents a spoof presentation.

### E. Training Procedure

Both models were trained using a common optimization protocol to provide consistent experimental conditions. Binary cross-entropy with logits (BCEWithLogitsLoss) was used as the classification objective. The Adam optimizer [41] was employed with an initial learning rate of $1 \times 10^{-4}$, weight decay of $1 \times 10^{-5}$, and a batch size of 32. A ReduceLROnPlateau scheduler reduced the learning rate by a factor of 0.5 when the validation loss did not improve for three consecutive epochs. Training was performed for a maximum of 25 epochs, with early stopping using a patience of six epochs based on validation ACER at a decision threshold of 0.5. The checkpoint achieving the best validation performance was retained for subsequent evaluation. A fixed random seed of 42 was used to improve reproducibility.

### F. Evaluation Metrics

Performance was evaluated using standard face PAD metrics which includes the Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER). APCER measures the proportion of presentation attacks incorrectly classified as bona fide:

$$\mathrm{APCER} = \frac{N_{\mathrm{attack}\rightarrow\mathrm{bona\ fide}}}{N_{\mathrm{attack}}} \tag{2}$$

BPCER measures the proportion of bona fide presentations incorrectly classified as attacks:

$$\mathrm{BPCER} = \frac{N_{\mathrm{bona\ fide}\rightarrow\mathrm{attack}}}{N_{\mathrm{bona\ fide}}} \tag{3}$$

ACER is calculated as the average of APCER and BPCER:

$$\mathrm{ACER} = \frac{\mathrm{APCER} + \mathrm{BPCER}}{2} \tag{4}$$

Equal Error Rate (EER) and Area Under the Receiver Operating Characteristic Curve (AUC) were additionally calculated to characterize error balance and threshold-independent discrimination. ROC curves and confusion matrices were generated to provide further insight into model behavior. The attack class was treated as the positive class for ROC/AUC computation.

### G. Cross-Dataset Generalization Protocol

The central cross-dataset experiment evaluates whether representations learned from OULU-NPU generalize to an unseen target domain. After model training and selection using OULU-NPU, the final SimpleCNN and ResNet18 models were evaluated directly on CASIA-FASD. No CASIA-FASD samples were used for training, validation, hyperparameter selection, or model selection, and no fine-tuning, domain adaptation, or threshold recalibration was performed before target-domain evaluation. Consequently, the OULU-NPU $\rightarrow$ CASIA-FASD experiment represents a strict zero-shot cross-dataset evaluation.

Predictions were generated independently for each CASIA-FASD frame, with no temporal aggregation across frames originating from the same video sequence. The same APCER, BPCER, ACER, EER, AUC, ROC, and confusion-matrix analyses used for the intra-dataset experiment were applied to the target-domain evaluation. This protocol enables direct assessment of the performance degradation associated with domain shift under a fixed source-trained model.

## IV. Results and Discussion

This section presents the experimental results obtained under both intra-dataset and cross-dataset evaluation settings. The analysis focuses not only on overall classification performance but also on the behavior of the two learning paradigms under domain shift. Performance is assessed using APCER, BPCER, ACER, EER, and AUC. Particular attention is given to the difference between performance on OULU-NPU Protocol 1 and zero-shot transfer to CASIA-FASD, since this difference provides a direct indication of the generalization capability of each model.

### A. Intra-Dataset Results (OULU-NPU Protocol)

The two models were first evaluated on 2,076 OULU-NPU frames obtained from 150 video sequences, comprising 449 bona fide and 1,627 presentation-attack frames. Predictions were generated independently at the frame level without temporal aggregation. Both architectures were evaluated under the same experimental protocol. Table III summarizes the resulting performance.

ResNet18 achieved substantially stronger overall intra-dataset performance than SimpleCNN. Its ACER decreased from 42.67% to 10.97%, representing a relative reduction of approximately 74.3%. ResNet18 also achieved a lower EER (10.22% versus 17.85%) and a higher AUC (0.967 versus 0.894), indicating stronger discrimination between bona fide and presentation-attack frames under the OULU-NPU evaluation conditions. However, APCER and BPCER reveal markedly different error distributions. SimpleCNN achieved a very low APCER of 0.49% but a BPCER of 84.86%, indicating that it correctly rejected most attacks while incorrectly rejecting a large proportion of bona fide frames. ResNet18 reduced BPCER to 0.67% but produced a higher APCER of 21.27%. Therefore, the improvement in overall ACER is accompanied by a substantial change in the balance between security-related and bona fide classification errors.

The substantial differences between APCER and BPCER indicate that the two models exhibit distinct error profiles at

TABLE III. FRAME-LEVEL INTRA-DATASET PERFORMANCE ON OULU-NPU PROTOCOL 1

| Model | APCER | BPCER | ACER | EER | AUC |
|---|---|---|---|---|---|
| SimpleCNN | 0.49 | 84.86 | 42.67 | 17.85 | 0.894 |
| ResNet18 | 21.27 | 0.67 | 10.97 | 10.22 | 0.967 |

the fixed decision threshold of 0.5. SimpleCNN favors attack detection at the expense of bona fide acceptance, whereas ResNet18 achieves considerably better bona fide recognition but accepts a larger proportion of attacks. These findings emphasize that ACER alone is insufficient to characterize the operational behavior of a PAD system. Moreover, because the models differ in architecture and initialization, their performance gap cannot be attributed exclusively to transfer learning.

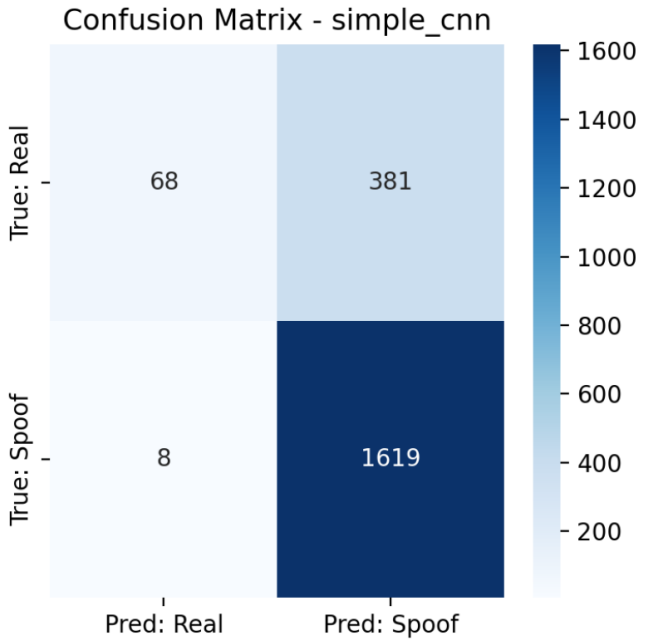


Fig. 2. Confusion matrix of the SimpleCNN model on OULU-NPU.

*1) Confusion Matrix Analysis:* Fig. 2 presents the SimpleCNN confusion matrix on OULU-NPU. Of the 1,627 presentation-attack frames, 1,619 were correctly classified as attacks and 8 were incorrectly accepted as bona fide. In contrast, only 68 of the 449 bona fide frames were correctly recognized, while 381 were incorrectly classified as attacks. This asymmetric prediction pattern explains the model's low APCER and high BPCER.

Fig. 3 shows the corresponding ResNet18 results. Of the 449 bona fide frames, 446 were correctly recognized and only 3 were incorrectly classified as attacks. Among the 1,627 attack frames, 1,281 were correctly detected, while 346 were incorrectly accepted as bona fide. ResNet18 therefore substantially improves bona fide recognition, although this improvement is accompanied by a higher attack acceptance rate. These results demonstrate the importance of interpreting ACER together with APCER and BPCER, since aggregate

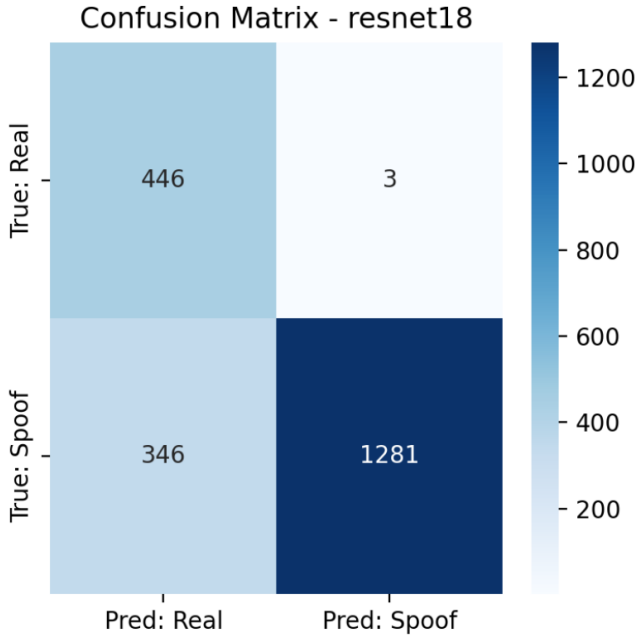


Fig. 3. Confusion matrix of the ResNet18 model on OULU-NPU.

performance alone may conceal substantially different error distributions.

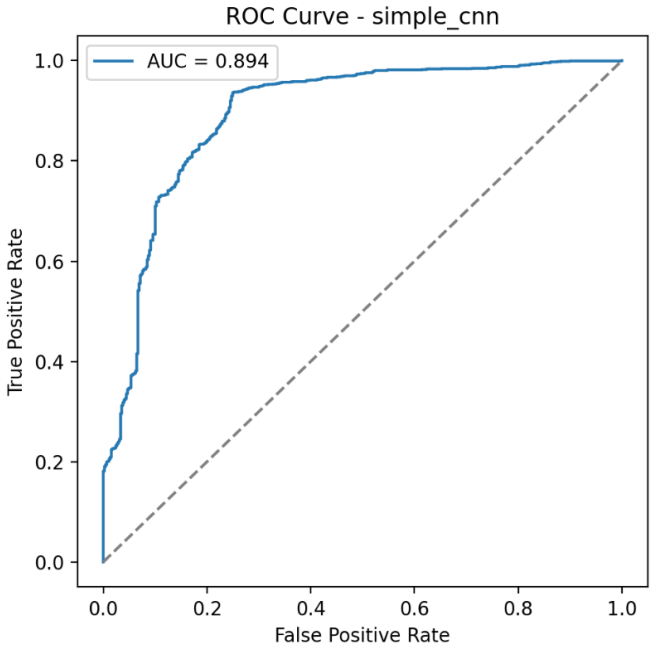


Fig. 4. ROC curve of the SimpleCNN model on OULU-NPU.

*2) ROC and AUC Analysis:* Receiver Operating Characteristic (ROC) curves for both models are shown in Fig. 4 and Fig. 5. The ROC analysis provides a threshold-independent assessment of the discrimination capability of the two architectures. ResNet18 achieved an AUC of 0.967, compared with 0.894 for SimpleCNN. The higher AUC indicates that

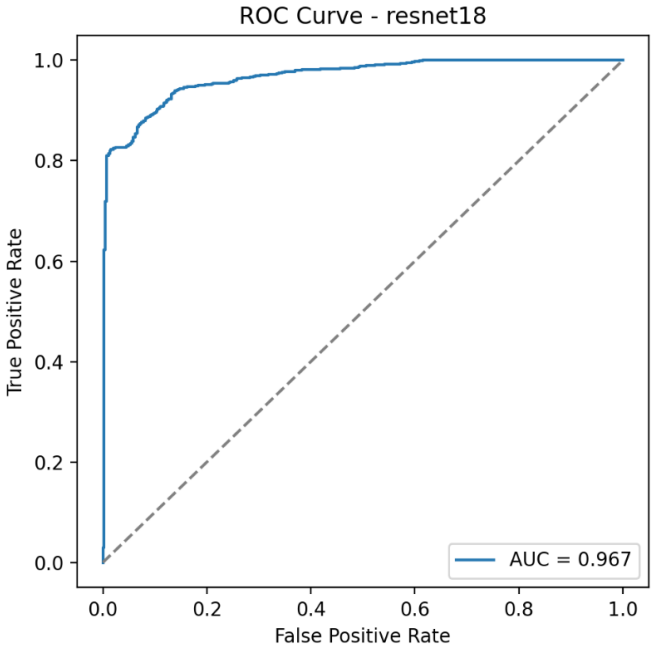


Fig. 5. ROC curve of the ResNet18 model on OULU-NPU.

ResNet18 provides stronger separation between bona fide and presentation attack score distributions across a range of operating thresholds.

The EER results support the same conclusion. ResNet18 achieved an EER of 10.22%, compared with 17.85% for SimpleCNN. Collectively, the ACER, EER, and AUC results show that the ImageNet-pretrained ResNet18 achieves substantially stronger discrimination than the scratch-trained SimpleCNN under the intra-dataset OULU-NPU setting.

### B. Cross-Dataset Results (CASIA-FASD)

To evaluate generalization to an unseen target domain, the OULU-NPU-trained models were applied directly to 65,786 CASIA-FASD frames extracted from 240 video sequences, comprising 10,128 bona fide and 55,658 presentation-attack frames. No CASIA-FASD data were used for training, validation, model selection, fine-tuning, domain adaptation, or threshold recalibration. Predictions were generated independently at the frame level without temporal aggregation. Table IV summarizes the resulting performance.

Cross-dataset evaluation reveals a substantial deterioration in both models. SimpleCNN produced an ACER of 50.80%, while ResNet18 achieved 50.43%, leaving only a 0.37 percentage-point difference between the two architectures. This contrasts sharply with their intra-dataset results, where ResNet18 substantially outperformed SimpleCNN.

The deterioration is particularly pronounced for ResNet18, whose ACER increased from 10.97% on OULU-NPU to 50.43% on CASIA-FASD, an absolute increase of 39.46 percentage points. SimpleCNN increased from 42.67% to 50.80%, an increase of 8.13 percentage points. Thus, the substantial source-domain advantage observed for ResNet18 does not persist under the OULU-NPU → CASIA-FASD domain shift.

TABLE IV. CROSS-DATASET PERFORMANCE: TRAINED OULU-NPU → TESTED CASIA-FASD

| Model | APCER | BPCER | ACER | EER | AUC |
|---|---|---|---|---|---|
| SimpleCNN | 1.60 | 100.00 | 50.80 | 57.97 | 0.403 |
| ResNet18 | 0.91 | 99.96 | 50.43 | 62.11 | 0.305 |

BPCER reached 100.00% for SimpleCNN and 99.96% for ResNet18, indicating that almost all bona fide CASIA-FASD frames were classified as attacks. In contrast, APCER remained low for both models. The resulting error distribution indicates a pronounced shift toward predicting the attack class in the unseen target domain.

*1) Cross-Dataset Confusion Matrix Analysis:* The confusion matrices provide further insight into the cross-dataset failure. For SimpleCNN, none of the 10,128 bona fide samples were correctly recognized; all were classified as presentation attacks as can be seen in Fig. 6. Among the attack samples, 54,768 were correctly detected, whereas 890 were incorrectly accepted as bona fide. This behavior indicates an extreme shift in the model's decision distribution toward the spoof class. ResNet18 demonstrated only a marginal improvement in bona fide recognition. Of the 10,128 bona fide samples, only 4 were correctly classified, while 10,124 were incorrectly rejected as attacks. Among the attack samples, 55,152 were correctly classified and 506 were incorrectly accepted as bona fide (see Fig. 7). Thus, although ResNet18 reduced the number of incorrectly accepted attacks relative to SimpleCNN, it did not recover meaningful bona fide recognition under the target-domain distribution.

The fact that two substantially different learning paradigms exhibit nearly the same failure pattern suggests that the observed degradation cannot be attributed solely to the capacity of a particular architecture. Instead, it indicates a systematic mismatch between the representations learned from OULU-NPU and the visual distribution encountered in CASIA-FASD

*2) ROC and AUC Analysis:* The cross-dataset ROC analysis provides additional evidence of poor transferability of the source-trained prediction scores (see Fig. 8 and Fig. 9). SimpleCNN achieved an AUC of 0.403, while ResNet18 achieved an AUC of 0.305 on CASIA-FASD. Both values fall below 0.5, indicating that the reported attack scores exhibit an unfavorable ranking of bona fide and attack samples under the target-domain evaluation. Such results require careful interpretation because they may reflect severe domain-dependent changes in score distributions, inconsistencies in score orientation, or implementation-related issues in label mapping and probability extraction. Therefore, the low AUC values should not be attributed solely to domain shift without independently verifying the positive-class definition, score direction, and evaluation implementation. The reported results are retained as observed, while this uncertainty is acknowledged as a limitation of the present analysis.

### C. Analysis of the Generalization Gap

The most significant finding of this study is the large discrepancy between intra-dataset and cross-dataset performance.

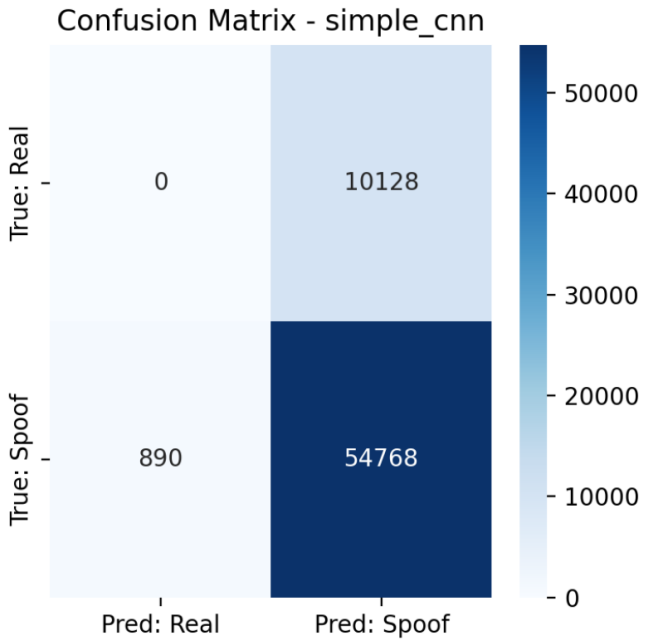


Fig. 6. Confusion matrix of the SimpleCNN model on CASIA-FASD.

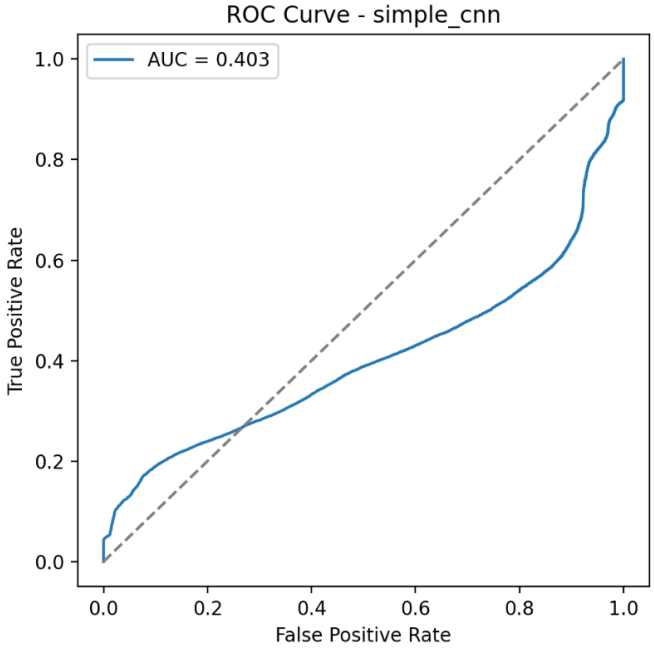


Fig. 8. ROC curve of the SimpleCNN model on CASIA-FASD.

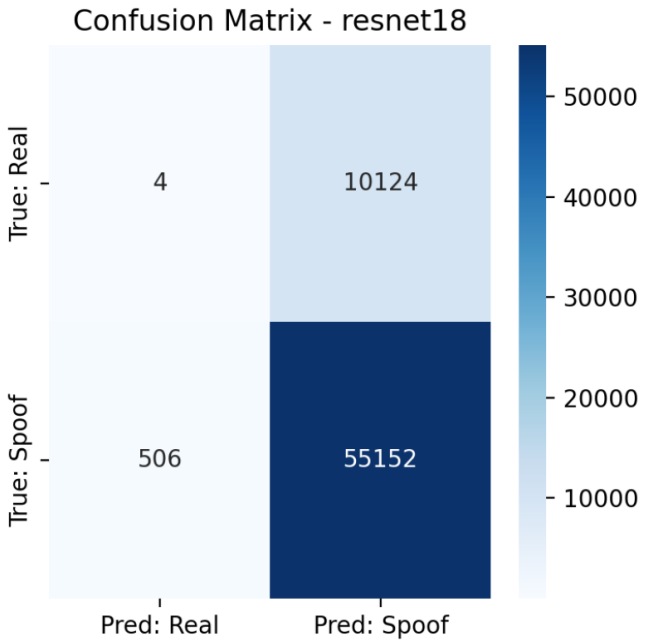


Fig. 7. Confusion matrix of the ResNet18 model on CASIA-FASD.

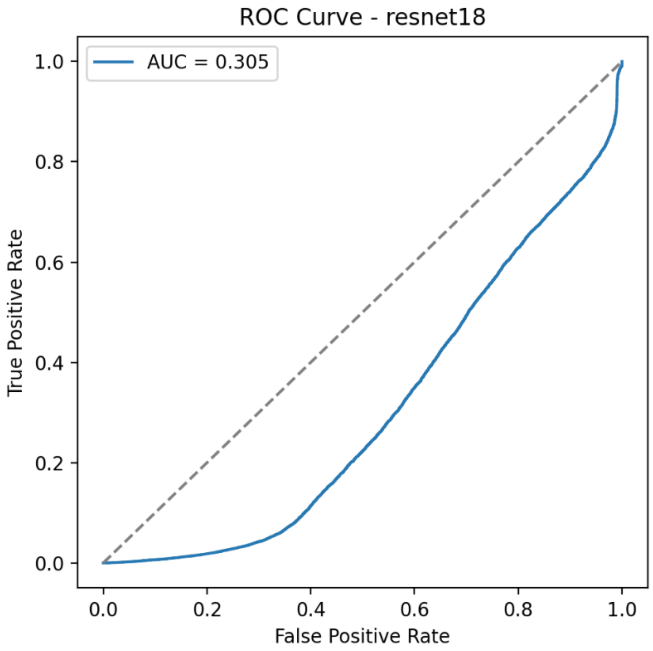


Fig. 9. ROC curve of the ResNet18 model on CASIA-FASD.

ResNet18 clearly outperformed SimpleCNN on OULU-NPU, demonstrating the strong intra-dataset performance of the pre-trained residual architecture under source-domain conditions. However, this advantage largely disappeared when both models were evaluated on CASIA-FASD.

This behavior suggests that the representations learned during OULU-NPU training contain substantial source-domain-specific information. Differences between OULU-NPU and CASIA-FASD include acquisition devices, environmental conditions, presentation media, image quality, illumination characteristics, and dataset construction procedures. Consequently, visual patterns that are predictive of bona fide and attack classes in OULU-NPU may not preserve the same relationship in CASIA-FASD.

The results are consistent with previous domain-generalization research showing that conventional face PAD

models can learn dataset-dependent representations that perform well on source-domain benchmarks but deteriorate considerably when exposed to unseen domains [22]–[25], [35]. Shao et al. explicitly demonstrated that conventional PAD representations can suffer from dataset bias and proposed multi-domain adversarial learning to improve transfer to unseen domains.

Similarly, methods such as CDCN attempt to learn fine-grained spoofing patterns rather than relying exclusively on conventional convolutional representations. Yu et al. reported both strong OULU-NPU Protocol 1 performance and improved cross-dataset behavior, supporting the importance of representations designed around more intrinsic spoof characteristics [14]. PatchNet likewise focuses on fine-grained local capture characteristics and reports evaluation across intra-dataset, cross-dataset, and domain-generalization settings.

The present results therefore should not be interpreted as evidence that transfer learning is ineffective. Rather, they show that ImageNet pretraining alone is insufficient to guarantee cross-dataset robustness for face PAD. ResNet18 achieves substantially stronger intra-dataset performance on OULU-NPU; however, the generic ImageNet-pretrained representation combined with binary PAD supervision does not prevent substantial performance degradation when acquisition characteristics change.

### D. Interpretation and Limitations

The experimental results highlight a clear distinction between source-domain discrimination and cross-dataset robustness. The ImageNet-pretrained ResNet18 substantially outperforms SimpleCNN within OULU-NPU, reducing ACER from 42.67% to 10.97%. However, this advantage almost disappears under zero-shot CASIA-FASD evaluation, where the two models obtain ACER values of 50.43% and 50.80%, respectively. These observations indicate that stronger intra-dataset performance does not necessarily imply greater robustness to unseen acquisition conditions.

Several limitations should be considered when interpreting these findings. First, evaluation is performed at the frame level without temporal aggregation. Frames originating from the same video sequence may therefore contain correlated information, and the reported metrics should not be interpreted as video-level PAD performance. Second, the evaluated frame distributions are imbalanced, and no class-weighted loss or balanced sampling strategy was employed. Although both models were evaluated under a common protocol, class imbalance may influence their error distributions.

Third, SimpleCNN and ResNet18 differ in both network architecture and initialization. Consequently, the observed performance difference cannot be attributed exclusively to ImageNet pretraining. A controlled comparison between randomly initialized and ImageNet-pretrained instances of the same ResNet18 architecture would more precisely isolate the contribution of pretraining. Finally, the cross-dataset analysis considers a single source-to-target direction, OULU-NPU $\rightarrow$ CASIA-FASD. Evaluation across additional datasets and bidirectional transfer settings would provide a broader assessment of generalization.

TABLE V. QUALITATIVE COMPARISON OF REPRESENTATIVE FACE PRESENTATION ATTACK DETECTION APPROACHES

| Method | Main Strategy | Auxiliary | Cross-Domain Focus |
|---|---|---|---|
| SimpleCNN (Ours) | CNN trained from scratch | None | Zero-shot OULU→CASIA |
| ResNet18 (Ours) | ImageNet transfer learning | None | Zero-shot OULU→CASIA |
| Liu et al. [13] | CNN | Depth + rPPG supervision | Limited |
| Shao et al. [22] | CNN | Multi-adversarial domain generalization | Yes |
| Jia et al. [24] | CNN | Single-side domain generalization | Yes |
| CDCN [14] | Central-difference CNN | Depth/fine-grained cues | Yes |
| PatchNet [30] | Patch-based CNN | Fine-grained local representation | Yes |
| DiVT [38] | Vision Transformer | Domain-invariant learning | Yes |

Table V positions the present study relative to representative face PAD approaches. Direct numerical comparison is not appropriate because the considered methods employ different source datasets, target datasets, evaluation protocols, supervision strategies, and domain-generalization settings. Therefore, the comparison is intended to clarify methodological positioning rather than claim performance superiority.

Compared with specialized state-of-the-art methods, the models investigated in this study deliberately exclude auxiliary supervision and explicit domain-generalization mechanisms. Their severe cross-dataset degradation therefore provides a useful controlled baseline demonstrating what occurs when conventional binary supervision and standard transfer learning are relied upon without additional mechanisms for learning domain-invariant spoofing cues.

The comparison supports a broader conclusion emerging from recent face PAD research: improving network capacity or source-domain accuracy alone is insufficient. Methods specifically designed to capture intrinsic spoofing characteristics or reduce domain dependence are more appropriate directions for achieving robust performance under unseen acquisition conditions. Domain-generalization work such as the multi-adversarial framework of Shao et al. was explicitly designed around this problem and demonstrated the value of learning a feature space shared across source domains.

## V. CONCLUSION

This study investigated the discrepancy between intra-dataset performance and cross-dataset generalization in face presentation attack detection through an empirical comparison of a lightweight CNN trained from scratch and an ImageNet-pretrained ResNet18. Both configurations were developed using OULU-NPU and evaluated under a strict zero-shot transfer protocol on CASIA-FASD without target-domain fine-tuning, adaptation, or threshold recalibration.

The results showed that ResNet18 substantially outperformed SimpleCNN on OULU-NPU, achieving an ACER

of 10.97% compared with 42.67%. However, both models experienced severe performance degradation on CASIA-FASD, with ACER values of 50.80% and 50.43% for SimpleCNN and ResNet18, respectively. The cross-dataset results were characterized by extremely high BPCER values and AUC scores below 0.5, indicating that the source-trained classifiers failed to maintain reliable target-domain discrimination under the evaluated protocol.

These findings demonstrate that strong intra-dataset performance does not necessarily translate into robust cross-dataset behavior. However, because the evaluated configurations differ in architecture and initialization, the present experiments do not isolate the effect of ImageNet pretraining. Furthermore, the conclusions are limited by the use of a single transfer direction, frame-level evaluation, and one random seed.

Future work should investigate controlled scratch-trained and pretrained versions of the same architecture, additional source–target dataset combinations, video-level score aggregation, class-balancing strategies, and repeated experiments with uncertainty estimates. Further examination of preprocessing differences, prediction-score distributions, and operating thresholds is also necessary to better characterize the causes of the observed cross-dataset degradation. Overall, this study reinforces the importance of rigorous and transparent cross-dataset evaluation when assessing the practical robustness of face PAD systems.

## DECLARATIONS

Generative AI tools were used for language refinement and editorial assistance during manuscript preparation. The author critically reviewed and edited all AI-assisted content and takes full responsibility for the accuracy, originality, and integrity of the manuscript.

## REFERENCES

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, 2004.

[2] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Comput. Surv.*, vol. 50, no. 1, Art. no. 8, pp. 1–37, 2017.

[3] Z. Ming, M. Visani, M. M. Luqman, and J.-C. Burie, "A survey on anti-spoofing methods for facial recognition with RGB cameras of generic consumer devices," *J. Imaging*, vol. 6, no. 12, Art. no. 139, 2020.

[4] D. Sharma and A. Selwal, "A survey on face presentation attack detection mechanisms: Hitherto and future perspectives," *Multimedia Syst.*, vol. 29, no. 3, pp. 1527–1577, 2023.

[5] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5609–5631, 2023.

[6] M. Sheikhfathollahi, S. Parkinson, R. Hill, and S. Khan, "A survey of deep learning for face presentation attack detection," *Neurocomputing*, vol. 652, Art. no. 131136, 2025.

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

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

[9] T. de Freitas Pereira, J. Komulainen, A. Anjos, J. M. De Martino, A. Hadid, M. Pietikäinen, and S. Marcel, "Face liveness detection using dynamic texture," *EURASIP J. Image Video Process.*, vol. 2014, Art. no. 2, 2014.

[10] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2015, pp. 2636–2640.

[11] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 8, pp. 1818–1830, 2016.

[12] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 746–761, 2015.

[13] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 389–398.

[14] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 5295–5305.

[15] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4680–4689.

[16] Y. Qin, C. Zhao, X. Zhu, Z. Wang, Z. Yu, T. Fu, F. Zhou, J. Shi, and Z. Lei, "Learning meta model for zero- and few-shot face anti-spoofing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 11916–11923, 2020.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[18] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, "Revisiting pixel-wise supervision for face anti-spoofing," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 3, pp. 285–295, 2021.

[19] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 10, pp. 2268–2283, 2016.

[20] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2021.

[21] Z. Ming, Z. Yu, M. Al-Ghadi, M. Visani, M. M. Luqman, and J.-C. Burie, "ViTransPAD: Video transformer using convolution and self-attention for face presentation attack detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2022, pp. 4248–4252.

[22] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 10023–10031.

[23] G. Wang, H. Han, S. Shan, and X. Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 6678–6687.

[24] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Single-side domain generalization for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8484–8493.

[25] R. Shao, X. Lan, and P. C. Yuen, "Regularized fine-grained meta face anti-spoofing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 11974–11981, 2020.

[26] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2017, pp. 612–618.

[27] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, 2012, pp. 26–31.

[28] J. Deng, W. Dong, R. Socher, J.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[29] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu, "Face anti-spoofing: Model matters, so does data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3507–3516.

[30] C.-Y. Wang, Y.-D. Lu, S.-T. Yang, and S.-H. Lai, "PatchNet: A simple face anti-spoofing framework via fine-grained patch recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 20281–20290.

[31] Z. Yu, Y. Qin, X. Li, Z. Wang, C. Zhao, Z. Lei, and G. Zhao, "Multi-modal face anti-spoofing based on central difference networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020.

[32] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2017, pp. 319–328.

[33] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *Proc. Int. Conf. Biometrics (ICB)*, 2019, pp. 1–8.

[34] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, "Face anti-spoofing with human material perception," in *Computer Vision–ECCV 2020*, Lecture Notes in Computer Science, vol. 12352, pp. 557–575, 2020.

[35] Y. Sun, Y. Liu, X. Liu, Y. Li, and W.-S. Chu, "Rethinking domain generalization for face anti-spoofing: Separability and alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 24563–24574.

[36] A. Dosovitskiy et al., "An image is worth $16 \times 16$ words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[37] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[38] C.-H. Liao, W.-C. Chen, H.-T. Liu, Y.-R. Yeh, M.-C. Hu, and C.-S. Chen, "Domain invariant vision transformer learning for face anti-spoofing," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2023, pp. 6098–6107.

[39] K. Srivatsan, M. Naseer, and K. Nandakumar, "FLIP: Cross-domain face anti-spoofing with language guidance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023.

[40] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, 2001, pp. I-511–I-518.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015.

[42] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, and Z. Lei, "Deep spatial gradient and temporal depth learning for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 5041–5050, doi: 10.1109/CVPR42600.2020.00509.

[43] Z. Wang, Z. Wang, Z. Yu, W. Deng, J. Li, T. Gao, and Z. Wang, "Domain generalization via shuffled style assembly for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 4113–4123, doi: 10.1109/CVPR52688.2022.00409.