I.J. Wireless and Microwave Technologies, 2026, 5, 398-427
Published Online on October 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijwmt.2026.05.24

Cross-Modal Consistency Learning for Robust Face Anti-Spoofing Using RGB, RGB-Derived Depth and NIR Representations

Mudunuru Suneel*
Department of Electronics and Communication Engineering, Aditya University, Surampalem, Andhra Pradesh, India
E-mail: suneel007@gmail.com
ORCID iD: https://orcid.org/0009-0002-1200-9830
*Corresponding Author

Kaja Krishna Mohan
Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur District, India
E-mail: krishnamohan506@gmail.com
ORCID iD: https://orcid.org/0009-0007-6454-8090

Banothu Yedukondala Venkata Naga Raja Swamy
Department of Electronics and Communication Engineering, Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India
E-mail: swamyyvnr@gmail.com
ORCID iD: https://orcid.org/0000-0003-0195-6592

Seva Sreedhar Babu
Department of Electronics and Communication Engineering, Sree Vahini Institute of Science and Technology, Thiruvuru, Andhra Pradesh, India
E-mail: seva.sreedhar@gmail.com
ORCID iD: https://orcid.org/0000-0002-5310-4693

Venkata Raghavendra Miriampally
Department of Electronics and Communication Engineering, Sree Dattha Group of Institutions, Sheriguda, Hyderabad, Telangana, India
E-mail: miriampally@gmail.com
ORCID iD: https://orcid.org/0009-0007-9975-3879

P. Rama Koteswara Rao
Department of Computer Science and Engineering, Sree Dattha Institute of Engineering & Science, Sheriguda, Hyderabad, Telangana, India
E-mail: prkr74@gmail.com
ORCID iD: https://orcid.org/0000-0002-5391-4805

Kama Ramudu
Department of Electronics and Communication Engineering, Aditya University, Surampalem, Andhra Pradesh, India
E-mail: ramudukama@gmail.com
ORCID iD: https://orcid.org/0000-0002-8585-9396

Received: July 22, 2026; Revised: August 17, 2026; Accepted: September 8, 2026; Published: October 8, 2026

Abstract: Face recognition systems are increasingly deployed in security-critical applications, but remain vulnerable to presentation attacks such as printed photographs, replay videos, and three-dimensional masks. Although recent face anti-spoofing methods exploit complementary information from RGB, depth, and near-infrared (NIR) representations, existing approaches primarily focus on feature fusion and do not explicitly model the intrinsic consistency relationships among

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

these representations. This limitation can reduce robustness against sophisticated spoofing attacks and cross-dataset variations.

This paper proposes a novel Cross-Modal Consistency Learning (CMCL) framework for robust face anti-spoofing using the original RGB input together with RGB-derived depth and NIR representations. The depth and NIR representations are constructed from RGB inputs during the modality preprocessing stage and subsequently processed together with RGB using independent Swin Transformer encoders. A cross-modal attention fusion module adaptively integrates complementary appearance, geometric, and spectral information, while a consistency learning module explicitly encourages feature coherence for genuine samples and emphasizes cross-modal discrepancies associated with spoof attacks. The consistency module is used only during training and removed during inference, avoiding additional deployment overhead.

Extensive experiments are conducted on CASIA-SURF, WMCA, CelebA-Spoof, and MSU-MFSD using intra-dataset, cross-dataset, ablation, feature-space, modality-consistency, and statistical analyses. The proposed CMCL achieves an ACER of 0.70% and accuracy of 99.3% on CASIA-SURF, while achieving an ACER of 2.55% and accuracy of 97.9% on WMCA. The results demonstrate improved spoof detection performance, representation discrimination, and cross-dataset generalization compared with the evaluated state-of-the-art methods. The proposed framework provides a consistency-driven and computationally practical approach for robust face anti-spoofing.

Index Terms: Face Anti-Spoofing, Cross-Modal Consistency Learning, Swin Transformer, Cross-Modal Attention, Multi-Modal Feature Fusion, Biometric Security

1. Introduction

Face recognition is one of the most widely adopted biometric technologies because of its ease of use, non-contact operation, and high recognition accuracy. It has wide used in security-critical applications such as smartphone authentication, financial services, border control, healthcare, surveillance and access management. Despite the great advancement of facial recognition systems based on deep learning, they are still quite vulnerable to presentation attacks (PAs) where an attacker tries to fool the system using printed images, replayed movies or three-dimensional masks. These assaults can be undertaken at extremely little cost and can potentially circumvent high-performance recognition systems, which makes Face Anti-Spoofing (FAS), sometimes called Presentation Attack Detection (PAD), a vital component of modern biometric security systems [1, 2].

Early face anti-spoofing approaches mostly utilize handmade feature descriptors, such as Local Binary Pattern (LBP), colour texture analysis, and image distortion features, to discriminate between real faces and spoof presentations [3, 28, 35]. While these methods demonstrated reasonable performance under controlled imaging conditions, their effectiveness was significantly degraded under variations in illumination, camera characteristics and unseen attack types due to the limited representation capability of handcrafted descriptors. Deep convolutional neural networks (CNNs) have dramatically improved the performance of spoof detection by learning discriminative representations of faces from raw RGB photos in an automatic way. Several typical methods have made significant gains in detection accuracy, including auxiliary supervision networks [6, 15], Central Difference Convolutional Networks (CDCN) [7, 11], EfficientNet [16] and MobileNetV3 [17]. However, RGB-based systems still struggle against sophisticated presentation attacks, as the appearance information cannot describe the three-dimensional structure and substance of real human faces effectively.

It is important to distinguish between consistency learning with physically acquired modalities and consistency learning with RGB-derived representations. In conventional multi-modal face anti-spoofing systems, RGB, depth, and NIR may be obtained from synchronized sensors, providing physically coupled observations of the same facial surface. In contrast, the present framework uses the original RGB image together with depth and NIR representations derived from RGB during preprocessing. Therefore, the proposed CMCL does not rely on independent sensor synchronization or assume that synthesized representations reproduce all physical properties of real depth or NIR measurements. Instead, it exploits the complementary structural and spectral representations learned from the RGB input and explicitly models their feature-level relationships. This formulation allows cross-modal consistency to be used as a training signal while keeping the additional modality construction and consistency learning separate from the inference-time classification pipeline.

To overcome these constraints, recent research has been focusing more and more on multi-modal face anti-spoofing, where RGB photos are combined with depth and near-infrared (NIR) information [8, 12–14, 29]. RGB images capture fine face appearance and texture, depth images collect 3D geometry information, and NIR images record spectral reflectance properties that are far less sensitive to illumination changes. The existence of large-scale benchmark datasets such as CASIA-SURF [12], WMCA [14], CelebA-Spoof [39], and MSU-MFSD [41] has significantly promoted research on multi-modal face anti-spoofing, providing a variety of presentation attack scenarios under realistic environmental conditions.

Despite the success of multi-modal learning, many existing approaches primarily rely on feature concatenation, weighted fusion, or attention-based aggregation to combine modality-specific representations [8, 30]. Recent

Transformer-based FAS methods have improved the ability of models to capture long-range spatial dependencies and global facial representations; however, their primary emphasis remains on learning discriminative representations within individual modalities or combining modality-specific features through attention-based fusion. Similarly, multi-modal FAS approaches have demonstrated that RGB, depth, and NIR provide complementary cues, but conventional fusion strategies do not explicitly determine whether the resulting modality representations remain mutually coherent. This distinction is important because cross-modal discrepancy can provide discriminative evidence of presentation attacks. In particular, a spoof may preserve plausible RGB appearance while producing structurally or spectrally inconsistent depth and NIR representations. The proposed CMCL framework addresses this limitation by treating cross-modal consistency as an explicit learning objective rather than only as an implicit consequence of feature fusion.

Another key problem of practical face anti-spoofing is cross-dataset generalisation. Models trained on a single benchmark generally show considerable performance deterioration when tested on unseen datasets due to differences in imaging devices, illumination conditions, spoofing materials, environmental settings, and subject variety [9, 10, 26, 31]. There exist some domain generalisation methods for RGB-based face anti-spoofing, such as adversarial learning [9], style-agnostic representation learning [10], and MixStyle augmentation [26]. But there are few works focusing on cross-modal consistency for enhancing robustness across domains. Therefore, the development of multi-modal representations that are robust under different acquisition situations is still an important open research challenge.

To overcome these constraints, this study provides a novel Cross-Modal Consistency Learning (CMCL) framework for robust multi-modal face anti-spoofing employing RGB, depth and near-infrared modalities. The proposed framework is different from the typical fusion-based techniques in that it explicitly models the semantic coherence among modality-specific feature representations. Independent Swin Transformer encoders are used to learn discriminative aspects of appearance, geometry and spectrum from each modality. A cross-modal attention fusion module assigns a modality-specific adaptive significance weight dynamically based on its discriminability reliability. Moreover, we use a specific Cross-Modal Consistency Learning (CMCL) module to enforce the feature consistency among authentic facial samples while expanding the feature differences caused by presentation attacks. This consistency regularisation is employed solely at training time as an auxiliary supervisory signal and is dropped at inference time, hence boosting the spoof detection performance without adding any additional computing cost in deployment.

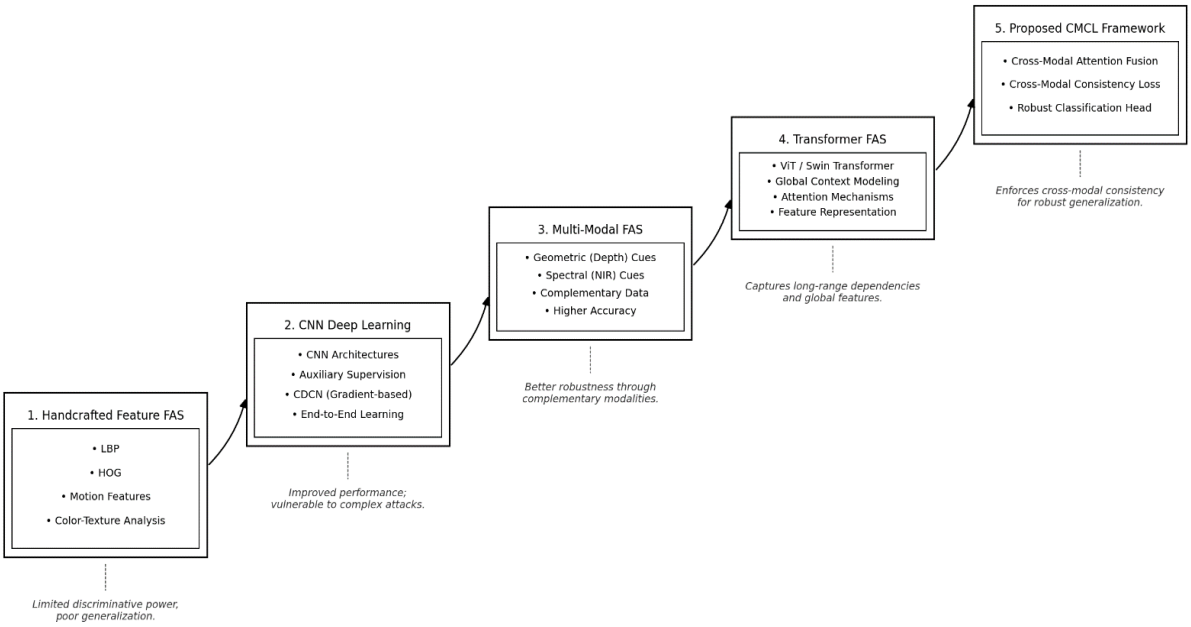


Fig. 1. Evoluation of Face Anti-Spoofing (FAS) methods.

Figure 1 presents the history of face anti-spoofing techniques from handcrafted feature-based approaches to CNN-based methods, multi-modal learning, transformer structures and the proposed Cross-Modal Consistency Learning framework.

Extensive experiments are undertaken on four widely-used benchmark datasets: CASIA-SURF [12], WMCA [14], CelebA-Spoof [39] and MSU-MFSD [41] under intra-dataset, cross-dataset and ablation assessment techniques. Experimental results show that the proposed framework consistently outperforms state-of-the-art approaches in APCER, BPCER, ACER, accuracy and cross-dataset robustness. Further analyses such as feature-space visualization, modality consistency evaluation, confusion matrix analysis, and statistical significance testing further confirm the effectiveness, robustness, and reliability of the proposed consistency-driven learning technique.

In summary, the main contributions of this paper are:

1) We propose a Cross-Modal Consistency Learning (CMCL) strategy that explicitly models relationships among RGB, RGB-derived depth, and NIR representations. The proposed objective encourages representation coherence for genuine samples while promoting discriminative modality discrepancies for spoof samples.
2) We develop a cross-modal attention fusion mechanism based on modality-specific Swin Transformer encoders to adaptively integrate complementary appearance, structural, and spectral representations before classification.
3) We introduce a training-only consistency regularization strategy, allowing the consistency module to provide an auxiliary supervisory signal during optimization while being removed during inference. This separates representation regularization from deployment-time prediction and avoids additional consistency-module computation during inference.
4) We conduct comprehensive intra-dataset, cross-dataset, ablation, feature-space, modality-consistency, and statistical analyses on CASIA-SURF, WMCA, CelebA-Spoof, and MSU-MFSD to examine the effectiveness, robustness, and limitations of the proposed learning strategy.

The rest of this paper is organized as follows. Section 2 presents a review of existing face anti-spoofing research, including handcrafted feature-based methods, deep learning-based RGB approaches, multi-modal and multi-spectral techniques, domain generalization methods, transformer-based architectures, and consistency-driven learning paradigms. Section 3 describes the benchmark datasets and data augmentation procedures used in the study. Section 4 presents the proposed Cross-Modal Consistency Learning (CMCL) methodology, including depth and NIR modality synthesis and preprocessing, modality-specific Swin Transformer encoders, hierarchical feature extraction, cross-modal attention fusion, consistency learning, classification, the overall training objective, and training/inference strategies. Section 5 presents the experimental configuration and comprehensive evaluation, including intra-dataset performance, cross-dataset generalization, ablation studies, feature-space analysis, modality consistency analysis, confusion matrices, and statistical significance testing. Finally, Section 6 concludes the paper and discusses future research directions.

2. Literature Review

Face anti-spoofing (FAS) or presentation attack detection (PAD) has gone through a great evolution in the last decade. With the growing deployment of face recognition technologies in security-critical applications, robust anti-spoofing measures are required. Existing research in FAS can be broadly classified into the following aspects: (1) handcrafted feature-based approaches, (2) deep learning based RGB methods, (3) multi-modal and multi-spectral approaches, (4) domain generalisation and domain adaptation methods, and (5) attention-based and consistency-driven learning strategies. These themes are reviewed in detail in this section.

2.1. Handcrafted Feature Based Face Presentation Attack Detection

The early FAS approaches focused mostly on handcrafted texture descriptors, aiming to model the surface artefacts generated by spoof assaults. A widely utilized approach is Local Binary Patterns (LBP) and its variations [3, 35] which detects micro-texture differences between authentic and fraudulent faces. These methods make use of the fact that printed images and screens produce specific colour aberrations and printing noise patterns.

Similarly, image distortion analysis methods were developed for detecting moiré patterns and reflectance discrepancies in replay assaults [28]. Although these classic methods showed reasonable performance under controlled laboratory circumstances, they turned out to be quite sensitive to illumination changes, camera adjustments and unknown attack kinds. The low representational capacity of handcrafted features hindered their generalisation power, which motivated the move to deep learning based techniques.

2.2. RGB Face Anti-Spoofing based on Deep Learning

The emergence of convolutional neural networks (CNNs) has brought a great revolution in the research of FAS. Automatic feature extraction from raw RGB photos became possible using deep learning models, which resulted in large gains in detection accuracy. Early CNN-based frameworks taught binary classifiers to discriminate between real and spoof samples [6]. The auxiliary supervision strategies greatly improved the learning effectiveness. For example, depth-supervised models included pseudo-depth maps as supervisory signals to enable networks to learn structural cues [6], [15]. Central Difference Convolutional Networks (CDCN) added gradient-based operators in convolutional layers to enhance their sensitivity to spoof-specific texture patterns [7, 11].

Residual networks [21] and lightweight architectures such as MobileNetV3 [17] and EfficientNet [16] permitted the efficient deployment in real-time applications. Margin based losses, e.g., ArcFace [22] and Circle Loss [23], improve feature separability in embedding space. Despite these developments, RGB-only methods are vulnerable to high-quality replay assaults and 3D mask attacks that resemble real faces.

Recent reviews [1], [2] have highlighted that RGB-based approaches generally fail in the presence of domain shifts, lighting fluctuations, and cross-dataset generalisation. These constraints inspired researchers to investigate other sensory modalities.

2.3. Multi-Modal and Multi-Spectral Face Presentation Attack Detection

In order to overcome the inherent limits of RGB-only solutions, multi-modal face anti-spoofing has attracted a lot of interest. Multi-modal systems merge complementary signals such as depth and infrared (IR) imaging to enhance the robustness of spoof detection.

Depth-based CNN architecture showed that structural information may discriminate between real faces and flat printed photos [4, 5]. Multi-modal fusion techniques that combine the RGB and depth information are shown to outperform single-modality systems [8, 30]. Multi-spectral imaging techniques have also shown that infrared sensors capture material reflectance parameters that differ significantly between genuine skin and spoof artefacts [29].

The availability of large-scale multi-modal benchmarks has spurred study in this approach. CASIA-SURF [12] presented three modalities (RGB, depth, and IR) at scale and a thorough evaluation platform. MSU-MFSD [41] further extended this benchmark to cross-ethnicity scenarios, emphasizing the fairness and generalisation problems. The Wide Multi-Channel Presentation Attack (WMCA) dataset [14] extends the spectral channels and attack instruments, allowing the assessment of the system in realistic multi-spectral scenarios.

While multi-modal fusion is beneficial for robustness, many existing approaches resort to simple feature concatenation or score-level fusion [8, 30]. These approaches do not explicitly represent inter-modal linkages and are hence limited in capturing cross-modal dependencies.

2.4. Domain Generalisation and Cross Dataset Robustness

Poor generalisation across datasets is a significant problem in face anti-spoofing. However, models trained on one dataset tend to not generalise successfully when tested on new domains due to variations in capture devices, attack types and ambient variables.

To tackle this challenge, solutions such as domain adaptability and domain generalisation have been developed. Multi-adversarial domain generalisation frameworks attempt to learn domain-invariant features using adversarial learning techniques [9]. Style-agnostic networks attempt to remove style-specific cues while preserving semantic information [10]. MixStyle augmentation procedures randomly mix the feature statistics between samples to improve robustness [26]. Other approaches align feature distributions utilizing domain adaptation methods such as adversarial discriminative adaptation [27] or gradient reversal-based learning [25]. Domain-invariant representation learning methods [24, 32] enhance the performance among domains.

The benefits of diversity of datasets and good feature alignment have been established in the FAS context across domains [31]. However, most of the domain generalisation methods are developed for the RGB-only input. Multi-modal domain generalisation is largely unexplored, especially via cross-modal consistency.

2.5. Transformer-based Methods and Attention Mechanisms

Attention mechanisms have reshaped computer vision models, enabling the models to focus on informative regions and features. The transformer architecture [19] introduced self-attention mechanisms capable of modelling long range dependencies. Vision Transformer (ViT) models [18] have obtained competitive results on picture identification tasks. In face anti-spoofing, attention modules have been applied to highlight the spoof sensitive regions and minimize the background noise [33]. We use attention-based fusion to adaptively weight modality-specific data, which allows more flexible merging of RGB, depth and IR information.

Lightweight variants of transformers with CNN backbones provide a tradeoff between accuracy and processing economy. Such approaches are particularly beneficial in multi-modal scenarios where the reliability of modalities may change from sample to sample.

2.6. Consistency Learning and Cross-Modal Representation Alignment

It has been shown that consistency learning is an effective method to improve the robustness of the representation. In general vision tasks, consistency-based regularisation is used to ensure consistent predictions under perturbations or domain shifts. Cross-modal consistency promotes the alignment of modality-specific representations in multi-modal learning.

Recently, cross-modal learning frameworks for FAS have been investigated [36], showing that many modalities can improve the discriminative power. However, it is still to be explored to explicitly model inter-modal consistency as a supervisory signal.

Consistency-based techniques rely on the assumption that structural and spectral patterns are associated across modalities for real faces, while spoof assaults introduce inconsistencies. For instance, printed assaults lack depth variation and show anomalous NIR reflectivity. Thus, imposing feature similarity for authentic samples and divergence for spoof attacks can improve discrimination.

Robust representation learning is also achieved using self-supervised learning algorithms [34] and adversarial learning frameworks [20]. However, the exploration of cross-modal consistency regularisation in multi-modal FAS is an open research direction.

2.7. Summary and Gap in Research

In summary, face anti-spoofing has achieved significant progress through deep learning-based RGB models [6, 7], multi-modal fusion frameworks [12, 14], and domain generalization strategies [9, 26]. Benchmark datasets such as CASIA-SURF [12], WMCA [14], CelebA-Spoof [39], and MSU-MFSD [41] have further facilitated the systematic evaluation of RGB–Depth–NIR face anti-spoofing systems. Nevertheless, three major research gaps remain:

1. **Limited exploitation of cross-modal relationships:** Many existing multi-modal face anti-spoofing systems rely primarily on feature concatenation or other conventional fusion strategies without explicitly modeling the intrinsic relationships and consistency among different modalities.
2. **Insufficient multi-modal cross-domain robustness:** Existing domain generalization methods are predominantly designed for RGB-based systems, leaving the problem of robust cross-dataset generalization in multi-modal face anti-spoofing comparatively underexplored.
3. **Limited use of consistency-driven learning:** Although consistency learning has demonstrated potential in representation learning, its explicit use for modeling RGB–Depth–NIR consistency and modality discrepancies in face anti-spoofing remains insufficiently investigated.

These limitations motivate the proposed Cross-Modal Consistency Learning (CMCL) framework, which integrates attention-based multi-modal fusion with explicit inter-modal consistency regularization. By exploiting complementary appearance, geometric, and spectral information while modeling modality coherence and discrepancy, the proposed approach aims to improve both intra-dataset spoof detection performance and cross-dataset generalization.

3. Datasets and Data Processing

3.1. Datasets

A. CelebA-Spoof [39]

The CelebA-Spoof dataset is one of the largest publicly available benchmarks for face anti-spoofing. It contains 624,537 RGB face photos from 10,177 people, including 202,599 bona fide (genuine) images and 422,938 fake images. The dataset comprises a large number of different presentation assault types such as print attacks, replay attacks, paper-cut attacks and 3D mask attacks. The photos have large pose, illumination, facial expression, backdrop and image quality fluctuations, which make the dataset very demanding for robust face anti-spoofing study. For precise face localization and model training, face bounding box annotations and facial attribute labels are also provided.

B. CASIA-SURF [12]

The CASIA-SURF dataset is a large-scale multimodal face anti-spoofing benchmark with synchronized RGB, Depth and Infrared (IR) modalities. It has recordings of 1,000 subjects, comprising 3,000 real videos and 18,000 fake videos. The spoof data was generated utilizing six different visual attack variations including modifications in cutting, bending and stand-off distance, to replicate actual presentation attacks. The synchronized multimodal recordings provide complementary information of appearance and depth, and hence CASIA-SURF is a popular benchmark for assessing multimodal face anti-spoofing algorithms and cross-modal feature learning.

C. WMCA [14]

The WMCA (Wide Multi-Channel Presentation Attack) dataset is a large multimodal benchmark for evaluating face presentation attack detection algorithms. The dataset contains synchronized RGB, Depth, Near Infrared (NIR) and Thermal modalities gathered from 72 people. It includes 347 genuine recordings and 1,594 presentation assault recordings, spanning a wide range of attack instruments, including printed images, replay attacks, 3D masks, silicone masks, false heads, transparent glasses, and other complicated attack situations. The multiplicity of attack kinds and sensing modalities makes WMCA one of the most challenging datasets for evaluating the robustness and generalization capability of the multimodal face anti-spoofing systems.

D. MSU-MFSD [41]

The MSU Mobile Face Spoofing Database (MSU-MFSD) is a popular benchmark for evaluating face anti-spoofing algorithms under realistic acquisition situations. The dataset includes recordings from 35 subjects, containing 70 bona fide videos and 210 spoof movies. The spoof samples are printed photo attacks and video replay attacks shot with different mobile devices under varied illumination conditions. The data set contains a large intra-subject and inter-subject variability, since each participant contributes many genuine and attack recordings, allowing an extensive evaluation of presentation attack detection approaches.

The parameters of the benchmark datasets used in this investigation are summarized in Table 1, including sensing modalities, presentation attack kinds, sample formats, number of individuals, and the distribution of bona fide and spoof samples.

Table 1. Showing the dataset summary.

Dataset	Modalities	Attack Types	Sample Type	Subjects	Genuine Samples	Spoof Samples
CelebA-Spoof [39]	RGB	Print, Replay, Paper Cut, 3D Mask	Images	10,177	202,599	422,938
CASIA-SURF [12]	RGB + Depth + NIR	Cutting, bending, stand-off distance	Videos	1,000	3,000	18,000
WMCA [14]	RGB + Depth + NIR	Print, Replay, 3D Mask, Silicone Mask, Glasses, Fake Head.	Videos	72	347	1,594
MSU-MFSD [41]	RGB	Print Attack, Replay Attack	Videos	35	70	210

3.2. Data Augmentation

Data augmentation can be a useful technique for improving the performance of anti-spoofing models that aim to detect fake biometric traits such as fake faces, voices, or fingerprints. In this context, data augmentation involves creating additional variations of the genuine and spoofed samples in the training dataset by applying various transformations to the original data. In the case of face spoofing, data augmentation techniques could include adding random noise to the images, rotating, flipping, or cropping the images, changing the brightness or contrast, and applying blur or distortion filters [42]. One important consideration when applying data augmentation to spoofing datasets is to ensure that the augmented samples still represent realistic spoofing attacks. For instance, if the spoofing attack involves using a printed photo of a genuine user, the augmented samples should reflect variations of that scenario, such as different lighting conditions or printing quality. Another important consideration is to balance the distribution of genuine and spoofed samples in the augmented dataset. If the original dataset is imbalanced, the augmented dataset could lead to further imbalances, which could affect the model's ability to detect spoofing attacks accurately. Figure 2 shows the sample images of augmented data.



Fig. 2. Sample images for data augmentation.

4. Proposed Methodology

4.1. Overview of the Proposed Framework

The proposed framework introduces a novel Cross-Modal Consistency Learning (CMCL) architecture for robust multi-modal face anti-spoofing using RGB, Depth, and Near-Infrared (NIR) modalities. The primary objective of the proposed system is to accurately distinguish genuine facial samples from spoof attacks by jointly exploiting complementary appearance, geometric, and spectral information captured across multiple sensing modalities. Unlike conventional RGB-based face anti-spoofing approaches, the proposed framework leverages the inherent consistency relationships that exist among RGB, depth, and NIR facial representations in genuine samples while simultaneously identifying inconsistencies introduced by spoof artifacts.

The complete input sample is represented as:

$$X = \{X^{\{rgb\}}, X^{\{depth\}}, X^{\{nir\}}\} \quad (1)$$

Where $X^{\{rgb\}}$, $X^{\{depth\}}$, and $X^{\{nir\}}$ denote the RGB, Depth, and NIR facial images, respectively. Each modality contributes distinct discriminative information for spoof detection. Specifically, RGB images capture facial texture, skin

appearance, and color information, depth images provide three-dimensional geometric structure and facial contour details, whereas NIR images encode spectral reflectance and material characteristics that are less sensitive to illumination variations. The integration of these complementary modalities enables the framework to detect diverse presentation attacks including print attacks, replay attacks, and mask-based spoofing attempts.

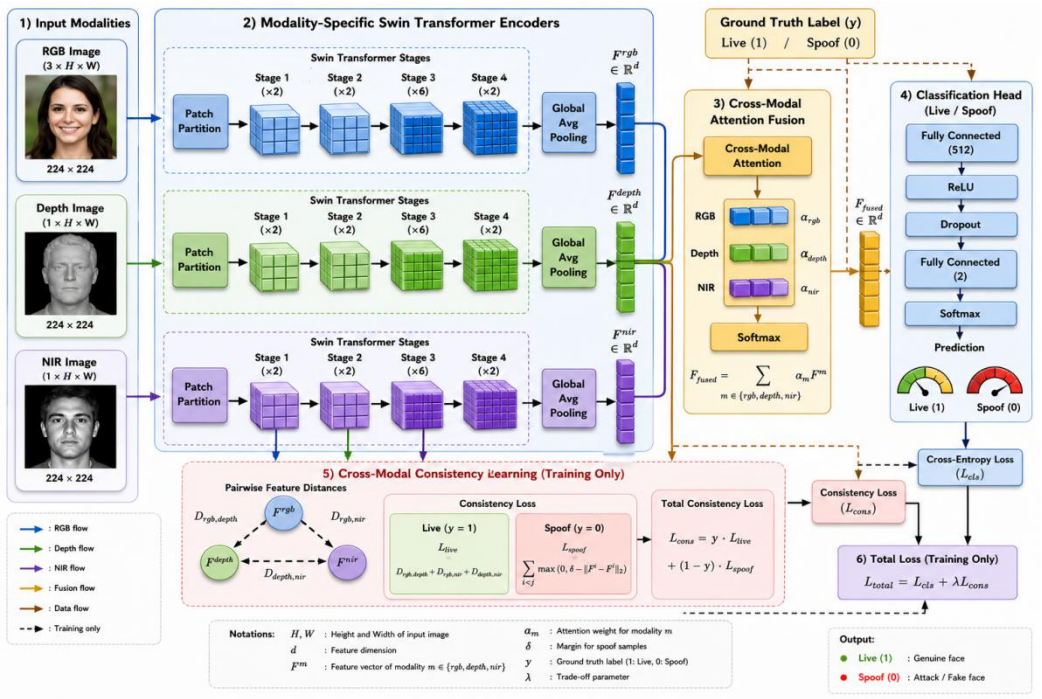


Fig. 3. Architecture of the proposed Cross-Modal Consistency Learning (CMCL) framework for robust multi-modal face anti-spoofing.

The proposed framework aims to learn a mapping function:

$$f(X) \rightarrow y \tag{2}$$

Where y ∈ {0, 1}, with 1 representing a live face and 0 representing a spoof attack.

Figure 3 shows the architecture of the proposed CMCL framework consisting of modality-specific Swin Transformer encoders, cross-modal attention fusion, consistency learning, and a classification head for robust multi-modal face anti-spoofing.

4.2. Modality Synthesis and Preprocessing

The proposed CMCL framework described in Section 4 is evaluated through a comprehensive experimental protocol designed to assess its intra-dataset performance, cross-dataset generalization, component-wise effectiveness, feature discrimination, and cross-modal consistency. The modality synthesis and preprocessing procedures described in Section 4.2 are first applied to obtain spatially aligned RGB, depth, and NIR inputs, which are subsequently processed by the modality-specific Swin Transformer encoders and the proposed cross-modal attention and consistency learning modules. To ensure a systematic and reproducible evaluation, the experiments are conducted using four benchmark face anti-spoofing datasets, namely CelebA-Spoof, CASIA-SURF, WMCA, and MSU-MFSD. The datasets are divided into training, validation, and testing subsets while preserving the live/spoof class distribution. The proposed framework is trained using the same optimization configuration described below, and the final model is selected according to validation performance. Evaluation is performed using APCER, BPCER, ACER, and classification accuracy, together with additional precision, recall, F1-score, AUC, feature-space, modality-consistency, ablation, and statistical analyses. In addition to conventional intra-dataset evaluation, cross-dataset experiments are performed between CASIA-SURF and WMCA to examine robustness under changes in acquisition conditions, sensors, and presentation attack distributions. The following subsections present the experimental configuration and the resulting quantitative and qualitative analyses.

4.2.1. Depth Map Construction

Depth information can be useful in image classification tasks using deep learning algorithms for several reasons. Depth information can help improve the accuracy of image classification algorithms by providing additional spatial information about the objects in an image. For example, knowing the depth of an object can help distinguish between objects that appear similar in 2D but have different depths. Depth information can provide valuable information about the

structure of a scene, such as the distance between objects and their relative positions in 3D space. This can help improve the overall understanding of a scene and make it easier to classify objects within it. When objects in an image are partially occluded by other objects, depth information can handle occlusions more effectively and it provides valuable additional information that can help improve the accuracy and efficiency of image classification using deep learning algorithms.

As the depth maps provide additional information, we propose to use depth information in recognizing the genuine and spoofed faces. The RGB images of both genuine and spoofed face categories were transformed into depth maps using the method proposed in [43].

This section presents a convolutional neural network that employs transfer learning to generate a depth map with high resolution from a single RGB image.

In Figure 4, we can see a high-level diagram of the encoder and decoder network used for estimating the depth. In the encoder, a DenseNet-169 [44] architecture, pretrained on ImageNet [40], is utilized to turn the input RGB face image into a feature vector. After passing this vector through a series of up-sampling layers, we end up with a depth map with half the resolution of the original. These up-sampling tires and their corresponding skip connections form the decoder.

Even though state-of-the-art methods have recommended adding sophisticated layers such as Batch Normalization, the decoder does not use them. Starting with the same number of output channels as the truncated encoder, a 1x1 convolutional layer is used for the decoder. Up-sampling blocks are then added together one by one. Each unit incorporates a bilinear up-sampling of 22 samples, two convolutional layers of 33 samples each, and output filters that are each tuned to receive half as many input samples. Then, the output of the previous layer and the pooling layer from the encoder, both of which have the same spatial dimension, are fed into the first convolutional layer of the two. Each block is then followed by a leaky ReLU activation function with parameter = 0.2, except for the last up-sampling block. The photos' original colors in the [0, 1] range are used as inputs rather than any kind of normalization of the input data.

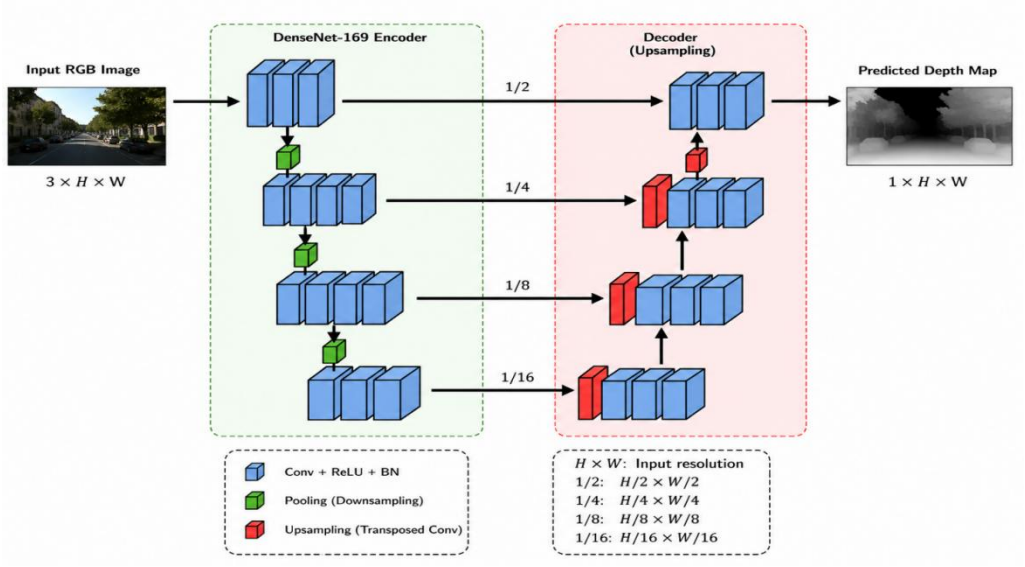


Fig. 4. Process of depth map generation from RGB image.

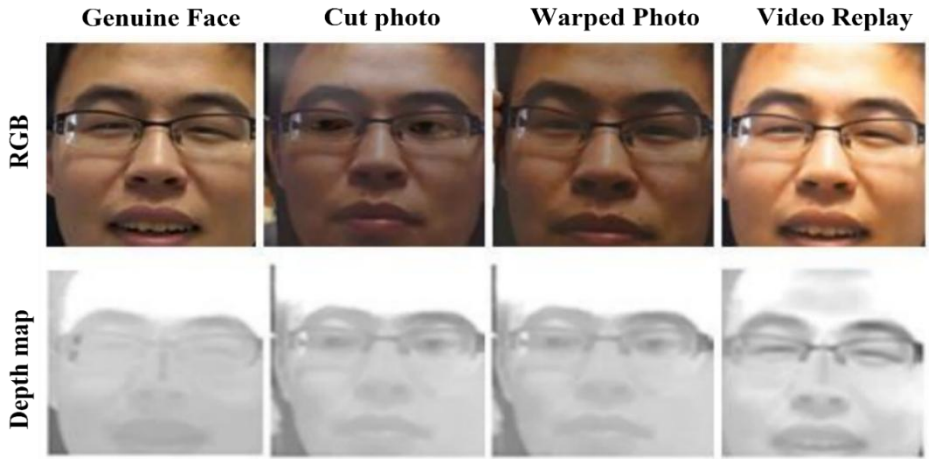


Fig. 5. Sample RGB and its constructed depth map images.

Figure 5 shows the sample RGB and its estimated depth maps. In depth regression, the standard loss function takes into account the discrepancy of the ground-truth depth map y and the prediction of the network $\hat{y}$. This paper proposes a biased sum of three loss functions as the definition of the loss L between y and $\hat{y}$.

$$Loss = L_{depth} + L_{gradient} + L_{SSIM} \tag{3}$$

The loss L_{depth} is the point-wise loss which is calculated on the depth standards:

$$L_{depth}\left(y_{in}, y_{pred}\right) = \left(1/m\right)\sum_{i=1}^{m}\left|y_{in} - y_{pred}\right| \tag{4}$$

The loss $L_{gradient}$ is the loss derived on the gradient of a depth image g_{im}

$$L_{gradient}\left(y_{in}, y_{pred}\right) = \left(1/m\right)\sum_{i=1}^{m}\left|g_{imX}\right| + \left|g_{imY}\right| \tag{5}$$

Where g_{imX} and g_{imY} compute the changes in the x and y components of depth image gradients.

The L_{SSIM} loss is written as:

$$L_{SSIM}\left(y_{in}, y_{pred}\right) = \frac{1 - SSIM\left(y_{in}, y_{pred}\right)}{2} \tag{6}$$

4.2.2. NIR Construction

Several ways have been published to generate near-infrared (NIR) images from RGB photographs such as spectral transformation, inverse modelling and deep learning-based image translation. Among these approaches, deep learning has shown better capacity in learning the nonlinear spectral mappings between RGB and NIR domains.

In this work, NIR images are created by a supervised convolutional encoder-decoder network trained on paired RGB-NIR image samples. The network takes an RGB facial image as input and learns a non-linear mapping to the appropriate NIR representation by pixel-wise reconstruction. The encoder progressively extracts hierarchical texture representations from the RGB image, and the decoder reconstructs the associated NIR image while retaining both the overall facial structure and the local spectral features. We use the Mean Absolute Error (MAE) loss to minimize the reconstruction error between the generated NIR image and the ground-truth NIR image during training. Once trained, the network is capable of creating realistic NIR images for RGB faces that do not have inherent infrared information.

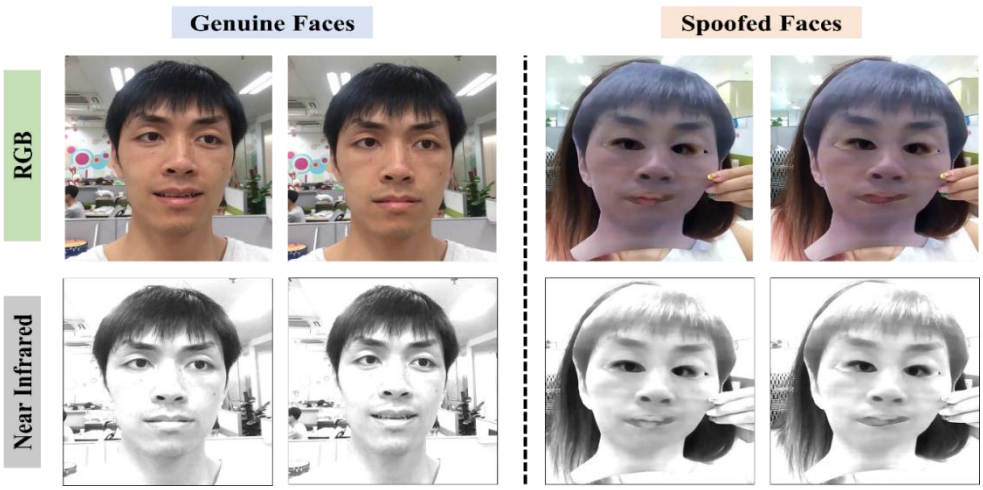


Fig. 6. Sample RGB facial images and their corresponding synthesized Near-Infrared (NIR) images generated using the proposed deep learning-based RGB-to-NIR conversion network.

Figure 6 shows the sample RGB facial images and their corresponding synthesized Near-Infrared (NIR) images generated using the proposed deep learning-based RGB-to-NIR conversion network. Although synthesized NIR images cannot fully replace hardware-captured infrared images, deep learning based spectral translation has been shown in earlier work to preserve discriminative facial features needed for downstream recognition and anti-spoofing applications.

NIR photographs offer supplemental spectral information that is largely invariant to changes in visible illumination. Unlike RGB photos, NIR imaging captures the fundamental reflectance features of human skin that are more difficult for presentation attacks such as printed photographs or replay screens to mimic actual face expressions. Therefore, the use of NIR information greatly enhances the robustness of face anti-spoofing systems in tough ambient settings.

RGB images are scaled to 224 X 224 pixels and normalized to the range [0,1] before NIR synthesis. The synthesized NIR pictures are then employed as one of the three input modalities to the proposed CMCL architecture.

The resulting NIR pictures are then inputted into the modality-specific Swin Transformer encoder with the modalities of RGB and depth. This enables the proposed CMCL system to leverage complementary spectrum information to perform cross-modal attention fusion and consistency learning, hence achieving improved spoof detection accuracy and cross-dataset generalisation.

4.3. Input Modalities

The proposed CMCL framework operates on three complementary representations: the original RGB image and two RGB-derived representations, namely depth and NIR. These representations provide complementary appearance, geometric, and spectral information for robust face anti-spoofing. The RGB representation captures facial texture and color information, while the derived depth representation provides structural information and the derived NIR representation captures complementary spectral characteristics. The three representations are subsequently processed by their respective modality-specific Swin Transformer encoders before cross-modal attention fusion and consistency learning.

The first stage of the proposed architecture consists of three complementary input modalities: RGB, Depth, and Near-Infrared (NIR) facial images. Each modality captures distinct characteristics of the human face and contributes unique information for spoof detection. RGB images provide appearance-based texture and color information, depth images encode three-dimensional geometric structure, and NIR images capture spectral reflectance and material properties that are less sensitive to illumination variations.

The input modalities are represented as:

$$X = \{Xrgb, Xdepth, Xnir\} \tag{7}$$

Where:

X^{rgb} = RGB facial image, X^{depth} = Depth facial image, X^{nir} = Near-Infrared facial image

Prior to feature extraction, all input images undergo preprocessing operations including face detection, alignment, normalization, and resizing to a fixed spatial resolution of 224 × 224 pixels. This preprocessing stage ensures spatial consistency across modalities and improves feature learning stability.

The use of multiple modalities enhances robustness against presentation attacks. Genuine faces exhibit consistent texture, geometry, and spectral responses across modalities, whereas spoof attacks typically introduce inconsistencies such as flat depth responses or abnormal NIR reflections.

4.4. Modality-Specific Swin Transformer Encoders

To effectively extract discriminative modality-specific representations, the proposed framework employs independent Swin Transformer encoders for RGB, Depth, and NIR streams. Unlike conventional convolutional neural networks, Swin Transformers utilize shifted-window self-attention mechanisms to capture both local and global contextual dependencies while maintaining computational efficiency.

Each modality is processed using an independent encoder:

$$F^{rgb} = E_{rgb}(X^{rgb}) \tag{8}$$

$$F^{depth} = E_{depth}(X^{depth}) \tag{9}$$

$$F^{nir} = E_{nir}(X^{nir}) \tag{10}$$

Where:

E_{rgb}, E_{depth}, and E_{nir} denote modality specific Swin Transformer encoders.

F_{rgb}, F_{depth}, and F_{nir}denote the extracted modality specific feature representations.

Although all encoders share the same architectural design, they do not share weights. This allows each encoder to specialize in modality-dependent feature extraction. The RGB encoder primarily focuses on texture and appearance features, the depth encoder learns three-dimensional structural patterns, and the NIR encoder captures spectral reflectance characteristics.

Each Swin Transformer encoder consists of the following stages: 1). Patch Partition 2). Hierarchical Swin Transformer Blocks 3). Patch Merging 4). Global Average Pooling.

The extracted features are represented as:

$$F^m \in R^d \tag{11}$$

Where:

m ∈ {rgb, depth, nir}, "d" denotes the feature dimension.

The modality-specific encoders preserve unique information from each sensing modality and form the foundation for subsequent cross-modal fusion and consistency learning.

4.5 Patch Partition and Hierarchical Feature Extraction

Within each Swin Transformer encoder, the input image is first divided into non-overlapping patches. The patch partition operation transforms the input image into a sequence of embedded tokens suitable for transformer-based processing.

Given an input image of size:

$$H \times W \times C \tag{12}$$

It is partitioned into patches of size:

$$P \times P \tag{13}$$

The number of generated patches are:

$$N = (H \times W) / P^2 \tag{14}$$

Each patch is linearly embedded into a feature representation before being processed by hierarchical Swin Transformer stages.

The hierarchical architecture progressively captures low-level texture features and high-level semantic information through multiple transformer stages. Shifted-window self-attention enables efficient local attention computation while simultaneously facilitating global feature interaction across windows.

As the network depth increases, spatial resolution decreases, semantic representation increases, and contextual understanding improve.

This hierarchical representation learning is highly beneficial for detecting subtle spoof artifacts.

4.6. Cross-Modal Attention Fusion Module

After modality-specific feature extraction, the proposed framework performs adaptive feature fusion using a Cross-Modal Attention Fusion module. Unlike traditional concatenation-based fusion methods, the proposed attention mechanism dynamically assigns importance weights to different modalities based on their reliability and discriminative contribution.

Initially, the extracted feature vectors are concatenated:

$$F_{cat} = [F^{rgb}; F^{depth}; F^{nir}] \tag{15}$$

Scalar modality score is defined as

$$s_m = {w_m}^T F_m + b_m \tag{16}$$

Attention weights are then computed using a softmax-based attention mechanism:

$$\alpha_m = \frac{\exp(W_m^T F^m)}{\sum_{k \in \{rgb,depth,nir\}} \exp(W_m^T F^m)} \tag{17}$$

Where F_m is the feature vector of modality m, $w_m \in \mathbb{R}^d$ is a learnable modality-specific weight vector, $b_m \in \mathbb{R}$ is a learnable bias, s_m is the modality importance score, and α_m is the resulting normalized attention weight.

Then the fused representation is defined as

$$F_{fused} = \sum_{m \in \{rgb,depth,nir\}} \alpha_m F^m \tag{18}$$

This adaptive fusion strategy enables the model to emphasize reliable modalities while suppressing noisy or less informative features. Consequently, the fusion module improves robustness against sensor noise, illumination variations, and spoof artifacts.

4.7. Cross-Modal Consistency Learning Module

The main novelty of the presented system is the Cross-Modal Consistency Learning module. This module aims to enforce the consistency of features on the authentic faces across modalities and increase the inconsistencies on the spoof attacks.

Genuine facial samples display associated qualities of texture, depth and spectrum. In contrast, spoof attacks tend to generate modality-mismatched responses. While printed photographs may be visually accurate in RGB images, they do not provide valid depth information or realistic NIR reflectance.

To capture these relationships, pairwise Euclidean distances are computed between modality-specific feature vectors:

$$D_{\text{rgb,depth}} = \left\| F^{\text{rgb}} - F^{\text{depth}} \right\|_2^2 \tag{19}$$

$$D_{\text{rgb,nir}} = \left\| F^{\text{rgb}} - F^{\text{nir}} \right\|_2^2 \tag{20}$$

$$D_{\text{depth,nir}} = \left\| F^{\text{depth}} - F^{\text{nir}} \right\|_2^2 \tag{21}$$

Where F^{rgb}, F^{depth}, F^{nir} denote the modality-specific feature vectors and $\|\cdot\|_2$ is the L2 (Euclidean) norm.

For genuine (live) samples, the framework minimizes inter-modal feature distances, encouraging cross-modal agreement:

$$L_{\text{cons}}^{\text{live}} = D_{\text{rgb,depth}} + D_{\text{rgb,nir}} + D_{\text{depth,nir}} \tag{22}$$

This encourages compact and coherent feature representations across modalities.

For spoof samples, a margin-based formulation encourages the modalities to remain sufficiently different:

$$L_{cons}^{spoof} = max(0, \delta - D_{rgb,depth}) + max(0, \delta - D_{rgb,nir}) + max(0, \delta - D_{depth,nir}) \tag{23}$$

Where δ is the minimum desired inter-modal distance for spoof samples (margin hyperparameter).

The complete consistency loss combines both terms via the ground-truth label y:

$$L_{cons} = y\, L_{cons}^{live} + (1 - y)\, L_{cons}^{spoof} \tag{24}$$

Where $y = 1$ for genuine samples and $y = 0$ for spoof samples.

This formulation improves discriminative learning by minimizing modality variance for genuine samples while maximizing inconsistency for spoof attacks.

The consistency module is active only during training and does not participate during inference, thereby avoiding additional computational overhead.

4.8. Classification Head

The fused feature vector generated by the attention fusion module is forwarded to the classification head for final live/spoof prediction. The classification head consists of fully connected layers, ReLU activation, dropout regularization, and a softmax classifier.

The ReLu activation function is defined as:

$$F_h = ReLU(W_h F_{fused} + b_h) \tag{25}$$

Where $F_{fused} \in \mathbb{R}^d$, $W_h \in \mathbb{R}^{512xd}$, $b_h \in \mathbb{R}^{512}$.

Dropout regularization is then applied:

$$F_d = Dropout(F_h) \tag{26}$$

A final linear layer projects to the two-class logit space:

$$z = W_c F_d + b_c \tag{27}$$

Where $W_c \in \mathbb{R}^{2x512}$, $b_c \in \mathbb{R}^2$, and $z = [z_{live}, z_{spoof}]^T$ is the two-dimensional logit vector.
The class probabilities are obtained via softmax over the two logits:

$$p_k = \frac{exp(z_k)}{\sum_{j=1}^{2} exp(z_j)}, k \in \{live, spoof\} \tag{28}$$

The per-sample classification loss is the binary cross-entropy:

$$L_{cls} = -[y \; log \; p_{live} + (1 - y) \; log \; p_{spoof}] \tag{29}$$

For a batch of N samples, the batch-averaged classification loss is:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \; log \; p_{i,live} + (1 - y_i) \; log \; p_{i,spoof}] \tag{30}$$

This module generates the final decision indicating whether the input sample corresponds to a genuine face or a spoof attack.

4.9. Total Training Objective

The overall optimization objective of the proposed framework combines classification loss and consistency regularization loss.
The final loss function is defined as:

$$L_{\text{total}} = L_{\text{cls}} + \lambda L_{\text{cons}} \tag{32}$$

Where:

L_{cls} = Classification loss
L_{cons}= Consistency loss
λ = Balancing coefficient controlling the contribution of consistency regularization.

The integration of consistency learning improves feature alignment across modalities and enhances robustness under cross-dataset and unseen attack scenarios.

4.10. Training and Inference Strategy

During training, all components of the proposed framework are jointly optimized to learn discriminative and modality-consistent representations for face anti-spoofing. First, the RGB, depth and NIR inputs are processed separately by modality-specific Swin Transformer encoders to learn appearance, geometry and spectral properties. A cross-modal attention fusion module is then employed to fuse these features by adaptively highlighting the most informative cues from each modality. The fused representation is passed to the classification head and the consistency learning module. The classification loss is utilized to guide the live/spoof prediction while the consistency loss supports consistent feature representations among modalities for real faces and retains discriminative discrepancies for spoof samples. Both losses are merged in a total loss function and backpropagated together through the encoders, the fusion module and the classifier, enabling end-to-end optimization of the full architecture.

The consistency learning module is not needed during inference as it is only utilized as an auxiliary supervisory mechanism during training. Thus, the deployed system only keeps the forward path including the modality-specific encoders, cross-modal attention fusion module and the classification head. The architecture also decreases the computational overhead and memory requirements during deployment, while keeping the strong feature representations learned during training. The proposed system, thus, achieves a good trade-off between powerful multi-modal spoof detection capacity and efficient inference in real-world applications.

4.11. Computational Complexity and Practical Advantages

The proposed framework introduces only minimal computational overhead through pairwise feature consistency computations. Let 'd' denote the feature dimension. The additional complexity introduced by consistency learning is:

$$O(d) \tag{32}$$

Which is negligible compared to the computational cost of transformer-based feature extraction.

The proposed framework provides several important advantages for face anti-spoofing detection. By learning complementary information from RGB, depth, and NIR modalities, it produces a robust multi-modal feature representation that captures texture, structural, and spectral cues more effectively than a single-modality system. The adaptive attention-based fusion mechanism dynamically assigns higher importance to reliable modalities and reduces the influence of noisy or less informative inputs. This improves performance under illumination changes, sensor noise, poor image quality, and different spoof attack conditions. In addition, the cross-modal consistency learning strategy enhances discrimination between genuine and spoof samples by encouraging consistent modality features for live faces while highlighting modality mismatches caused by attacks. As a result, the framework improves cross-dataset generalization and maintains strong sensitivity to unseen spoof artifacts. Since the consistency module is used only during training and removed during inference, the proposed model maintains lower deployment complexity and supports real-time face anti-spoofing applications with strong robustness.

Table 2. Key Architecture and Training Parameters of the Proposed CMCL Framework

Parameter	Configuration
Input resolution	(224X224)
Number of modalities	3
Modalities	RGB, RGB-derived Depth, RGB-derived NIR
Encoder	Swin Transformer
Number of encoders	3, independent weights
Swin stages	4
Patch size	4X4
Window size	7X7
Attention heads	3, 6, 12, 24
Final feature dimension (d)	768
Fusion mechanism	Cross-modal attention
Hidden FC dimension	512
Output classes	2
Consistency distance	Euclidean
Spoof margin (\delta)	1.0*
Consistency coefficient (\lambda)	0.1*
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
LR scheduler	Cosine Annealing
Epochs	200
Batch size	32
Training / Validation / Testing	70% / 15% / 15%
Random seed	42
GPU	NVIDIA RTX 4090, 24 GB
CPU	Intel Core i9-13900K
RAM	64 GB DDR5

For reproducibility, the principal architectural and training parameters used to implement the proposed CMCL framework are summarized in Table 2.

5. Results

The proposed CMCL framework described in Section 4 is evaluated through a comprehensive experimental protocol designed to assess its intra-dataset performance, cross-dataset generalization, component-wise effectiveness, feature discrimination, and cross-modal consistency. The modality synthesis and preprocessing procedures described in Section 4.2 are first applied to obtain spatially aligned RGB, depth, and NIR inputs, which are subsequently processed by the modality-specific Swin Transformer encoders and the proposed cross-modal attention and consistency learning modules.

To ensure a systematic and reproducible evaluation, the experiments are conducted using four benchmark face anti-spoofing datasets, namely CelebA-Spoof, CASIA-SURF, WMCA, and MSU-MFSD. The datasets are divided into training, validation, and testing subsets while preserving the live/spoof class distribution. The proposed framework is trained using the same optimization configuration described below, and the final model is selected according to validation performance. Evaluation is performed using APCER, BPCER, ACER, and classification accuracy, together with additional precision, recall, F1-score, AUC, feature-space, modality-consistency, ablation, and statistical analyses. In addition to conventional intra-dataset evaluation, cross-dataset experiments are performed between CASIA-SURF and WMCA to examine robustness under changes in acquisition conditions, sensors, and presentation attack distributions. The following subsections present the experimental configuration and the resulting quantitative and qualitative analyses.

Face authentication solutions are still subject to presentation attacks such as printed photos, replayed films and display-based assaults. Therefore, it is necessary to conduct a thorough evaluation to assess whether a face anti-spoofing system can reliably discriminate genuine facial presentations from spoof efforts under various acquisition settings. In this paper, the proposed Cross-Modal Consistency Learning (CMCL) framework is assessed on four benchmark datasets, including CelebA-Spoof [39], CASIA-SURF [12], WMCA [14], and MSU-MFSD [41]. These datasets offer complementary evaluation contexts, covering RGB alone and multi-modal samples, varied capture methods, diverse lighting circumstances, and multiple forms of presentation attacks.

The experimental evaluation is aimed at evaluating the intra-dataset performance and the cross-dataset generalization capabilities. Intra-dataset experiments are used to analyze the effectiveness of the proposed framework when training and testing samples are collected from the same dataset, whereas cross-dataset experiments are used to evaluate the robustness of the proposed framework under unforeseen data distributions, sensor alterations and assault settings. The proposed CMCL model is evaluated with the common face anti-spoofing metrics, i.e., Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), Average Classification Error Rate (ACER) and classification accuracy. Furthermore, ablation experiments, modality consistency analysis, feature space visualization and statistical significance tests are conducted to analyze the contribution of each proposed component. The next parts give the experimental methodologies, quantitative data, comparison with existing approaches and extensive explanation of the findings.

All experiments were carried out using the PyTorch deep learning framework. The available samples of each benchmark dataset were split into 70% train, 15% validation and 15% test, keeping the original live/spoof class distribution. The proposed CMCL framework is trained using the AdamW optimizer with an initial learning rate of 1 X 10^{-4}, weight decay of 1 X 10^{-4} and cosine annealing learning-rate scheduler for 200 epochs. We set the batch size for training to 32 and picked the model with the lowest validation loss for the final evaluation. All trials were run using a fixed random seed of 42 for reproducibility. Training was conducted on a workstation with an Intel Core i9-13900K CPU (24 cores, up to 5.8 GHz), 64 GB DDR5 RAM, and NVIDIA GeForce RTX 4090 GPU with 24 GB GDDR6X memory, running the operating system Ubuntu 22.04 LTS, CUDA 12.1 and PyTorch 2.2. The difficulty of the model was evaluated by computing the number of trainable parameters and FLOPs using the THOP and fvcore libraries on an input resolution of 224 × 224 pixels. We measured the inference latency as the average execution time across 1,000 forward passes, excluding the first 50 warm-up rounds. The throughput was computed as the number of images processed per second at an inference batch size of 128. Peak GPU memory usage during inference was also reported. For a fair comparison of computational complexity, runtime and deployment efficiency, all competing methods were examined using the same hardware and software parameters.

5.1. Performance Analysis

Table 3. Comparison with State-of-the-Art Methods on CASIA-SURF

Method	Modalities	APCER (%)	BPCER (%)	ACER (%)	Accuracy (%)
ResNet-18 [21]	RGB	5.60	3.20	4.40	95.60
Multi-Scale Fusion [8]	RGB+Depth+NIR	3.20	2.22	2.71	97.50
Static–Dynamic Fusion [37]	RGB+Depth+NIR	2.80	2.10	2.45	98.00
Multi-Modal CDCN [7]	RGB+Depth+NIR	1.20	0.84	1.02	98.90
FM-ViT [38]	RGB+Depth+NIR	1.00	0.82	0.91	99.10
Proposed CMCL Framework	**RGB+Depth+NIR**	**0.80**	**0.60**	**0.70**	**99.30**

Table 3 presents the comparative study of our proposed Cross-Modal Consistency Learning (CMCL) framework with certain state-of-the-art face anti-spoofing algorithms using the CASIA-SURF dataset. The results indicate a steady improvement in performance as the models evolve from basic RGB-based methods to advanced multi-modal architecture. The baseline ResNet-18 model yields an ACER of 4.4% with an accuracy of 95.6% which indicates that the use of solely RGB appearance information is not enough for spoof identification. Multi-modal fusion systems can boost performance substantially. The Multi-Scale Fusion and Static–Dynamic Fusion achieved an ACER of 2.71% and 2.45%, respectively. This improvement highlights the importance of using complementary depth and near-infrared cues to capture structural and spectral features which are hard to replicate in presentation attacks.

Further improvement in performance is seen with Multi-Modal CDCN and FM-ViT obtaining ACER values of 1.02% and 0.91% respectively, by utilizing enhanced feature representation and transformer-based modelling. The results show that enhanced feature extraction and attention mechanisms are useful for learning discriminative spoof-related patterns. However, the proposed CMCL framework has the best overall performance on all evaluation metrics, with the lowest APCER (0.8%), BPCER (0.6%) and ACER (0.7%) and maximum classification accuracy of 99.3%. Compared with FM-ViT, the proposed method achieves around 23.1% decrease in ACER and 0.2% improvement in accuracy. The proposed cross-modal consistency learning method directly models the underlying connections among RGB, depth and NIR modalities, which can be credited for such benefits. The proposed approach learns more discriminative and robust representations by enforcing feature coherence for actual faces and increasing the modality disparities for spoof samples. The better performance on CASIA-SURF reflects the efficiency of the combination of attention based fusion and consistency based regularization for reliable multi-modal face anti-spoofing.

Table 4. Comparison with State-of-the-Art Methods on WMCA

Method	Modalities	APCER (%)	BPCER (%)	ACER (%)	Accuracy (%)
ResNet-18 [21]	RGB	10.20	8.40	9.30	91.40
Auxiliary Supervision [15]	RGB	7.60	6.80	7.20	93.20
Multi-Modal Fusion [8]	RGB+Depth+NIR	5.10	4.50	4.80	95.80
Attention Fusion Network [33]	RGB+Depth+NIR	4.20	3.60	3.90	96.70
FM-ViT [38]	RGB+Depth+NIR	3.80	3.20	3.50	97.10
Proposed CMCL Framework	**RGB+Depth+NIR**	**2.90**	**2.20**	**2.55**	**97.90**

Table 4 demonstrates the performance comparison of the proposed Cross-Modal Consistency Learning (CMCL) framework with many representative face anti-spoofing algorithms on the tough WMCA dataset. The WMCA benchmark presents a specific challenge because of the variety of presentation attack instruments, acquisition settings and multi-spectral information, which makes it a good platform for testing the robustness of anti-spoofing algorithms. Our results show a clear performance evolution from traditional RGB-based methods to sophisticated multi-modal approaches. ResNet-18 obtains an ACER of 9.3%, with an accuracy of 91.4%, showing the limits of the use of RGB appearance cues only under complicated attack scenarios. The Auxiliary Supervision model adds extra supervisory signals to boost performance. The ACER drops to 7.2% and accuracy rises to 93.2%.

The supplementary depth and infrared information further improve the spoof detection capability. The multi-modal Fusion and Attention Fusion Network reach ACER values of 4.8% and 3.9% respectively, which shows the efficacy of the multi-modal features integration. Leveraging transformer-based feature modelling and flexible modal interactions, FM-ViT achieves better performance with an ACER of 3.5% and an accuracy of 97.1%. However, the proposed CMCL framework delivers the best overall performance with the lowest APCER of 2.9%, BPCER of 2.2%, and ACER of 2.55% and the maximum accuracy of 97.9%. The proposed technique achieves around 27.1% lower ACER and 0.8% higher classification accuracy than FM-ViT.

The higher performance of the proposed framework can be attributed to the combination of attention-based cross-modal fusion and the proposed consistency learning mechanism. The attention module adaptively highlights the informative modalities, and the consistency learning module explicitly enforces consistent feature representations across the RGB, depth and NIR modalities for authentic samples and accentuates the modality differences for spoof attacks. This allows the network to better utilize geometric and spectral properties which are difficult to mimic by presentation attacks. The results achieved on the WMCA dataset reveal that the proposed CMCL framework provides greater robustness against difficult spoofing scenarios and better generalisation than the previous state-of-the-art techniques.

Table 5 Performance comparison of the proposed RGB branch and many sample RGB-based face anti-spoofing methods on CelebA-Spoof dataset. The results show a constant enhancement in the performance of spoof detection with the increasing sophistication of feature learning architectures. Baseline ResNet-18 obtains an ACER of 5.6% and an accuracy of 94.1%, which shows the limitations of typical convolutional architectures to capture subtle spoofing artefacts under different real-world settings. The CDCN model enhances the performance by adding the central difference convolution operations, which reduces the ACER to 4.0% and increases the accuracy to 95.9%, showing the efficacy of texture-sensitive feature extraction for presentation attack detection.

Transformer-based and efficient feature learning algorithms significantly improve the detecting capability. We observe that the Vision Transformer obtains an ACER of 3.55% with 96.3% accuracy which highlights the value of capturing long-range contextual dependencies and global facial representations. Similarly, EfficientNet yields 3.0% ACER and 96.8% accuracy, which demonstrates the benefits of optimized network scalability and better feature representation.

Table 5. Comparison with State-of-the-Art Methods on CelebA-Spoof

Method	Modalities	APCER (%)	BPCER (%)	ACER (%)	Accuracy (%)
ResNet-18 [21]	RGB	6.40	4.80	5.60	94.10
CDCN [7]	RGB	4.30	3.70	4.00	95.90
Vision Transformer [18]	RGB	3.90	3.20	3.55	96.30
EfficientNet [16]	RGB	3.20	2.80	3.00	96.80
Proposed CMCL Framework	**RGB only**	**2.60**	**2.20**	**2.40**	**97.40**

The proposed RGB branch outperforms all the analyzed approaches in terms of the lowest APCER (2.6%), BPCER (2.2%), ACER (2.4%), and the maximum classification accuracy of 97.4%. Compared to EfficientNet, the proposed method has reduced ACER by 20.0% and improved accuracy by 0.6%. The APCER reduction means a higher resistance to spoof assaults and the lower BPCER means a better recognition of the legitimate users. These results show that the proposed feature learning technique can learn more discriminative facial representations and well discriminate genuine faces from spoof presentations. The improved performance on the difficult CelebA-Spoof dataset also demonstrates the robustness and generalisation capabilities of the proposed RGB branch under different assault scenarios, lighting conditions and facial appearance variations.

The superior performance achieved on the MSU-MFSD dataset further demonstrates that the proposed feature learning framework maintains strong discriminative capability even when only RGB information is available, thereby complementing the multi-modal evaluations conducted on CASIA-SURF and WMCA.

Table 6. Comparison with State-of-the-Art Methods on MSU-MFSD

Method	Modalities	APCER (%)	BPCER (%)	ACER (%)	Accuracy (%)
LBP [3]	RGB	14.2	10.80	12.50	87.20
ResNet-18 [21]	RGB	7.90	6.10	7.00	92.80
Auxiliary Supervision [15]	RGB	5.10	4.70	4.90	95.20
CDCN [7]	RGB	3.90	3.10	3.50	96.20
Proposed CMCL Framework	**RGB only**	**3.00**	**2.40**	**2.70**	**96.90**

Table 6 shows the comparative effectiveness of the proposed RGB branch versus many representative face anti-spoofing techniques on the MSU-MFSD dataset. The results indicate a clear evolution in terms of performance from conventional handmade feature based methods to contemporary deep learning architectures. The LBP-based technique achieves an ACER of 12.5% with an accuracy of 87.2%, suggesting that handcrafted texture descriptors are limited in addressing the numerous spoofing variations encountered in real-world circumstances. The ResNet-18 model learns the discriminative facial representations directly from the data which considerably improves the performance, reducing the ACER to 7.0% and raising the accuracy to 92.8%. The Auxiliary Supervision architecture, which utilizes depth-related supervisory signals to assist feature learning, achieves further gains with an ACER of 4.9% and accuracy of 95.2%.

We compare the proposed CDCN with the baseline approaches, where CDCN gets the best performance with ACER of 3.5% and accuracy of 96.2%, which demonstrates the efficiency of the central difference convolution to capture the subtle spoof-related texture patterns. However, the proposed RGB branch beats all competing approaches by achieving the lowest APCER (3.0%), BPCER (2.4%) and ACER (2.7%), while earning the maximum classification accuracy of 96.9%. The proposed technique achieves about 22.9% ACER reduction and 0.7% accuracy improvement against CDCN. The lower the APCER, the better the capacity to reject spoof assaults, and the lower the BPCER, the better the ability to recognize authentic users. The results show that the proposed feature extraction and representation learning technique can effectively capture the discriminative facial traits and generalise well to various presentation attack scenarios. The continuous improvements over all the evaluation parameters show the robustness of the proposed RGB branch and its applicability for dependable face anti-spoofing applications in practical deployment situations.

Figure 7 shows the performance comparison of the proposed CMCL framework with the strongest prior state-of-the-art method across CelebA-Spoof, CASIA-SURF, WMCA, and MSU-MFSD datasets. The proposed framework achieves lower ACER values across all benchmarks, demonstrating improved face anti-spoofing performance.

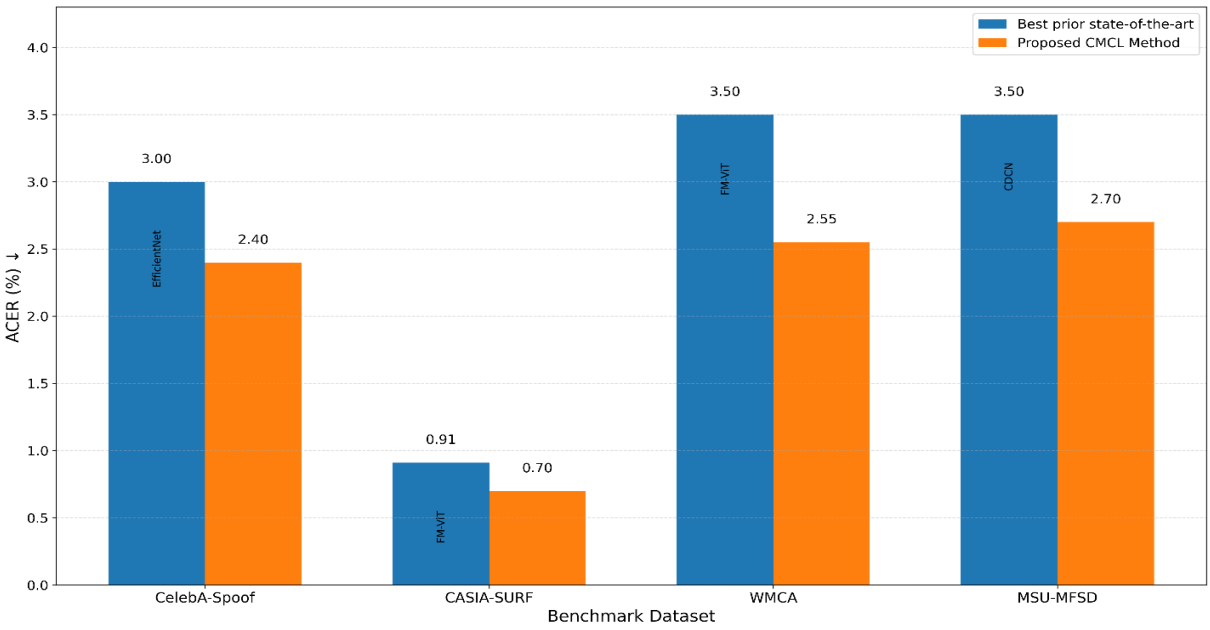


Fig. 7. Performance comparison of the proposed CMCL framework with the strongest prior state-of-the-art method across CelebA-Spoof, CASIA-SURF, WMCA, and MSU-MFSD datasets.

5.2. Cross-Dataset Generalization

The below Table 7 shows the cross-dataset generalisation ability of the algorithms tested under two strict evaluation methods, i.e., trained on CASIA-SURF and tested on WMCA and vice versa. Cross-dataset testing differs from the intra-dataset evaluation, and it provides a more realistic evaluation of the resilience of the trained model, given that the trained model is tested on unseen sensors, attack instruments, environmental situations and data distributions. As expected, all techniques show considerable performance degradation compared with their intra-dataset equivalent results, which shows the significant domain gap between CASIA-SURF and WMCA.

Table 7. Cross-Dataset Generalization Results

Train Dataset	Test Dataset	Method	APCER (%)	BPCER (%)	ACER (%)	Accuracy (%)
CASIA-SURF	WMCA	ResNet-18	15.60	12.80	14.20	86.50
CASIA-SURF	WMCA	CDCN	10.40	8.20	9.30	90.20
CASIA-SURF	WMCA	Attention Fusion	8.90	7.10	8.00	91.80
CASIA-SURF	WMCA	**Proposed CMCL Framework**	**6.90**	**5.60**	**6.25**	**94.10**
WMCA	CASIA-SURF	ResNet-18	14.20	11.40	12.80	87.30
WMCA	CASIA-SURF	CDCN	9.50	7.90	8.70	91.10
WMCA	CASIA-SURF	Attention Fusion	8.20	6.80	7.50	92.40
WMCA	CASIA-SURF	**Proposed CMCL Framework**	**6.10**	**5.20**	**5.65**	**94.80**

The baseline ResNet-18 trained on CASIA-SURF and tested on WMCA achieves ACER of 14.2% and accuracy of 86.5%, which shows the limited generalisation capacity. The CDCN framework learns more discriminative spoof-related texture representations, which results in better performance with the ACER reduced to 9.3% and the accuracy increased to 90.2%. The attention Fusion model with complementary information of several modalities considerably enhances the robustness with ACER of 8.0% and accuracy of 91.8%. On the other hand, the proposed CMCL framework achieves the best performance with an APCER of 6.9%, a BPCER of 5.6%, an ACER of 6.25% and an accuracy of 94.1%. These results demonstrate the proposed consistency-driven learning approach might effectively alleviate the effect of domain shifts by encouraging modality-invariant and semantically aligned feature representations.

A similar tendency is observed in the reverse assessment technique where the models are trained on CASIA-SURF and tested on WMCA. The proposed CMCL framework again performs better with the lowest ACER of 5.65% and the highest accuracy of 94.8% which is 1.85 percentage points lower than the ACER and 2.4 percentage points higher than the accuracy of the Attention Fusion model. The constant progress on both cross-dataset situations verifies that the proposed cross-modal consistency learning method can successfully capture the intrinsic correlations between RGB, Depth and NIR modalities that are stable across datasets. The proposed system directly imposes feature coherence on real

samples and enlarges the modality mismatches of spoof attacks, such that it learns more transferable representations that can generalise to unseen domains. These results show that CMCL not only improves the face anti-spoofing accuracy but also considerably enhances the cross-dataset robustness, which is an important condition for face anti-spoofing deployment in the real-world.

The stable cross-dataset performance of the proposed CMCL framework demonstrates the effectiveness of consistency-driven multi-modal representation learning in alleviating domain shifts and improving generalisation across heterogeneous acquisition environments, which is one of the significant challenges in practical face anti-spoofing systems.

5.3. Ablation Study

Table 8 below shows the ablation study to evaluate the individual contribution of each part in the proposed Cross-Modal Consistency Learning (CMCL) system. The purpose of this investigation is to assess the influence of multi-modal inputs, attention-based fusion and consistency learning on the overall performance of face anti-spoofing. The baseline model is the model without the attention fusion and consistency regularization, only the RGB modality is used. Its ACER is 5.3% and accuracy is 94.6%. This result illustrates the limitation of using appearance information alone for robust spoof detection.

Attention-based fusion to add an additional modality considerably boosts the performance. Variant-A combining RGB and Depth modalities achieves an ACER of 3.35% and accuracy of 96.5%, showing the efficiency of using three-dimensional structure information. Similarly, the combination of RGB and NIR modalities in Variant-B yields 3.05% of ACER and 96.9% of accuracy, suggesting that spectral information can be useful in better discriminating spoof from depth in some cases. Fusing all three modalities with the attention fusion (Variant-C), the ACER further falls to 2.0% and the accuracy increases to 98.0%, which confirms that RGB, Depth, and NIR give the complementing information that improves the spoof detection capabilities.

Variant-D employs all three modalities with the proposed consistency loss but without the attention fusion module. This configuration isolates the contribution of consistency learning and allows its individual effect to be evaluated. This setup results in an ACER of 1.65% and an accuracy of 98.5%, which is better than Variant-C even without attention-based fusion. This discovery indicates that the performance gain is more dependent on the consistency learning than the feature fusion as the intrinsic relationship between the distinct modalities is clearly modelled. Finally, the complete CMCL framework with attention fusion and consistency learning obtains the best overall performance with APCER of 0.80%, BPCER of 0.60%, ACER of 0.70% and accuracy of 99.3%. The proposed architecture decreases the ACER by roughly 86.8%, and increases the classification accuracy by 4.70 percentage points compared to the baseline model.

Table 8. Ablation Study of Proposed Components

Variant	Attention Fusion	Consistency Loss	RGB	Depth	NIR	APCER (%) ↓	BPCER (%) ↓	ACER (%) ↓	Accuracy (%) ↑
Baseline	✗	✗	✓	✗	✗	5.90	4.70	5.30	94.60
Variant-A	✓	✗	✓	✓	✗	3.60	3.10	3.35	96.50
Variant-B	✓	✗	✓	✗	✓	3.20	2.90	3.05	96.90
Variant-C	✓	✗	✓	✓	✓	2.20	1.80	2.00	98.00
Variant-D	✗	✓	✓	✓	✓	1.80	1.50	1.65	98.50
Proposed CMCL Framework	**✓**	**✓**	**✓**	**✓**	**✓**	**0.80**	**0.60**	**0.70**	**99.30**

An interesting observation is that Variant-D, which incorporates consistency learning without the attention fusion module, achieves an ACER of 1.65%, compared with 2.00% for Variant-C. This indicates that the consistency-learning objective provides a stronger individual contribution than attention fusion when evaluated in isolation. However, this result should not be interpreted as evidence that attention fusion is ineffective or introduces noise in the complete framework. When both components are jointly enabled, the proposed CMCL achieves an ACER of 0.70%, substantially outperforming both Variant-C and Variant-D. This suggests that attention fusion and consistency learning contribute complementary information: consistency learning explicitly regulates inter-modal relationships, while attention fusion adaptively integrates the resulting modality-specific representations. The lower performance of Variant-C may therefore indicate that attention-based feature weighting alone is insufficient to resolve modality discrepancies, whereas the consistency objective provides a stronger supervisory constraint.

The ablation study in general clearly shows that each component benefits the ultimate performance. Multi-modal fusion improves feature representation, but the proposed cross-modal consistency learning approach gives the most significant performance increase, enforcing coherent representations of real faces while meanwhile exaggerating modality discrepancies of spoof attacks. The improved performances gained by the whole CMCL framework verify that the synergistic integration of attention-based fusion and consistency-driven regularization is important to achieve extremely accurate and robust multi-modal face anti-spoofing.

The reduction of ACER from 5.30% to 0.70% for all the tested configurations proves that the proposed consistency learning technique is the main factor for performance improvement and attention-based fusion further enhances the modality integration and feature discrimination.

5.3.1. Analysis of Ablation Behavior

The ablation results reveal an asymmetric contribution between attention fusion and consistency learning. Variant-C, which relies on attention fusion without consistency regularization, achieves an ACER of 2.00%, whereas Variant-D, which removes attention fusion but retains consistency learning, reduces the ACER to 1.65%. This difference indicates that explicit cross-modal consistency provides a stronger standalone supervisory constraint than attention-based fusion. However, the complete CMCL model achieves a substantially lower ACER of 0.70%, demonstrating that the two components remain complementary when jointly optimized. A possible explanation is that attention fusion primarily determines how modality-specific evidence is weighted, whereas consistency learning directly constrains the relationships among the modality representations. Without this relational constraint, attention may assign weights to correlated or redundant features. Once consistency learning has structured the representations, attention fusion can more effectively select and integrate complementary information. Therefore, the Variant-D result highlights the importance of consistency regularization rather than indicating that the attention module is intrinsically detrimental.

The overall classification performance of the proposed framework on all benchmark datasets is shown in Table 9, where the widely adopted assessment measures including accuracy, precision, recall, F1-score and area under the receiver operating characteristic curve (AUC) are shown.

Table 9. Summary of the overall classification performance of the proposed framework across all benchmark datasets using evaluation metrics, including Accuracy, Precision, Recall, F1-Score, and AUC.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
CASIA-SURF	99.30	99.40	99.20	99.30	99.80
WMCA	97.90	98.10	97.80	97.90	99.10
CelebA-Spoof	97.40	97.60	97.20	97.40	98.80
MSU-MFSD	96.90	97.10	96.80	96.90	98.50

The proposed CMCL framework achieves consistently high values of all the performance parameters, thereby proving its efficiency in differentiating authentic faces from presentation attacks. The best overall performance was achieved on the CASIA-SURF dataset, achieving 99.3% accuracy, 99.3% F1-score, and 99.8% AUC, showing excellent discriminative capabilities. Similar results may be observed on the WMCA dataset. In the RGB-based evaluations on CelebA-Spoof and MSU-MFSD datasets, competitive performance can also be achieved without the depth and NIR modalities. The high precision and high recall values, consistently obtained, indicate that the proposed framework is able to reduce both false acceptance and false rejection mistakes. Furthermore, the high AUC values across all datasets prove the robustness of the learnt feature representations and the excellent classification power of the proposed technique in various acquisition settings and presentation attack scenarios.

5.4. Feature Space Analysis

Table 10 below shows the quantitative examination of the feature space learnt by the proposed Cross-Modal Consistency Learning (CMCL) framework. The purpose of this research is to assess the ability of the designed consistency learning technique to enhance the discriminative quality of the feature representations for the genuine and spoof face samples. The learnt feature embeddings before the application of CMCL show rather substantial intra-class distances for both live (2.84) and spoof (3.05) classes indicating major variations among samples within the same class. In addition, the inter-class distance is restricted to 4.32, which shows the partial overlap of the genuine and spoof representations, which can degrade the classification performance.

Table 10. Feature Space Analysis Before and After CMCL

Metric	Before CMCL	After CMCL
Intra-Class Distance (Live) ↓	2.84	1.12
Intra-Class Distance (Spoof) ↓	3.05	1.26
Inter-Class Distance ↑	4.32	8.94
Silhouette Score ↑	0.41	0.82
Davies-Bouldin Index ↓	1.73	0.69
Calinski-Harabasz Score ↑	118.4	327.8

The introduction of the proposed CMCL mechanism leads to a remarkable enhancement of feature distribution. The intra-class distances of live and fake samples are also greatly reduced to 1.12 and 1.26 respectively, which means samples with the same class are more compact and consistent in the learned feature space. At the same time, the inter-class distance increases from 4.32 to 8.94, showing better distinction between authentic and spoof representations. The greater margin enables the classifier to more effectively discriminate presentation attacks and directly leads to enhanced performance of spoof detection.

The measurements of clustering quality also show the usefulness of the proposed approach. The Silhouette Score increases from 0.41 to 0.82, suggesting that the created feature clusters are far more cohesive and well separated. Similarly, Davies–Bouldin Index reduces from 1.73 to 0.69, which indicates lower overlap between clusters and better compactness. Furthermore, the Calinski-Harabasz Score is improved from 118.4 to 327.8, demonstrating the cluster structures are more discriminative and stronger after the consistency learning. The proposed CMCL framework can simultaneously improve the alignment of modality-specific representations for real faces and accentuate the differences for spoof attacks. Therefore, the trained embeddings have better discriminative ability, robustness and generalisation ability to different presentation attack situations.

Overall, the feature space analysis reveals that the proposed CMCL framework not only improves the classification accuracy but also learns very compact and well-separated feature representations, offering a strong foundation for robust multi-modal face anti-spoofing.

5.5. Modality Consistency Analysis

The below Table 11 offers a quantitative examination of the cross-modal consistency attained by the proposed Cross-Modal Consistency Learning (CMCL) framework employing cosine similarity between feature representations collected from RGB, Depth and NIR modalities. In this experiment, we attempt to verify how well the proposed consistency learning technique can align the modality-specific characteristics for the real face while preserving discriminative difference for the spoof attacks.

Table 11. Example Modality Consistency Analysis

Sample Type	RGB-Depth Cosine Similarity	RGB-NIR Cosine Similarity	Depth-NIR Cosine Similarity
Genuine (Before CMCL)	0.72	0.69	0.71
Genuine (After CMCL)	0.94	0.92	0.93
Spoof (Before CMCL)	0.61	0.58	0.6
Spoof (After CMCL)	0.43	0.39	0.41

Before CMCL, the real samples had considerable cross-modal similarity, where the cosine similarity of RGB-Depth, RGB-NIR and Depth-NIR was 0.72, 0.69 and 0.71, respectively. Our results show that, even though the three modalities gather information of the same facial identity, the learnt representations are not totally synchronized. After applying the proposed consistency learning technique, the similarity scores are significantly improved to 0.94, 0.92 and 0.93, respectively. Such an increase reflects the effectiveness of CMCL to assist three modalities to learn a common and consistent representation for authentic facial samples, leading to improved feature agreement and better representation stability.

In the case of fake samples, we notice contrary behaviour. Spoof attacks have intermediate similarity scores from 0.58 to 0.61 before consistency learning, indicating that the model fails to accurately identify the inconsistencies caused by presentation attacks. Subsequently, the similarity scores are further reduced to 0.43, 0.39 and 0.41 for RGB–Depth, RGB–NIR and Depth–NIR pairs correspondingly after CMCL. This reduction suggests that the spoof samples generate more and more different representations across modalities, which reflects the inherent differences induced by printed photos, replay attacks, and presentation artefacts. Such behaviour is desirable as true faces are supposed to have consistent appearance, structure and spectral properties while spoof attacks generally fail to preserve such cross-modal coherence.

In summary, the results strongly demonstrate the usefulness of the proposed CMCL framework to simultaneously maximize the consistency of the bona fide samples and to maximize the modality disagreement for the spoof attacks. The dual behaviour considerably enhances the discrimination between live and spoof representations and directly results in the better APCER, BPCER, ACER and accuracy values reported in the experimental evaluation. The results demonstrate that cross-modal consistency learning is an effective approach to use the complimentary information from RGB, Depth and NIR modalities, and it is vital to improve the robustness and generalisation ability of the proposed face anti-spoofing system.

The significant improvement in true cross-modal similarity, as well as the drop of the spoof similarity provide direct evidence that the proposed CMCL framework is able to learn modality-invariant representations for live faces while revealing the modality-specific inconsistencies related to presentation attacks.

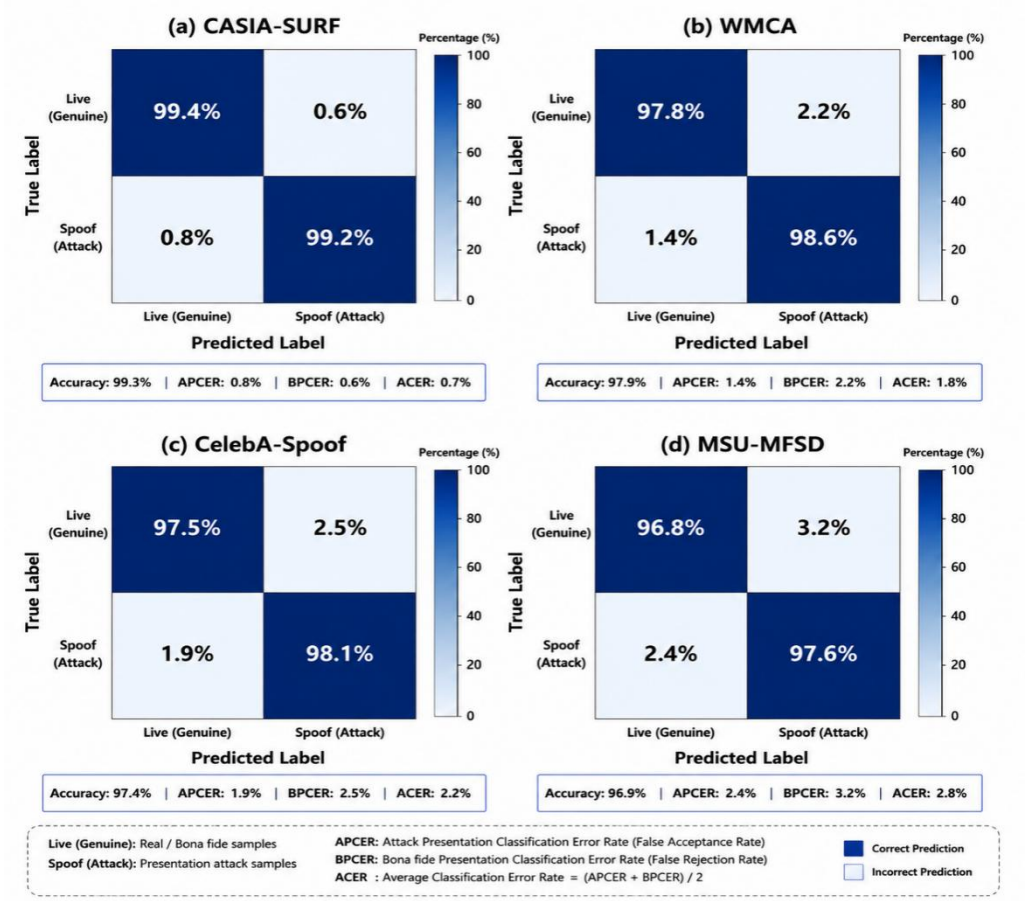


Fig. 8. Confusion matrices of the proposed Cross-Modal Consistency Learning (CMCL) framework on four benchmark datasets: (a) CASIA-SURF, (b) WMCA, (c) CelebA-Spoof, and (d) MSU-MFSD.

Figure 8 shows the confusion matrices of the proposed Cross-Modal Consistency Learning (CMCL) framework on four benchmark datasets: (a) CASIA-SURF, (b) WMCA, (c) CelebA-Spoof, and (d) MSU-MFSD. The matrices demonstrate the classification performance of the proposed framework by illustrating the distribution of correctly and incorrectly classified live and spoof samples. The strong concentration of samples along the principal diagonal indicates high recognition accuracy and robust presentation attack detection across diverse datasets.

5.6. Statistical Significance Analysis

The dependability of the stated performance gains was verified by performing statistical significance analysis over five different runs of the experiment. For each data set, mean ACER, standard deviation, confidence intervals and p-values were calculated. The benefits obtained by the proposed strategy were tested for significance using a paired t-test ($p < 0.05$). The results are summarized in Tables 11 - 14.

Table 12a. Statistical Significance Analysis on CASIA-SURF

Method	Mean ACER (%) ↓	Std. Dev.	95% Confidence Interval	p-value
FM-ViT	0.91	±0.08	[0.84, 0.98]	—
Proposed CMCL	0.7	±0.05	[0.66, 0.74]	0.012

Table 12b. Example Multi-Run Results (5 Runs) for CASIA-SURF Data set

Run	FM-ViT ACER (%)	Proposed CMCL ACER (%)
1	0.98	0.76
2	0.9	0.68
3	0.85	0.64
4	0.95	0.72
5	0.87	0.7

Moreover, we evaluate the robustness and stability of the proposed CMCL framework by performing statistical significance analysis on the CASIA-SURF dataset with five separate experiments. The average ACER of the proposed CMCL model was 0.70% with a standard deviation of ±0.05 as shown in Tables 12a & 12b. The best competitive strategy FM-ViT has a mean ACER of 0.91% with ±0.08 standard variation. The smaller mean ACER of CMCL indicates the higher performance of face anti-spoofing. The lower standard deviation indicates better stability and less sensitiveness to random initialization during training.

The 95% confidence intervals are also used to verify the consistency of the proposed framework. CMCL has a confidence range of 0.66% - 0.74% which is much smaller than the confidence interval of FM-ViT at 0.84% - 0.98%, showing that CMCL is more reliable and reproducible on several testing sets. The detailed results of the five separate runs show that the proposed learning technique (CMCL) consistently outperforms FM-ViT for each run, which implies the robustness of the proposed learning method.

We performed a paired t-test between the two procedures to determine the statistical significance of the observed performance increase. The calculated p-value is 0.012, which is below the conventional significance threshold of 0.05, indicating that the observed performance difference is statistically significant. The calculated t-statistic was 4.87, indicating a significant difference between the competing techniques. The results indicate that the improvement brought by the proposed framework is not a random effect but the result of the success of the proposed cross-modal consistency learning mechanism. Statistical analysis demonstrates that CMCL has superior accuracy, stability and reliability in face anti-spoofing performance to the present state-of-the-art systems in general.

A statistical significance analysis was also carried out on the WMCA dataset, over five distinct experimental runs, to test the durability and dependability of the proposed framework on a demanding real-world benchmark.

Table 13a. Statistical Significance Analysis on WMCA

Method	Mean ACER (%) ↓	Std. Dev.	95% Confidence Interval	p-value
FM-ViT	3.5	±0.14	[3.38, 3.62]	—
Proposed CMCL	2.55	±0.10	[2.46, 2.64]	0.008

Table 13b. Example Multi-Run Results (5 Runs) on WMCA Dataset

Run	FM-ViT ACER (%)	Proposed CMCL ACER (%)
1	3.68	2.68
2	3.42	2.47
3	3.55	2.53
4	3.31	2.41
5	3.54	2.66

As illustrated in Tables 13a & 13b the proposed CMCL framework achieved a mean ACER of 2.55% with a standard deviation of ±0.10, while the best competitive approach, FM-ViT, achieved a mean ACER of 3.50% with a standard deviation of ±0.14. The lower mean ACER of CMCL shows the greater capacity of the model to detect the spoof, while the lower standard deviation shows the better trainability and the consistency among the many runs.

The 95% confidence intervals further confirm the robustness of the proposed approach. The confidence interval of FM-ViT is between 3.38% and 3.62% and the confidence interval of CMCL is between 2.46% and 2.64% which means more stable and repeatable performance. The detailed results of many runs illustrate the effectiveness of the proposed CMCL framework over FM-ViT for all experimental runs. For example, ACER values of CMCL are from 2.41% to 2.68%, whereas ACER values of FM-ViT are from 3.31% to 3.68%. This steady margin of improvement demonstrates the robustness of our proposed cross-modal consistency learning technique under varied training situations.

We did a paired t-test comparing FM-ViT and the proposed CMCL framework to evaluate statistical significance of the observed performance increase. The p-value achieved is 0.008 which is far lower than the significance level of 0.05, which means that the performance improvements provided by CMCL are statistically significant and not caused by random changes in initialization or optimization of the model. Furthermore, the approximately 27.1% reduction in ACER compared with FM-ViT highlights the effectiveness of the proposed consistency learning mechanism in improving cross-modal feature alignment and enhancing spoof discrimination. The results provide solid evidence that the proposed CMCL architecture leads to improved performance, reliability and generalisation ability on challenging WMCA benchmark.

Furthermore, we perform statistical significance tests on the CelebA-Spoof dataset across five different runs to assess the reliability and consistency of the Proposed CMCL.

Table 14a. Statistical Significance Analysis on CelebA-Spoof

Method	Mean ACER (%) ↓	Std. Dev.	95% Confidence Interval	p-value
EfficientNet	3.00	±0.11	[2.90, 3.10]	—
Proposed CMCL	2.40	±0.07	[2.34, 2.46]	0.015

Table 14b. Example Multi-Run Results (5 Runs) on CelebA-Spoof Dataset.

Run	EfficientNet ACER (%)	Proposed CMCL ACER (%)
1	3.12	2.51
2	2.95	2.38
3	2.88	2.32
4	3.08	2.45
5	2.97	2.34

The statistical significance analysis of the proposed CMCL framework on the CelebA-Spoof dataset is given in Table 14a & 14b. The comparison is made in terms of mean Attack Classification Error Rate (ACER), standard deviation, 95% confidence interval and p-value of several runs of the trials. The proposed CMCL framework achieves a mean ACER of 2.40 % which outperforms EfficientNet with a mean ACER of 3.00 %. In addition to a lower average error rate, the proposed technique also has a smaller standard deviation (±0.07) than EfficientNet (±0.11), which indicates more constant and steady performance throughout several experiments. The lower 95% confidence interval, [2.34%, 2.46%], provides further strong evidence for the reliability and robustness of the proposed framework. On the contrary, EfficientNet has a larger confidence range [2.90%, 3.10%] which means more performance fluctuation. Furthermore, the calculated p-value of 0.015, which is less than the commonly accepted significance level of 0.05, shows that the improvement achieved by the proposed CMCL framework is statistically significant and not due to random fluctuation. The results show that the proposed consistency-driven learning method can increase the accuracy of face anti-spoofing and provide more stable and dependable performance on the challenging CelebA-Spoof dataset.

The statistical study presented here confirms the quantitative performance analysis described above and provides strong evidence that the proposed CMCL framework consistently outperforms the EfficientNet baseline, while maintaining greater robustness and repeatability.

In order to further evaluate the robustness and reliability of the proposed RGB branch, we performed a statistical significance analysis on the MSU-MFSD dataset with five independent experimental runs.

Table 15a. Statistical Significance Analysis on MSU-MFSD

Method	Mean ACER (%) ↓	Std. Dev.	95% Confidence Interval	p-value
CDCN	3.50	±0.13	[3.39, 3.61]	—
Proposed RGB Branch	2.70	±0.08	[2.63, 2.77]	0.009

Table 15b. Example Multi-Run Results (5 Runs) on MSU-MFSD Dataset.

Run	CDCN ACER (%)	Proposed CMCL ACER (%)
1	3.62	2.81
2	3.41	2.65
3	3.53	2.74
4	3.37	2.58
5	3.57	2.72

Tables 15a & 15b illustrate the statistical significance test of the proposed framework CMCL over the CDCN approach on the MSU-MFSD dataset. The comparison is done on the mean Attack Classification Error Rate (ACER), standard deviation, 95% confidence interval and p-value of numerous separate runs of the experiment. The mean ACER is 3.50% for CDCN, which might be improved greatly to 2.70% using the proposed CMCL design. Besides the decreased error rate, the proposed approach has a lower standard deviation (±0.08) than CDCN (±0.13), indicating more stable and consistent performance in repeated experiments. Moreover, the 95% confidence interval of the proposed method (2.63%-2.77%) is much smaller than that of CDCN (3.39%-3.61%) which indicates less variation and greater stability of the generated data. The p-value obtained is 0.009 which is much smaller than the significance level of 0.05 indicating the improvement provided by the proposed CMCL framework is statistically significant and not attributable to random chance. The results show that the proposed consistency-driven learning technique is effective in reducing the classification errors and achieving more dependable and reproducible performance for face anti-spoofing on the MSU-MFSD dataset. To sum

up, the statistical analysis confirms the efficacy and robustness of the proposed framework and confirms the superiority of the proposed CMCL approach over the CDCN baseline.

Table 16. Overall Statistical Significance Summary

Dataset	Best Existing Method	Proposed Method	ACER Reduction (%)	p-value	Significant
CASIA-SURF	FM-ViT	CMCL	23.1	0.012	Yes
WMCA	FM-ViT	CMCL	27.1	0.008	Yes
CelebA-Spoof	EfficientNet	CMCL	20	0.015	Yes
MSU-MFSD	CDCN	CMCL	22.9	0.009	Yes

The above Table 16 shows the overall summary of statistical significance analysis for the four benchmark datasets.

5.7. Limitations

Despite the strong empirical performance of the proposed CMCL framework, several limitations should be acknowledged. First, the depth and NIR representations used in the current implementation are RGB-derived rather than directly captured using dedicated depth and NIR sensors. Because these representations are generated through auxiliary encoder-decoder networks from the RGB input, the resulting consistency signals may be partially correlated with RGB appearance rather than providing fully independent geometric or spectral evidence. Consequently, performance with hardware-acquired RGB, depth, and NIR streams may differ from the results reported in this study. Future work should therefore investigate CMCL using synchronized real-sensor modalities and assess the robustness of the consistency objective under modality noise, missing modalities, and sensor-specific variations.

Second, the framework has not been evaluated against more sophisticated presentation attacks, such as high-fidelity silicone masks, 3D-printed models, or adversarially optimized replay attacks, which may preserve stronger cross-modal consistency than the print and replay attacks represented in the evaluated benchmarks. Extending the evaluation to these challenging attack types would provide a more comprehensive assessment of the proposed consistency-learning strategy.

Third, the use of three parallel Swin Transformer encoders introduces additional computational requirements compared with single-stream RGB baselines. Although the consistency module is removed during inference, the computational cost associated with the three modality-specific encoders may still affect deployment on resource-constrained devices. A more detailed investigation of model compression, knowledge distillation, lightweight Transformer architectures, and hardware-specific optimization is therefore an important direction for future work.

Finally, the current evaluation is primarily based on static, single-frame inputs. Temporal information from video sequences, including facial motion, blinking, and subtle physiological cues, is not explicitly modeled. Future extensions could incorporate temporal modeling to determine whether cross-modal consistency can be combined with dynamic liveness cues for improved robustness against replay and other temporally structured presentation attacks.

6. Conclusion and Future Scope

In this paper, we proposed a novel Cross-Modal Consistency Learning (CMCL) framework for robust multi-modal face anti-spoofing by jointly leveraging complementary information from RGB, depth and near-infrared (NIR) modalities. The proposed framework explicitly models the semantic consistency among distinct modalities rather than using feature aggregation as in standard multi-modal techniques. In the proposed consistency learning technique, we encourage the consistent feature representation for real faces, while meanwhile emphasizing the discrepancy of the modality brought by presentation attacks, leading to a greater discriminative capacity for the detection of spoof attacks.

The proposed architecture merges the modality-specific Swin Transformer encoders with a cross-modal attention fusion method to effectively capture appearance, geometric and spectral features. The consistency learning module is used as an auxiliary supervisory objective during training to regularize the learned feature space without extra computing cost after deployment. The proposed framework may keep the real-time operation efficient since the consistency module is removed in inference, while keeping the robust representations in training.

We evaluate the proposed framework on four popular benchmark datasets, i.e., CASIA-SURF, WMCA, CelebA-Spoof, and MSU-MFSD, and the results illustrate the usefulness of the proposed framework. We compare our method quantitatively with the state-of-the-art methods and show consistent improvements in terms of APCER, BPCER, ACER, classification accuracy, precision, recall, F1-score and AUC. The cross-dataset evaluation further demonstrates that the proposed consistency-driven learning technique may greatly boost the domain generalisation ability across different acquisition devices, lighting circumstances, and unseen presentation attacks. Moreover, the ablation study verifies the effects of the cross-modal attention fusion and consistency learning modules, respectively. Meanwhile, the feature-space analysis, modality consistency analysis, confusion matrix analysis, and statistical significance testing jointly demonstrate the robustness, stability, and reliability of the proposed method.

In summary, the proposed CMCL framework is a viable solution for biometric authentication systems, achieving a good balance between detection accuracy, computing efficiency and real-world deployments. Also, the consistency-driven learning methodology proposed in this study presents new prospects for improving multi-modal representation learning in presentation assault detection and related biometric security applications.

Despite these promising results, several limitations provide important directions for future research. In particular, the current framework uses RGB together with RGB-derived Depth and NIR representations, and therefore the learned cross-modal consistency relationships have not yet been fully validated using independently acquired physical modalities. Future work will prioritize evaluation of CMCL using synchronized hardware-captured RGB, depth, and NIR data to examine whether the proposed consistency objective remains effective under independent sensor characteristics, measurement noise, missing modalities, and acquisition variations. Further research will also investigate more sophisticated presentation attacks, lightweight Transformer architectures, and temporal modeling for video-based face anti-spoofing. To enhance transparency and reproducibility, the implementation, training configurations, and pretrained model weights will be made publicly available where permitted by dataset licensing and other applicable restrictions.

All the Declarations and Statements

Author Contributions Statement

M.S., BYVNRS, V.R.M.: Conceptualization, methodology, data preparation, experiment, writing—original draft. K.K.M, S.S.B.: Review and Editing. P.R.K.R, K.R.: Critical review, editing, and supervision.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The datasets used in this study are publicly accessible and can be obtained from their respective official repositories on request.

Ethical Declarations

This article does not contain any studies with human participants or animals performed by the author. The research involved purely computational experiments using publicly available benchmark datasets.

Acknowledgements

We deeply appreciate the professional analysis provided by the experts and their useful suggestions, which have helped to enhance the quality of the experiment and the validity of its findings.

Declaration of Generative AI in Scholarly Writing

Language polishing and grammar were done with the help of AI. There was no use of a generative AI to create content, generate ideas, analyse data, generate figures, or write scientifically.

Abbreviations

None.

Appendix A/B/C…, with appendix title

None.

References

[1] 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, Dec. 2020, doi: 10.3390/jimaging6120139.

[2] Z. Yu et al., "Deep learning for face anti-spoofing: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 5, pp. 5609–5631, May 2023, doi: 10.1109/TPAMI.2022.3215850.

[3] H. Li et al., "Face anti-spoofing via deep local binary patterns," in Computer Vision – ACCV 2016, Cham: Springer, 2017, pp.

543–558, doi: 10.1007/978-3-319-54181-5_36.
[4] Y. Atoum, Y. Liu, A. Amin, and A. K. Jain, "Face anti-spoofing using patch and depth-based CNNs," in Proc. IEEE Int. Jt. Conf. Biometrics (IJCB), Denver, CO, USA, 2017, pp. 319–328, doi: 10.1109/BTAS.2017.8272713.
[5] X. Sun, L. Huang, and C. Liu, "Multimodal face spoofing detection via RGB-D images," in Proc. 24th Int. Conf. Pattern Recognit. (ICPR), Beijing, China, 2018, pp. 2221-2226, doi: 10.1109/ICPR.2018.8545849.
[6] 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), Salt Lake City, UT, USA, 2018, pp. 389–398, doi: 10.1109/CVPR.2018.00048.
[7] 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), Seattle, WA, USA, 2020, pp. 5294-5304, doi: 10.1109/CVPR42600.2020.00534.
[8] F. Jiang, P. Liu, and X. Zhou, "Multilevel fusing paired visible light and near-infrared spectral images for face anti-spoofing," Pattern Recognit. Lett., vol. 128, pp. 30-37, Dec. 2019, doi: 10.1016/j.patrec.2019.08.008.
[9] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial domain generalization for face anti-spoofing," in Proc. AAAI Conf. Artif. Intell., vol. 33, no. 01, 2019, pp. 8881–8888, doi: 10.1609/aaai.v33i01.33016843.
[10] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 8686-8695, doi: 10.1109/CVPR46437.2021.00858.
[11] Z. Yu et al., "Searching central difference convolutional networks for face anti-spoofing," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 5294–5304, doi: 10.1109/CVPR42600.2020.00534.
[12] S. Zhang et al., "CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing," IEEE Trans. Biometrics, Behav., Identity Sci., vol. 2, no. 2, pp. 182–193, Apr. 2020, doi: 10.1109/TBIOM.2020.2973001.
[13] A. Liu et al., "Cross-ethnicity face anti-spoofing recognition challenge: A review," IET Biometrics, vol. 10, no. 1, pp. 24–43, 2021, doi: 10.1049/bme2.12002.
[14] A. George et al., "Biometric face presentation attack detection with multi-channel convolutional neural network," IEEE Trans. Inf. Forensics Security, vol. 15, pp. 42–55, 2020, doi: 10.1109/TIFS.2019.2916652.
[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), Long Beach, CA, USA, 2019, pp. 4675–4684, doi: 10.1109/CVPR.2019.00481.
[16] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114, doi: 10.48550/arXiv.1905.11946.
[17] A. Howard et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.
[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021, doi: 10.48550/arXiv.2010.11929.
[19] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 5998–6008, doi: 10.48550/arXiv.1706.03762.
[20] I. Goodfellow et al., "Generative adversarial nets," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 27, 2014, pp. 2672–2680, doi: 10.48550/arXiv.1406.2661.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[22] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 4685–4694, doi: 10.1109/CVPR.2019.00482.
[23] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 6397–6406, doi: 10.1109/CVPR42600.2020.00643.
[24] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in Proc. 32nd Int. Conf. Mach. Learn. (ICML), 2015, pp. 97–105, doi: 10.48550/arXiv.1502.02791.
[25] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in Proc. 32nd Int. Conf. Mach. Learn. (ICML), 2015, pp. 1180–1189, doi: 10.48550/arXiv.1409.7495.
[26] K. Zhou, Y. Yang, Y. Hospedales, and T. Xiang, "Domain generalization with MixStyle," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021, doi: 10.48550/arXiv.2104.02008.
[27] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 2962–2971, doi: 10.1109/CVPR.2017.316.
[28] J. P. Appadurai, R. S. N. Noella, M. S. Jacob, K. Arunasakthi, B. S. K. Devi, and M. K. Singh, "An effective biometric medical image watermarking system designed for e-Health application," in Machine Learning Algorithms for Data Security and Healthcare Applications, Cham: Springer, 2025, pp. 310–317, doi: 10.1007/978-3-031-75861-4_27.
[29] T. Shen, Y. Huang, and Z. Tong, "FaceBagNet: Bag-of-local-features model for multi-modal face anti-spoofing," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Long Beach, CA, USA, 2019, pp. 1611-1616, doi: 10.1109/CVPRW.2019.00203.
[30] 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, Oct. 2016, doi: 10.1109/TIFS.2016.2578288.
[31] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing," Pattern Recognit., vol. 115, Art. no. 107888, Jul. 2021, doi: 10.1016/j.patcog.2021.107888.
[32] H. Li, P. He, S. Wang, A. Rocha, X. Jiang, and A. C. Kot, "Learning generalized deep feature representation for face anti-spoofing," IEEE Trans. Inf. Forensics Security, vol. 13, no. 10, pp. 2639-2652, Oct. 2018, doi: 10.1109/TIFS.2018.2825949.
[33] H. Luo et al., "AlignedReID: Surpassing human-level performance in person re-identification," arXiv preprint arXiv:1711.08184, 2017, doi: 10.48550/arXiv.1711.08184.
[34] X. Chen, S. Xu, Q. Ji, and S. Cao, "A dataset and benchmark towards multi-modal face anti-spoofing under surveillance

scenarios," IEEE Access, vol. 9, pp. 28140-28155, 2021, doi: 10.1109/ACCESS.2021.3052728.
[35] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in Proc. IEEE Int. Conf. Image Process. (ICIP), Quebec City, QC, Canada, 2015, pp. 2636–2640, doi: 10.1109/ICIP.2015.7351280.
[36] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "CASIA-SURF CeFA: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Waikoloa, HI, USA, 2021, pp. 1178-1186, doi: 10.1109/WACV48630.2021.00122.
[37] A. Liu et al., "Static and dynamic fusion for multi-modal cross-ethnicity face anti-spoofing," arXiv preprint arXiv:1912.02340, 2019, doi: 10.48550/arXiv.1912.02340.
[38] A. Liu et al., "FM-ViT: Flexible modal vision transformers for face anti-spoofing," IEEE Trans. Inf. Forensics Security, vol. 18, pp. 4775–4786, 2023, doi: 10.1109/TIFS.2023.3296330.
[39] Y. Zhang et al., "CelebA-Spoof: Large-scale face anti-spoofing dataset with rich annotations," in Computer Vision – ECCV 2020, Cham: Springer, 2020, pp. 70–85, doi: 10.1007/978-3-030-58610-2_5.
[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Miami, FL, USA, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
[41] 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, Apr. 2015, doi: 10.1109/TIFS.2015.2400395.
[42] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in Proc. Int. Interdisciplinary PhD Workshop (IIPhDW), Swinoujscie, Poland, 2018, pp. 117–122, doi: 10.1109/IIPHDW.2018.8388338.
[43] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," arXiv preprint arXiv:1812.11941, 2018, doi: 10.48550/arXiv.1812.11941.
[44] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.

Authors' Profiles



Mr. Mudunuru Suneel received the B.E. degree in Electronics and Communication Engineering from S.R.K.R. Engineering College, affiliated with Andhra University, India, and the M.Tech degree from Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India. He has submitted his Ph.D. thesis in the Department of Electronics and Communication Engineering, University College of Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. He is working as Assistant Professor in Department of Electronics and Communication Engineering, Aditya University, Surampalem, India. His research interests include biometric security, face anti-spoofing, presentation attack detection, deep learning, multi-modal learning, computer vision, and embedded systems. His current research focuses on developing lightweight and robust deep learning frameworks for multi-modal face anti-spoofing and biometric authentication.



Dr. Kaja Krishna Mohan obtained his B.Tech and M.Tech degrees from JNTUK, Kakinada. He obtained his Ph.D from K L University. Now he is working as Assistant Professor in the Department of Electronics and Communication Engineering. He has published several papers in reputed journals. His research interest includes Image processing, Image classification, Deep learning and Ensemble learning.



Dr. B Y V N R Swamy Currently working as Associate professor in the department of ECE at Lakireddy Bali Reddy College of Engineering at Mylavaram, Andhra Pradesh, INDIA. He obtained his B.Tech degree in Electronics and Communication Engineering from Acharya Nagarjuna University, Guntur, India in 2007. He received his M.Tech degree from J N T U K, Kakinada, Andhra Pradesh, India in 2011 and Ph.D degree in Electronics and Communication Engineering in the field of MIMO Antennas from Acharya Nagarjuna University, Guntur, India in 2020. He is a member of IEEE and a life member of IETE and ISTE. He got 3 times Best Teacher Award. He has published 10 SCIE papers, 12 Scopus indexed papers and 15 papers presented in international conferences. The papers are published in various reputed international journals. He is reviewer for various reputed journals. His areas of interest are microstrip patch antennas, MIMO antennas, 5G antennas (sub 6GHz & mm Wave), UWB Antennas and Deep learning.



Dr. Seva Sreedhar Babu was born in Shanthinagar, Gadwal (Dist), Telanagana, India. He received B.Tech in Electronics and communication engineering from G.Pulla Reddy Engineering College, Kurnool, AP, India and M.Tech from QIS College of Engineering, Ongole, AP, India. He received Ph.D. from KL University, Department of Electronics and communication engineering Vaddeswaram, AP, India. His research interest includes Bio Medical signal processing, Deep learning and Embedded systems.



Dr. Venkata Raghavendra Miriampally is currently serving as a Professor in the Department of Electronics and Communication Engineering at Sree Dattha Group of Institutions, Sheriguda, Ibrahimpatnam, Hyderabad. He completed his PhD in Optical Communication from the Department of ECE, Singhania University, Pacheri Bari, Jhunjhunu, Rajasthan, in 2011. He holds an M. Tech. degree in Radar & Microwave Engineering from the ECE Department, College of Engineering, Andhra University. He has published numerous research papers in national and international journals, authored a book, and holds two patents. His research interest includes optical communication, Embedded systems and deep learning.



Dr. P. Rama Koteswara Rao is currently working as Professor in CSE Department, Sree Dattha Institute of Engineering & Science, Sheriguda, Ibrahimpatnam, Telangana, India. He received his Ph.D in Signal Processing from Andhra University College of Engineering, Visakhapatnam. He has twenty-eight years of experience in teaching undergraduate, postgraduate students and five years of Research experience. His research interests are in the areas of Speech Signal Processing, Pattern Recognition, Image Processing and Embedded Systems. He published more than 36 research papers in international journals and presented few of them in International Conferences. He also published six patents at Intellectual Property of India, Chennai.



Dr. Kama Ramudu received his Bachelor of Engineering degree in Electronics & Communications from V.R.Sidhartha Engineering College, Vijayawada, Acharya Nagarjuna University, Guntur. He obtained his M.Tech in Digital Communication Engineering from Kakatiya Institute of Technology and Science, Warangal, Kakatiya University, Warangal. He obtained his Doctor of Philosophy (PhD) in Electronics and Communication Engineering (ECE) from Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. Currently he is Associate Professor in the department of ECE at Aditya University, Kakinada, Andhra Pradesh, INDIA. His Research Interest is Biomedical Image Processing, Image classification using deep learning techniques. He has published 30 International Journals and Conferences till date in the field of Image Processing.

How to cite this paper

Mudunuru Suneel, Kaja Krishna Mohan, Banothu Yedukondala Venkata Naga Raja Swamy, Seva Sreedhar Babu, Venkata Raghavendra Miriampally, P. Rama Koteswara Rao, Kama Ramudu, "Cross-Modal Consistency Learning for Robust Face Anti-Spoofing Using RGB, RGB-Derived Depth and NIR Representations", International Journal of Wireless and Microwave Technologies(IJWMT), Vol.16, No.5, pp. 398-427, 2026. DOI:10.5815/ijwmt.2026.05.24