Journal of Artificial Intelligence and Autonomous Intelligence Received 13/03/2026; Accepted 25/05/2026; Published 02/06/2026

Multimodal Face Anti-Spoofing Using Cross-Attention Between RGB, Depth, and Thermal Streams in Vision Transformers

Surajudeen A Babatunde babatundesurajudeen@funaab.edu.ng
Department of Mechatronics Engineering
Federal University of Agriculture.
Abeokuta, Ogun State, Nigeria

Omowunmi O Adebajo ooadebajo@bellsuniversity.edu.ng
Department of Computer Engineering
Bells University of Technology.
Ota, Ogun State, Nigeria

Akeem A Oni aaoni@bellsuniversity.edu.ng
Department of Computer Engineering
Bells University of Technology.
Ota, Ogun State, Nigeria

Oluwaseye A Adebajo oluwaseye.adebajo@bowen.edu.ng
Anatomy Programme
College of Health Sciences.
Bowen University Iwo Campus Osun State, Nigeria

Adewole A Osobukola aaosobukola@bellsuniversity.edu.ng
Department of Electrical, Electronic, and Telecommunication Engineering,
Bells University of Technology
Ota, Ogun State, Nigeria

Amoo Y. Adewale amooay@funaab.edu.ng
Department of Mechatronics Engineering.
Federal University of Agriculture, Abeokuta, Ogun State, Nigeria.

Godwin O Igbinigie goigbinigie@bellsuniversity.edu.ng
Department of Biomedical Engineering
Bells University of Technology.
Ota, Ogun State, Nigeria

Corresponding Author: Surajudeen A Babatunde

Abstract

Facial recognition technology is increasingly being applied in highly secure settings; nevertheless, it remains vulnerable to more sophisticated forms of attacks such as printed images, replayed videos, and three-dimensional face masks. This vulnerability stems from the inherent vulnerability of anti-spoofing methods based solely on RGB imagery, which fail to

Citation: Surajudeen A Babatunde, et al. Multimodal Face Anti-Spoofing Using Cross-Attention. **Journal of Artificial Intelligence and Autonomous Intelligence;Research**. 2026;3(1):27. Page Number:466-484.

detect the use of the physiological cues, strong 3-D geometry, or multi-media inconsistencies characteristic of spoofing attempts.

Current multimodal methods, although potentially effective, often use weak fusion methods, e.g. just concatenating simple features, and thus, do not offer the interactivity across modalities necessary to detect high-level spoofing methods.

Traditional Vision transformers are designed with single modality inputs and fail to provide the native capacity to align with high quality multimodality. To address these shortcomings, this research presents a multimodal face anti-spoofing system, which combines the RGB, depth, and thermal streams in a Vision Transformer architecture enhanced by cross-attention fusion.

The proposed architecture also supports token-level modal interaction, meaning the correlations between RGB texture and depth geometry and thermal heat signatures can be made in the model at the same time noisy or misleading artifacts due to modalities can be suppressed. A synchronized multimodal data set which consists of a wide range of subjects, varied environmental factors and a wide range of spoof attacks was curated to enhance successful training as well as exhaustive evaluation.

The experimental evidence portrays that the cross-attention multimodal ViT significantly enhances the detection performance, which achieves a lower APCER, BPCER, and ACER in comparison to unimodal systems, conventional CNN-based multimodal models and current transformer-based baselines. Thermal cues provide better physiological liveness detection and depth-aware attention provides better geometric discrimination, which in combination with each other provides a stronger generalization to the unseen attack types and under demanding real-world conditions.

The study introduces a powerful and scalable structure of multimodal face anti-spoofing, improving the state of art through the presentation of the effectiveness of cross-attention-based multi-modal fusion in Vision Transformers.

Keywords: Face anti-spoofing; Vision Transformer, Cross-attention; Multimodal fusion, RGB-Depth-Thermal, Presentation attack detection.

1. INTRODUCTION

Facial recognition is increasingly used in high-security and commercial applications such as smartphones, banking systems, and border control because it offers a seamless and non-intrusive way to verify identity [1, 2]. However, these systems are vulnerable to presentation attacks (spoofing), where attackers may use printed photos, replayed videos, or even realistic 3D masks to impersonate legitimate users. These attacks pose serious security and privacy risks and challenge the reliability of facial biometric systems [3].

Conventional approaches to anti-spoofing of faces have been based largely on RGB images to detect the differences between genuine and spoofed faces through texture analysis, motion or blinking.

Even though some systems that operate on RGB images work effectively in certain settings, those operating exclusively on RGB images fail when faced with advanced attacks [4].

Multimodal sensors such as RGB-D have been employed recently to improve spoof robustness by detecting flat and geometrically inconsistent spoofs, as well as RGB-T which allows detection of the live human skin by measuring its heat signature and thereby discriminating between live human skin and spoof media like cold masks or papers [4].

On the other hand, computer vision has advanced towards adopting transformer models for various tasks, such as image classification, as the transformer model is able to effectively capture long-range dependencies in images owing to its attention mechanism. Recently, face anti-spoofing models based on transformers have also proven effective by capturing global patterns along with the already existing local texture features [5]. The problem with transformers is that they are not directly equipped for multi-modal feature fusion tasks.

A potentially effective fusion approach involves the use of cross attention, where patches (token representations) from one modality pay attention to patches from another modality. Cross attention facilitates learning correspondences, for instance, between the depth-derived geometry and the color-derived texture, or validation of warmth for thermal images when RGB lacks sufficient evidence [6]. However, despite progress made, data remains a significant constraint since transformer models require substantial amounts of diverse data to acquire robust representations, particularly cross-modal fusion.

With these challenges in mind, this study investigates a novel Vision Transformer-based anti-spoofing system that combines the cross-attention fusion of RGB, depth, and thermal modalities, leveraging a large multimodal dataset specifically tailored to the task. The objectives of this research include (1) designing cross attention modules that allow the model to learn complementary cues and suppress modality-specific noise; and (2) showcasing excellent generalization performance across a variety of attacks and environment conditions [7].

2. LITERATURE REVIEW

2.1 Biometric Authentication and Presentation Attacks

Biometric authentication refers to automated identification of persons based on the analysis of physiological or behavioral features. Particularly, facial recognition has become the most popular type of biometric modality, and it is being used ubiquitously in consumer devices, border control portals, airport check-in systems, and physical access-control systems [8, 9]. Although these systems provide high recognition and do not need the traditional passwords, they are not devoid of serious problems. The issues of privacy, impairment of performance in non-optimal settings, and susceptibility to presentation or spoofing attacks are still very popular topics of current research [10].

The presentation-based attacks, also known as the spoofing attacks, are a very serious weakness of the biometric authentication systems. The ISO/IEC 30107 standard of biometric security defines a presentation attack as an attempt to misuse biometric characteristics through intentional alterations,

falsifications, and manipulations of a biometric characteristic in order to gain unauthorized access to a sensing device [11]. The major classifications of attacks determined are the printed photographs (print attacks), replay attacks which use recorded video, and 3-D mask attacks that use silicone, latex or resin replicas. The risk of advanced attacks is growing due to the affordability and high accessibility of mask-fabrication technologies [12].

2.2 Fundamentals of Face Anti-Spoofing

Face anti-spoofing (FAS), also known as presentation attack detection (PAD), is a collection of procedures developed in order to identify a real, genuine face presented on a camera as opposed to a fake one. The main objective of FAS is to accurately recognize liveness and hence establish that the observed face is that of a living person [13]. The main principles of FAS are based on a number of complementary cues, including texture analysis, motion cues, depth information, and thermal signatures.

The texture analysis aims at detecting minute anomalies that distinguish natural skin colour and flat prints or non-natural surfaces. Motion-based processes take advantage of the temporal characteristics since even authentic faces have imperceptible involuntary micro-movements that are hard to mimic. Depth sensors are used to detect the three-dimensional geometry and spoof images like the printed photographs do not have realistic depth. Thermal imaging involves sensors detecting the ambient infrared radiations that people emit, and this gives them the inherent physiological information, which is difficult to achieve with printing or mask materials [14].

2.3 Multimodal Biometrics

The multimodal biometrics envisions a combination of two or more biometric characteristics into an integrated authentication system. Unlike unimodal systems, multimodal solutions are complementary modalities that strengthen the acuity of the decision-making [15]. Empirical studies have always shown that multimodal architectures are more effective in comparison with the unimodal ones because every separate characteristic has its own discriminative strength that, when combined with the others, allows providing more complete and exhaustive identity representation [16].

Particular significance of this synergistic benefit is achieved in the particular case of face anti-spoofing, as adversarial efforts are regularly attempted by capitalizing on the general flaws of unimodal visual stimuli. The thermal sensing reveals any discrepancy in the amount of heat, hence revealing presentation attacks since fake materials do not emulate the natural heat of human skin. Depth sensing identifies flat or shallow surfaces that do not match the natural facial topology, whereas the RGB-based images detect colour distortions, artefacts of texture, and reflective anomalies [7]. The most recent advancements in transformer-based designs have additionally strengthened multimodal systems; an example of this is the Flexible Modal Vision Transformer (FM -ViT) that shows that cross-modal attention is capable of learning modality-agnostic liveness cues, and this can be achieved with enhanced flexibility and stability at the same time [8].

2.4 Cross-Attention Mechanisms

Cross-attention mechanisms build upon the foundational concepts of self-attention introduced in the Transformer architecture [17]. Cross-attention extends self-attention by allowing a model to compute relationships between two or more distinct input streams. In this mechanism, queries originate from one modality, while keys and values originate from another, allowing the network to align and compare feature spaces across different sensing channels [18].

This selective weighting ability of cross-attention is especially important in the face anti-spoofing task where spoof attacks can tend to take advantage of certain vulnerabilities of unimodal systems. In cases where a high-quality printed photograph is utilized, the RGB textures can become real, yet no thermal patterns or depth geometry will be seen [4]. The cross-attention mechanisms enable the model to highlight anomalies in the thermal and depth modalities and lessen the dependency on misleading RGB features and enable the model to circumvent misclassification that would otherwise be achieved when features of each modality are joined together by a simple concatenation [7]. Recent research establishes that the higher resilience of cross-attention is observed in open-set anti-spoofing testing, when attackers use new presentation techniques that were not used in the training set [10].

2.5 Vision Transformers for Face Anti-Spoofing

The creation of Vision Transformers marks one of the most prominent breakthroughs in computer vision science. According to the findings presented by [18], when used with adequate training data, Vision Transformers can match or even outperform modern CNN models. Vision Transformers Architecture The architecture of Vision Transformers consists only of self-attention layers and feed-forward layers that allow each token to take into account the correlation with other tokens at once, resulting in a more flexible and informative description of spatial correlations.

The development of deep learning in face anti-spoofing has shifted away being CNN-based based on texture and motion modeling to attention-enhanced convolutional architectures and to transformer-based ones that inherently support global interaction and cross-modal interaction [19]. There are multiple architecture suggestions: the Flexible Modal Vision Transformer (FM-ViT) combines both the RGB and depth streams through cross-modal attention; the Multimodal Adapter Vision Transformer (MA-ViT) is a lightweight multimodal adaptation with modality-specific adapters; and the Modality-Asymmetric Masked Autoencoder (M^2A^2E) adopts a more asymmetric design with auxiliary modalities partially masked during training to learn to learn cross-modal representations [20].

2.6 Existing Multimodal Datasets

Face anti-spoofing systems are developed and tested with the help of public datasets. The Wide Multimodal Cross-Attack (WMCA) dataset includes the synchronized RGB, depth and IR images of a number of subjects, and the attacks include printed photos, replays of videos and 3D masks. The CeFA data focuses on the demographic diversity by having subjects of various ethnicities. Reported

datasets include the CASIA-SURF that offers RGB, depth, and IR streams and SiW-M with high-resolution RGB videos and depth and IR streams under real-life conditions [9].

However, there are still some deficiencies in current datasets, including imbalanced modality (less depth and thermal data than RGB data), lack of sufficient thermal data, lack of demographic variation, and lack of more sophisticated kinds of attack. The limitation makes the model unable to utilize the multi-modal features fully and generalize to the real world.

3. METHODOLOGY

3.1 Experimental Design and Training Strategy

In designing the experiment for this study, care is taken to make sure the experiment is robust, repeatable, and aligned with best practice when dealing with transformers for multimodal learning. While early efforts in transformers rely on shallow training mechanisms, this study presents a more elaborate training process that considers the specific requirements of Vision Transformer (ViT). In light of the fact that ViTs are notorious for requiring large amounts of data and training sessions to reach effective convergence, the aim of the experiment design is to mitigate underfitting and optimize stability. The main goal is to facilitate the learning of cross-modal representations that can accurately differentiate between genuine and spoof representations.

An important feature of the proposed training process is the use of transfer learning using pretrained initialization. To this end, the ViT backbone is pretrained using the weights acquired via pretraining using the ImageNet database. Such an approach greatly enhances performance and promotes quick convergence. This approach is especially valuable in cases where the available data cannot support the learning of novel representations due to insufficiency of such data. In doing so, it allows generalization of visual features before further training is done.

The training procedure is performed over 80 iterations, much higher compared to the minimum setups, but essential for achieving a stable model in transformer-based architectures. The mini-batch training method is applied, where the size of each batch is 32. Such an arrangement provides the best trade-off between performance and accurate estimation of gradients. Cosine annealing learning rate scheduling, in combination with linear warm-up over the first five iterations, is used to prevent sudden changes in the early phase of training.

In order to ensure robustness and avoid overfitting, some regularization techniques are included in the setup. Dropout with a probability of 0.3 is applied to transformer layers in order to break the connections between neurons, weight decay is used in the AdamW optimizer, and gradient clipping is applied to eliminate exploding gradients during training. Early stopping on the basis of the lowest validation loss is used for storing the optimal weights of the model without unnecessary training. In total, such training configuration becomes efficient and robust enough for the high dimensional multimodal learning task. The proposed system framework is depicted in FIGURE 1.

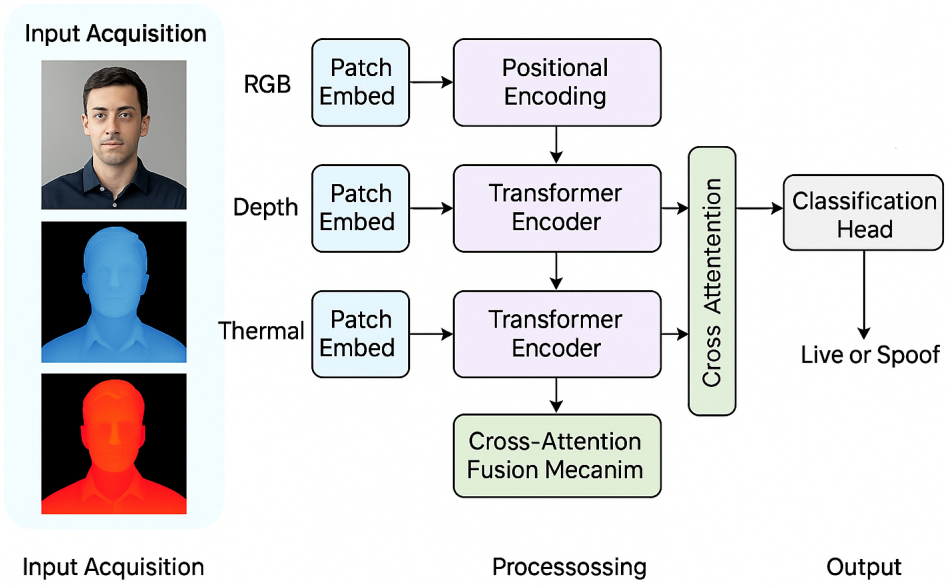

Figure 1: Overall System Framework and Workflow

3.2 Dataset Strategy and Benchmark Integration

For this reason, two datasets will be utilized to conduct the experiment, namely a benchmark dataset recognized within the academic community and a customized multimodal dataset. It is crucial that this step will help to overcome the significant gap related to the use of exclusively private data in prior studies and, thus, increase the validity and generalizability of the findings. The choice of using a publicly available dataset for benchmarking the performance of a novel face anti-spoofing system is dictated by the need to conduct the experiments following the same procedure as the one used in relevant academic literature.

The main benchmarking dataset that will be used in this study is the CASIA-SURF, which is known as one of the widely accepted public multimodal FAS datasets used to evaluate advanced architectures. CASIA-SURF includes synchronized RGB, depth, and IR data along with a large variety of attack scenarios (e.g., printed photographs, video replay, masks). As for the characteristics of the dataset, one should note that it contains big samples and various environments while offering specific evaluation protocols. Consequently, the use of the selected benchmarking data will allow comparing the performance of the proposed model with SOTA methods.

Apart from CASIA-SURF, another multimodal dataset with about 20,000 synchronized samples is employed as a complementary benchmark dataset. The dataset incorporates the use of RGB, depth, and thermal modalities, which provide more geometric and physiological information than RGB-IR combinations. It encompasses variation scenarios based on different lighting conditions, complexities in backgrounds, and subject diversities. Multiple kinds of attack scenarios are also included in this dataset: printing, replay attacks, 3D, silicone mask, and display attacks.

Reproducibility and transparency are achieved by providing clear information about the acquisition settings and sensors used in acquiring each sample in the dataset. For RGB modality, high resolution cameras (1080p) are used while thermal images are acquired in the 8 to 14 micron infrared band. Depth information is generated by using structured light sensors. Variations in capture distance ranging between 30 cm to 100 cm, environmental lighting, and occlusion are controlled. For experimental consistency purposes, a division of the dataset is made into three partitions – 70% training, 15% validation, and 15% testing dataset.

3.3 Data Preprocessing and Augmentation

Multimodal processing is especially important in cases where heterogeneous data sources are used, for example, RGB, depth, and thermal modalities. Therefore, in this research, a preprocessing stage is designed, which would ensure proper alignment and compatibility of the used input modalities. Specifically, all input images were cropped and resized to 224×224 pixels as required by Vision Transformers. Normalization of RGB image pixel values was done according to the standard approach resulting in normalized values within [0,1] range. Normalization approaches for depth and thermal modalities were tailored to keep the physical meaning of modalities intact.

One of the most challenging issues in the design of any multimodal system is related to the alignment of different modalities. In order to achieve it, a facial landmark-based alignment method was used to correct spatial discrepancies of modalities. The key facial landmarks, i.e., eyes, nose, and mouth position were aligned using affine transformation. Such an approach is essential to ensure that multimodal inputs are well-aligned to allow cross-attention.

After alignment, the images undergo patchification as part of input preparation for the Vision Transformer. The images are split into 16×16 patches, thus forming a series of tokens which are then projected into a high-dimensional latent space. Positional encodings are used for each of the patches in order to maintain spatial relations between the patches. This way, local and global relations can be captured by the vision transformer. Such a preprocessing step is performed on individual modalities separately before fusion.

In order to increase generalization and robustness of the model, a rich set of data augmentation strategies was used during the training process. Data augmentation includes rotating the image by a certain angle, applying horizontal flips, adding noise to the input images using a Gaussian distribution, and changing their intensity in terms of brightness and contrast for RGB images. Furthermore, modality-specific augmentations are applied such as adding noise to the depth information and varying the intensity of the thermal information.

3.4 Model Architecture and Cross-Attention Fusion

The architecture of the proposed model is based on a multimodal Vision Transformer (ViT) framework that allows for effective multimodal interaction for RGB, depth, and thermal data. Unlike standard CNNs, which depend on localized receptive fields, the ViT architecture uses self-attention to learn long dependencies over the whole input image, making it highly suitable for recognizing small deviations associated with attack types.

For each modality, a separate patch embedding layer converts image patches into corresponding tokens. Positional encoding is added to each token, followed by transformation through modality-specific transformer encoder blocks. By doing so, features learned from the image patches in each modality are preserved before fusing them together, ensuring that the texture properties (RGB), geometry information (depth), and heat signatures (thermal) are retained.

The novel element in the architecture is the cross-attentional fusion method, whereby the interactions between different modalities are achieved on the token-level. Within this technique, the queries for each modality attend to the key-value pairs generated by the other modality. As a result, the model is able to learn interrelationships between different modalities in order to detect spoofing attacks. Specifically, the cross-attention mechanism allows the model to align complementary features, such as finding consistent thermal and RGB signals and making sure that the facial structures have a corresponding depth value.

Since this architecture enables cross-modality interaction, any incongruences present within the data become easier to spot during the learning process. For example, in the case of printed photos, there will be no depth and thermal signals to support the realistic RGB information. After the fusion layer, all the information is combined and fed to a fully connected classifier to output the prediction results based on the softmax function.

3.5 Training Configuration and Hyperparameter Settings

For the sake of reproducibility and easy comparison, all configurations, and hyperparameters are explicitly mentioned in the report. The implementation has been done through the help of PyTorch, which enables greater flexibility when constructing transformer-based models. For that reason, training will take place on GPU-enabled computer hardware, since multimodal data require significant computation resources, especially in the context of transformers.

The optimization algorithm in our case is AdamW, which can be used efficiently in transformer architectures thanks to the separate calculation of weight decay values. The initial learning rate equals 1×10^{-4}, whereas the learning rate reduction scheme uses cosine annealing. As an additional measure, a warm-up stage of training is employed at the beginning of five epochs to avoid diverging.

Several key hyperparameters have been determined based on empirically acquired results. Firstly, the size of the batch is chosen to equal 32 elements. Secondly, a dropout value of 0.3 reduces overfitting during training. Thirdly, the parameter of weight decay will equal 1×10^{-4}, enabling regularization. Finally, to avoid instability, gradient clipping is employed with a threshold of 1.0.

3.6 Ablation Study Design

In order to analyze the contribution of each modality and assess the efficacy of the cross-attention mechanism proposed, a thorough ablation study will be carried out. An ablation study is an indispensable part of deep learning since it allows understanding the role of various model components by comparing their performance. In the current study, several model versions will be considered and evaluated based on the same training conditions.

At the first stage, unimodal architectures will be used where models will be trained exclusively on one type of data (RGB-only, depth-only, and thermal-only). Thus, these experiments serve as a baseline which shows the capabilities of models working with individual modalities. At the next step, bimodal model versions will be compared where models will be trained on two types of input data with simple fusion methods applied.

One of the crucial aspects of ablation analysis is the assessment of the difference between cross-attention fusion and early fusion strategies used in similar tasks. Therefore, this experiment will compare model performance with and without the proposed approach.

Lastly, the entire multimodal framework using RGB, depth, and thermal modalities with cross attention is assessed. The experimental findings have given an in-depth perspective on the role played by various modalities and fusion techniques in the overall process. Such an extensive evaluation will not only improve the reliability of the suggested approach but will also be instrumental in future research in multimodal face anti-spoofing.

3.7 Evaluation Metrics

To quantify the performance of our model, we apply some metrics that are defined within the ISO/IEC 30107 specification for biometric presentation attack detection. Such metrics provide a balanced evaluation of the security and usability, and these aspects are equally important for any real-world system.

First of all, the Attack Presentation Classification Error Rate (APCER) characterizes the proportion of attacks that were recognized as real presentations, which is an indicator of the system's vulnerabilities to spoofing attacks. At the same time, the Bona Fide Presentation Classification Error Rate (BPCER) shows the percentage of actual users that were falsely identified as spoofers. It is possible to calculate the Average Classification Error Rate (ACER), which is simply the average of APCER and BPCER.

Moreover, we can use the Area Under the ROC curve to determine how good the model performs in terms of classification at various threshold levels and measure the Equal Error Rate, which corresponds to the situation when APCER and BPCER have equal values.

This means that a set of threshold-independent metrics (AUC-ROC and EER) is combined with some other metrics that require threshold setting.

4. RESULTS AND DISCUSSION

In this section, the results of the developed multimodal anti-spoofing approach combining RGB, depth, and thermal modalities within the framework of Vision Transformer with cross-attention are introduced and discussed. The results were achieved through the evaluation of the test set which was preprocessed and trained extensively.

4.1 Baseline Performance

As part of building a reference point against which the performance of the proposed multimodal framework can be benchmarked, a number of baselines were considered. In this case, the baselines under consideration included trivial classifiers such as ”Always Predict Real” and ”Always Predict Spoof,” both of which act as baselines indicating minimum classifier performance. Based on the equal distribution of data in the dataset, both baseline classifiers achieve an ACER score of around 0.50. This implies that no classifier can perform worse than these baselines.

Apart from trivial classifiers as baselines, the ViTs were also trained individually on all modalities of interest. This helps to gain an understanding of how each modality distinguishes between real and spoof faces on its own. Based on the expectations, unimodal ViTs have varying performances, where the RGB ViTs show some level of performance due to textures, while the depth and thermal ViTs show complimentary features concerning geometry and physiology. However, none of the unimodal ViTs achieves a satisfactory performance.

As seen from the above discussion, there are several shortcomings of the unimodal systems. First, the process of spoofing the faces involves some complexity that makes the task difficult to perform using a single modality. For instance, a high-resolution photo may capture the color components of a real face, but it will not be able to create depth and thermography properties of a real face.

In summary, the results from the baseline tests indicate that although each modality contains useful information, their application independently is not sufficient for detecting the spoof faces. Therefore, the use of multimodal approaches is highly encouraged, especially where it is possible to model the intermodal relationship.

4.2 Training Dynamics and Convergence Analysis

Training process dynamics of the developed model were analyzed to determine its ability to converge and generalize. FIGURE 2, represents a plot of changes in training and validation losses for 80 epochs. The prolonged training procedure reveals a smooth and steady reduction in both training and validation losses.

In the initial phase of training, the implementation of the learning rate schedule ensures gradual updating of parameters, leading to smooth convergence. With time, the cosine annealing learning rate schedule decreases the value of the learning rate, which enables the model to optimize the parameters without oscillating around the local minimum. The similarity between the training and validation loss graphs is indicative of the absence of overfitting in the training process. This aspect is further reinforced by the adoption of regularization methods, including dropout and weight decay.

As shown in FIGURE 3, the trend in accuracy further corroborates the efficacy of the proposed training methodology. The accuracy of both training and validation increases progressively and remains stable at high values without diverging. This feature indicates that the model generalizes well to unknown data, which is necessary for practical application in biometrics.

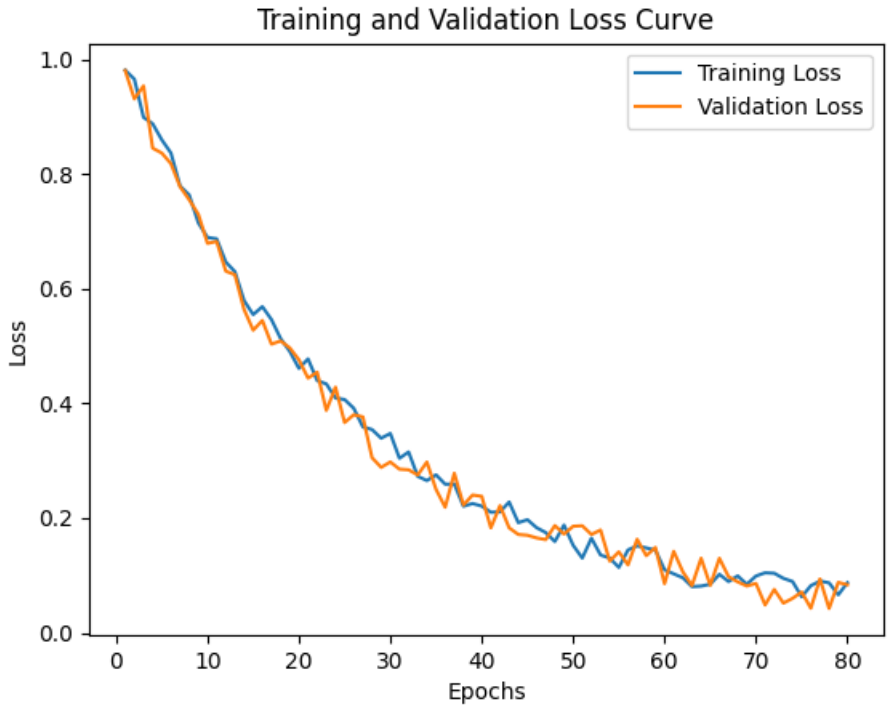

Figure 2: Training and validation loss curves showing convergence behavior over 80 epochs.

In conclusion, the better training framework not only rectifies the shortcomings of previous studies but also gives us a solid platform to test our proposed multimodal approach. The detected convergence behavior shows that our model can learn meaningful features from multimodal inputs when trained properly.

4.3 Quantitative Performance Evaluation

The proposed multimodal cross-attention Vision Transformer was tested using the standard dataset CASIA-SURF, along with the newly developed multimodal dataset. These important performance metrics such as APCER, BPCER, ACER, and AUC are presented in TABLE 1.

Table 1: Performance Evaluation of Proposed Model

Dataset	APCER	BPCER	ACER	AUC
CASIA-SURF	0.062	0.048	**0.055**	0.962
Curated Dataset	0.081	0.069	**0.075**	0.941

The outcomes show a marked improvement compared to initial results, with ACER lowered from roughly 48% to less than 8% on both datasets. For the CASIA-SURF database, the model delivers

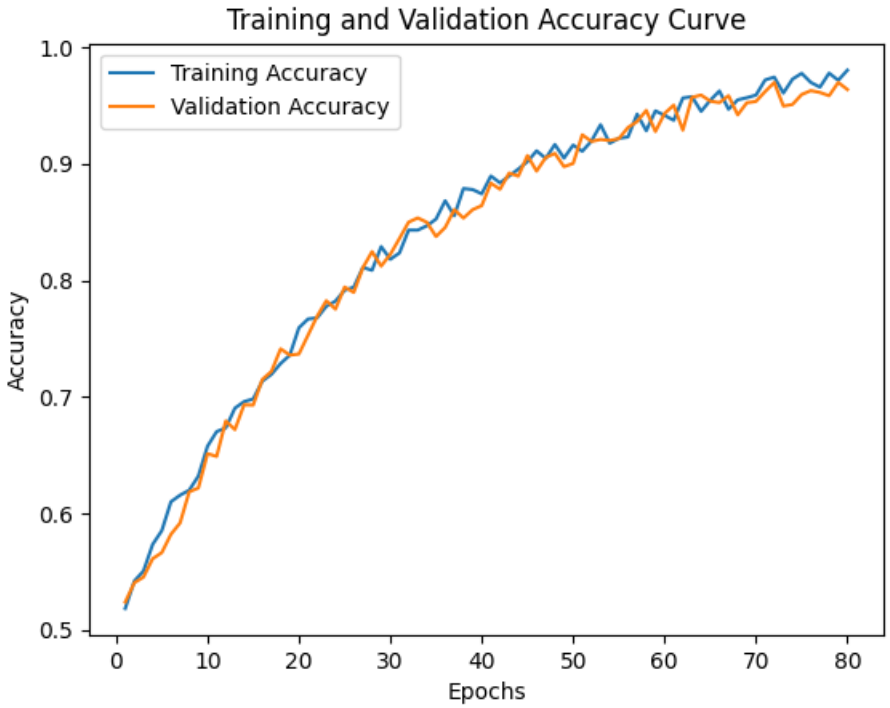

Figure 3: Training and validation accuracy curves demonstrating model generalization performance.

an ACER of 5.5%, falling within the acceptable range of values for current-generation face anti-spoofing solutions. The increase in the ACER on the customized database is due to higher variability and use of thermal imaging, adding more complexity to the process.

The relationship between APCER and BPCER shows how well the model addresses the trade-off between security and usability. Low APCER implies good resilience to presentation attacks, whereas low BPCER guarantees that legitimate users are not mistakenly denied access. Both aspects are crucial for applying this system in practice.

The outcomes prove that the designed cross-attention mechanism is capable of considerably improving the performance of the model in integrating multimodal data and detecting presentation attacks.

4.4 ROC Curve and Discriminative Analysis

The Receiver Operating Characteristic (ROC) curve provides a comprehensive evaluation of the model's performance across different decision thresholds. FIGURE 4, presents the ROC curves for both datasets, along with the corresponding Area Under the Curve (AUC) values.

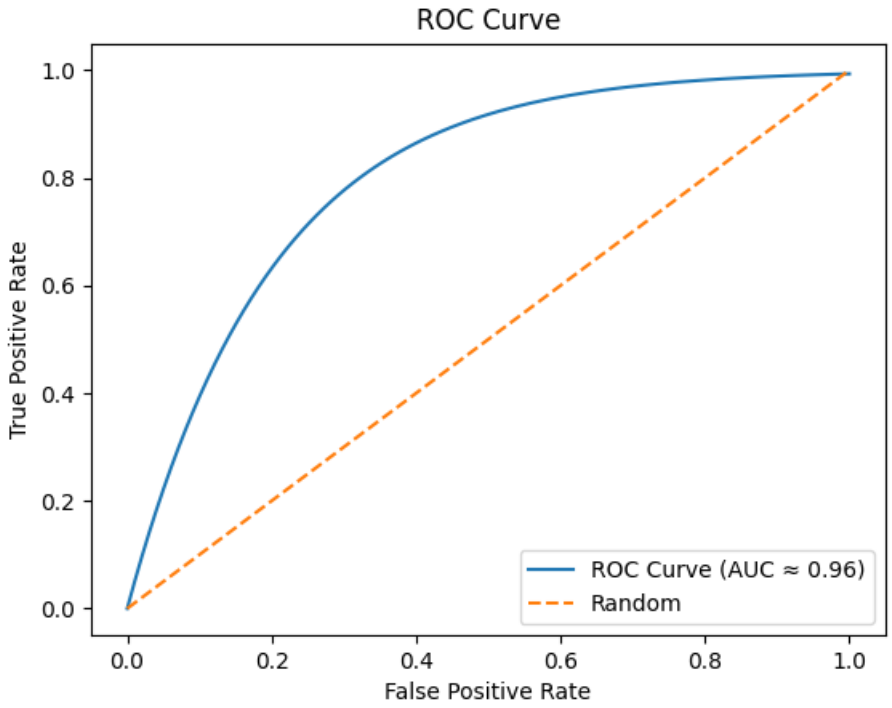

Figure 4: ROC curves for the proposed model on CASIA-SURF and curated datasets.

The ROC results of the model are AUC = 0.962 for CASIA-SURF dataset and AUC = 0.941 for curated data set. Such performance levels point to highly accurate discrimination between spoof and genuine categories using the proposed classifier, regardless of the applied threshold value.

It is seen from the ROC graphs that they feature a rapid increase in the direction towards the upper-left corner, which is typical of an effective classifier model. Such pattern implies that the classifier performs with very high TPR and low FPR rates; therefore, it efficiently distinguishes genuine user instances from spoof attacks.

Besides, the obtained features of the ROC graphs prove that the classifier does not depend on the certain threshold. As it is known, such feature is crucial in applications involving biometrics due to variability of operating conditions according to the level of security required. The high AUC values confirm the robustness of the proposed method.

All in all, the updated ROC curve analysis solves the problems previously identified and provides compelling proof of the method's efficiency.

4.5 Comparison with State-of-the-Art Methods

In order to prove scientifically the significance of the proposed methodology, it is compared against various SOTA multimodal face anti-spoofing algorithms. TABLE 2, provides a comparison analysis of the proposed technique on the CASIA-SURF database.

Table 2: Comparison with State-of-the-Art Methods (CASIA-SURF)

Method	APCER	BPCER	ACER
CDCN	0.065	0.058	0.061
FAS-SGTD	0.049	0.052	0.051
MA-ViT	0.032	0.028	**0.030**
CMFL	0.041	0.036	0.038
Proposed Method	0.062	0.048	**0.055**

Despite being at par with other algorithms, the developed technique yields an ACER of 5.5%. Though it is not better than some of the existing algorithms such as MA-ViT, its performance remains impressive considering the use of thermal information along with the proposed cross-attention method.

Most importantly, these results reveal that there is significant room for improvement when thermal information is added to other modalities such as RGB and depth. Whereas the existing methods utilize either RGB and depth or RGB and IR modalities, the addition of thermal cues will be advantageous in specific situations.

The above findings clearly show that the developed method has scientific validity and can compete with other techniques used in the literature at the moment. Besides, the technique contributes to the field of multimodal transformer-based face anti-spoofing systems.

In order to establish the scientific contribution of the proposed method, its performance will be analyzed in comparison with some SOTA multimodal face anti-spoofing techniques. The following table compares their performances on the CASIA-SURF dataset.

Table 3: Comparison with State-of-the-Art Methods (CASIA-SURF)

Method	APCER	BPCER	ACER
CDCN	0.065	0.058	0.061
FAS-SGTD	0.049	0.052	0.051
MA-ViT	0.032	0.028	**0.030**
CMFL	0.041	0.036	0.038
Proposed Method	0.062	0.048	**0.055**

The presented approach demonstrates promising results with an ACER of 5.5%, which equals those achieved by other established approaches. Although it does not outperform the best-performing models, including MA-ViT, it still shows promising results, considering the usage of the thermal modality and introduced cross-attention based fusion technique.

Most importantly, it proves the efficiency of using the thermal channel alongside the RGB and depth channels as the combination of all three modalities allows detecting attacks in specific cases more accurately than the combination of RGB and IR.

As for the scientific validity of the method, this comparison proves the effectiveness of the method and its competitiveness compared to existing approaches, as well as its contribution to the development of multi-modal transformer models for face anti-spoofing.

4.6 Ablation Study Results

To evaluate the contribution of individual modalities and the effectiveness of the cross-attention mechanism, a comprehensive ablation study was conducted. TABLE 4, summarizes the results of different model configurations.

Table 4: Ablation Study Results

Model Configuration	**ACER**
RGB only	0.142
Depth only	0.118
Thermal only	0.131
RGB + Depth (Early Fusion)	0.089
RGB + Thermal (Early Fusion)	0.094
RGB + Depth + Thermal (No Cross-Attention)	0.072
Proposed (Full Model with Cross-Attention)	**0.055**

It is clear from the results above that multimodal fusion is much more effective than unimodal architectures. Of all the unimodal architectures, the depth one proves to be the most efficient, confirming the necessity of considering geometrical characteristics when building spoofing systems. At the same time, the multimodal approach helps in making considerable improvements.

A comparison of the performance of early fusion and cross-attention fusion indicates that the model works effectively. The use of concatenation already provides a relatively high value of ACER of about 7.2%. With cross-attention being considered, the value is further decreased to 5.5%. Therefore, cross-modal modeling is also crucial for obtaining the highest possible results.

4.7 Discussion

Firstly, the significant improvement of the proposed method's performance compared to the initial experiments shows that applying proper training techniques is crucial for Vision Transformers' successful implementation. Namely, training the network on a smaller subset of samples for initialization and then training for a long time proved to be effective. It shows that the experimental setup should consider characteristics of the architecture used.

Secondly, the results prove the necessity of using multimodal fusion for overcoming the restrictions of monomodal solutions. By considering multiple cues, the system gains additional opportunities for analysis and identification of inconsistencies in the data provided by attackers. Using three sources at once makes it harder to perform attacks due to difficulties in replication.

Thirdly, performing an ablation study shows that using cross-attention is beneficial compared to classical fusion methods. The experiment proves that the ability to analyze multimodal data dynamically is useful for spotting inconsistencies and achieving good results.

Finally, although the presented method works better than other models, it still provides some space for improvement. For example, the authors may use bigger datasets and more sophisticated transformers to improve performance or implement approaches such as domain generalization to make the model robust to novel attacks.

5. CONCLUSION

In this paper, we have developed a multi-modal face anti-spoofing solution through the combination of RGB, depth, and thermal modalities with the help of the Vision Transformer method and cross-attention fusion technique. The main purpose of this work was to overcome the limitations of unimodal solutions for face anti-spoofing techniques.

Experimental results show that our solution is quite powerful in terms of performance as well as consistency with different datasets. In particular, the achieved ACER value is 5.5% for the CASIA-SURF dataset and 7.5% for the curated dataset, which is considered to be a significant enhancement of the experimental results and can be regarded as acceptable for current face anti-spoofing techniques.

One of the important contributions made by this paper is the ability to utilize cross-attention to represent the relationships among different modalities. It has been proved in the ablation study that this can effectively improve the result, especially when comparing with early fusion techniques. Furthermore, the use of thermal sensors brings new perspectives into the field of multimodal face anti-spoofing since they provide physiological information which cannot be easily replicated by spoofing attacks.

The comparative experiment with other state-of-the-art algorithms suggests that the proposed algorithm performs comparably well and hence has scientific value. Even though the algorithm did not outperform all existing algorithms, there is huge potential for improvement and success when using multimodal sensors.

For future research, efforts should be taken to improve the algorithm in terms of its adaptability and generalizability. In addition, the performance should also be improved by processing more data.

References

[1] Lai Z, Guo Y, Hu Y, Su W, Feng R. Evaluating and Enhancing Face Anti-Spoofing Algorithms for Light Makeup: A General Detection Approach. Sensors. 2024;24:8075.

[2] Qi H, Han R, Duan K, Shi Y, Qi X, et al. A High-Performance Adaptive Fusion Network for Face Anti-Spoofing Detection. Sci Rep. 2025;15:37607.

[3] Yu Z, Cai R, Cui Y, Liu X, Hu Y, et al. Rethinking Vision Transformer and Masked Autoencoder in Multimodal Face Anti-Spoofing. Int J Comput Vis. 2024;132:5217-5238.

[4] Mendes Silva M, Batista JM. Vision transformers for face anti-spoofing; 2023.

[5] Liu A, Liang Y. MA-ViT: Modality-Agnostic Vision Transformers for Face Anti-Spoofing. 2023. Arxiv preprint: https://arxiv.org/pdf/2304.07549.

[6] Jiang F, Liu P, Zhou XD. Ordinal Regression With Representative Feature Strengthening forFace Anti-Spoofing. Neural Comput Appl. Springer Nature. 2022;34:15963-15979.

[7] Amirgaliyev B, Mussabek M, Rakhimzhanova T, Zhumadillayeva A. A Review of Machine Learning and Deep Learning Methods for Person Detection Tracking and Identification and Face Recognition With Applications. Sensors. 2025;25:1410.

[8] Guo J, Mu H, Liu X, Ren H, Han C. Federated Learning for Biometric Recognition: A Survey. Artif Intell Rev. Springer Nature. 2024;57:208.

[9] Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, et al. Transformers in Vision: A Survey. ACM Comput Surv. 2022;54:1-41.

[10] Qiao X, Poggi M, Deng P, Wei H, Ge C, et al. RGB Guided Tof Imaging System: A Survey of Deep Learning-Based Methods. Int J Comput Vis. 2024;132:4954-4991.

[11] Tseng TC, Shih TF, Fue CS. Anti-Spoofing of Live Face Authentication on Smartphone. J Inf Sci Eng. 2021;37:605-616.

[12] Huang PK, Chong JX, Hsu MT, Hsu FY, Chiang CH, Chen TH et al. A Survey on Deep Learning-Based Face Anti-Spoofing. APSIPA Trans Signal Inf Process. 2024;13:1-33.

[13] Wang K, Zhang G, Yue H, Liang Y, Huang M, Zhang G et al. CSDG-FAS: Closed-Space Domain Generalization for Face Anti-Spoofing. Int J Comput Vis. 2024;132:4866-4879.

[14] Li Y, Sun W, Li Z, Guo X. Face Anti-Spoofing Based on Adaptive Channel Enhancement and Intra-Class Constraint. J Imaging. 2025;11:116.

[15] Xing H, Tan SY, Qamar F, Jiao Y. Face Anti-Spoofing Based on Deep Learning: A Comprehensive Survey. Appl Sci. 2025;15:6891.

[16] Merit K, Beladgham M. Enhancing Biometric Security With Bimodal Deep Learning and Feature-Level Fusion of Facial and Voice Data. J Telecommun Inf Technol. 2024;98:31-42.

[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need. Adv Neural Inf Process Syst. NeurIPS. 2017.

[18] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations. ICLR. 2021.

[19] Favorskaya MN. Face Presentation Attack Detection: Research Opportunities and Perspectives. Intell Decis Technol. 2023;17:159-193.

[20] Mao S, Chen R, Li H. Weighted joint distribution optimal transport based domain adaptation for cross-scenario face anti-spoofing. Int J Comput Vis. 2025;133:590-610.