Gesture Command Recognition Using Multi-Modal Attention Fusion from RGB and Thermal Image Streams

Padmavathi B.[1], Aarthi Elaveini M.[2], Kapileswar N.[3], Judy Simon[4], Reshma P Vengaloor[5*]

[1,2,5]Department of Electronics and Communication Engineering, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, India.

[3,4]Department of Electronics and Communication Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai, India.

E-mail:[1]padhmavb@srmist.edu.in, [2]aarthim@srmist.edu.in, [3]kapileswarn.sse@saveetha.com, [4]judysimon.sse@saveetha.com, [5*]reshmav1@srmist.edu.in

Abstract

Gesture recognition serves as a vital interface in human-machine communication, enabling systems to interpret and respond to user commands through natural body movements, particularly hand gestures. In the development of smart environments, assistive systems, and augmented reality applications, accurate and real-time gesture interpretation is essential. However, gesture recognition faces several challenges, including variations in lighting, background complexity, hand occlusions, and the temporal dynamics of human gestures. Existing approaches primarily depends on RGB data, making them susceptible to environmental noise and fluctuations in illumination. Additionally, some existing methods are ineffective in modeling temporal dependencies, resulting in decreased recognition reliability. To address these limitations, this research proposes a novel architecture, DMT-GAFNet, designed to enhance gesture command recognition by integrating dual-modality encoding with a guided attention fusion model. The model incorporates parallel encoders for RGB and thermal streams, alongside a modality confidence estimator that dynamically weights features based on input reliability. A lightweight GRU-based temporal encoder ensures effective

* **Corresponding author**

DOI: https://doi.org/10.36548/jiip.2025.2.007
Received: 15.04.2025, received in revised form: 14.05.2025, accepted: 28.05.2025, published: 10.06.2025

sequential modeling of gestures. The system was experimentally validated on a dataset combining HaGRID RGB data and Zenodo thermal data, encompassing six gesture classes and diverse visual conditions. Comparative analysis with existing deep learning models, including CNN-LSTM, MobileNetV2, ResNet18, EfficientNetB0, and VGG16, demonstrates that the proposed model outperforms these alternatives, achieving a precision of 0.9399, recall of 0.9484, F1-score of 0.9493, specificity of 0.9523, and accuracy of 97.05%. The proposed method not only achieves high classification accuracy under varying conditions but also exhibits significant potential for deployment in real-time gesture-based interaction systems.

Keywords: Multimodal Gesture Recognition, Guided Attention Mechanism, Thermal-RGB Fusion, Temporal Sequence Modeling, Dual-Encoder Architecture, Modality Confidence Estimation.

1. Introduction

Gesture recognition has emerged as a fundamental paradigm in the field of human-computer interaction (HCI), which enables machines to interpret human motions and respond accordingly without the need for physical contact or wearable sensors. Among the various modalities, hand gesture recognition is considered a useful method for controlling devices in smart environments, gaming, robotics, sign language translation, and augmented or virtual reality systems [1]. Its non-intrusive nature and control make it especially suitable for assistive technologies for elderly or disabled users. The basic gesture recognition process analyzes the spatial and temporal patterns of hand movements and classifies them into predefined categories [2]. Recent trends in computer vision allow systems to recognize gestures using image sequences, skeletal tracking, and even electromyography signals [3]. RGB cameras and infrared sensors are widely used as input sources for visual gesture recognition. However, the reliability of gesture detection still depends on the quality of the input stream, particularly in dynamic environments. Another challenge is the susceptibility of RGB-based models. Illumination changes and background clutter significantly affect recognition accuracy in real-world applications [4-5].

Traditional gesture recognition techniques are broadly classified into handcrafted feature-based methods and deep learning-based approaches. Traditional approaches depend on shape descriptors, motion vectors, optical flow, and skin color segmentation to extract relevant features from gesture sequences [6]. These features are then classified using machine learning algorithms such as support vector machines (SVM), hidden Markov models (HMM), k-nearest neighbors (KNN), and decision trees [7]. Although these methods are computationally efficient, they are highly sensitive and lack generalization across diverse gesture styles and environments. To overcome this, deep learning-based methods have been developed, which perform automated feature extraction and achieve improved recognition accuracy. Specifically, convolutional neural networks (CNNs) are widely utilized to extract spatial features from individual frames [8], while recurrent neural networks (RNNs), specifically long short-term memory (LSTM) and gated recurrent units (GRUs), excel at modeling temporal dependencies [9]. Hybrid models have become popular in recent times as they combine the strengths of multiple modules to achieve temporal-spatial modeling [10, 11]. However, these models generally utilize RGB data and exhibit limited performance due to lighting variations. Transformer-based models have also been explored for gesture recognition, as they provide parallel processing and long-term sequence modeling. However, their high complexity and resource requirements prevent deployment in real-time environments [12].

Multimodal learning is an emerging research area in gesture recognition that combines complementary sources of information to improve robustness and accuracy. Multimodal techniques incorporate data from RGB, depth, infrared, or thermal cameras. In some cases, inertial measurement units, audio signals, and electromyography sensors are also used. Different fusion strategies are followed in earlier and recent hybrid models [13-15]. The main objective of this research is to design a novel lightweight and context-adaptive gesture recognition model using the complementary features of RGB and thermal streams. The goal is to maintain high recognition accuracy across varied environments while preserving computational efficiency suitable for real-time applications.

To achieve the above objective, this research work introduces a novel architecture, Dual-Modality Transformer with Guided Attention Fusion Network (DMT-GAFNet), which is specifically developed for gesture command recognition. The proposed model consists of two lightweight parallel encoders designed for RGB and thermal image processing. A modality confidence estimation block is also incorporated to compute frame-level reliability scores for each stream. These scores dynamically guide the fusion module to highlight the informative modality under varying conditions. The fused features are then passed through a gated recurrent unit (GRU)-based temporal encoder, which captures sequential dependencies that are essential for gesture recognition. The proposed DMT-GAFNet introduces guided attention fusion, which provides context-aware weighting and improved adaptability. The key contributions of this research are as follows:

- A novel dual-encoder architecture is proposed for robust multimodal gesture recognition using RGB and thermal inputs. Additionaly, a guided attention fusion module is incorporated to provide dynamic weighting based on modality confidence scores. Finally, lightweight temporal modeling is presented using GRU for efficient and effective sequence recognition.
- A hybrid dataset is constructed, combining RGB and thermal samples from benchmark datasets HaGRID and Zenodo to validate the proposed model performances through various metrics.
- An extensive experimental analysis is presented to demonstrate superior performance across all metrics compared to conventional deep learning models. The proposed model achieves high recognition accuracy while maintaining computational efficiency, making it suitable for real-time gesture recognition systems.

The remaining discussions in the research work are arranged in the following order: Section 2 provides a detailed analysis of existing research works. Section 3 provides the mathematical model for the proposed work. Section 4 presents the experimental results and discussion and section 5 presents the conclusion of the research work.

2. Related Works

Recent gesture recognition methodologies utilize manually extracted features as well as automatically extracted features. However, attaining better recognition performance is still a work in progress due to modality changes and environmental variations. A detailed literature review is presented in this section, considering existing approaches to spatial encoding, temporal modeling, and multimodal fusion for gesture interpretation. The comprehensive gesture recognition analysis presented in [16] incorporates both machine learning and deep learning approaches. ML models like SVM and LSTM are trained on a custom dataset that contains grayscale gesture images of different classes. Initially, a unique vector-based mathematical model is used to extract gesture features from 3D key point coordinates. Further, the features are then classified into their respective categories. Experimental results exhibit the LSTM model superior performance over SVM with reduced cross-entropy loss.

The machine learning-based ensemble model presented in [17] for static hand gesture recognition integrates edge detection and robust classification strategies to improve accuracy across multiple datasets. The presented model includes segmentation using the Canny edge detector, feature extraction using Histogram of Oriented Gradients (HOG), and classification using an ensemble model that includes DT, LR, NB, KNN, and SVM algorithms. A majority voting mechanism is applied to aggregate the predictions for final classification. The dynamic hand gesture recognition model presented in [18] incorporates the Leap Motion (LM) sensor and machine learning techniques. The presented model extracts both static and dynamic features from hand geometry, such as velocity, direction, angle, and finger distances. Furthermore, to improve computational efficiency and remove redundant data, optimal features are selected using a random forest with the Gini Index. The resultant optimal features are then fed into an SVM classifier with an RBF kernel for classification [19].

The multi-head deep neural network (DNN) model presented in [20] for gesture-based control systems is designed for visually impaired individuals using mobile devices. The model utilizes a two-stage approach in which the first stage extracts visual features using a backbone CNN, Darknet,followed by specialized heads for gesture classification, object localization, image captioning, and zoom control. The proposed system mainly depends on RGB input, and accurate gesture recognition requires specific retraining of specific heads. The hand gesture recognition model presented in [21] incorporates a hybrid approach using the Spotted Hyena and Sine-Cosine Chimp Optimization algorithms. The model combines the optimization algorithm with a deep neural network (DNN) to attain enhanced recognition performance. The Sine-Cosine Chimp Optimization is employed in the proposed work for dimensionality reduction. The optimized feature subsets are then processed through DNN [22]. The experimental results of the presented approach exhibit the model’s better recognition accuracy over existing optimization models.

A TinyML-based gesture recognition model reported in [23] captures sensor data and preprocesses it through rasterization to convert continuous gesture trajectories into 2D grid-based images. Finally, classification is performed using a lightweight CNN model. However, the dataset used for the presented model experimentation was limited in diversity, as it was collected from a single user. This raises concerns about generalization across various style recognition. The human action and gesture recognition model presented in [24] utilizes skeleton-based modeling and a 3D pose estimation technique to acquire accurate body and hand joint coordinates. These skeletons are processed through multiple Shift-GCN models, which handle a subset of joints like body and hands and are combined through an ensemble averaging model for final classification [25]. The results show that the ensemble approach, which combines the body and single hand models, attains improved recognition accuracy over traditional methods.

The dynamic hand gesture recognition model presented in [26] incorporates a 3D separable CNN for real-time human-computer interaction. The model utilizes an appearance-based approach and employs frame differencing as a lightweight preprocessing technique to convert RGB video into grayscale motion vectors. A customized 3D separable CNN model is then incorporated to improve generalization, feature extraction, and recognition efficiency. The gesture recognition model presented in [27] utilizes an adaptive cross-modal weighting approach to improve multi-modal feature fusion from RGB-D data. The methodology integrates spatial and temporal feature fusion strategies, which are embedded into different layers of deep network backbones like C3D and 3D ResNet-50 to exhibit the recognition efficiency of the presented model.

2.1 Research Gap

A comprehensive analysis presented above exhibits the limitations and research gaps in existing gesture recognition systems and highlights the need for a more adaptive and robust approach. Most existing works mainly depend either on RGB data or inertial signals, which are often sensitive to illumination changes, occlusions, or physical sensor placement, limiting their applicability in real-world environments. While few researchers have made progress with attention mechanisms and fusion strategies, they primarily target static or unimodal gesture inputs and neglect dynamic sequence modeling across heterogeneous modalities. Furthermore, several gesture recognition models are computationally complex and unsuitable for real-time systems. Methods employing thermal or depth inputs are underutilized and often lack context-aware fusion strategies, while dynamic gestures, which are critical for command interpretation, remain underexplored. This highlights a significant gap in developing lightweight, multimodal architectures that can dynamically adapt to modality reliability and support dynamic gestures in practical settings. Addressing these challenges forms the major motivation for the proposed DMT-GAFNet framework.

3. Proposed Work

The proposed novel deep learning model, DMT-GAFNet, is designed for efficient and robust gesture command recognition utilizing RGB and thermal image streams. The architecture of the model integrates a dual-path encoder that employs a lightweight CNN Transformer for each modality, accompanied by a guided attention-based fusion mechanism and a temporal sequence processing module. The inclusion of transformer-based modules in the encoder is driven by their capacity to capture long-range spatial dependencies, thereby enhancing gesture recognition, particularly in complex backgrounds. To mitigate computational complexity in the temporal dimension, the model utilizes a Gated Recurrent Unit (GRU)-based temporal gesture encoder, as opposed to traditional transformer layers. The complete process flow of the proposed model is illustrated in Figure 1. Initially, the input streams undergo preprocessing, which includes resizing, normalization, and modality alignment to ensure spatial coherence. Subsequently, each modality is processed through its respective encoder to extract high-level features. These features are then evaluated by the guided modality confidence estimator, which generates dynamic confidence scores based on ambient conditions, guiding the subsequent cross-attention fusion process. The cross-attention fusion facilitates mutual attention exchange between RGB and thermal features, enabling the selective enhancement of modality-relevant patterns. The fused features across time steps are then forwarded to the GRU, which encodes the temporal dynamics of gestures. Finally, a compact classification head, comprising two fully connected layers with GELU activation and a SoftMax output, generates the predicted gesture command.

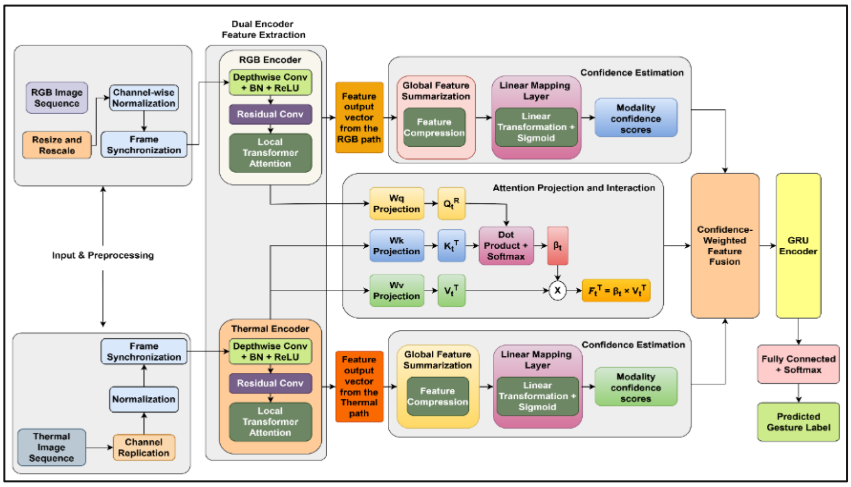
Dual Encoder Feature Extraction
RGB Image Sequence
Channel-wise Normalization
Resize and Rescale
Frame Synchronization
Input & Preprocessing
Frame Synchronization
Normalization
Thermal Image Sequence
Channel Replication
RGB Encoder
Depthwise Conv + BN + ReLU
Residual Conv
Local Transformer Attention
Thermal Encoder
Depthwise Conv + BN + ReLU
Residual Conv
Local Transformer Attention
Feature output vector from the RGB path
Feature output vector from the Thermal path
Confidence Estimation
Global Feature Summarization
Feature Compression
Linear Mapping Layer
Linear Transformation + Sigmoid
Modality confidence scores
Attention Projection and Interaction
Wq Projection
Q_t^R
Wk Projection
K_t^T
Dot Product + Softmax
β_t
Wv Projection
V_t^T
X
$F_t^T = \beta_t \times V_t^T$
Confidence Estimation
Global Feature Summarization
Feature Compression
Linear Mapping Layer
Linear Transformation + Sigmoid
Modality confidence scores
Confidence-Weighted Feature Fusion
GRU Encoder
Fully Connected + Softmax
Predicted Gesture Label

Figure 1. Process Flow of Proposed Model

3.1 Input Acquisition and Preprocessing

At each time step t , the model receives a pair of images captured simultaneously from RGB and thermal sensors. These inputs are represented as $I_t^R \in R^{H\times W\times 3}$ and $I_t^T \in R^{H\times W\times 1}$, where H and W indicates the respective spatial dimensions of the image, and the last dimension represents the number of channels. I_t^R includes three color channels (red, green, blue), while I_t^T consists of a single-channel thermal intensity map. To maintain dimensional consistency and enable fusion, the thermal frame is converted to a three-channel format using channel-wise replication which is mathematically expressed as

$$\tilde{I}_t^T = \phi(I_t^T) \in R^{H\times W\times 3} \tag{1}$$

where $\phi(\cdot)$ indicates the duplication function that copies the single thermal channel across three identical planes and creates a pseudo-color image. This alignment ensures compatibility between modalities during parallel feature extraction and preserves the original thermal information. Following this, channel adjustment is done in which both RGB and thermal images are normalized to enhance training stability and convergence speed. Normalization is carried out using the mean and standard deviation values calculated over the input which is mathematically formulated as

$$\hat{I}_t^R = \frac{I_t^R - \mu_R}{\sigma_R} \tag{2}$$

$$\hat{I}_t^T = \frac{\tilde{I}_t^T - \mu_T}{\sigma_T} \tag{3}$$

where $\mu_R \in R^3$ and $\sigma_R \in R^3$ indicates the per-channel mean and standard deviation for the RGB images. Similarly, $\mu_T \in R^3$ and $\sigma_T \in R^3$ are derived from the thermal data and are broadcasted across the replicated channels. The resulting normalized tensors, $\hat{I}_t^R \in R^{H\times W\times 3}$ and $\hat{I}_t^T \in R^{H\times W\times 3}$ are then fed into the dual-encoder modules of the proposed architecture.

3.2 Dual-Modality Feature Encoding

Once the normalized input tensors $(\widehat{I_t^R} \in R^{H\times W\times 3})$ and $(\widehat{I_t^T} \in R^{H\times W\times 3})$ are obtained, they are processed through two parallel feature encoding modules. Each encoder is designed with a hybrid lightweight convolutional layers and local transformer attention for extracting modality-specific semantic and spatial information with better efficiency. The initial stage in both encoders applies a series of depthwise separable convolutional layers to reduce spatial resolution and extract necessary patterns. Mathematically, the process is formulated as

$$F_t^{R,1} = \delta\left(\mathrm{BN}\left(W_1^R * \widehat{I_t^R} + b_1^R\right)\right) \tag{4}$$

$$F_t^{T,1} = \delta\left(\mathrm{BN}\left(W_1^T * \widehat{I_t^T} + b_1^T\right)\right) \tag{5}$$

where W_1^R, W_1^T indicates the convolutional kernels for the first convolutional layer in the RGB and thermal branches, respectively: b_1^R, b_1^T indicates the bias terms; '$*$' indicates the convolution operation, BN indicates the batch normalization to stabilize training, $\delta(\cdot)$ indicates the non-linear activation function, $(F_t^{R,1}, F_t^{T,1} \in R^{H' \times W' \times C_1})$ indicates the intermediate feature maps after the first convolutional block, where $(H' < H)$, $(W' < W)$, and (C_1) indicates the intermediate channel size. This step captures edges, textures, and basic spatial gradients, which are essential for early visual representation in both modalities. Furthermore, to compress the feature maps and introduce semantic abstraction, additional convolutional layers and residual blocks are incorporated which is mathematically formulated as

$$F_t^{R,2} = \delta\left(F_t^{R,1} + \psi\left(W_2^R * F_t^{R,1} + b_2^R\right)\right) \tag{6}$$

$$F_t^{T,2} = \delta\left(F_t^{T,1} + \psi\left(W_2^T * F_t^{T,1} + b_2^T\right)\right) \tag{7}$$

where W_2^R, W_2^T indicates the convolutional weights for the second stage, $\psi(\cdot)$ indicates the dropout used for regularization. The residual connection encourages gradient flow and reduces information loss during deeper layer propagation. The output $F_t^{R,2}, F_t^{T,2} \in R^{H'' \times W'' \times C_2}$ is further downsampled with better spatial representation and semantic information. This stage is essential for transforming raw spatial features into high-level abstractions. To capture broader spatial dependencies, a local attention transformer is integrated by flattening the intermediate tensor into non-overlapping patches, which is mathematically formulated as

$$P_t^R = \rho\left(F_t^{R,2}\right) \in R^{N \times D} \tag{8}$$

$$P_t^T = \rho\left(F_t^{T,2}\right) \in R^{N \times D} \tag{9}$$

where $\rho(\cdot)$ indicates the patch embedding operation that flattens spatial feature maps into tokens, N indicates the number of patches per frame, D indicates the token embedding dimension. These patches are then passed through a self-attention layer which is formulated as

$$Z_t^R = \text{MSA}(P_t^R) + P_t^R \tag{10}$$

$$Z_t^T = \text{MSA}(P_t^T) + P_t^T \tag{11}$$

where MSA$(\cdot)$ indicates the multi-head self-attention module. The residual connection helps to preserve low-level spatial details and the transformer-based attention allows each patch to attend to others across its local window which refines the modality-specific spatial encoding. The final spatial representations Z_t^R and Z_t^T are then passed through global average pooling and linear projection to compress the temporal embeddings. Mathematically it is formulated as

$$F_t^R = W_f^R \cdot \text{GAP}(Z_t^R) + b_f^R \tag{12}$$

$$F_t^T = W_f^T \cdot \text{GAP}(Z_t^T) + b_f^T \tag{13}$$

where GAP indicates the global average pooling over all spatial tokens, $W_f^R, W_f^T \in R^{d\times D}$ indicates the final projection matrices for each modality, $b_f^R, b_f^T \in R^d$ and indicates the bias terms, $F_t^R, F_t^T \in R^d$ indicates the final feature vectors representing RGB and thermal inputs. These vectors are modality-specific and form the core representation used for cross-modal attention and downstream temporal fusion. As each frame $t \in \{1, 2, \dots, T\}$ is processed independently, the above process generates two synchronized temporal sequences of deep feature embeddings which are expressed as

$$\mathcal{F}^{\mathcal{R}} = \{F_1^R, F_2^R, \dots, F_T^R\} \tag{14}$$

$$\mathcal{F}^{\mathcal{T}} = \{F_1^T, F_2^T, \dots, F_T^T\} \tag{15}$$

These sequences are now structurally and dimensionally aligned, ready to be fused using guided attention in the next stage. A simple illustration of the dual modality feature encoder is presented in Figure 2 for better understanding the process.

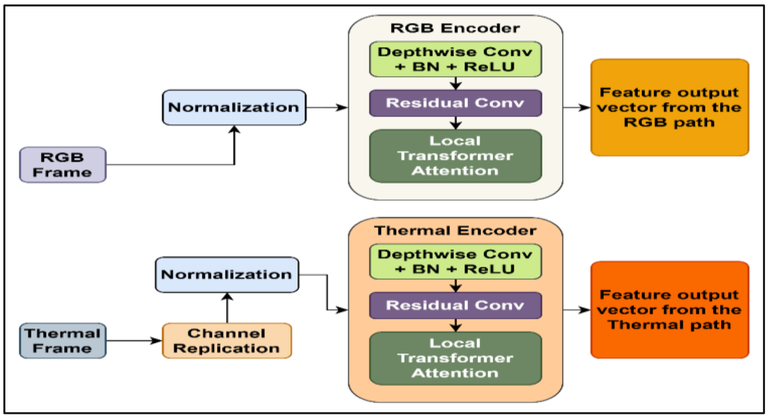
RGB Encoder
Depthwise Conv + BN + ReLU
Residual Conv
Local Transformer Attention
Normalization
RGB Frame
Feature output vector from the RGB path
Thermal Encoder
Depthwise Conv + BN + ReLU
Residual Conv
Local Transformer Attention
Normalization
Channel Replication
Thermal Frame
Feature output vector from the Thermal path

Figure 2. Dual-Modality Feature Encoding

3.3 Guided Modality Confidence Estimation

After generating spatial feature representations for both RGB and thermal frames through the dual encoders, it is processed by the Guided Modality Confidence Estimation (GMCE) module. It dynamically evaluates the reliability and informative strength of each modality at every time step and enables the system to adaptively prioritize modality contributions in the fusion process. The GMCE plays an important role in environments where RGB or thermal information may become unreliable due to illumination changes, occlusion, or sensor noise. For each time step t, the output feature vectors from the RGB and thermal encoders are denoted by $F_t^R \in R^d$ and $F_t^T \in R^d$, respectively. To compress these vectors into a suitable form a global pooling operation is applied which is mathematically formulated as

$$v_t^R = \mathrm{GAP}(F_t^R) \in R^d \tag{16}$$

$$v_t^T = \mathrm{GAP}(F_t^T) \in R^d \tag{17}$$

where $\mathrm{GAP}(\cdot)$ indicates a global average pooling function. The resulting vectors v_t^R and v_t^T serve as summarized modality-aware representations for time step t. The GMCE module evaluates the relevance of each modality through a learnable transformation followed by a sigmoid activation function. This results in modality-specific scalar confidence scores which is mathematically formulated as

$$\alpha_t^R = \sigma(W_c^R \cdot v_t^R + b_c^R) \tag{18}$$

$$\alpha_t^T = \sigma(W_c^T \cdot v_t^T + b_c^T) \tag{19}$$

where $\alpha_t^R, \alpha_t^T \in [0,1]$ indicates the confidence weights for the RGB and thermal modalities, $W_c^R, W_c^T \in R^{1\times d}$ indicates the trainable weight vectors that project the pooled features into a scalar, $b_c^R, b_c^T \in R$ indicates the scalar bias terms, $\sigma(\cdot)$ indicates the sigmoid function that ensures the output scores are in the range between 0 and 1. These confidence scores act as dynamic gates and modulates the relative influence of each modality during the fusion stage. A higher value of α_t^R indicates the greater trust in the RGB input at that frame, while a higher α_t^T reflects the stronger reliability of thermal data. To maintain alignment with the sequence of feature vectors across the gesture frames, confidence values are aggregated over time which is mathematically formulated as

$$\alpha^R = \{\alpha_1^R, \alpha_2^R, \dots, \alpha_T^R\} \tag{20}$$

$$\alpha^T = \{\alpha_1^T, \alpha_2^T, \dots, \alpha_T^T\} \tag{21}$$

These two sequences represent the per-frame modality confidence distributions and are used in the next stage to guide attention-based fusion. A simple illustration of guided modality confidence estimation is presented in Figure 3.

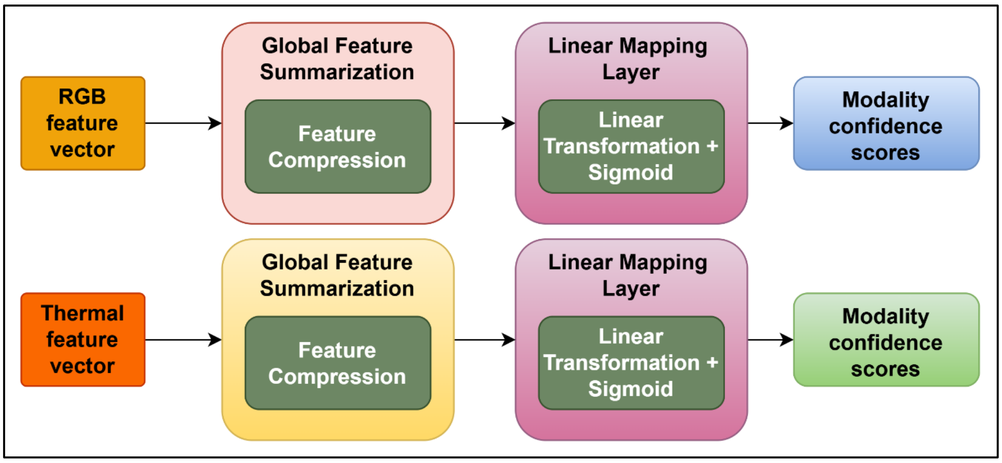
RGB feature vector
Global Feature Summarization
Feature Compression
Linear Mapping Layer
Linear Transformation + Sigmoid
Modality confidence scores
Thermal feature vector
Global Feature Summarization
Feature Compression
Linear Mapping Layer
Linear Transformation + Sigmoid
Modality confidence scores

Figure 3. Guided Modality Confidence Estimation

3.4 Cross-Attention Fusion

Following by modality-specific encoding and confidence estimation, the next process in the proposed DMT-GAFNet is the cross-attention fusion which takes the responsible for merging the RGB and thermal features in a context-sensitive manner. Thus, it allows each modality to selectively contribute information based on its relevance to the other. Instead of simply concatenating the features, this fusion block utilizes the attention principles to generate a deep, interdependent representation that captures the complementary characteristics of both inputs. For the linear projection into attention space, let $F_t^R \in R^d$ and $F_t^T \in R^d$ be the encoded RGB and thermal feature vectors which are first projected into query, key, and value spaces using learned linear transformations. Mathematically it is formulated as

$$Q_t^R = W_q F_t^R, \quad K_t^T = W_k F_t^T, \quad V_t^T = W_v F_t^T \tag{22}$$

where $Q_t^R \in R^{d_a}$ indicates the Query vector derived from RGB features, $K_t^T, V_t^T \in R^{d_a}$ indicates the Key and value vectors derived from thermal features, $W_q, W_k, W_v \in R^{d_a \times d}$ indicates the trainable projection matrices that map d dimensional input features to an attention-specific dimension d_a. These projections prepare the two modalities for interaction and allows RGB to consider the information that is important from the thermal modality based on learned relevance. Further the RGB query vector interacts with the thermal key vector to compute an attention coefficient β_t. This indicates how strongly the RGB stream should attend to the thermal features. Mathematically the attention coefficient is expressed as

$$\beta_t = softmax(Q_t^R \cdot K_t^T) \tag{23}$$

where $Q_t^R \cdot K_t^T$ indicates the dot product between query and key vectors, $SoftMax(\cdot)$ ensures the attention score present in the range [0, 1] so that it can be interpreted as a normalized weight. This scalar score reflects the semantic alignment between the two modalities at that particular frame. A higher value of β_t indicates that the thermal features closely match with RGB query and hence it should be highlighted. Further for thermal feature refinement through attention the computed attention score β_t is used. The thermal value vector is modulated to produce an aligned and filtered version that is compatible with the RGB stream. Mathematically it is formulated as

$$\tilde{F}_t^T = \beta_t \cdot V_t^T \tag{24}$$

where $\tilde{F}_t^T \in R^{d_a}$ indicates the attention-enhanced thermal feature vector. The multiplication process selectively highlights or suppresses the thermal features based on their utility to RGB. This operation is considered as a query-driven feature extraction in which RGB determines which parts of thermal data are most helpful. Finally, both the original RGB feature and the attention-modified thermal vector are integrated to produce a fused representation. This integration is performed based on the confidence scores $\alpha_t^R, \alpha_t^T \in [0,1]$ which is generated earlier. Mathematically the process is formulated as

$$F_t^F = \alpha_t^R \cdot F_t^R + \alpha_t^T \cdot \tilde{F}_t^T \tag{25}$$

where $F_t^F \in R^d$ indicates the final fused feature representation for time step t. The weighted sum balances the influence of each modality and provides more weight to the most reliable features at that frame. The fusion result is a compact, information rich embedding that captures spatial details from RGB and thermal cues modulated by relevance and confidence. This process is repeated for all frames in a gesture sequence to obtain a series of fused features which is formulated as

$$\mathcal{F}^{\mathcal{T}} = \{F_1^F, F_2^F, \ldots, F_T^F\} \tag{26}$$

This sequence retains temporal structure and will be processed in the next stage for recognition.

3.5 Temporal Modeling Using Efficient Temporal Gesture Encoder (ETGE)

After generating fused per-frame feature vectors through the cross-attention mechanism, the next objective in the proposed model is to model the temporal dynamics of gestures. Gestures are fundamentally sequential actions, meaning their interpretation mainly depends on the order and progression of frames. To capture these dependencies efficiently and without excessive computational cost, the proposed model incorporates a Gated Recurrent Unit (GRU) based encoder ETGE. This module summarizes the entire gesture sequence into a suitable temporally-aware embedding for final classification. In the previous fusion stage, each frame $t \in \{1, 2, \ldots, T\}$ contributes a fused feature vector $F_t^F \in R^d$. These vectors are collected to form the input sequence for temporal modeling which is mentioned in Equation (26). This sequence maintains the temporal order and contains semantically rich representations that reflect both modality contributions and spatial information from each time step. To process this sequence, the model utilizes a GRU cell, which captures short to medium range dependencies with minimized parameters. At each time step t, the GRU takes the current feature F_t^F and the hidden state from the previous step h_{t-1} and produces an updated hidden state h_t which is formulated as

$$h_t = \text{GRU}(F_t^F, h_{t-1}) \tag{27}$$

where $h_t \in R^{d_h}$ indicates the hidden state at time t, d_h indicates the GRU's internal state dimensionality which is selected based on memory and performance trade-offs, $\mathrm{GRU}(\cdot)$ indicates the standard gated recurrent unit update mechanism which has reset and update gates. This step is repeated for all T frames to obtain a sequence of hidden states which is mathematically expressed as

$$\mathcal{H} = \{h_1, h_2, \dots, h_T\} \tag{28}$$

The GRU operates sequentially and summarizing the sequence in a recurrent manner without any additional attention or self-attention layers. After processing all frames, the final hidden state h_T is selected as the representative temporal encoding for the entire gesture which is mathematically expressed as

$$F^G = h_T \tag{29}$$

where $F^G \in R^{d_h}$ indicates the global gesture embedding that incorporates all temporal cues from the input sequence. This vector captures both modality fusion and time evolving gestures provides a computationally lightweight and highly expressive means of modeling sequential gesture patterns.

3.6 Gesture Classification Head

Once the complete gesture sequence has been extracted into a single, temporally-aware embedding vector by the GRU-based Efficient Temporal Gesture Encoder (ETGE), the next and final task is to classify the gesture command categories. This is attained through a compact fully connected (FC) classifier which is optimized to maintain low computational overhead with strong discriminative performance. This classification head consists of a two-layer structure with a non-linear activation function between them, and ends with a SoftMax layer for multi-class probability. Consider $F^G \in R^{d_h}$ be the final output of the temporal modeling stage, which captures spatial, modality, and temporal features for a complete gesture instance. This embedding is passed through a dense projection layer to increase representational richness before the final decision which is formulated as

$$z = \delta(W_1 F^G + b_1) \tag{30}$$

where $W_1 \in R^{h \times d_h}$ indicates the weight matrix for the first fully connected layer, $b_1 \in R^h$ indicates the bias vector, h indicates the size of the hidden layer, $\delta(\cdot)$ indicates the non-linear activation function GELU (Gaussian Error Linear Unit) which is selected due to its smooth activation and improved convergence in deep networks, $z \in R^h$ indicates the intermediate feature vector with enhanced discrimination capability. The transformed vector z is then projected into a space with dimensionality equal to the number of gesture classes C, and a softmax function is applied to generate class probabilities which is mathematically formulated as

$$\hat{y} = \text{softmax}(W_2 z + b_2) \tag{31}$$

where $W_2 \in R^{C \times h}$ indicates the weight matrix for the output layer, $b_2 \in R^C$ indicates the bias vector, $\hat{y} \in R^C$ indicates the probability distribution across all gesture categories, where each element $\hat{y}_i \in [0,1]$ and $\sum_{i=1}^{C} \hat{y}_i = 1$. The SoftMax activation ensures that the output vector $\hat{y}$ can be interpreted as a categorical probability distribution. The predicted gesture is then determined by the class with the highest probability. Mathematically it is formulated as

$$Predicted\ Gesture = arg\,\max_{i} \hat{y}_i \tag{32}$$

The classification head in the proposed DMT-GAFNet architecture is designed to convert the gesture embedding into a class prediction efficiently. The summarized pseudocode for the proposed model for gesture recognition is presented as follows.

Pseudocode for the Proposed DMT-GAFNet for Gesture Recognition

```
Input:
        $\mathcal{I}^{\mathcal{R}} = \{I_1^R, I_2^R, \dots, I_T^R\} \in R^{T \times H \times W \times \mathbb{3}}$ - RGB frame sequence, $\mathcal{I}^{\mathcal{T}} = \{I_1^T, I_2^T, \dots, I_T^T\} \in R^{T \times H \times W \times \mathbb{1}}$ - Thermal frame sequence, $T$-Total number of frames, $d$ - Feature dimension, $d_a$ - Attention dimension, $d_h$- GRU hidden state size, $C$-Number of gesture classes
Output:
        $\hat{y} \in R^C$ - Predicted gesture class probabilities, Label $= \arg\max_{i} \hat{y}_i$ - Final classification output
Begin
For each frame $t = 1$ to $T$
    Normalize $I_t^R$ using channel-wise mean and std → $\hat{I}_t^R$
        Replicate thermal channel → $\tilde{I}_t^T \in R^{H \times W \times \mathbb{3}}$
            Normalize replicated thermal frame → $\hat{I}_t^T$
                Encode RGB $F_t^R = \mathcal{E}_{\mathcal{R}}(\hat{I}_t^R) \in R^d$
                Encode thermal $F_t^T = \mathcal{E}_{\mathcal{T}}(\hat{I}_t^T) \in R^d$
                    Compute per-modality weights
                        $\alpha_t^R = \sigma(W_c^R \cdot F_t^R + b_c^R)$
                        $\alpha_t^T = \sigma(W_c^T \cdot F_t^T + b_c^T)$
              Project into attention space
```

```
                $Q_t^R = W_q \cdot F_t^R, K_t^T = W_k \cdot F_t^T, V_t^T = W_v \cdot F_t^T$
            Compute attention weight
                $\beta_t = \text{Softmax}\left(Q_t^R \cdot {K_t^T}^{\top}\right)$
            Apply attention
                $\tilde{F}_t^T = \beta_t \cdot V_t^T$
          Fuse features
                $F_t^F = \alpha_t^R \cdot F_t^R + \alpha_t^T \cdot \tilde{F}_t^T$
        Append $F_t^F$ to sequence list $\mathcal{F}^{\mathcal{F}}$
      Initialize hidden state $h_0 = \vec{0} \in R^{d_h}$
        For $t = 1$ to $T$
            $h_t = \text{GRU}(F_t^F, h_{t-1})$
        End loop
       Final gesture representation $F^G = h_T \in R^{d_h}$
         Hidden projection $z = \text{GELU}(W_1 \cdot F^G + b_1)$
           Class probabilities $\hat{y} = \text{Softmax}(W_2 \cdot z + b_2)$
              Return $\hat{y}$, Label $= \arg\max_i \hat{y}_i$
           End
         End
      End
  End
```

4. Results and Discussion

The experimentation for the proposed DMT-GAFNet model was conducted using Python with the PyTorch tool and GPU acceleration to ensure efficient training and evaluation. The dataset includes RGB and thermal gesture sequences and is preprocessed to ensure spatial alignment, channel compatibility, and temporal synchronization. Each modality was independently normalized, with the thermal images replicated to match the three-channel RGB format to ensure input consistency across encoders. The model was trained and validated in a ratio of 80% training and 20% testing samples. The simulation hyperparameters used for the proposed model experimentation are presented in Table 1.

Table 1. Simulation Hyperparameters

S.No	**Parameter**	**CNN-LSTM**	**MobileNetV2**	**ResNet18**	**EfficientNetB0**	**VGG16**	**Proposed DMT-GAFNet**
1	Learning Rate	0.0005	0.001	0.001	0.001	0.007	0.001
2	Optimizer	Adam	RMSprop	SGD	RMSprop	Adam	Adam
3	Batch Size	32	32	32	32	32	32

4	Epochs	50	50	50	50	50	50
5	Dropout Rate	0.4	0.25	0.3	0.3	0.3	0.3
6	Loss	Cross-Entropy	Cross-Entropy	Cross-Entropy	Cross-Entropy	Cross-Entropy	Cross-Entropy

The experimentation of the proposed DMT-GAFNet model was carried out using a novel multi-modal dataset, constructed by combining two distinct benchmark datasets such as HaGRID [28] for RGB gesture sequences and the Zenodo Thermal Gesture Dataset [39] for thermal imagery. These datasets are independently better but lack a unified format for multi-modal gesture learning. To address this limitation and enable better validation of the proposed model's diversity, a new RGB-Thermal gesture dataset was prepared by carefully aligning the semantic gesture classes across both sources, which are suitable for dual-stream processing.

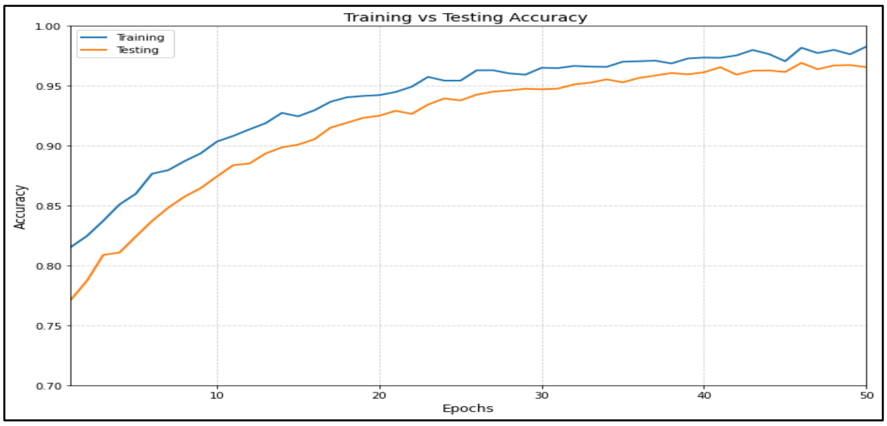

Figure 4. Accuracy Analysis

The accuracy analysis presented in Figure 4 reflects the learning ability of the proposed DMT-GAFNet model across 50 training epochs. The results demonstrate consistent improvement in both training and testing phases. The model begins with a training accuracy of 80.2% and attains a maximum of 97.8% by the final epoch. The testing accuracy starts from 76.5% and reaches a maximum of 97.05%, which closely aligns with the training accuracy. The marginal performance gap between the training and testing curves, nearly 2% after 25 epochs, indicates the model's strong generalization with no signs of overfitting. Similarly the training and testing loss depicted in Figure 5 highlights the convergence behavior of the proposed DMT-GAFNet model across 50 epochs, indicating the proposed model's progressive error minimization. This result confirms that the proposed model has successfully learned modality-aware and temporally discriminative representations for gesture recognition.

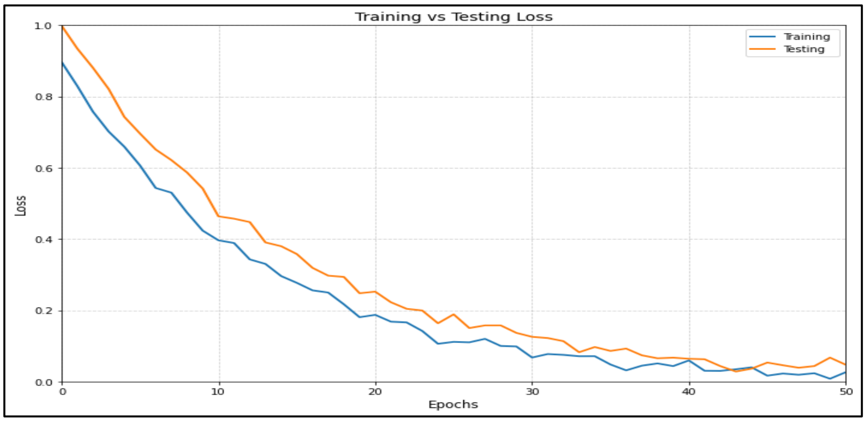

Figure 5. Loss Analysis of Proposed Model

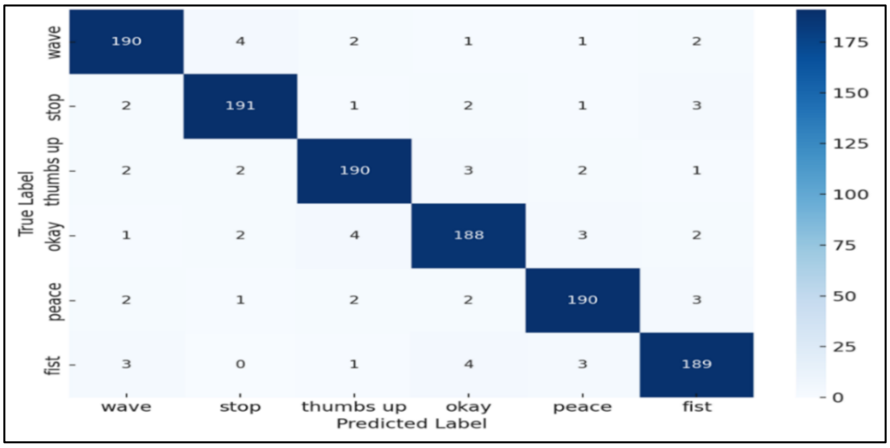

Figure 6. Confusion Matrix of Proposed Model

The test confusion matrix depicted in Figure 6 represents the classification performance of the proposed DMT-GAFNet model across six gesture categories.

Table 2. Performance Metrics of Proposed Model

Metric	Training	Testing
Precision	0.947	0.9399
Recall	0.9594	0.9484
F1-Score	0.9601	0.9493
Specificity	0.9613	0.9523

TPR	0.9519	0.9456
Accuracy	0.985	0.9705

The performance metrics of the proposed DMT-GAFNet model are presented in Table 2 for both training and testing phases. The proposed model attains a training accuracy of 98.5% and a testing accuracy of 97.05%, which clearly demonstrates its strong generalization ability. The training precision of 0.947 and the testing precision of 0.9399 indicate the proposed model's consistency in predicting true positives with minimal false positives across seen and unseen data. Overall, the proposed model results demonstrate its reliability in multimodal gesture recognition and ensure its balanced and stable classification performances.

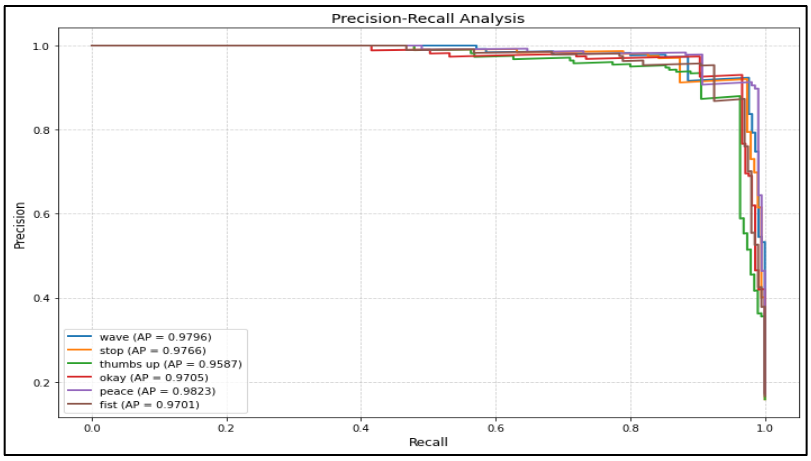

Figure 7. Precision-Recall Analysis of Proposed Model

The precision-recall curve presented above illustrates the class-wise prediction quality of the proposed DMT-GAFNet model using the Average Precision (AP) scores for each gesture. The precision curves across recall thresholds of the proposed model further confirm the model's ability to minimize false positives while preserving recall across all six gesture classes in multimodal gesture recognition.

4.1 Ablation Study

To validate the adaptive behavior of the Guided Modality Confidence Estimation (GMCE) module, a series of ablation experiments were conducted by introducing controlled variations in the RGB and thermal input streams. In the first process, Gaussian noise with varying standard deviations (σ = 0.01 to 0.05) was applied selectively to the RGB frames without introducing any changes in the thermal input. Furthermore, the confidence scores were generated for the RGB stream and compared against baseline scores obtained under noise-free conditions. The second stage of experimentation introduced random occlusions in thermal frames by masking central and edge regions of the images during inference. The confidence scores obtained for the normal as well as noise-included RGB and occlusion-included thermal streams are presented in Table 3.

Table 3. Ablation Study

Noise Type	Noise Level (σ or Mask %)	RGB Confidence	Thermal Confidence	Overall Accuracy (%)
None	0	0.72	0.76	97.05
Gaussian (RGB)	σ = 0.01	0.64	0.76	93.48
	σ = 0.03	0.52	0.76	88.76
Occlusion (Thermal)	10% masked	0.70	0.60	91.34
	30% masked	0.68	0.43	86.52

From the results given in Table 3, it can be observed that the confidence for the thermal path reduced from 0.76 to 0.60 at a 10% noise level. Furthermore, when increasing the noise level to 30%, the confidence decreases to 0.43. However, the model exhibits better accuracy by utilizing the confidence of the RGB path in both cases. These findings confirm that the GMCE module functions as planned, dynamically adjusting the weighting of input modalities based on the response to signal degradation. Thus, it supports robust gesture recognition even in noisy or uncertain environments.

To quantify the responsiveness of the Guided Modality Confidence Estimator (GMCE) to input degradation, a numerical factor called the Confidence Sensitivity Index (CSI) is formulated. This index is defined as the ratio of the proportional change in confidence score relative to the increase in noise intensity applied to a specific input modality. It is used to assess how effectively the GMCE adapts to the loss of data fidelity in either RGB or thermal streams. Let C_{mod}^{clean} and C_{mod}^{noisy} indicates the average confidence scores assigned to a modality under clean and perturbed conditions. Let ΔN indicates the incremental level of applied noise. Then, the Confidence Sensitivity Index (CSI) is obtained as follows

$$CSI_{mod} = \frac{C_{mod}^{clean} - C_{mod}^{noisy}}{\Delta N} \tag{33}$$

This formulation measures the rate at which confidence drops as the signal quality degrades. Higher CSI values indicate greater estimator sensitivity, which is desirable for maintaining robust feature fusion in dynamic environments. To evaluate CSI, the RGB and thermal inputs were independently subjected to noise with increasing severity. For the RGB stream, Gaussian noise was applied with standard deviations $\sigma = 0.01, 0.03, 0.05$. For the thermal stream, structured occlusion was introduced by masking 20%, and 30% of the frame area. For each setting, the confidence scores were recorded across the entire validation set and averaged per condition.

Table 4. Confidence Sensitivity Index (CSI) Under Controlled Perturbation

Modality	**Perturbation Type**	**Clean Confidence**	**Noisy Confidence**	**Noise Level (ΔN)**	**CSI**
RGB	Gaussian (σ = 0.05)	0.72	0.44	0.05	5.60
RGB	Gaussian (σ = 0.03)	0.72	0.52	0.03	6.67
Thermal	Occlusion (30%)	0.76	0.43	0.30	1.10
Thermal	Occlusion (20%)	0.76	0.58	0.20	0.90

The CSI values given in Table 4 depict that the GMCE responds sharply to subtle changes in RGB quality, with CSI exceeding 5.5 across noise settings. This demonstrates high sensitivity, enabling the fusion module to reduce reliance on compromised RGB frames. For thermal inputs, the CSI exhibits moderate sensitivity, which is useful when occlusion partially affects image structure but leaves temperature gradients. These results confirm that the GMCE module is not only adaptive but also responsive to signal degradation. The dynamic adjustment of modality weights ensures the system's adaptability to real-world variations and enhances the overall robustness of gesture recognition in multimodal environments.

To examine the influence of confidence-guided fusion on the quality of fused representations and overall recognition accuracy, a comparative analysis was conducted between two fusion strategies such as static fusion and the proposed confidence-guided fusion.

The static fusion method performs equal-weighted concatenation of RGB and thermal features, whereas the confidence-guided fusion utilizes dynamically generated modality weights that adapt based on the reliability of the input data. This adaptive behavior is crucial for constructing robust feature embeddings under variable environmental and sensor conditions.

To quantify the impact on feature reconstruction, a metric termed the Feature Reconstruction Alignment Score (FRAS) is introduced. This metric evaluates how closely the fused representation F aligns with the true gesture class center G_c in the latent feature space. The FRAS is computed using cosine similarity, which is mathematically formulated as

$$FRAS = \frac{F \cdot G_c}{|F| \cdot |G_c|} \tag{34}$$

where F indicates the final feature vector after fusion, and G_c indicates the centroid vector of all training samples belonging to class c. Higher FRAS values indicate that the fused features are more class-discriminative and well-aligned for classification. In addition to FRAS, standard classification metrics such as accuracy, precision, recall, and F1-score were recorded to evaluate the downstream recognition performance. Table 5 summarizes the results of both fusion approaches.

Table 5. Comparative Analysis of Fusion Methods

S.No	Metric	Static Fusion	Confidence-Guided Fusion
1	Classification Accuracy (%)	91.34	97.05
2	Precision	0.8825	0.9399
3	Reall	0.8698	0.9484
4	F1-Score	0.8710	0.9493
5	Feature Reconstruction Alignment Score (FRAS)	0.792	0.925

The results clearly demonstrate that the confidence-guided fusion mechanism significantly improves both representation quality and classification performance. Specifically, the proposed method achieves a 6.1% increase in classification accuracy, and the FRAS value increases by over 16%. This indicates better alignment between fused features and true class semantics. This improvement is obtained from the guided fusion strategy of the proposed model, which reduces the impacts of unreliable modality features and highlights the more informative input at each time step. The confidence-guided fusion selectively weights the clean input, thereby preserving the structural and temporal cues essential for accurate gesture recognition.

4.2 Statistical Validation of Performance Improvement of Confidence-Guided Fusion

To establish that the improvements observed in recognition performance are statistically attributable to the incorporation of confidence-guided fusion, hypothesis-driven statistical analysis was performed. Two fusion strategies were compared: the static fusion model, which equally combines RGB and thermal features without adaptive weighting, and the proposed guided fusion model, which uses confidence scores to modulate the contribution of each modality. Both models were trained and evaluated under identical conditions, and the classification metrics were recorded for all six gesture categories across five independent trials.

To examine whether the difference in accuracy and F1-score between the two models is statistically significant, a paired sample t-test was performed. Let X_i and Y_i denote the classification metric for the guided and static fusion models respectively, across n matched samples. The difference $D_i = X_i - Y_i$ is computed for each class. The null hypothesis H_0 assumes no significant mean difference which is expressed as $H_0\colon \mu_D = 0$ vs $H_1\colon \mu_D \neq 0$. The test statistic is mathematically expressed as

$$t = \frac{\bar{D}}{s_D/\sqrt{n}} \tag{35}$$

where $\bar{D}$ indicates the sample mean of differences, s_D and indicates the standard deviation of D_i, n indicates the number of paired observations. This formulation tests whether the guided fusion model yields a significant improvement in performance metrics over static fusion.

In addition to pairwise comparisons, a repeated measures ANOVA was conducted to assess whether the fusion strategy significantly influences multiple dependent metrics. This test evaluates whether the mean performance varies significantly across fusion methods within the same dataset configuration, while controlling for within-subject variance. The test is particularly suited here as the same set of gesture classes is evaluated across both models, thereby satisfying the assumption of related groups. A significance level of $\alpha = 0.05$ was used as the threshold for rejecting the null hypothesis.

Table 6. Statistical Test Results Comparing Static vs. Confidence-Guided Fusion

Metric	Static Fusion	Guided Fusion	Mean Difference	t-value	p-value	Significance
Accuracy (%)	91.34	97.05	5.71	4.57	0.0012	Yes
F1-Score	0.8710	0.9493	0.0783	4.21	0.0021	Yes

The test presented in Table 6 indicate that the improvements in both accuracy and F1-score introduced by the guided fusion module are statistically significant ($p < 0.05$), with confidence intervals excluding zero. Moreover, the repeated measures ANOVA revealed that the overall variation across all evaluation metrics is significantly influenced by the choice of fusion strategy, further reinforcing the value of the confidence-guided approach. These statistical results confirm that the observed performance gains are not due to random variation or dataset bias but stem directly from the adaptive fusion mechanism that modulates feature contribution based on input reliability.

4.3 Comparative Analysis with Existing Methods

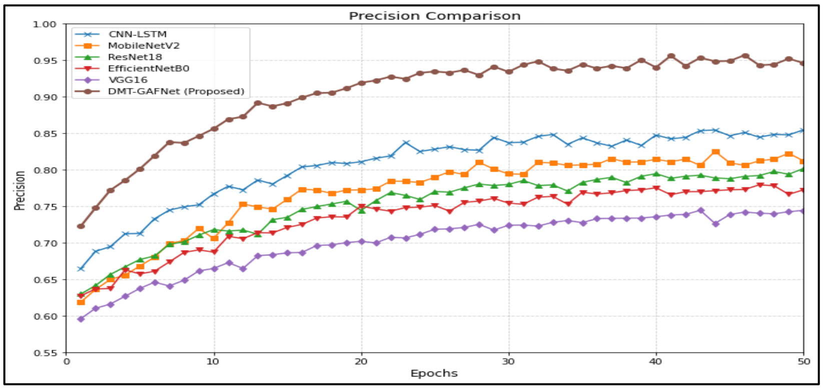

Figure 8. Precision Comparative Analysis

The precision comparison analysis presented in Figure 8 highlights the proposed model's superiority over five existing approaches. The proposed model's precision begins at

72% and quickly outperforms all others by epoch 15, reaching a maximum of 95.2% in the final epochs. The recall comparison given in Figure 9 illustrates the proposed model's ability to maintain high recall without overfitting highlights its robustness in diverse conditions, making it highly effective for gesture command recognition.

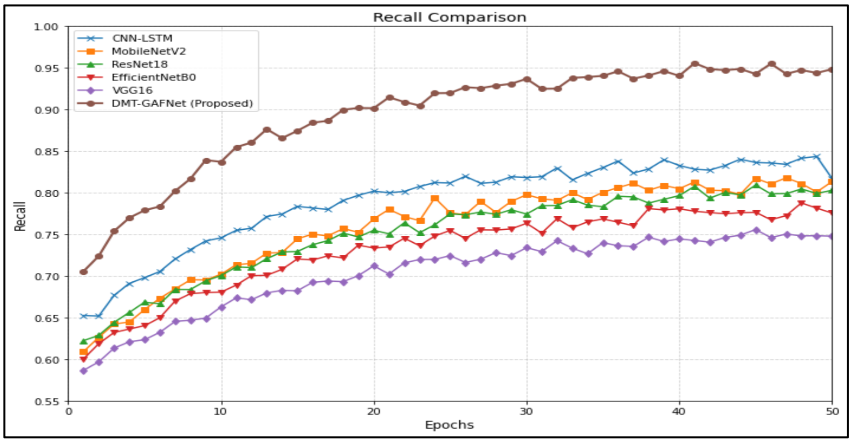

Figure 9. Recall Comparative Analysis

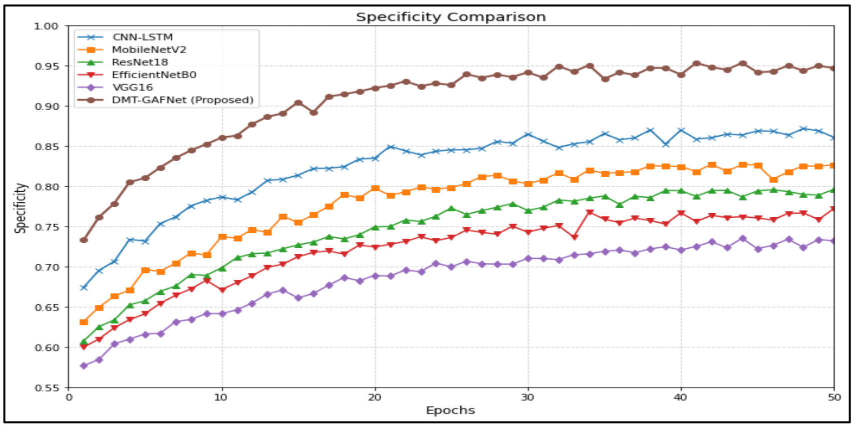

Figure 10. Specificity comparative analysis

The specificity comparison given in Figure 10 depicts the proposed model and existing models' ability to correctly identify negative gesture instances and reduce false positive predictions. The proposed DMT-GAFNet consistently outperforms other models with a maximum specificity of 95.2% by the 50th epoch. The existing CNN-LSTM exhibits 86.2%, which is lower than the proposed model. Following are MobileNetV2 and ResNet18, which exhibit 83.1% and 79.6%, respectively. EfficientNetB0 stabilizes near 77.5%, while VGG16 performs the least, exhibiting a maximum of 74.2%.

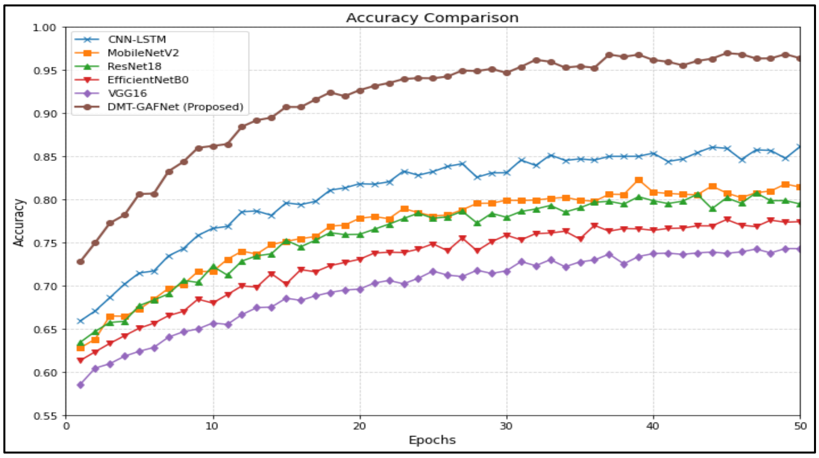

Figure 11. Accuracy Comparative Analysis

Figure 11 presents the accuracy comparison analysis of the proposed model and existing models for 50 epochs. The proposed model's accuracy increases from 72% and converges to 97%, which is better compared to existing models. The existing CNN-LSTM, which exhibits the next best performance, stabilizes around 85%, followed by MobileNetV2, which reaches 81.6%, and ResNet18 at 79.9%. The EfficientNetB0 model achieves an accuracy of 77.4%, while VGG16 lags behind at 74%, which is much lower compared to the proposed model.

Table 7. Overall Performance Metrics Analysis

Metric	**CNN-LSTM**	**MobileNet V2**	**ResNet18**	**EfficientNet B0**	**VGG16**	**Proposed DMT-GAFNet**
Precision	0.8542	0.8126	0.7942	0.7751	0.7432	0.9399
Recall	0.8369	0.8087	0.8046	0.766	0.7521	0.9484
F1-Score	0.8333	0.8107	0.7927	0.7742	0.7277	0.9493
Specificity	0.8671	0.8291	0.7956	0.7598	0.7348	0.9523
TPR	0.8511	0.8284	0.8074	0.7746	0.7464	0.9456
Accuracy	0.8556	0.816	0.7997	0.7737	0.7398	0.9705

The overall comparative analysis presented in Table 7 provides an evaluation of the proposed DMT-GAFNet model against existing methods using key performance metrics. The proposed model significantly outperforms all existing approaches, achieving a precision of 0.9399, recall of 0.9484, F1-score of 0.9493, specificity of 0.9523, TPR of 0.9456, and accuracy of 0.9705, thereby demonstrating its superiority resulting in robust classification and minimal false predictions across all gesture categories.

5. Conclusion

This research presents a novel multimodal gesture recognition model that utilizes dual-modality encoding and confidence-guided attention fusion to effectively process RGB and thermal image streams. The proposed model integrates lightweight convolutional encoders for both modalities and employs a GRU-based temporal modeling unit to capture sequential dependencies within gesture patterns. Experimental evaluations were conducted using a newly created dataset that amalgamates the HaGRID and Zenodo thermal datasets, ensuring robust training and validation across six common gesture classes. The model achieved a testing accuracy of 97.05%, with precision of 0.9399, recall of 0.9484, F1-score of 0.9493, and specificity of 0.9523, outperforming existing models such as CNN-LSTM, MobileNetV2, ResNet18, EfficientNetB0, and VGG16. Despite these superior performance metrics, the proposed model is constrained by computational overhead associated with dual-stream processing and limited testing under real-time deployment conditions. Future research directions may focus on incorporating transformer-based temporal encoders for enhanced long-range dependency modeling, as well as exploring additional modalities such as depth or radar, to facilitate the deployment of the model in real-time systems for gesture-based control in assistive and industrial environments.

References

[1] Abdirahman Osman Hashi, Siti Zaiton Mohd Hashim, and Azurah Bte Asama, “A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024,” IEEE Access, vol. 12, 2024, 143599- 143626.

[2] Meng, Yuting, Haibo Jiang, Nengquan Duan, and Haijun Wen. "Real-Time Hand Gesture Monitoring Model Based on MediaPipe’s Registerable System." Sensors 24, no. 19 (2024): 6262.

[3] Rahman, Md Mijanur, Ashik Uzzaman, Fatema Khatun, Md Aktaruzzaman, and Nazmul Siddique. "A comparative study of advanced technologies and methods in hand gesture analysis and recognition systems." Expert Systems with Applications (2024): 125929.

[4] Sarma, Debajit, and Manas Kamal Bhuyan. "Methods, databases and recent advancement of vision-based hand gesture recognition for hci systems: A review." SN Computer Science 2, no. 6 (2021): 436.

[5] Brenner, Martin, Napoleon H. Reyes, Teo Susnjak, and Andre LC Barczak. "RGB-D and thermal sensor fusion: A systematic literature review." IEEE Access 11 (2023): 82410-82442.

[6] Qi, Jing, Li Ma, Zhenchao Cui, and Yushu Yu. "Computer vision-based hand gesture recognition for human-robot interaction: a review." Complex & Intelligent Systems 10, no. 1 (2024): 1581-1606.

[7] Bhushan, Shashi, Mohammed Alshehri, Ismail Keshta, Ashish Kumar Chakraverti, Jitendra Rajpurohit, and Ahed Abugabah. "An experimental analysis of various machine learning algorithms for hand gesture recognition." Electronics 11, no. 6 (2022): 968.

[8] Reddy, Veluru Karthik, and Vanapalli Durga Prasanth. "Hand Gesture Recognition Using Convolutional Neural Networks." (2024).

[9] Toro-Ossaba, Alejandro, Juan Jaramillo-Tigreros, Juan C. Tejada, Alejandro Peña, Alexandro López-González, and Rui Alexandre Castanho. "LSTM recurrent neural network for hand gesture recognition using EMG signals." Applied Sciences 12, no. 19 (2022): 9700.

[10] Ur Rehman, Muneeb, Fawad Ahmed, Muhammad Attique Khan, Usman Tariq, Faisal Abdulaziz Alfouzan, Nouf M Alzahrani, and Jawad Ahmad. "Dynamic hand gesture recognition using 3D-CNN and LSTM networks." Computers, Materials & Continua 70, no. 3 (2021).

[11] Kapileswar, Nellore, Judy Simon, Kota Sirisha, Bezawada Raja Pujitha, Lekkala Charan Sai Kumar, and Chappagadda Harish. "Enhanced Agricultural Monitoring Through Hyperspectral Imaging and Advanced Machine Learning Techniques." In 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), IEEE, 2024, 1495-1502.

[12] Padmavathi, B., K. R. Sushkrutha, Judy Simon, M. Aarthi Elaveini, and N. Kapileswar. "Implementation of a Health Monitoring Sytem using Sensors and RedTacton." In 2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), IEEE, 2023, 384-390.

[13] Oleh, Ugonna, Roman Obermaisser, and Abu Shad Ahammed. "A Review of Recent Techniques for Human Activity Recognition: Multimodality, Reinforcement Learning, and Language Models." Algorithms 17, no. 10 (2024): 434.

[14] Zhang, Zhi-Yuan, Hao Ren, Hao Li, Kang-Hui Yuan, and Chu-Feng Zhu. "Static gesture recognition based on thermal imaging sensors." The Journal of Supercomputing 81, no. 4 (2025): 1-21.

[15] Kumar, Ushus S., Judy Simon, Reshma P. Vengaloor, and M. Aarthi Elaveini. "Image Processing Techniques in Thermal and Non-thermal Images." In Second International Conference on Image Processing and Capsule Networks: ICIPCN 2021 2, Springer International Publishing, 2022, 533-544.

[16] Mukhanov, Samat, Raissa Uskenbayeva, Abd A. Rakhim, Akbota Akim, and Symbat Mamanova. "Gesture recognition of the Kazakh alphabet based on machine and deep learning models." Procedia Computer Science 241 (2024): 458-463.

[17] Alteaimi, Amal, and Mohamed Ben Othman. "Robust Interactive Method for Hand Gestures Recognition Using Machine Learning." Computers, Materials & Continua. 72 (2022): 577-595.

[18] Shin, Jungpil, Md Al Mehedi Hasan, Md Maniruzzaman, Taiki Watanabe, and Issei Jozume. "Dynamic Hand Gesture-Based Person Identification Using Leap Motion and Machine Learning Approaches." Computers, Materials & Continua 79, no. 1 (2024).

[19] Kapileswar, N., Judy Simon, K. Kavitha Devi, Phani Kumar Polasi, Dasari Naga Vinod, and Chappagadda Harish. "An Intelligent Emotion Recognition System based on Speech Terminologies using Artificial Intelligence Assisted Learning Scheme." In 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), IEEE, 2024, 1-7.

[20] Alashhab, Samer, Antonio Javier Gallego, and Miguel Ángel Lozano. "Efficient gesture recognition for the assistance of visually impaired people using multi-head neural networks." Engineering Applications of Artificial Intelligence 114 (2022): 105188.

[21] Mohyuddin, Hassan, Syed Kumayl Raza Moosavi, Muhammad Hamza Zafar, and Filippo Sanfilippo. "A comprehensive framework for hand gesture recognition using hybrid-metaheuristic algorithms and deep learning models." Array 19 (2023): 100317.

[22] Oloyede, Muhtahir O., Gerhard P. Hancke, and Nellore Kapileswar. "Evaluating the effect of occlusion in face recognition systems." In 2017 IEEE AFRICON, IEEE, 2017, 1547-1551.

[23] Lamaakal, Ismail, Khalid El Makkaoui, Ibrahim Ouahbi, and Yassine Maleh. "A TinyML model for gesture-based air handwriting Arabic numbers recognition." Procedia Computer Science 236 (2024): 589-596.

[24] Terreran, Matteo, Leonardo Barcellona, and Stefano Ghidoni. "A general skeleton-based action and gesture recognition framework for human–robot collaboration." Robotics and Autonomous Systems 170 (2023): 104523.

[25] Kapileswar, Nellore, Palepu V. Santhi, Vijay KR Chenchela, and CH Venkata Siva Prasad. "A fast information dissemination system for emergency services over vehicular ad hoc networks." In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), IEEE, 2017, 236-241.

[26] Rizwan, Muhammad, Sana Ul Haq, Noor Gul, Muhammad Asif, Syed Muslim Shah, Tariqullah Jan, and Naveed Ahmad. "Appearance Based Dynamic Hand Gesture

Recognition Using 3D Separable Convolutional Neural Network." Computers, Materials & Continua 76, no. 1 (2023).

[27] Zhou, Benjia, Jun Wan, Yanyan Liang, and Guodong Guo. "Adaptive cross-fusion learning for multi-modal gesture recognition." Virtual Reality & Intelligent Hardware 3, no. 3 (2021): 235-247.

[28] https://www.kaggle.com/datasets/kapitanov/hagrid.

[29] https://zenodo.org/records/10393655.