electronics

MDPI

Article

TriViT-Lite: A Compact Vision Transformer–MobileNet Model with Texture-Aware Attention for Real-Time Facial Emotion Recognition in Healthcare

Waqar Riaz [1,2], Jiancheng (Charles) Ji [1,*] and Asif Ullah [1,2]

[1] Institute of Intelligent Manufacturing, Shenzhen Polytechnic University, 4089 Shahe West Road, Shenzhen 518055, China; riazwaqar@szpu.edu.cn (W.R.); asifkh@szpu.edu.cn (A.U.)
[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
* Correspondence: jcji20@szpu.edu.cn

Abstract

Facial emotion recognition has become increasingly important in healthcare, where understanding delicate cues like pain, discomfort, or unconsciousness can support more timely and responsive care. Yet, recognizing facial expressions in real-world settings remains challenging due to varying lighting, facial occlusions, and hardware limitations in clinical environments. To address this, we propose TriViT-Lite, a lightweight yet powerful model that blends three complementary components: MobileNet, for capturing fine-grained local features efficiently; Vision Transformers (ViT), for modeling global facial patterns; and handcrafted texture descriptors, such as Local Binary Patterns (LBP) and Histograms of Oriented Gradients (HOG), for added robustness. These multi-scale features are brought together through a texture-aware cross-attention fusion mechanism that helps the model focus on the most relevant facial regions dynamically. TriViT-Lite is evaluated on both benchmark datasets (FER2013, AffectNet) and a custom healthcare-oriented dataset covering seven critical emotional states, including pain and unconsciousness. It achieves a competitive accuracy of 91.8% on FER2013 and of 87.5% on the custom dataset while maintaining real-time performance (~15 FPS) on resource-constrained edge devices. Our results show that TriViT-Lite offers a practical and accurate solution for real-time emotion recognition, particularly in healthcare settings. It strikes a balance between performance, interpretability, and efficiency, making it a strong candidate for machine-learning-driven pattern recognition in patient-monitoring applications.

Keywords: facial emotion recognition; pattern recognition; lightweight deep learning; cross-attention; healthcare in AI; human–robot interaction

Academic Editors: Verónica Vasconcelos and Maryam Abbasi

Received: 24 June 2025
Revised: 23 July 2025
Accepted: 25 July 2025
Published: 16 August 2025

Citation: Riaz, W.; Ji, J.; Ullah, A. TriViT-Lite: A Compact Vision Transformer–MobileNet Model with Texture-Aware Attention for Real-Time Facial Emotion Recognition in Healthcare. *Electronics* **2025**, *14*, 3256. https://doi.org/10.3390/electronics14163256

1. Introduction

Over the last couple of decades, Facial Emotion Recognition (FER) has played a key role initially in image-processing-based applications and now in computer vision (CV) and artificial intelligence (AI), facilitating domains such as healthcare, security monitoring, and human–computer interaction, etc. Emotions are fundamental features of humans that play an important role in social communication, recognizing emotions from any cues is a cumbersome task as expressing emotions depends on the interlocutors' cultures [1,2]. The field of FER initially relied on handcrafted feature extraction methods, such as Principal Component Analysis (PCA) and Local Binary Patterns (LBP). PCA was effective for dimensionality reduction, but it struggled to capture unique emotions

under real-world conditions, while LBP performed well in analyzing facial textures [3]. Human express emotion in various ways, including facial expression, speech, and body language [4–7]. With the rise of deep learning, Convolutional Neural Networks (CNNs) have revolutionized the domain of FER. CNN architectures like ResNet achieved higher accuracy by learning features directly from raw data [8,9] Vision Transformers (ViTs) overcame many limitations of CNNs by using self-attention mechanisms to capture both local and global dependencies across images [10]. Unlike CNNs, ViTs treat each image as a sequence of patches, enabling them to better analyze relationships between distant facial regions and it is interesting that, in the case of facial expressions, the lips and eyes provide the most cues, whereas other facial features, such as the ears and hair, play a small role.

The advancement of Smart Healthcare System demands a real-time FER to monitor patients' emotional states, particularly for detecting pain or unconsciousness in individuals who may not be able to communicate verbally, and brings highly needed novelty towards this area. Facial expressions in the form of frames/images were used as an input modality for recognizing the emotional state of a human emotion, and we used state-of-the-art datasets of FER2013 [11] and AffectNet [12], along with our own custom datasets collected according to the setting of patient's facial expression. To address the computational demands of ViTs, Liu et al. introduced the benefits of global feature extraction while reducing the computational burden and achieved notable performance on datasets like FER2013 and AffectNet [13].

This work presents an advancement of our previously published work on publicly available datasets [9], and the current version is one part of our complete smart healthcare mobile robot system, in which monitoring patient's emotional state is implemented at the first stage; hardware for the mobile pickup robot can be seen in Figure 1. With this motivation in our study, we introduced TriViT-Lite, which combines MobileNet for local feature extraction with Vision Transformer components to capture global dependencies and LBP as handcrafted features to capture the high- and low-level points of patient emotions beneficial for healthcare use [9,14,15]. This integration enhances the model's ability to capture fine-grained local details (e.g., eyebrow curvature, mouth movements) alongside global spatial relationships (e.g., coordination between the eyes and mouth), which is critical for accurate recognition of unique facial emotion, i.e., painful and unconscious. MobileNet is specifically chosen for its ability to efficiently capture fine-grained, localized features, such as eyebrow curvature and subtle muscle movements, which are essential for distinguishing delicate emotions. Its depthwise separable convolutions significantly reduce computation, enabling the model to operate efficiently. LBP and HOG add robustness in challenging conditions like low light or partial occlusion by capturing textural and edge details that might be overlooked by deep features alone. LBP encodes micro-patterns in facial textures, while HOG describes shape and orientation information. ViT contributes by modeling global dependencies across distant facial regions through self-attention mechanisms. This is critical for recognizing emotions that involve coordinated changes in multiple facial areas (e.g., the simultaneous movement of eyes and mouth for fear). A key factor of TriViT-Lite is its features fusion for cross-attention mechanism [16], which integrates local and global along with handcrafted features to enhance the recognition of complex emotions, such as painful or unconscious states. The cross-attention mechanism plays a key role in focusing on the most relevant information across multiple feature types. Apart from publicly available datasets, we also created a custom healthcare dataset consisting of seven emotions altogether, with two unique emotions of painful and unconscious solely advancing the healthcare industry.

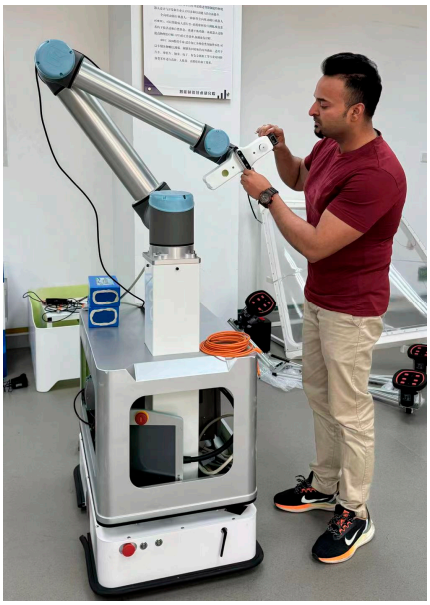

Figure 1. Hardware Structure of Mobile Robot for proposed smart healthcare system [9].

The remainder of this paper is organized as follows: Section 2 reviews related work in FER and transformer-based architectures. Section 3 describes the proposed model. Section 4 outlines the experimental setup and the results and ablation study, while Section 5 offers Discussions and Conclusions and suggestions for future research.

2. Related Work

The foundational method in facial recognition was Principal Component Analysis (PCA), developed by Turk et al. via introducing the "Eigenfaces" technique [17]. Subsequent to PCA, further texture-based techniques such as Local Binary Patterns (LBP), introduced by Ahonen et al., enhanced the robustness of FER systems, specifically for controlled contexts. LBP functioned by encoding local texture patterns from facial pictures into binary sequences, providing improved classification of facial emotions across varying illumination conditions [15]. As FER systems required the handling of significantly larger datasets and the implementation of real-time performance, the deficiencies of conventional machine learning became apparent, prompting a transition to deep learning methodologies. Krizhevsky et al. demonstrated the power of CNNs with the introduction of AlexNet, which achieved groundbreaking results on large-scale datasets like ImageNet [18]; later, CNNs were quickly applied to FER tasks, where they outperformed traditional methods. Building on the success of AlexNet, Simonyan and Zisserman et al. developed VGGNet, a deeper architecture that enabled more granular feature extraction, resulting in improved performance for facial recognition tasks [19]. The deeper layers in VGG-Net captured more detailed facial features, which is especially important in FER, where unique emotions often require a higher degree of feature granularity. He et al. introduced ResNet, addressing the vanishing gradient problem through residual connections. ResNet's ability to train deeper networks while maintaining accuracy was crucial in handling complex FER data, allowing it to achieve state-of-the-art performance on benchmark datasets like FER2013 and AffectNet [8].

To overcome CNN limitations, Dosovitskiy et al. built Vision Transformers (ViTs) for visual tasks by treating images as sequences of patches utilizing attention mechanisms, unlike CNNs' convolutional filters [9]. In tasks demanding long-range relationship modeling, such as face Emotion Recognition (FER), where unique emotions are often spread across numerous face areas, Vision Transformers (ViTs) outperform CNNs. Khan et al. found

that Vision Transformers (ViTs) outperformed Convolutional Neural Networks (CNNs) in distinguishing unique responses that require both diligent local features information and extensive global details [20]. ViTs are ideal for practical FER activities because they can capture local and global correlations. However, the global attention mechanism employed by ViTs augment computational overhead, presenting ViTs in increased computational cost, making scalability and practicality in real-time FER systems challenging. To address the computational limitations of ViTs, Liu et al. proposed the Swin Transformer, which reduces computational costs through a hierarchical architecture and shifted window attention [13]. The Swin Transformer divides an image into smaller windows and applies attention within these windows, shifting them across layers to efficiently capture both local and global dependencies. Afterwards, this type of configuration makes the ViTs well-suited for real-time in volatile settings, i.e., healthcare, FER applications require continuous patient monitoring. Early on video-based face emotion recognition techniques used CNNs with RNNs or LSTM networks to record emotion temporal progression. The Former-DFER model by Zhao et al. addresses these concerns by combining Convolutional Spatial Transformers (CST) and Temporal Transformers (TT) [21]. Baltrusaitis et al. highlight models which perform well in video-based face classification tasks, especially in recognizing evolving emotions like fear, anger, and disgust that develop over time. Such multimodal approach improves the accuracy of FER, especially in dynamic environments where visual cues alone may be insufficient [22]. Gowda et al. developed the FE-Adapter, a real-time FER model specifically designed for healthcare environments [23]. By adapting pre-trained emotion classifiers for video-based FER, the FE-Adapter enables continuous monitoring of patients' emotional states with minimal computational resources.

Although the abovementioned architectures have demonstrated strong results in facial emotion recognition, they face limitations in real-time applications where computational efficiency is critical. In contrast, our TriViT-Lite combines MobileNet's efficiency with Vision-Transformer-based global feature extraction, achieving state-of-the-art results while maintaining low computational overhead, making it ideal for proposed tasks in the recognition of patients' emotions in healthcare industry.

3. Materials and Methods

3.1. Dataset

To evaluate the methodology of this study, we used three different types of datasets, and two of them are publicly available, namely FER2013 and AffectNet. We also collected a custom dataset, and the details are discussed in subsequent sections.

3.1.1. Data Collection and Preprocessing

The custom dataset was created from real-time video recordings in accordance with the patient situation, where it is taken into consideration that in real time, the patients were continuously monitored using high-definition cameras. Keeping in mind that patients could be from diverse demographic backgrounds, representing various ethnicities, genders, and ages diversity, an example of real-time video frames can be seen in Figure 2 (with consent of the focal people). To simulate real-world monitoring scenarios, we intentionally introduced variability in facial features, such as different skin tones, facial hair, large and small eyes, and varying facial structures (i.e., bearded and shaved, etc.). Additionally, the data were collected under a range of lighting conditions, from well to low-light environments. The emotional states of the patients were recorded across seven categories, including neutral, happy, sad, disgust, fear, painful, and unconsciousness. These emotions were annotated and then ensured by the expert in healthcare for the accuracy of the labels used for training. To enhance the diversity and generalization of the dataset, real-time

collected data were merged with some of the publicly available data available on various platforms, but just including the training samples, and results were obtained on a separate list of testing datasets for our custom dataset, FER2013 and AffectNet. To further diversify the data, emotions were captured under different head poses, lighting conditions and facial expressions. In total, almost 500 video segments were collected, with each segment having 30 frames per second, which were later increased to a larger amount after data augmentation and are discussed in a further subsection.

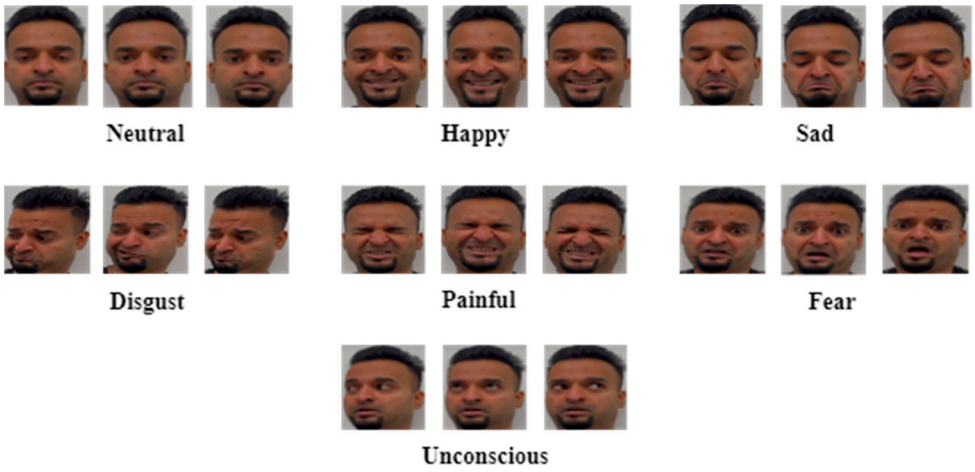

Figure 2. Shows the change in frames in a real-time video feed for the samples for face of main author.

3.1.2. FER2013

The FER2013 [11] dataset consists of 35,887 images, originally compiled for facial emotion recognition tasks. To fit our custom dataset, we chose the categories of happy, sad, disgust, fear, and neutral for our study. For evaluating the model in controlled settings with illumination and facial occlusion, FER2013 provides a high benchmark. About 34% of the total photos fall into the happy category, compared to for the sad, disgust, fear and neutral categories, respectively. We can assess the model's capacity to manage uneven class distributions according to this minimal imbalance. This dataset supplemented our data in categories such as happiness, sadness, fear, and neutrality, sample shown in Figure 3 [11].

Figure 3. Shows random selection of emotions from FER2013 dataset [11].

3.1.3. AffectNet

There are more than a 0.4 million manually labelled photos of facial emotions in the AffectNet dataset [12]. In order to improve our testing and training procedures, we chose 50,000 images from the categories of happy, sad, disgusted, fear, and the neutral sample shown in Figure 4 [12]. This dataset is ideal for testing the robustness of the model since it includes a wide range of emotions in uncontrolled settings with different lighting, occlusion, and head positions. Compared with custom dataset where all the emotions are

distributed rather evenly, although there is minor unbalancing with AffectNet, with neutral, sad, and happy emotions being more common here with a greater number of samples. This variation enables us to evaluate the model's capacity to manage imbalanced data also.

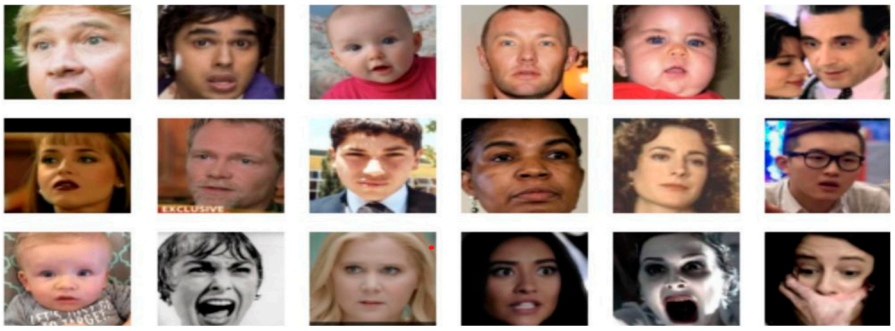

Figure 4. Shows random selection of emotions from AffectNet dataset [12].

3.1.4. Custom Dataset

The custom dataset focused on delicate and clinical expressions (such as painful and unconscious) and complements the FER2013 and AffectNet datasets, which contain broader emotional categories captured under more controlled and uncontrolled settings, respectively. The custom dataset consists of nearly 500 video segments collected in authentic clinical settings, each recorded at 30 frames per second. We utilized data augmentation techniques and data synthesis to increase the effective dataset size, hence ensuring rigorous model training. Furthermore, the collected data were thoroughly balanced to guarantee equitable representation, i.e., every person equally contributed to collect all seven emotions keeping care of their data privacy. Following data augmentation and frame selection procedures, the final dataset comprises a total of 13,200 annotated face images distributed across six emotional categories as follows: neutral (2900), happy (2300), sad (2000), fear (1800), pain (2100), and unconscious (2100). This distribution reflects the genuine prevalence of each emotion observed during data collection rather than an artificially balanced sample. Stratified sampling was used to partition the dataset, with approximately 80% of frames from each category assigned to the training set and 20% to the test set, thus preserving the natural proportions across both subsets. Additional details on the augmentation process and final per-class sample counts are provided in the subsequent subsection. To achieve cross-domain learning, we used domain adaptation techniques to align the feature distributions between the datasets.

- Feature Alignment: During training, a shared embedding space was learned to map samples from both the custom dataset and public datasets into a unified representation space.
- Loss Function Regularization: A weighted cross-entropy loss was applied to encourage the model to balance its performance across domains. The modified loss function is given by the following:

$$L_{total} = \alpha.L_{custom} + (1 - \alpha).L_{public} \quad (1)$$

where L_{custom} and L_{public} are the individual losses for the custom and public datasets, and $\alpha \in [0, 1]$ controls the contribution of each dataset to the total loss. This approach ensures that the model benefited from the strengths of each dataset while learning to generalize across both healthcare and general-use cases.

By merging the real-time collected data with these publicly available datasets, we ensured that the FER system would generalize well across diverse facial features, lighting conditions, and environmental contexts. The FER2013 and AffectNet datasets expanded

the range of emotions and facial expressions for model training, allowing the system to identify complex emotional states across ethnicities.

3.1.5. Preprocessing Pipeline

The custom dataset was thoroughly preprocessed to ensure uniformity and reduce noise. The following tasks were carried out:

- Face Detection: Multi-task Cascaded Convolutional Networks (MTCNNs) were implemented to detect and obtain facial regions from video frames. The MTCNN was optimal for preprocessing as it can effectively handle variation in head posture and lighting changes. Only the face was extracted to retain solely the facial area from each frame while eliminating background interference.
- Normalization: All facial frames were scaled to 224 × 224 pixels for model input uniformity. Pixel values were standardized to [0, 1] for faster and more reliable training.
- Data Augmentation: To increase the variability of the dataset, augmentation techniques such as random cropping, rotation, flipping, and brightness/contrast adjustments were applied. These augmentations account for natural variations in facial poses and lighting, particularly in healthcare settings where lighting may vary. This step is critical to ensuring the model's robustness in real-world conditions.
- Emotion Labelling: A semi-supervised approach was employed for annotating the real-time dataset. Initially, the two unique emotions were labelled carefully based on experts' considerations and their observations, and these annotations were further refined using active learning techniques. Other common emotions like in public datasets that were pre-labelled and integrated with the real-time data through consistent label alignment.
- Temporal Alignment: To capture the evolution of emotions over time, frames were segmented into continuous sections where transitions between emotional states (all seven considered emotions) were explicitly labelled. This temporal annotation ensures that the model can recognize unique changes in facial expressions as patients' emotional states shift during video monitoring.

Finally, dataset consisting of frames and including both real-time and publicly available data, evenly distributed across seven emotion categories is preprocessed. The integration of real-time data and established FER datasets guarantees that the system can handle a wide variety of facial expressions and conditions, making it suitable for real-world applications in healthcare.

3.2. Proposed TriViT-Lite Model

In the following subsections, all the four important constituents of our proposed TriViT-Lite model are discussed in detail.

3.2.1. Proposed Architecture Overview

The TriViT-Lite model is developed to address the challenges faced by facial emotion recognition (FER) systems in real-world applications. FER systems often need to operate under varying conditions, such as changes in lighting, head pose, or partial occlusion of the face, while maintaining high accuracy and real-time performance. Many existing models struggle with these limitations, particularly when emotions are delicate or when the system is operating in resource-constrained environments, like embedded systems or mobile devices [15–22]. TriViT-Lite uses a framework that involves several key components combining local, global, and texture-based features in order to achieve a high degree of accuracy and efficiency. For a given face extracted from the video feed, represented by an RGB image I_{RGB} of dimensions H × W × 3, we first derive its local and texture-specific

features using the feature extraction block of MobileNet. The MobileNet framework employs two distinct backbones as one processes the RGB input, and the other handles the corresponding texture image, which is represented as H × W × 1. We first obtain its local features, and then we secondly obtained the texture feature with the image size of H × W × 1. In our framework, $I_{handcrated}$ are the initial feature maps generated from LBP and HOG applied to each input face image, representing handcrafted texture and edge information. Meanwhile, $f_{handcrafted}$ are the deeper feature representations obtained by passing $I_{handcrated}$ through a convolutional network branch, capturing more abstract patterns from the handcrafted features with the size of H × W × 3 for both, respectively. Particularly, we employ the Depthwise and Pointwise convolution for local features and the first four phases of MobileNet as the backbone to extract feature maps of texture image as $f_{handcrafted}$.

On the parallel side, for the Vision Transformer (ViT) block, recent studies demonstrate that incorporating CNN-based downsampling at the initial stage of the Transformer network significantly enhances performance. Hence, the input undergoes a CNN-based stem layer, which performs downsampling via convolutional and max-pooling operations, reducing the spatial resolution to half of the input size while capturing shallow image information efficiently. Other than that, the position embeddings at initialization time carry no information about the patches' 2D positions. To address this, we create the input sequence from the feature maps produced by the CNN backbone. The embedding of this feature map is obtained by adding positional embeddings to a linear projection of the CNN output. To obtain valuable results in capturing global features, which are valuable in recognizing all seven facial emotions, we further enhance the entire network performance by using pre-trained weights from the MS-Celeb-1M dataset [24], ensuring robust feature learning from large-scale face recognition data.

To ensure that the model's three streams local, global, and handcrafted features do not operate in isolation, we designed a dedicated fusion block to unify and exploit their complementary strengths. After individual extraction, feature maps from MobileNet (local features), ViT (global features), and the handcrafted LBP and HOG streams are concatenated along the channel dimension. Later, the output of ViT and MobileNet is passed through the 1 × 1 conv-layer and processes into the feature fusion block, while without having loss of generalization, the feature fusion block utilized to combine the features extracted from the local, global, and handcrafted features. By dynamically adjusting the feature weights, the fusion block directs the network's attention towards discriminative features that are critical for enhancing emotion recognition accuracy. The fusion weights are derived from a global–local representation mechanism, which effectively combines global context with localized information to refine expression recognition. Afterwards, we feed these features from global and local representation into the 1 × 1 conv-layer where the output of these two features is added with handcrafted features $f_{handcrafted}$, which in the end are flattened to a linear projection and a learnable classification token is added. Positional embeddings are also added at this stage to retain spatial information within the representation. The resulting embeddings are further passed into a multi-layer encoder, where a cross-attention mechanism executed, which is described in detail in subsequent sections, facilitating the effective unification of multi-scale features. Finally, a fully connected layer followed by a softmax activation function is employed to generate probability distribution over facial expression classes, as accordingly shown in shown in Figure 5. The further subsections provide a more detailed breakdown of each part of the architecture.

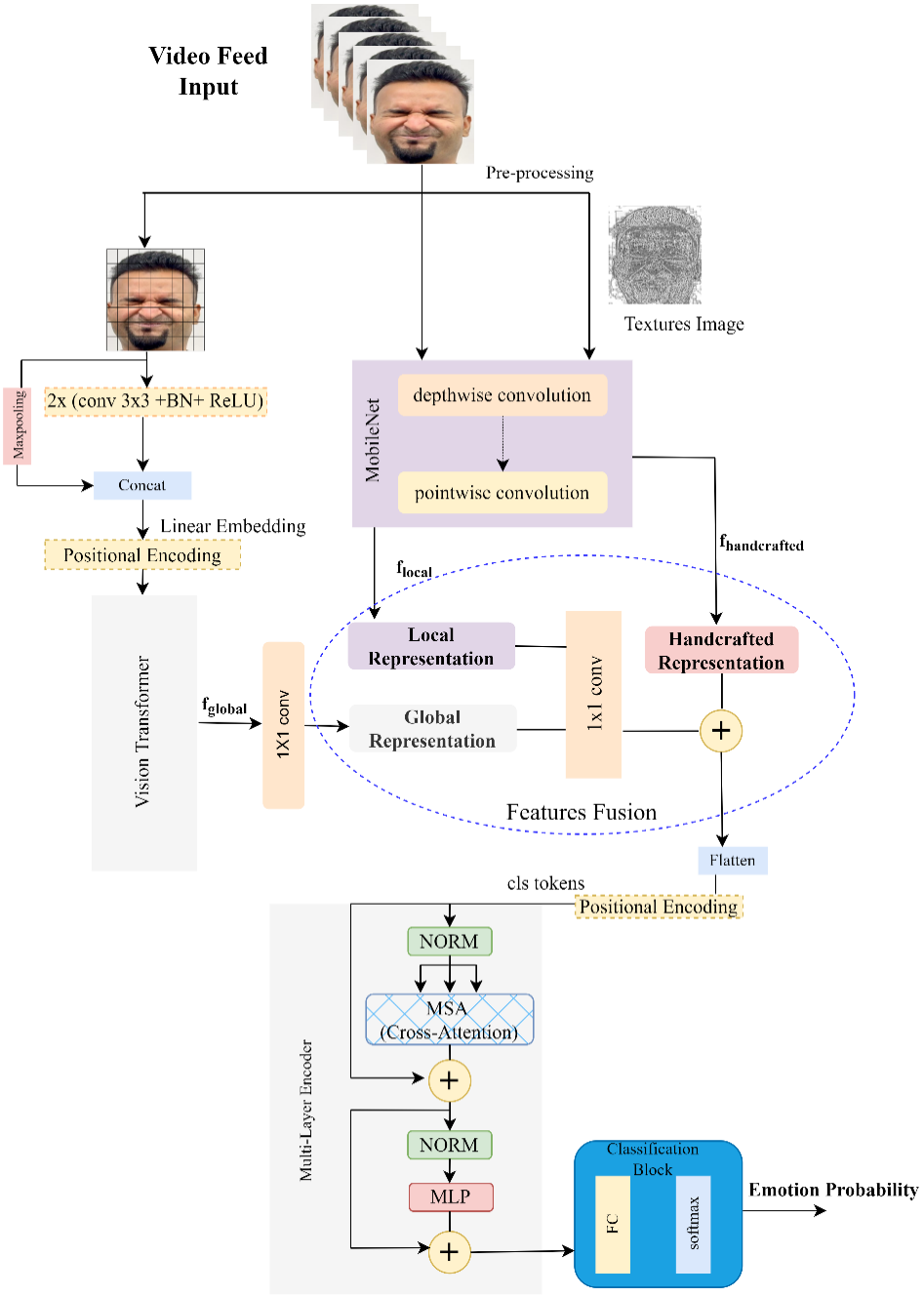

Figure 5. Shows the architecture of proposed model. The architecture of MobileNet and Cross-Attention is discussed in further subsections as an input facial video feed of main author.

3.2.2. Local Feature Extraction with MobileNet

The first component of the TriViT-Lite model is the MobileNet [14], a lightweight CNN that is designed to efficiently extract local features from facial images; configuration is shown in Figure 6. CNNs are well-known for their ability to capture local structures, such as the edges of the mouth or the curve of the eyebrows, which are important for recognizing facial expressions. However, traditional CNNs can be computationally expensive, especially when used in deep architectures with many layers. MobileNet addresses this issue by using a technique called depthwise separable convolutions, which splits the standard convolution operation into two steps:

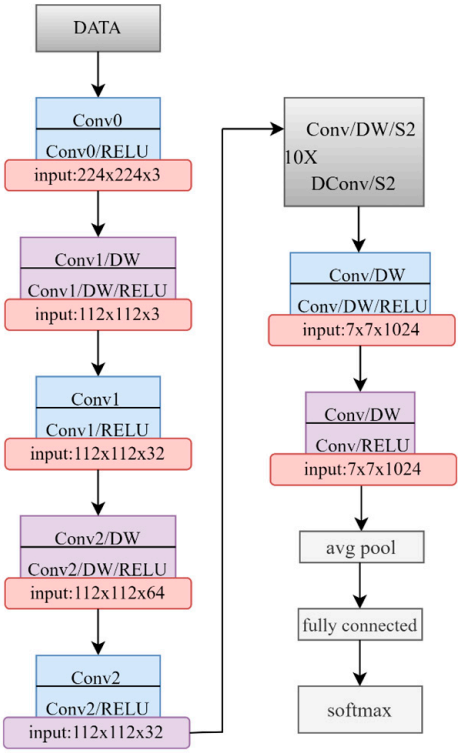

Figure 6. Shows the layer configuration for MobileNet.

- Depthwise Convolution: In the depthwise convolution operation, a single convolutional filter is applied independently to each input channel, allowing the model to capture spatial features in each channel separately. Let the input feature map be denoted by $X \in \mathbb{R}^{H \times W \times C_{in}}$, where H, W, and C_{in} are the height, width, and the number of input channels, respectively. The depthwise convolution applies a unique filter $W_{depth}^{(c)} \in \mathbb{R}^{k \times k}$ of kernel size k × k to each channel c. The output of the depthwise convolution at position (i, j) in the c-th channel of the resulting feature map is given by the following:

$$Y_{depth}^{(c)}(i,j) = \sum_{u=0}^{k-1} W_{depth}^{(c)}(u, v).X^{(c)}(i+u, j+v) \quad (2)$$

where $X^{(c)}$ is the c-th input channel of X, and $W_{depth}^{(c)}(u, v)$ represents the value at position (u, v) of the c-th filter. In this configuration, an additional layer is introduced, which performs a linear combination of the outputs from the depthwise convolution using a 1×1 convolution. This layer is essential for creating new features by effectively integrating information across the filtered channels.

- Pointwise Convolution: Following depthwise convolution, pointwise convolution is used to combine information across channels. For each output channel m, the pointwise convolution is formulated as follows:

$$Y_{point}^{(m)}(i,j) = \sum_{c=0}^{C_{in}-1} W_{point}^{(m,c)}.Y_{depth}^{(c)}(i,j) \quad (3)$$

Here, $Y_{point}^{(m)}$ represents the output of the pointwise convolution for the m-th output channel, while $W_{point}^{(m,c)} \in \mathbb{R}^{1 \times 1}$ is the 1×1 convolutional filter weight connecting the c-th input channel to the m-th output channel. The depthwise layer is primarily responsible for filtering individual channels, while the pointwise layer combines the resulting feature maps across different channels. This decomposition of operations significantly reduces the

computational cost compared to standard convolution. Using N convolutional kernels, feature maps are produced. However, this approach incurs a computational cost as outlined in [14]. The combination of these two steps greatly reduces the number of parameters and computations needed to process the input, making the MobileNet backbone highly efficient while retaining the ability to learn important local features. Formally, if the input image is $I \in \mathbb{R}^{H \times W \times C}$, where H, W and C represent the height, width, and number of channels, the depthwise convolution can be expressed as follows:

$$D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F \tag{4}$$

Depthwise separable convolution consists of two consecutive operations: a depthwise convolution and a pointwise convolution. Together, these operations significantly reduce computational complexity, with the total computational cost expressed as follows:

$$D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F \tag{5}$$

Equation (4) gives the computational cost for a standard convolutional layer, while Equation (5) represents the sum of the costs for depthwise and pointwise convolutions as used in MobileNet's depthwise separable convolution, where D_K represents the kernel size of the convolutional filter, M denotes the number of input channels to the convolution layer, N is the number of output channels, and D_F indicates the spatial dimension (height or width) of the feature map after convolution. The computational efficiency of depthwise separable convolution compared to standard convolution can be expressed by the following ratio:

$$\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2} \tag{6}$$

The combined operation of depthwise and pointwise convolutions in MobileNet reduces computational complexity compared to standard convolution, as the computational cost.

3.2.3. Handcrafted Features with LBP and HOG

The integration of LBP and HOG features into TriViT-Lite is performed through a multi-step process to ensure alignment and maximize the information extracted from facial images. For each preprocessed face image, we compute LBP and HOG descriptors using standard algorithms. The LBP map captures local texture variations by encoding the relationship of each pixel to its neighbors, while the HOG map summarizes gradient orientations to highlight structural edges. Both the LBP and HOG maps are resized and spatially aligned to match the dimensions of the original input image. This ensures that each pixel in the texture maps corresponds precisely to the same spatial location as in the RGB image. The original RGB image, along with its corresponding LBP and HOG maps, are each passed through separate but structurally identical MobileNet-based streams. This parallel extraction allows each channel type color, local texture, and edge structure to contribute unique feature representations without interference, which is explained in detail subsequently.

- Histogram of Oriented Gradients (HOG): The Histogram of Oriented Gradients (HOG) descriptor characterizes features by analyzing the distribution of gradient orientations within specific regions of an image. It achieves this by dividing the image into a grid of cells and computing the gradients within each cell. Originally proposed by Dalal and Triggs [25], HOG operates on the intensity (grayscale) function, denoted as I, which represents the image under analysis. In Figure 7a, the image is divided up into K × K pixels cells. For the x-axis: $G_x = I(x+1,\ y) - I(x-1,\ y)$; and for the y-axis:

$G_y = I(x, y+1) - I(x, y-1)$, by the following equation, gradient magnitude can be denoted as follows:

$$M_{(x,y)} = \sqrt{G_x^2 + G_y^2} \tag{7}$$

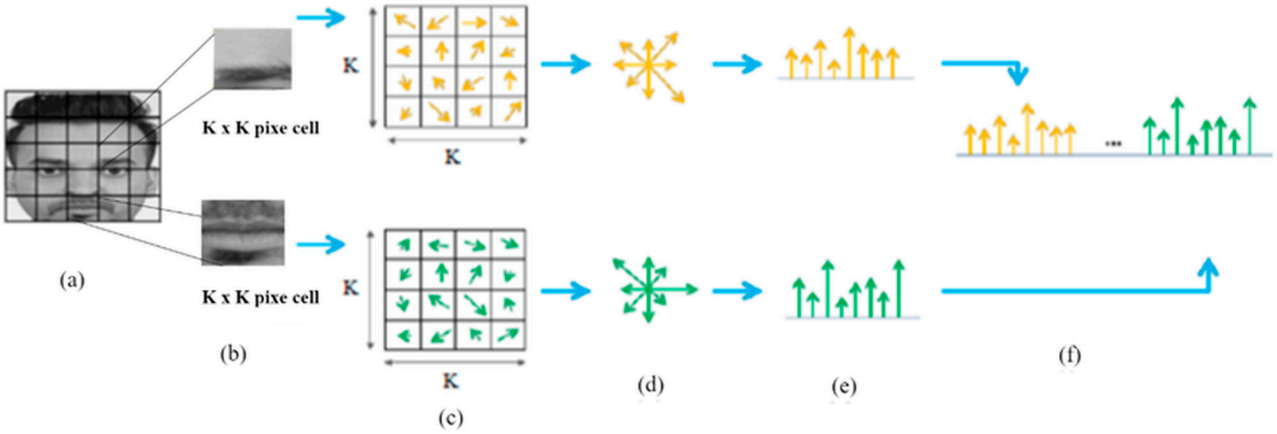

Figure 7. Illustrates the process of extracting Histogram of Oriented Gradients (HOG) features from face of main author (**a**) The original image; (**b**) the division of the image into cells; (**c**) computation of gradient orientations; (**d**) accumulation of orientation values; (**e**) generation of the cell histogram; and (**f**) concatenation of all histograms to form the final feature vector.

And the gradient's Q (x, y) orientation is calculated in each pixel as in Figure 7b,c:

$$Q_{(x,y)} = \tan^{-1}\left(\frac{G_x}{G_y}\right) \tag{8}$$

An M-bin histogram of orientations is constructed by calculating the gradient orientation for each pixel and accumulating these values accordingly, as illustrated in Figure 7d,e. To form the final feature vector, the histograms from all cells are concatenated into a single vector, as depicted in Figure 7f [15]. Studies have shown that HOG is effective for detecting fine facial features in image processing and computer vision algorithms [26].

- Local Binary Patterns (LBP): This is a texture descriptor, which is useful for analyzing the texture of images, as described in [27]. Local Binary Patterns (LBP) encode the relationship of each pixel to its surrounding neighbors by comparing the pixel's value with the central pixel. This comparison is expressed as a binary number. The binary codes are then concatenated in a clockwise direction, starting from the top-left neighbor, forming a binary string. The resulting binary sequence is converted into its corresponding decimal value, which is used for labeling the pixel [28]. In decimal form, the resulting (LBP code) is as follows:

$$LBP_{m,b} = \sum_{m=0}^{m-1} y(j_m - j_i)2^m \tag{9}$$

In this context, j_m denotes the intensity value of a neighbouring pixel, while j_i represents the intensity of the central pixel. The variable m indicates the total number of pixels within the circular neighborhood, and b defines the radius of this neighbourhood. The thresholding function is then applied as follows:

$$y(v) = \begin{cases} 1 & if\ v \geq 0 \\ 0 & if\ v < 0 \end{cases} \tag{10}$$

For subsequent analysis, the histogram of these binary codes is utilized. As illustrated in Figure 8, the LBP encoding process begins by transforming the image into its correspond-

ing LBP representation. The encoded image is then divided into smaller patches, and an LBP histogram is computed for each patch. All patches' LBP histograms are concatenated to create the final feature vector. It is possible to use LBP to describe facial expressions, which provides good results in face recognition. Both LBP and HOG provide texture information that aids in FER under occluded or low-light conditions, complementing the local and global features [29].

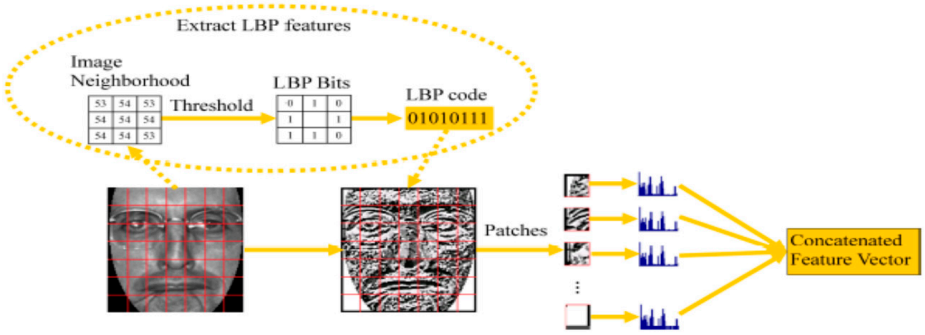

Figure 8. The LBP encoding method [29].

3.2.4. Global Feature Extraction with Vision Transformer

While CNNs are effective at capturing local features [30], they often struggle to model global dependencies, the relationships between different regions of the face, such as the simultaneous raising of both eyebrows or the widening of the mouth and eyes during expressions of fear. To overcome this limitation, Vision Transformers (ViTs) are used to extract global features from the facial image. Vision Transformers use self-attention to model relationships between spatially distant regions of an image, capturing global context across the entire input. The process begins by the input undergoing a CNN-based stem layer, which performs downsampling via convolutional and max-pooling operations and later dividing the features into patches, each processed individually but contextually linked via self-attention.

- Patch Embedding: In the ViT module, the input features, such as image $I \in \mathbb{R}^{H\times W\times 3}$ is divided into $N = \frac{H\times W}{p^2}$, a grid of non-overlapping patches $p \times p$. Each patch $P_i \in \mathbb{R}^{p\times p\times 3}$ is then flattened and projected into a higher-dimensional space using a linear embedding layer, transforming it into a sequence of vectors:

$$Z_0 = [PatchEmb(P_1),\ PatchEmb(P_2),\ \ldots,\ PatchEmb(P_N)] + E_{pos}] \tag{11}$$

where $P_i \in \mathbb{R}^{p\times p\times 3}$ represents the i-th patch of the image, PatchEmb$(P_i) \in \mathbb{R}^d$, is the embedding of the patch, and E_{pos} is the positional encoding that retains spatial information about the patches. This sequence of embedded patches is then passed through the transformer layers, where self-attention is applied to model the relationships between the patches.

- Self-Attention Calculation for Feature Representation: Within each transformer layer, the self-attention mechanism calculates relationships between patches by transforming the input sequence $z \in \mathbb{R}^{T\times(N+1)\times D}$ at time t into queries Q, keys K, and values V through learned projections, such as the following:

$$Q = z_t W_Q,\ K = z_t W_K, \text{ and } V = z_t W_V \tag{12}$$

where W_Q, W_K, and $W_V \in \mathbb{R}^{D\times D}$ are the weight matrices that map the input to the query, key, and value spaces, respectively, at layer l, and d_k is the dimensionality of the key vectors. These are used to compute the i-th attention head respective to the time t as follows:

$$head_i^{(t)} = a\left(Q_i^{(t)}, K_i^{(t)}\right) V_i^{(t)} \in \mathbb{R}^{(N+1)\times \frac{D}{h}} \tag{13}$$

When $t = 1$, then $i = 1 \ldots\ldots\ldots, h$, the similarity between queries and keys is calculated to produce attention scores a, normalized by the dimension d_k to stabilize gradients. The attention mechanism allows the transformer to capture long-range dependencies in the image and models how different regions of the face interact to produce a coherent emotional expression. Here, the softmax function scales the attention scores to sum to 1 across each patch, which helps highlight the most relevant patches for each query:

$$a(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{14}$$

The self-attention output produces an updated representation for each patch, enabling the transformer to model relationships across spatially distributed features. The output z_{l+1} at layer $l + 1$ incorporates both the new self-attended features and a residual connection from the previous layer, preserving learned information:

$$z_{l+1} = a(Q, K, V) + z_l \tag{15}$$

$$Q^{(t)} = \left[Q_1^{(t)}, \ldots, Q_i^{(t)} \ldots, Q_h^{(t)}\right] \tag{16}$$

where $Q_i^{(t)} \in \mathbb{R}^{T\times(N+1)\times D/h}$ and is the part of $Q(t)$ corresponding to i-th head, the same applies for $K_i^{(t)}$ and $V_i^{(t)}$, and in the Multi-Head Self-Attention mechanism, the patches within the t-th frame are attended to only by other patches within the same frame at time t. Consequently, this process does not involve any temporal interactions between frames:

$$MSA(z_t) = head_i^{(t)}, \ldots, head_h^{(t)} \tag{17}$$

By combining attention and residual connections, the transformer effectively captures both immediate and long-range dependencies, which is particularly useful in FER tasks where emotions are expressed through various spatially distant facial regions. Additionally, the transformer's self-attention mechanism ensures each patch attends all other patches, capturing the global dependencies needed to identify unique and complex facial expressions. This ability to model interdependencies across all patches gives the ViT a significant advantage over purely convolutional architectures in handling global context, especially for tasks where emotions involve multiple facial regions.

3.2.5. Feature Fusion with Cross-Attention

Dealing with scenarios where facial features are occluded or affected by changes in illumination is a significant challenge in facial emotion recognition. We use a multi-feature fusion method in the TriViT-Lite model to combine features from many sources, which are then included in a cross-attentional feature block. It processes the handcrafted features like Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) with global (ViT) features and local (MobileNet) features. Mathematically, it can be shown as follows:

$$f_{fusion} = \left(f_{local},\ f_{global},\ f_{handcrafted}\right) \tag{18}$$

A global feature map $f_{global} \in \mathbb{R}^{N \times d}$ is produced by the transformer, where d indicates the embedding dimension and N is the number of patches. In parallel, texture-based features produce $f_{handcrafted}$, whereas MobileNet produces f_{local}. The global feature map provides information about the facial structure, which is crucial for identifying complex emotions. The handcrafted nature of LBP allows it to detect unique changes in texture associated with emotions such as disgust or unconsciousness.

This information is crucial for recognizing the unique changes in facial texture caused by these emotions. This model can recognize emotions even when the face is partially obscured or distorted by combining handcrafted features and deep-learning-based features. Feature fusion block integrates the local, global, and texture-based features into a 1×1 Conv layer, followed by summation to create a unified feature map that is robust to variations in the input data. In the final stage, the extracted features are divided into patches to process through a cross-attention mechanism. This block computes attention across frames, enabling patches in the t-th frame to also attend to patches from the preceding frame at $t-1$ and the subsequent frame at $t+1$, and it can be carried with shift operations [16]. After computing the query (Q), key (K), and value (V) representations for each frame, these representations are exchanged with neighboring frames. The selection of which components, Q, K, or V, to shift determines the interaction mechanism. A common approach is to shift K and V, allowing the query from the current frame to attend to key-value pairs from adjacent frames. This process is mathematically represented as follows:

$$head_i^{(t)} \begin{cases} a\left(Q_i^{(t)}, K_i^{(t-1)} V_i^{(t-1)}\right) & 1 \le i < h_b \\ a\left(Q_i^{(t)}, K_i^{(t+1)} V_i^{(t+1)}\right) & h_b \le i < h_b + h_f \\ a\left(Q_i^{(t)}, K_i^{(t)} V_i^{(t)}\right) & h_b + h_f \le i \le h \end{cases} \tag{19}$$

Here, we have the first h_b heads with the backward shift, the next h_f heads with the forward shift, and the rest heads with no shift. Figure 9 shows the diagram of the shift operation for the cross-attention mechanism we used in this study, inspired by [16]. Initially, queries, keys, and values are computed for each frame, followed by selective shifting of some components prior to attention computation. Solid arrows represent the key-value shifts from adjacent frames $t-1$ and $t+1$. Similarly, dotted arrows illustrate the key-value shifts originating from the current frame t to the neighboring frames $t-1$ and $t+1$. This shifting process occurs simultaneously across all frames in a consistent manner.

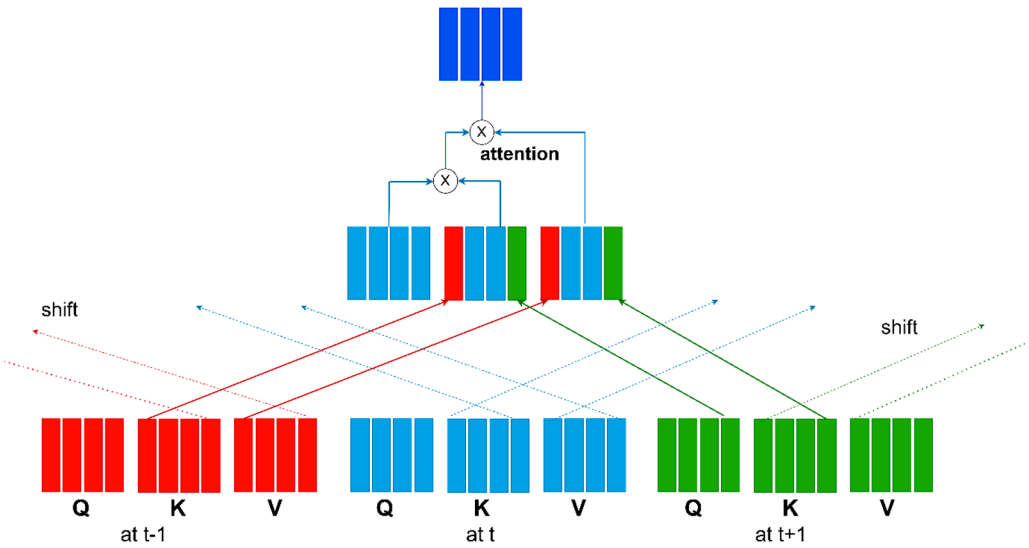

Figure 9. Shows the shift operation for cross-attention mechanism.

4. Results

The TriViT-Lite model underwent training and testing on a high-performance system equipped with an NVIDIA A100 Tensor Core GPU, an Intel Xeon 32-core CPU, and 128 GB of RAM. PyTorch 1.13.1 adaptability and dynamic computing structure led to its selection as the primary platform for designing models and training. Additionally, PyTorch's automatic differentiation feature enabled the backpropagation during training to be simpler and allowed the Vision Transformer and MobileNet deep layers' gradients to be computed. To make sure that the model receives high-quality input for both training and testing, data preparation is a crucial step.

- Resizing: A 224 × 224 pixel scale was applied to all input frames to provide consistency across the collection. The input dimensions were chosen based on the requirements of MobileNet and the Vision Transformer, which work best with photos of this size. Because larger images would require significantly more memory and processing power, resizing allowed for more effective GPU use.
- Normalization: A common technique in deep learning, the mean and standard deviation values from the ImageNet dataset [31] were applied to each image. By preventing too large or small gradients during training, this normalization process brings the pixel values into a uniform range. The mean and standard deviation are calculated across all three color channels (RGB). The normalizing formula used is as follows:

$$\text{Normalized Image} = \frac{\text{Image} - \text{mean}}{\text{standard deviation}} \tag{20}$$

- Video Frame Extraction: A representative sample of facial expressions throughout time was obtained by extracting frames from video sequences at a rate of 10 frames per second (FPS). By lowering the frame rate, we were able to remove unnecessary frames that would interfere with training and yet catch the most notable changes in facial expression.
- Preprocessing Pipeline: Painful and unconscious emotions were preprocessed with an emphasis on facial alignment to enhance classification. As discussed earlier, to achieve precise face detection, we used MTCNN (Multi-Task Cascaded Convolutional Networks) for preprocessing, assuring consistency in aspects like jaw alignment and eye positioning. This stage is critical because uneven expressions (such as synchronous eye movement or drooping of one side eye, etc.) are important indicators of unconsciousness.
- Temporal Alignment for Overlapping Emotions: Recognizing transitions between sad and pain require monitoring temporal dependency across frames. We utilized a sliding window approach to facilitate the smoothing of transitions across multiple time steps, while also assuring the precise distinction between the painful and sad by analyzing adjacent frames. The temporal alignment is depicted as follows:

$$f(t) = \max\Big(P_{sad}(t-1), P_{painful}(t)\Big) \tag{21}$$

4.1. Training and Optimization

The training strategy was carefully designed to ensure that the model achieved optimal performance while minimizing computing cost. Each component of the training pipeline was carefully chosen to maximize resource efficiency, stability, and convergence rate. We outline the key elements of the training methodology and explain how each chosen design affected the model's performance. After conducting a thorough grid search across a variety of values, the initial learning rate was determined to be 5×10^{-5}. This learning rate was

selected because it strikes the best possible balance between stability and convergence speed as shown in Figure 10.

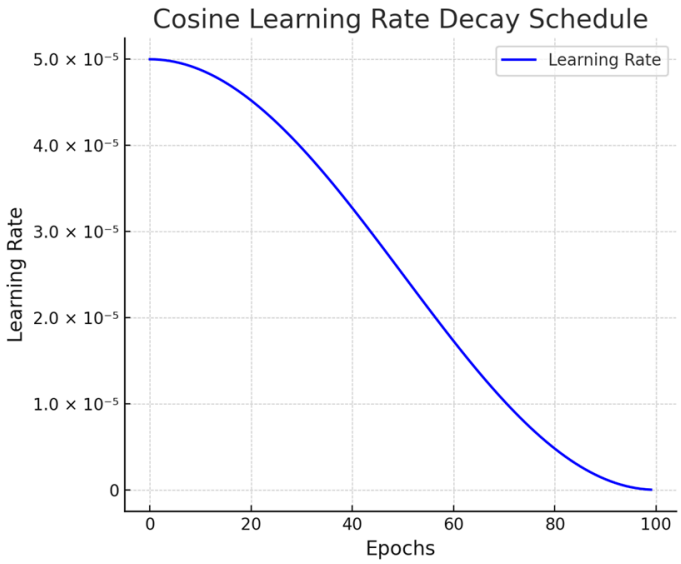

Figure 10. Shows the learning rate is gradually reduced allowing the model to make significant progress in the early epochs and fewer growth as it approaches convergence.

We used a cosine learning rate decay schedule to increase convergence. The learning rate is gradually reduced by this schedule, allowing the model to make significant progress in the early epochs and fewer growth as it approaches convergence. The formula used to express the cosine learning rate is as follows:

$$\eta_t = \eta_0 \times \frac{1 + \cos\left(\frac{\pi t}{T}\right)}{2} \tag{22}$$

where η_0 refers to the initial learning rate, t denotes the current epoch, and T denotes the total number of epochs.

To ensure optimal convergence and fair evaluation, we conducted a systematic grid search over combinations of batch size and learning rate using the Adam optimizer. For each setting, we monitored both training and validation loss curves. As shown in Figure 11, TriViT-Lite consistently achieved stable convergence and low final training loss across batch sizes of 16, 32, and 64, with minimal variation in learning dynamics. This indicates that the model is robust to reasonable variations in these hyperparameters, supporting our final selection of batch size 32 and an initial learning rate of 0.001. Furthermore, we used the Adam optimizer to train the model. The Adam optimizer integrates the advantages of momentum-based optimization and adaptive learning rates, making it particularly effective for training deep neural networks such as ViTs. Adam's ability to adaptively adjust the learning rate for each parameter helped the model converge faster, particularly when training on large datasets like AffectNet. The update rule for the Adam optimizer is given by the following:

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t \tag{23}$$

where θ_t is the updated model parameters, η is the learning rate, $\hat{m}_t$ is the biased first moment estimate (mean of the gradients), $\hat{v}_t$, is the biased second moment estimate (uncentered variance of the gradients), and ϵ is small constant added for numerical stability.

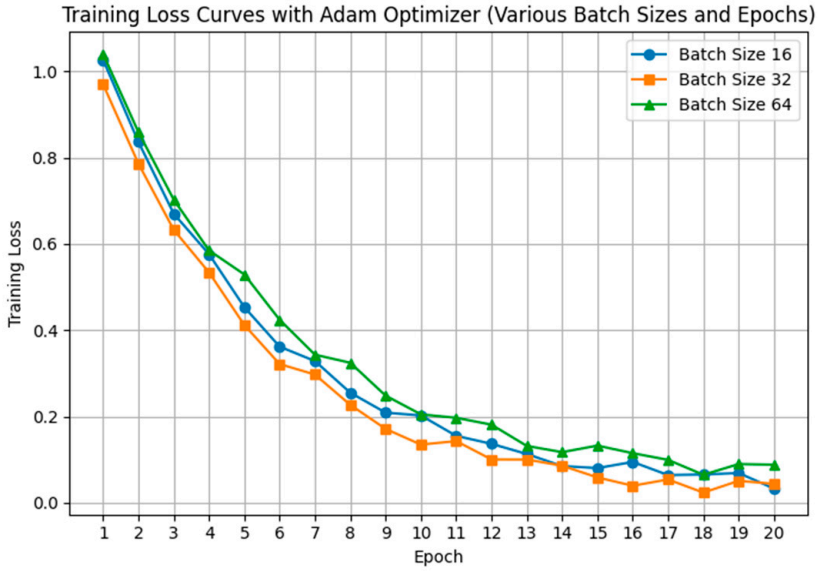

Figure 11. Shows the batch sizes selection comparison with batch size of 32 provided the most stable gradient updates.

As shown in Figure 12 above, we observe how the loss decreases smoothly over epochs, especially during the initial stages, showing Adam's ability to speed up convergence. Such optimization behavior ensures that the model effectively minimizes the loss early on, making it a reliable choice for models like TriViT-Lite.

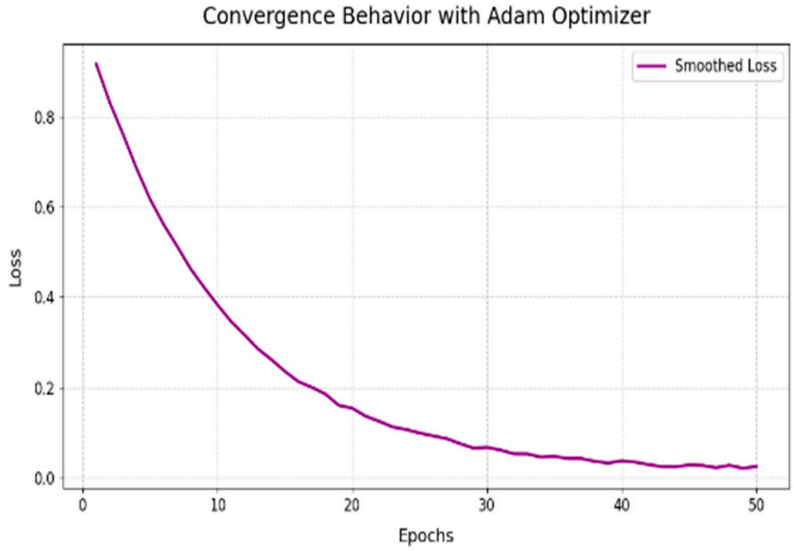

Figure 12. Shows smooth convergence with respect to epochs for Adam.

To avoid the effect of overfitting, we employed an early stopping mechanism, which terminates training if the validation loss fails to improve for 50 consecutive epochs. This approach ensures that the model halts training once it achieves optimal performance on the validation set, preventing unnecessary computation and over-optimization. Thus, saving computational resources and preventing overfitting to the training data.

Figure 13 illustrates the training and validation loss curves, showing how early stopping was triggered when the validation loss plateaued for 50 consecutive epochs. The gradual convergence of the two loss curves demonstrates the model's ability to learn effectively without overfitting. During training, dropout randomly deactivates a subset of activations by setting them to zero, compelling the network to develop more robust and generalized feature representations. We applied dropout with a rate of 0.2 in the fully

connected layers of the ViT component, ensuring that the model could generalize well to unseen global data.

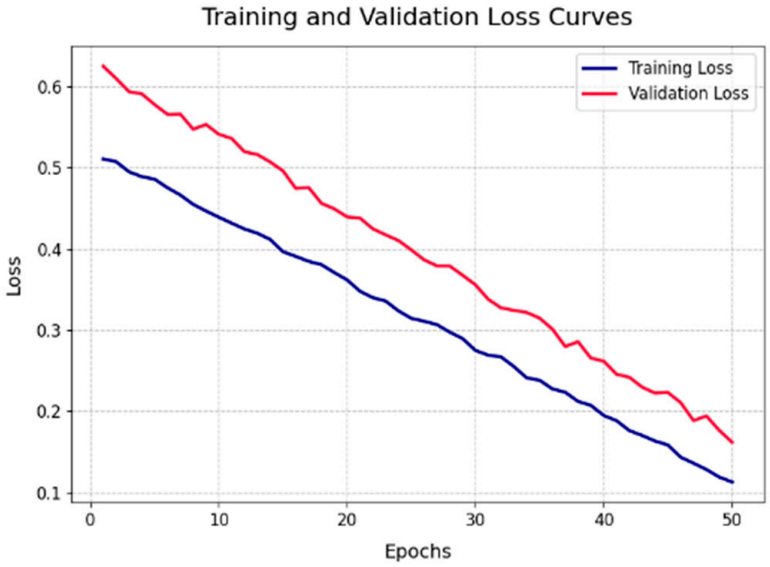

Figure 13. Shows the training and validation loss curves for 50 consecutive epochs ensuring that the model halts training once it achieves optimal performance.

4.2. Evaluation Metrics

We evaluated standard classification metrics and performed a comprehensive error analysis to look at specific failure cases to fully assess the TriViT-Lite model's performance. The metrics include accuracy, precision, recall, and the F1-score, which provide information on many aspects of the model's performance.

- Accuracy: The simplest straightforward metric is accuracy, which provides a thorough evaluation of the model's predicted reliability:

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total predictions}} \tag{24}$$

Even though accuracy offers an extensive view, it can be misleading when class distributions are unbalanced, requiring the employment of additional measurements.

- Precision: The observed ratio of positive predictions to all expected positives is known as precision. Precision is crucial in this study because false positives could lead to unnecessary interventions (such as incorrectly classifying a patient to be unconscious) proving precision to be important:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{25}$$

A high precision means that there are not excessive amounts of false positives and the model correctly identifies situations like painful.

- Recall: Recall, or sensitivity, measures the proportion of true positives out of all actual positive cases. For example, missing a case of painful or unconscious emotion can be critical for patient care, so recall provides insight into how well the model identifies all relevant instances:

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{26}$$

High recall ensures that the model does not miss important cases, which is essential for our applications.

- F1-Score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both:

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recal}} \tag{27}$$

This metric is critical for assessing performance in classes like painful and unconscious, where precision and recall must both be high.

- Confusion Matrix: We generated confusion matrices for each dataset to highlight places where the model struggled. For example, in some circumstances, the model confuses sad with painful states, particularly in video sequences with minimal facial movements, as discussed in more depth in the following section. This insight is useful for understanding the model's limitations and will be explored further in future study.

4.3. Experimental Evaluation

The TriViT-Lite model was evaluated on three datasets, the custom dataset, FER2013, and AffectNet. The custom dataset played an essential role in evaluating the model's ability to detect unique emotional states in real-world scenarios. In particular, the model's performance on emotions, like painful and unconscious, demonstrated the effectiveness of combining local feature extraction (MobileNet) with global feature extraction (ViT). In the subsequent sections, we provide a detailed analysis of the results on the custom dataset followed by the public dataset, evidencing a comprehensive analysis of the model's performance, presenting results on key metrics, as well as an ablation study to quantify the impact of each component of the model. Additionally, we compare the model against several state-of-the-art architectures to highlight the effectiveness of our approach.

The model achieved competitive accuracy across all seven emotion categories in the custom dataset. The following subsections discuss the model's strengths and areas for improvement.

4.3.1. Class-Level Performance Analysis

To better understand the performance of the TriViT-Lite model across different emotional categories, Table 1, below breaks down the precision, recall, and F1-scores for each of the seven emotions on the custom dataset. The happy, sad, and neutral categories exhibit high precision and recall, reflecting the model's effectiveness in recognizing these well-defined emotions. However, the model's performance on painful and unconscious emotions were slightly lower but still remains impressive given the unique facial cues associated with these states. For example, painful expressions are often characterized by small changes in muscle tension of the cheeks and eyebrows or slight frowning, which makes them more difficult for models to detect. Similarly, unconscious states can be easily confused with disgust, particularly when facial movements are minimal and only the upper area of face is in movement with the eyes and forehead. Nevertheless, the model's ability to distinguish these emotional demonstrates its capability in healthcare settings where accurate detection of patient distress is crucial. Figure 14 shows emotion probability graphs reflecting softmax outputs for individual samples.

Table 1. Shows the performance metric on custom dataset.

Emotion	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Happy	95	93	93	93
Sad	92	89	90	89
Neutral	94	91	92	91
Fear	90	87	86	86
Disgust	89	84	85	84
Painful	76	69	70	68
Unconscious	77	71	72	71

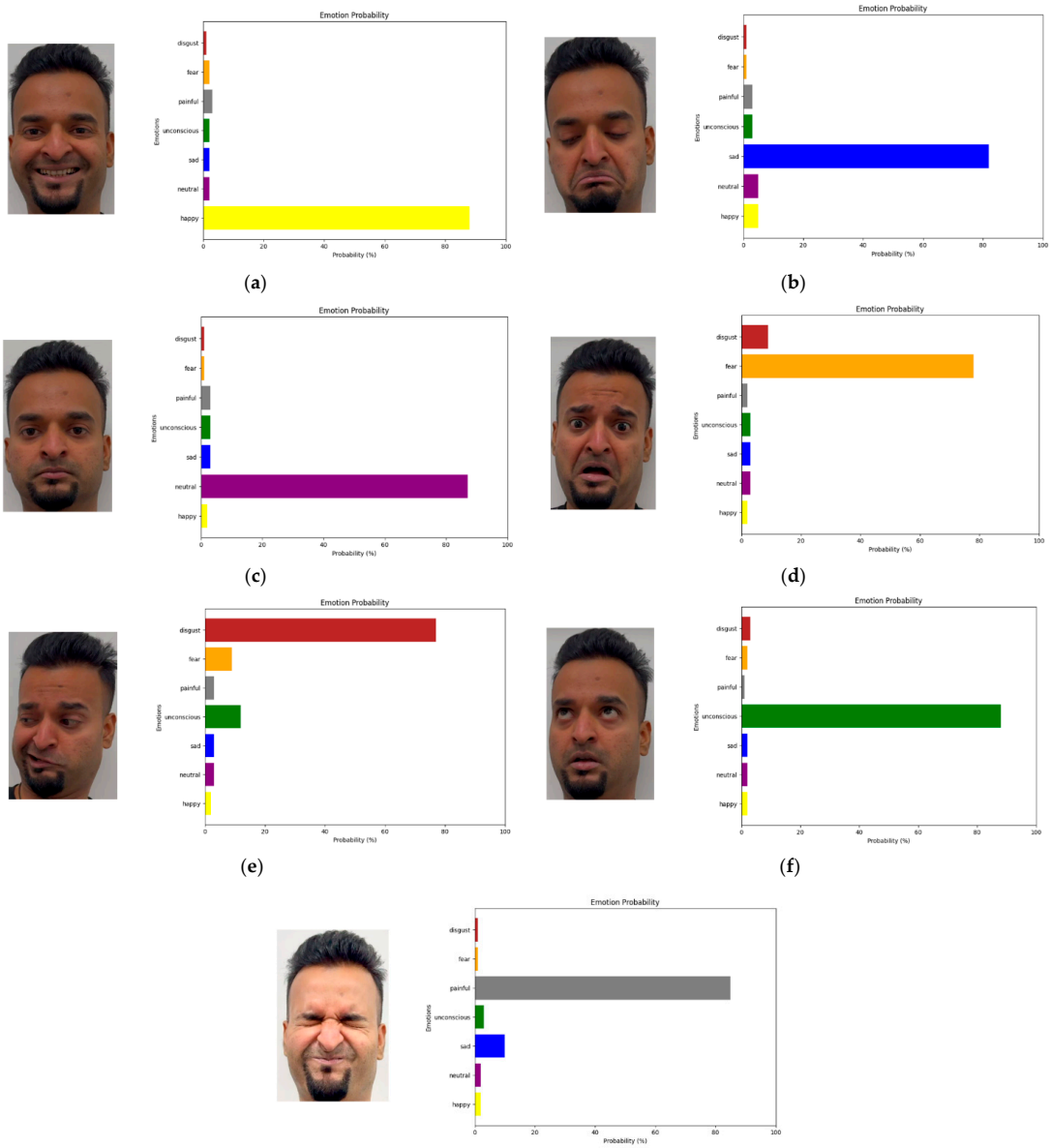

Figure 14. Shows emotion probability graphs reflecting softmax outputs for face of main author, from (**a**) Happy, (**b**) Sad, (**c**) Neutral, (**d**) Fear, (**e**) Disgust, (**f**) Unconscious, and (**g**) Painful.

4.3.2. Confusion Matrix Evaluation

A detailed confusion matrix was generated for our custom collected dataset to analyze the errors made by the TriViT-Lite model. The confusion matrix provides insights into the model's misclassifications, highlighting areas where it struggles, which will be in our considerations for future work. Our analysis revealed that unconscious expressions were just occasionally misclassified as disgust emotions. This confusion can be attributed to the fact that disgust and unconscious, as well as sad and painful, faces often exhibit very minor differences, particularly in static images. Figure 15 below illustrates graphically the confusion matrix for the seven emotions, showing the true and predicted classes. It can be observed that while most emotion categories achieve high correct classification rates, the "painful" and "unconscious" classes are most frequently confused with one another. Specifically, a substantial number of samples from the "painful" category are misclassified as "unconscious," and vice versa. This pattern indicates that the facial cues distinguishing these two states are particularly delicate and often challenging for the model. Improving discrimination between these categories remains a key direction for future work. Future work could focus on improving this classification by incorporating more texture-based features that highlight the caretaking differences between these expressions.

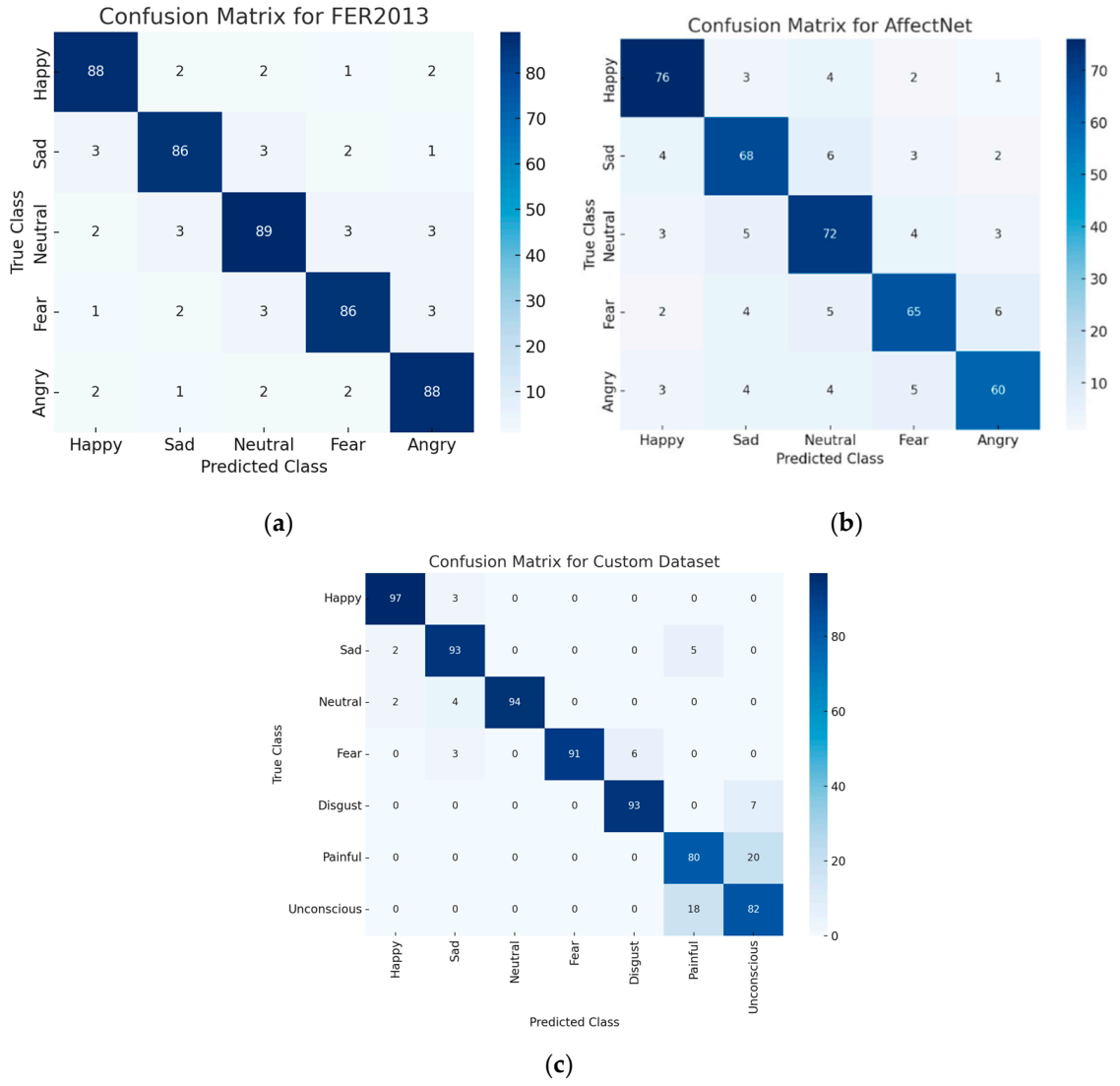

Figure 15. Shows the confusion matrix for (**a**) FER2013, (**b**) AffectNet, and (**c**) Custom datasets.

4.3.3. Overall Accuracy and Baseline Comparison Analysis

The TriViT-Lite model demonstrated strong performance across all datasets, particularly in the custom patient emotions dataset, where it successfully identified complex and unique emotional states of painful and unconscious which are of our main focus.

To address the comprehensive evaluation, we conducted comparative analyses against several state-of-the-art architectures, including ResNet-50 [8], VGG-16 [19], Swin Transformer [13], Res2Net [32], and ViT [10]. Detailed performance indicators, such as accuracy, precision, recall, F1-score, memory usage, and FLOPs are reported, in Table 2. These comparisons illustrate TriViT-Lite's superior balance between computational efficiency and accuracy, especially significant for real-time FER applications. The result indicates the model's strong ability to generalize with real-world data, even when the facial cues are unique, as seen in categories like painful and unconscious. The model also performed well on FER2013, which features more controlled and curated facial images, and AffectNet, where the data are sourced from a variety of uncontrolled environments, but these datasets are without painful and unconscious emotions. This performance highlights the versatility of the TriViT-Lite model, as it handles both controlled and real-world datasets effectively. Table 3 shows comparisons of the performance analysis, including specialized FER models, clearly illustrating the competitive advantage of our TriViT-Lite model across standard FER datasets, including FER2013 and AffectNet.

Table 2. Shows the baseline comparison through a range of methodologies.

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Memory Usage (MB)	FLOPs (G)
VGG-16	FER2013	85.30	84	83	84	528	15.47
ResNet-50	FER2013	87.90	87	86	87	177	4.12
ViT	FER2013	88.00	88	87	87	330	7.23
Res2Net	FER2013	89.50	89	89	89	150	3.95
Swin Transformer	FER2013	91.00	90	89	90	290	4.51
TriViT-Lite	FER2013	91.80	91	90	91	128	2.89
VGG-16	AffectNet	64.00	63	62	63	528	15.47
ResNet-50	AffectNet	65.80	65	64	65	177	4.12
ViT	AffectNet	70.00	69	68	68	330	7.23
Res2Net	AffectNet	71.50	70	70	70	150	3.95
Swin Transformer	AffectNet	73.00	72	72	72	290	4.51
TriViT-Lite	AffectNet	74.00	73	72	73	128	2.89
VGG-16	Custom Dataset	78.40	78	77	78	528	15.47
ResNet-50	Custom Dataset	81.20	80	79	80	177	4.12
ViT	Custom Dataset	82.00	81	80	81	330	7.23
Res2Net	Custom Dataset	83.80	83	82	83	150	3.95
Swin Transformer	Custom Dataset	85.20	84	84	84	290	4.51
TriViT-Lite	Custom Dataset	87.50	87	85	87	128	2.89

Table 3. Shows the performance comparison across specialized FER models.

Authors	Model	FER2013 Accuracy (%)	AffectNet Accuracy (%)
Roy et al. [33]	ResEmoteNet	79.79	72.39
Georgescu et al. [34]	Local Learning	75.42	63.31
Hasani et al. [35]	BReG-NeXt-50	71.53	68.50
Zhang et al. [36]	HFE-Net	71.69	58.55
Kolahdouzi et al. [37]	FaceTopoNet	72.66	70.02
Ours	TriViT-Lite	91.80	74.00

In the context of FER applications that require real-time performance, such as healthcare, the subsequent experimentational details provide an analysis of TriViT-Lite's inference speed and resource effectiveness. We ran the model tests on a wide range of devices, from desktop GPUs to mobile GPUs with certain limitations. We conducted thorough FPS performance evaluations on different GPU platforms, an NVIDIA A100 desktop GPU achieved ~85 FPS, an NVIDIA RTX 3090 delivered ~78 FPS, and the mobile GPU NVIDIA Jetson AGX Xavier provided ~15 FPS, which confirms TriViT-Lite's suitability for real-time FER. These findings validate TriViT-Lite's capacity to manage real-time requirements, obtaining a frame rate of approximately 15 FPS, which is vital for FER tasks in dynamic settings.

The custom dataset yields particularly impressive results, as evidenced by Tables 2 and 3, which emphasizes TriViT-Lite's superior performance across datasets. TriViT-Lite performed optimally on FER2013 (91.8%) and AffectNet (74.0%) and achieved prominent accuracy on the custom dataset (87.5%) by combining the cross-attention fusion mechanism, as well as the MobileNet and Vision Transformer modules. It also outperformed its competitor models on all other datasets. This enhancement is due to the architecture's distinctive design, which incorporates MobileNet for the extraction of fine-grained local features, Vision Transformer for the acquisition of global relationships, and handcrafted features that optimize texture and edge details. By this combination technique through a cross-attention method, the model performs well even in difficult situations like occlusion and varying lighting. To quantitatively assess the robustness of TriViT-Lite under real-world variations, we conducted additional experiments using test sets with controlled lighting changes and various occlusion scenarios. As shown in Table 4, the model maintains high performance across all challenging conditions.

Table 4. Shows the performance under different scenarios like occlusion and varying lighting.

Scenario	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Standard (normal conditions)	87.5	87	85	87
Low Lighting	81.6	83	82	82
High Lighting (overexposed)	84.4	83	81	82
Mild Occlusion (e.g., glasses)	86.8	85	84	84

Figure 16 illustrates these findings and validates that TriViT-Lite is a reliable and robust approach for these emotion recognition tasks, surpassing conventional transformer-based models and CNN-based architectures. Its generalization and adaptability to a wide range of real-world settings can be verified by its consistent performance across datasets, which sets a novel standard in facial expression recognition. As shown in Figure 17, TriViT-Lite achieves the best balance between accuracy and efficiency, with the lowest parameter count, fastest inference, and lowest memory usage among all compared baselines. Its real-

time capability is validated across multiple hardware platforms. This validation robustly supports our claim that TriViT-Lite is not only accurate but also among the most lightweight and deployable models currently available for facial emotion recognition.

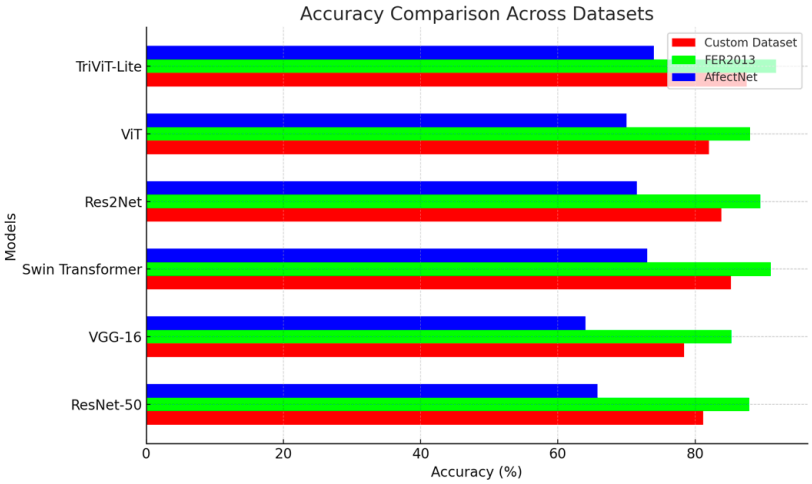

Figure 16. Shows the accuracy comparison across all datasets.

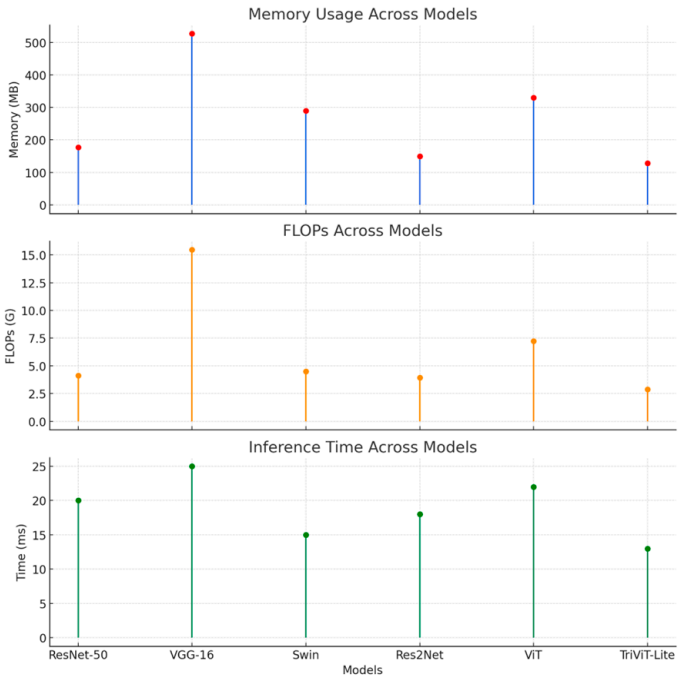

Figure 17. Validation comparison across baseline for TriViT-lite.

4.4. Ablation Study

To thoroughly understand the contributions of various components of the TriViT-Lite model, we conducted an extensive ablation study as shown in Table 5. To determine which aspects of the model contribute most to its performance and will be taken into consideration for future research, the goal of this study is to quantify the impact of eliminating or altering certain architectural components. Aiming to identify practical characteristics of four critical components, the ablation experiments targeted four critical mechanisms as cross-attention fusion, handcrafted features, Vision Transformer (ViT), and MobileNet. As follows, using the model configuration, we analyzed four distinct configurations.

Table 5. For the ablation study, the performance metrics for each configuration on the custom dataset are presented in the table.

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
TriVit-Lite (complete model)	87.5	87	85	87
Without MobileNet (ViT only)	81.7	79	77	79
Without ViT (MobileNet only)	84.5	82	80	83
Without Handcrafted Features	86	84	83	85

4.4.1. Full Model (TriViT-Lite)

The complete version of TriViT-Lite comprises all three elements: MobileNet for local feature extraction, the ViT for global feature extraction, handcrafted features for edge and texture information, which are fused together for the cross-attention fusion mechanism that efficiently integrates various feature sets. This configuration obtained the highest overall performance, with an F1-Score of 87% and an accuracy of 87.5%. The ability of the entire model to capture both global and contextual associated with fine-grained local features in the face is responsible for its outstanding performance. MobileNet prioritizes local facial traits, such as the contour of the mouth, eyes wrinkles, and eyebrow corners, which are crucial for identifying distinct emotions, such as sad or pain. In contrast, the Vision Transformer captures global relationships throughout the face, such as the simultaneous rising of both brows or the widening of the mouth and eyes during a fear look. Handcrafted features assist with texture edge information, which is essential for the real-time deployment of input video frames. These complimentary feature sets are successfully combined by the features fusion to perform the cross-attention fusion mechanism, which enables the model to produce precise predictions even under difficult circumstances like partial occlusion or changing lighting, etc.

4.4.2. Without MobileNet

The MobileNet backbone to extract local features was eliminated in this variant, leaving only the Vision Transformer to extract both local and global features. The accuracy of this setup decreased to 81.7%, followed by the F1-Score of 79%. This finding emphasizes how essential it is to have a separate module for local feature extraction. The Vision Transformer is extremely good at collecting global relationships across multiple parts of the face, but it struggles to extract the fine-grained local information required to properly detect individual changes in facial expression. MobileNet, which is meant to work well with local features, is an important part of capturing fine details like the tightness in the muscles across the forehead, the narrowing of the eyes, or the slight curving of the corners of the mouth, etc. In the absence of MobileNet, the model must depend exclusively on the Vision Transformer for the extraction of both local and global features. Although the ViT is still capable of capturing local dependencies, it is not architecturally specialized enough to detect delicate, localized changes. This limitation becomes especially perceptible in identifying the delicate emotions, such as unconscious or distinguishing between sad and painful, when slight changes in facial muscles can make a huge impact.

4.4.3. Without Vision Transformer

In third case, the Vision Transformer was replaced with fully connected layers, the model's accuracy decreased to 84.5%, with an F1-Score of 82%, and only MobileNet was left for local and global feature extraction. This arrangement performs worst, showing the necessity of global feature extraction. The MobileNet backbone is quite effective at capturing local features, but it struggles to recognize relationships between different

portions of the face. Happy and sad emotions often involve various facial changes. For instance, a happy expression can incorporate both the expansion of the mouth and the elevation of the cheeks and eyes. Without the Vision Transformer, the model cannot capture global relationships, hence reducing performance. This restriction is most noticeable when identifying complicated emotions, such as sad or painful, where it is essential to understand how different facial regions interact. MobileNet excels at identifying localized features (e.g., corners of the mouth), but without the Vision Transformer, it misses the broader context of facial movement even with having handcrafted features, leading to more frequent misclassifications.

4.4.4. Without Handcrafted Features

When we removed the handcrafted features from the model, leaving only the MobileNet and Vision Transformer for feature extraction, the accuracy dropped to 86% with an F1-Score of 84%. This change underscores how important handcrafted features are in improving performance, especially when the model needs to work in real-time scenarios. While MobileNet and ViT do a great job at capturing both local and global features, they cannot always pick up on these unique and sensitive textures and edge details that handcrafted features provide. These extra features are crucial, especially in challenging conditions like low light or when parts of the face are hidden. For instance, when differentiating between a comparable emotion, such as disgust and unconsciousness, the model must identify minute differences in facial textures, which would be challenging to recognize without these supplementary features. The model's capacity to distinguish between emotions that are closely related is compromised in their absence.

4.4.5. Without Cross-Attention

The cross-attention mechanism was eliminated in the last setup, leading to a model that extracts handcrafted as well as local and global characteristics with summation, followed by flattening and softmax without explicit integration of the cross-attention mechanism. Accuracy decreased to 80.4% as a result of the removal of the fusion mechanism, which also resulted in a decrease in F1-Score to 78%. While both local and global information is still available, the absence of this block makes the fusion step meaningless, so that the model struggles to utilize all feature sets appropriately to make optimal predictions. This is particularly problematic for emotions like unconscious or painful, where these delicate facial changes (mostly captured by local features) need to be interpreted in the context of the overall facial structure (mostly captured by global features). Without proper fusion, the model misses out on the full context, leading to more misclassifications. For example, when detecting painful expressions, small, localized wrinkles or the tightening of facial muscles need to be considered alongside the broader context of the face. Without cross-attention fusion, these fine-grained features are not combined with the overall facial context, leading to increased confusion with other emotions, such as sad or fear.

An ablation study was conducted by removing the cross-attention fusion module to assess its impact. As shown in Table 6, excluding this component led to a 7.1% reduction in accuracy and a 9% decline in F1-score in occluded conditions, underscoring the cross-attention mechanism's role in maintaining high performance by combining local, global, and texture-based information. This analysis highlights the importance of cross-attention in fusing distinct feature types, thereby enhancing the model's robustness across various challenging scenarios.

Table 6. Emphasizing the importance of cross-attention mechanisms.

Model Configuration	Accuracy (%)	F1-Score (%)	Occlusion Accuracy (%)
TriViT-Lite (full)	87.5	87	86.8
TriViT-Lite (without cross-attention)	80.4	78	79

The ablation study indicates that each component of the TriViT-Lite model, including MobileNet, Vision Transformer, handcrafted features, and the cross-attention fusion mechanism plays an important role in overall performance. In particular, the cross-attention fusion method is crucial for efficiently combining local and global features, and the MobileNet and ViT components are required for capturing global context and fine-grained details, respectively. The research emphasizes the beneficial effects of integrating local and global feature extraction methods. Removing any of these components significantly reduces performance, showing that the TriViT-Lite model's strength is its ability to evaluate local and global information simultaneously and combine them meaningfully.

5. Conclusion and Future Work

In this study, we presented the TriViT-Lite model, a novel approach, to the best of our knowledge, for patient facial emotion recognition that leverages the local and global, as well as handcrafted feature through the cross-attention architecture. The model's cross-attention fusion mechanism allows for an effective merging of these features, resulting in improved accuracy, especially in recognizing complex emotions like painful and unconscious. Our experiments demonstrated that TriViT-Lite outperforms several state-of-the-art models across multiple datasets, including our self-collected data, where the accurate detection of emotional states is critical. While the model showed great promise, there are still areas for further improvement.

Future research may concentrate on addressing the model's infrequent misclassification of delicate emotions and improving its capacity to manage dynamic expressions. An optimistic anticipation may be the incorporation of spatio-temporal in term of LSTM with our cross-attention mechanism. Another approach would be integrating multi-modal data, which represents a significant improvement. Combining facial emotion recognition with physiological signals, such as pulse rate, EEG signals, or respiratory patterns, can offer a more comprehensive comprehension of a patient's emotional state, particularly in the context of pain. Nonetheless, the TriViT-Lite model provides a robust and efficient solution for real-time emotion recognition, with potential applications in a wide range of disciplines, from healthcare to surveillance and human–computer interaction.

Author Contributions: W.R. and A.U. have contributed equally to this work and are the first coauthors. Conceptualization, W.R. and A.U.; methodology, W.R.; software, W.R.; validation, W.R. and A.U.; formal analysis, W.R.; investigation, W.R.; resources, J.J.; data curation, W.R.; writing—original draft preparation, W.R.; writing—review and editing, J.J.; visualization, J.J. and A.U.; supervision, J.J.; project administration, W.R. and J.J.; funding acquisition, W.R. and J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the school-level project of SZPU Grants (No. 6021330010K and No. 6023310026K), Post-doctoral Later-stage Foundation Project of Shenzhen Polytechnic University (No. 6021271014K), Shenzhen Polytechnic University Scientific Research Start-up Project (No. 6022312028K), and the Post-doctoral Foundation Project of Shenzhen Polytechnic University (No. 6024331008K).

Data Availability Statement: AffectNet and FER2013 datasets are publicly available datasets. The other data presented in this study are available on request from the main author.

Acknowledgments: We thank all the members taking part in the datasets collection including healthcare expert for their valuable time for annotation confirmations as well as all the co-authors for their research contributions.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Zaman, K.; Zengkang, G.; Zhaoyun, S.; Shah, S.M.; Riaz, W.; Ji, J.C.; Hussain, T.; Attar, R.W. A Novel Emotion Recognition System for Human–Robot Interaction (HRI) Using Deep Ensemble Classification. *Int. J. Intell. Syst.* **2025**, *2025*, 6611276. [CrossRef]
2. Lim, N. Cultural differences in emotion: Differences in emotional arousal level between the east and the west. *Integr. Med. Res.* **2016**, *5*, 105–109. [CrossRef] [PubMed]
3. Shan, C.; Gong, S.; McOwan, P.W. Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
4. Rasool, A.; Aslam, S.; Hussain, N.; Imtiaz, S.; Riaz, W. nBERT: Harnessing NLP for Emotion Recognition in Psychotherapy to Transform Mental Health Care. *Information* **2025**, *16*, 301. [CrossRef]
5. Avila, A.R.; Akhtar, Z.; Santos, J.F.; O'Shaughnessy, D.; Falk, T.H. Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild. *IEEE Trans. Affect. Comput.* **2021**, *12*, 177–188. [CrossRef]
6. Soleymani, M.; Pantic, M.; Pun, T. Multimodal emotion recognition in response to videos. *IEEE Trans. Affect. Comput.* **2012**, *3*, 211–223. [CrossRef]
7. Noroozi, F.; Marjanovic, M.; Njegus, A.; Escalera, S.; Anbarjafari, G. Audio-visual emotion recognition in video clips. *IEEE Trans. Affect. Comput.* **2019**, *10*, 60–75. [CrossRef]
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
9. Riaz, W.; Ji, J.; Zaman, K.; Zengkang, G. Neural Network-Based Emotion Classification in Medical Robotics: Anticipating Enhanced Human–Robot Interaction in Healthcare. *Electronics* **2025**, *14*, 1320. [CrossRef]
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Wien, Austria, 7–11 May 2024.
11. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing, Berlin, Germany, 3–7 November 2013; pp. 117–124.
12. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]
13. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 10–17 October 2021; pp. 10012–10022.
14. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
15. Wang, H.; Wei, S.; Fang, B. Facial expression recognition using iterative fusion of MO-HOG and deep features. *J. Supercomput.* **2020**, *76*, 3211–3221. [CrossRef]
16. Hashiguchi, R.; Tamaki, T. Temporal cross-attention for action recognition. In Proceedings of the 16th Asian Conference on Computer Vision (ACCV) Workshops, Macao, China, 4–8 December 2022; pp. 283–294. [CrossRef]
17. Turk, M.; Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **1991**, *3*, 71–86. [CrossRef]
18. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. In Proceedings of the Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
20. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv.* **2023**, *54*, 1–41. [CrossRef]
21. Zhao, Z.; Liu, Q. Former-DFER: Dynamic facial expression recognition transformer. In Proceedings of the 29th ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–2 November 2023.
22. Baltrusaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef]

23. Gowda, S.N.; Gao, B.; Clifton, D.A. FE-Adapter: Adapting image-based emotion classifiers to videos. In Proceedings of the 18th International Conference on Automatic Face and Gesture Recognition (FG), Istanbul, Turkey, 27–31 May 2024.
24. Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 87–102.
25. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; pp. 886–893. [CrossRef]
26. Chen, J.; Chen, Z.; Chi, Z.; Fu, H. Facial expression recognition based on facial components detection and HOG features. In Proceedings of the International Workshops on Electrical and Computer Engineering Subfields, Yogyakarta, Indonesia, 20–21 August 2014; pp. 884–888.
27. Sedaghatjoo, Z.; Hosseinzadeh, H. The use of the symmetric finite difference in the local binary pattern (symmetric LBP). *arXiv* **2024**, arXiv:2407.13178. [CrossRef]
28. Huang, D.; Shan, C.; Ardabilian, M.; Wang, Y.; Chen, L. Local binary patterns and its application to facial image analysis: A survey. *IEEE Trans. Syst. Man Cybern.* **2011**, *41*, 765–781. [CrossRef]
29. Ren, J.; Jiang, X.; Yuan, J. Face and facial expressions recognition and analysis. In *Context Aware Human-Robot and Human-Agent Interaction*; Springer International Publishing: Cham, Switzerland, 2015; pp. 3–29. [CrossRef]
30. Butt, M.H.F.; Li, J.P.; Ji, J.C.; Riaz, W.; Anwar, N.; Butt, F.F.; Ahmad, M.; Saboor, A.; Ali, A.; Uddin, M.Y. Intelligent tumor tissue classification for Hybrid Health Care Units. *Front. Med.* **2024**, *11*, 1385524. [CrossRef] [PubMed] [PubMed Central]
31. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 25–29 June 2009; pp. 248–255.
32. Gao, S.; Cheng, M.; Zhao, K.; Zhang, X.; Yang, M.; Torr, P.H.S. Res2Net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef]
33. Roy, A.K.; Kathania, H.K.; Sharma, A.; Dey, A.; Ansari, M.S.A. ResEmoteNet: Bridging Accuracy and Loss Reduction in Facial Emotion Recognition. *arXiv* **2024**, arXiv:2409.10545. [CrossRef]
34. Georgescu, M.-I.; Ionescu, R.T.; Popescu, M. Local Learning with Deep and Handcrafted Features for Facial Expression Recognition. *arXiv* **2018**, arXiv:1804.10892. [CrossRef]
35. Hasani, B.; Negi, P.S.; Mahoor, M.H. BReG-NeXt: Facial Affect Computing Using Adaptive Residual Networks With Bounded Gradient. *arXiv* **2020**, arXiv:2004.08495. [CrossRef]
36. Zhang, X.; Huang, Y.; Liu, H.; Gao, W.; Chen, W. HFE-Net: Hybrid Feature Extraction Network for Facial Expression Recognition. *PLoS ONE* **2023**, *18*, e0312359.
37. Kolahdouzi, M.; Sepas-Moghaddam, A.; Etemad, A. FaceTopoNet: Facial Expression Recognition using Face Topology Learning. *arXiv* **2022**, arXiv:2209.06322. [CrossRef]