Computing&AI Connect
ISSN: 3104-4719
2025, Article ID. 2025.0019, Cite as: https://doi.org/10.69709/CAIC.2025.102030

SCIFINITI

Research Article

OPEN ACCESS

AI Enabled Facial Emotion Recognition Using Low-Cost Thermal Cameras

James Thomas Black*, Muhammad Zeeshan Shakir

School of Computing, Engineering & Physical Sciences, University of the West of Scotland, Paisley PA1 2BE, UK

Article History
Submitted: March 11, 2025 Accepted: June 30, 2025 Published: July 21, 2025

Abstract
While expensive hardware has historically dominated emotion recognition, our research explores the viability of cost-effective alternatives by utilising IoT-based low-resolution cameras with Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs). In this work, we introduce a novel dataset specifically for thermal facial expression recognition and conduct a comprehensive performance analysis using ResNet, a standard ViT model developed by Google, and a modified ViT model tailored to be trained on smaller dataset sizes. This allows us to compare the efficacy of the more recent ViT architecture against the traditional CNN. Our findings reveal that not only do ViT models learn more swiftly than ResNet, but they also demonstrate superior performance across all metrics on our dataset. Furthermore, our investigation extends to the Kotani Thermal Facial Emotion (KTFE) test set where we evaluate the generalisation capability of these models when trained using a hybrid approach that combines our dataset with the KTFE dataset. Both ResNet and the ViT model by Google achieved high performance on the KTFE test samples, suggesting that leveraging diverse data sources can significantly strengthen model robustness and adaptability. This study highlights three critical implications: the promising role of accessible and affordable thermal imaging technology in emotion classification; the potential of ViT models to redefine state-of-the-art approaches in this domain; and the importance of dataset diversity in training models with greater generalisation power. By bridging the gap between affordability and sophistication, this research contributes valuable insights into the fields of emotion recognition and affective computing.

Keywords:
vision transformers; convolutional neural network; emotion recognition; thermal camera; facial expression recognition; internet of things

1. Introduction

Facial expression recognition (FER) is a crucial area of research within the fields of affective computing and computer vision. It involves the identification of human emotions based on facial cues or poses, which are essential for non-verbal communication. Starting in the 1970s, Paul Ekman and Wallace V. Friesen categorised expressions into six basic emotions: happiness, sadness, anger, fear, disgust, and surprise [1]. Nowadays, FER has become much more sophisticated by incorporating specialised hardware and a variety of techniques, including local feature analysis [2], holistic analysis [3], and deep learning techniques such as Convolutional Neural Networks (CNNs) [4].

Although the CNN architecture has predominantly dominated computer vision tasks such as image classification and object recognition [5], recent advancements in machine learning, particularly the introduction of Vision Transformers (ViTs) [6], have enabled FER systems to become more sophisticated and robust. Early research shows that ViTs often demonstrate superior performance over CNNs due to their lack of locality inductive bias, which allows them to capture global dependencies more effectively and leverage self-attention mechanisms to provide a more holistic representation of input data [7].

The intersection of emotion recognition technology and affordable hardware solutions has sparked significant interest within the research community. Traditionally, facial expression recognition has required expensive, spe-

* **Corresponding Author:**
James Thomas Black, School of Computing, Engineering & Physical Sciences, University of the West of Scotland, Paisley PA1 2BE, UK, james.black@uws.ac.uk

cialised equipment, often only available in a commercial environment. However, the advent of affordable hardware capable of delivering comparable performance to costly alternatives is reshaping the landscape of machine learning applications [8]. Moreover, the rapid development of IoT technology has inspired the utilisation of low-resolution cameras as a viable, cost-effective solution for different types of computer vision tasks [9]. Low-resolution imagery, once considered a hindrance, can now be harnessed effectively thanks to advancements in machine learning architectures. However, images in the visible spectrum pose privacy and security risks and often do not meet the ethical requirements for real-world applications. Therefore, in this work, we utilise a thermal camera, which is considered non-invasive.

Thermal imaging offers several key advantages over visible-spectrum cameras in FER tasks. First, thermal cameras are not affected by lighting variations, ambient light conditions, or shadows [10]. Second, thermal imaging captures more nuanced details related to physiological signals, such as heat distribution and blood flow, which correlate with emotional states and provide additional visible cues not present in RGB cameras [11]. Third, thermal data protects identity more effectively than RGB images due to its inability to capture fine-grained texture and colour information, making it a more privacy-conscious choice for real-world applications [12].

This research explores the potential of these technological developments by introducing a new dataset for thermal emotion classification and performing a comprehensive performance comparison using ResNet50 [13], a state-of-the-art CNN model, against two ViT models; one by Google [14] and another modified architecture tailored for smaller dataset sizes [15]. We also form a hybrid dataset with the state-of-the-art KTFE [16] to provide insights into the generalisation potential of these technologies. To the best of our knowledge, no research papers include ViT models when conducting emotion classification using images in the thermal spectrum, and no thermal camera dataset exists where data is collected using an IoT-based thermal camera. Our key contributions are as follows:

- Introduction of a novel thermal facial expression recognition dataset, University of the West of Scotland Thermal Faces (UWSTF), collected using low-cost thermal cameras
- Evaluation of low-cost, low-resolution thermal hardware for emotion recognition, assessing its performance using state-of-the-art models
- A side-by-side comparison of CNNs and ViT models on both our dataset and the hybrid dataset, addressing the question of bias produced by CNN models with vision transformers

The rest of the paper is organised as follows: Section 1.1 summarises related work, Section 2 describes the data collection methodology, data preprocessing steps, and model architecture, Section 3 details the results and describes the evaluation metrics and general discussion before the paper concludes in Section 5.

1.1. Related Work

In this section, we highlight various research articles related to thermal camera datasets and thermal facial expression recognition. We briefly summarise each paper and reflect on certain limitations or drawbacks of each approach, including how they could potentially be improved.

1.1.1. Thermal Camera Datasets

Thermal camera datasets have become increasingly significant in facial expression recognition. Although not as prevalent as visible spectrum datasets, thermal datasets serve as crucial resources for training and testing algorithms operating under varying conditions, poses, and expressions. A summary of prominent thermal camera datasets can be seen in Table 1.

A popular dataset among researchers is the KTFE dataset introduced by [17]. It focuses on various facial expressions; however, the demographic is limited to Vietnamese, Japanese, or Thai nationals, which could adversely affect models trained using this dataset. The same authors also released KTFE v2 [16], this time integrating different modalities for a more comprehensive analysis. Similar issues exist in this dataset, as the same range of nationalities was used.

Another popular dataset is NVIE, introduced by [18], which incorporates both visible and thermal spectrum images. This dataset is a comprehensive collection with different variations in angles of poses, glasses on and off, and various expressions. However, the dataset's complexity may pose challenges in terms of processing, and it is stored on a Chinese data storage platform, making it difficult for non-Chinese speakers to access. Other notable mentions of popular datasets include IRIS [19] and the NIST/Equinox database by Equinox Corporation (http://www.equinoxsensors.com/products/HID.html).

The Tufts Face Database, as described by [20], offers a comprehensive collection of images across multiple modalities, which include images in the thermal spectrum. This dataset encompasses over 10,000 images from a diverse group of individuals, and although the inclusion of other modalities allows for cross-comparison and benchmarking of algorithms, the dataset's complexity may pose challenges in terms of processing and analysis.

The authors in [21] introduced a dataset containing various posed expressions across different angles under controlled conditions. A notable weakness of their dataset is the limited diversity, as it primarily features a narrow demographic, which may affect the generalisability of algorithms developed using this data. Wesley et al. [22] also collected their data under controlled conditions and conducted a comparative analysis between thermal and visual modalities. This data may be limited, as environmental factors could influence the results of future inference by models trained using it, which may not be accounted for in all scenarios.

Kopaczka et al. [23] introduced a thermal dataset including rigorous annotation of facial expressions in their approach. The comprehensive nature of the annotations allows for detailed analysis of facial expressions. However, similar to [21], the demographic representation may be limited, as it is not explicitly addressed, which could affect the performance of models in real-world applications where diversity is a critical factor.

This literature highlights significant advancements in the development of thermal facial expression databases. While these studies provide valuable data and methodologies, they also share common weaknesses such as limited demographic diversity and environmental influences. Furthermore, these studies use expensive cameras, often unavailable to general consumers, to collect their data, which is a gap in the literature that this work intends to fill.

1.1.2. Thermal Facial Expression Recognition

Due to advancements in thermal datasets, there has been a significant impact on the performance of facial expression recognition models that utilise different architectures and preprocessing approaches. The authors in [24] introduced the InfraRed Facial Expression Network (IRFacExNet), a deep learning model that capitalises on thermal imaging features, such as low-light conditions where visible spectrum cameras struggle. Their model outperformed conventional methods. Similarly, focusing on low-light conditions, work in [25] utilises a CNN architecture and bases their analysis on four regions of the face. While their results were promising, the study's limited focus on only four facial regions may overlook the complexity of emotional expressions that involve broader facial movements. In [26], the authors present TIRFaceNet, with an architecture tailored for robustness against varying environmental conditions, and report admirable accuracy on both the DHU and DHUFO datasets. However, to the best of our knowledge, there are no insights into the model's limitations, and limited information or no reference is given for the dataset.

The availability of multimodal datasets has enabled researchers to conduct comparative analyses between thermal and visible spectrum images. Using the Gray Level Co-occurrence Matrix (GLCM) method to extract statistical features, Sathyamoorthy et al. [27] employed a custom CNN alongside SVM classifiers. Their findings indicate that thermal images outperformed visible images by achieving higher accuracy, although the proposed approach's reliance on statistical features may not fully capture the nuances of emotional expressions. Unlike statistical features, the authors in [28] proposed an architecture based on a modified version of ResNet152 to enhance feature extraction capabilities. Their study reported metrics; however, due to the complexity of the architecture, challenges may arise in terms of computational efficiency and applications operating in real-time.

Many popular models based on the CNN architecture exist, and often approaches use the base model and conduct transfer learning. Kamath et al. [29] present TERNet—an architecture adopting the VGG-Face CNN model. Their findings highlight the potential of transfer learning in enhancing facial expression recognition accuracy while bypassing the need to train complex models from scratch. The authors in [30] conducted a comprehensive analysis of feature-based facial expression recognition, which provided valuable insights into feature extraction techniques both locally and globally. Contrary to [29], their findings reported that features extracted using the VGG CNN network performed poorest on the KTFE and NVIE datasets. This suggests that different models and approaches are better suited to different datasets. Also adopting a transfer learning approach, the authors in [31] utilise AlexNet for feature extraction and classification. They enhanced the architecture by creating a hybrid model that combined AlexNet and SVM to yield the best results, with thermal images' accuracy outperforming visible images.

The reviewed work highlights the many approaches that can be taken utilising thermal images, machine learning classifiers, deep learning models, and transfer learning, which show promise in improving recognition accuracy. However, while deep learning models have shown significant improvements, issues related to overfitting, computational efficiency, and real-time applicability remain under-explored. Furthermore, existing work tends to use datasets where data has been collected using expensive and often commercial thermal cameras. In this work, we intend to demonstrate model performance using accessible hardware combined with deep learning techniques.

Table 1: Comparison of thermal facial expression datasets.

Dataset	Modality	Expressions Included	Demographic Diversity	Elicitation Method	Limitations
KTFE [17]	Thermal	anger, disgust, fear, happy, sad, surprise, neutral	Low	Spontaneous	Limited demographic diversity; restricted to specific nationalities (Vietnamese, Japanese, Thai)
KTFE v2 [16]	Multimodal	anger, disgust, fear, happy, sad, surprise, neutral	Low	Spontaneous	Same demographic limitations as KTFE (Vietnamese, Japanese, Thai); limited generalisability
NVIE [18]	Multimodal	anger, disgust, fear, happy, sad, surprise	Moderate	Posed, Spontaneous	Complex data structure and preprocessing; hosted on a Chinese platform, making access difficult for non-Chinese users
IRIS [19]	Multimodal	anger, laughing, surprise	Moderate	Posed	Lacks detailed documentation on expressions, demographics, and elicitation methods
NIST Equinox	Thermal	smiling, frowning, surprised	Unknown	Posed	Limited public documentation; commercial dataset with restricted access; unclear participant demographics and methodology
Tufts [20]	Multimodal	neutral, smile, closed eyes, shocked	High	Posed	Complex multimodal structure not tailored specifically for facial expression recognition; thermal data can be difficult to isolate; higher processing overhead
Kowalski et al. [21]	Thermal	smiling, sad, surprise, angry	Limited	Posed	Demographic details such as age, gender, and ethnicity not reported; limited subject diversity
Wesley et al. [22]	Multimodal	surprise, fear, sadness, disgust, anger, happiness	Moderate	Posed	Small sample size; limited scalability to real-world environments with natural variability
Kopaczka et al. [23]	Thermal	neutral, happiness, sadness, surprise	Limited	Posed	Small participant pool with limited demographic diversity; demographic details (age, gender, ethnicity) not specified; expressions limited to a subset of basic emotions
University of the West of Scotland Thermal Faces (UWSTF)	Thermal	anger, fear, happy, neutral, sad, surprise	Moderate	Posed, Spontaneous	Small sample size; lacks in-the-wild data

2. Materials and Methods

In this section, we explain the data collection methodology by providing information on how participants were recruited, how emotions were stimulated, what equipment was used, what the environment setup looked like, how the dataset was curated, what the preprocessing stages entailed, and information on the architecture of the CNN and ViT models.

2.1. Participant Recruitment

Fourteen male participants, aged between 21 and 42, were recruited. They were either post-doctorate researchers or PhD students at the University of the West of Scotland in the UK. The participants hailed from various countries, including Scotland, England, China, Romania, and Pakistan. They were recruited through internal communications within the research group. The study received ethical approval from the University of the West of Scotland's ethics board under project #18323. None of the participants wore glasses.

Each participant was assigned a specific date and time for data collection. Upon arrival, they were provided with the necessary information sheets and consent forms, which were signed before any exercises commenced.

In addition to recording age, gender, and nationality, participants were asked whether they had any known affective disorders or prior experience with tasks involving facial expression recognition. None reported any affective history or prior exposure to such tasks. While the participant pool consisted of individuals in a research setting, this background was not associated with specialised training in affective computing. These measures were taken to minimise potential selection bias and enhance the generalisability of the dataset within the scope of this exploratory study.

2.2. Stimuli

A series of video clips was selected by the authors to elicit spontaneous facial expressions, comprising 2 clips each for anger, fear, happiness, sadness, and surprise, and 9 clips for the neutral expression. Videos have been effective stimuli for eliciting a wide range of emotions [32,33]. Short videos, such as YouTube Shorts and TikTok, are extremely popular and were used to keep participants engaged [34].

The selection process was informed by prior work in similar thermal datasets such as KTFE [16] and NVIE [18], and validated through participant self-reports immediately following each video clip. In KTFE, the stimuli were selected by the authors based on cultural relevance, and participants self-reported their emotional responses using standardized scales. In NVIE, emotional videos were used to elicit spontaneous expressions, and self-reported responses were used to confirm the appropriateness of each clip, with mismatches excluded.

Similarly, we chose videos based on cultural relevance and public viewer comments indicating strong emotional responses. After each clip, participants reported how the video made them feel, and only clips where the intended emotion was confirmed by participant feedback were retained in the dataset. While most emotions during spontaneous elicitation were successfully captured—particularly happiness and surprise—some emotions proved more difficult to evoke, as also reported in previous work [35].

For the posed facial expression data, emoticons representing each facial expression were displayed to prompt participants. Rather than imitating the emoticons directly, participants were asked to produce expressions that felt natural to them for each emotion category. Emoticons were selected for their simplicity and cultural familiarity, which have been shown to support intuitive emotional interpretation and expression [36].

2.3. Equipment

A TOPDON TC001 thermal camera was used in this study to collect thermal images. The TC001 operates in a spectral range of 8-14 micrometres and has a resolution of 256 × 192 pixels. The camera can record at a rate of up to 25 frames per second and has a temperature range of -20 to 150 degrees Celsius with a temperature accuracy of ±2 degrees Celsius. The camera also features a Noise Equivalent Temperature Difference of 40 milli-Kelvin at 25 degrees Celsius and operates with a power consumption of 0.35W. The TC001 weighs 30 g and measures 71 × 42 × 14 mm.

The camera was connected to an ASUS Vivobook laptop running Windows 11, which had the TCView software installed. A table was used to place the laptop, and a 50-inch Samsung TV was used to show video clips and images of the emoticons. A tripod was employed to hold the camera in position, and a traditional office chair was provided for participants to sit on.

2.4. Environment Setup

The data was collected inside a lab at the University of the West of Scotland. The room was an approximately 24 square metre L-shaped area, featuring four windows on one wall and five windows on the other side. The room was not empty, as there were various desks with computer monitors situated around the perimeter. A small 3 × 3 metre area in the room was used as the data collection area, the arrangement of which can be seen in Figure 1.

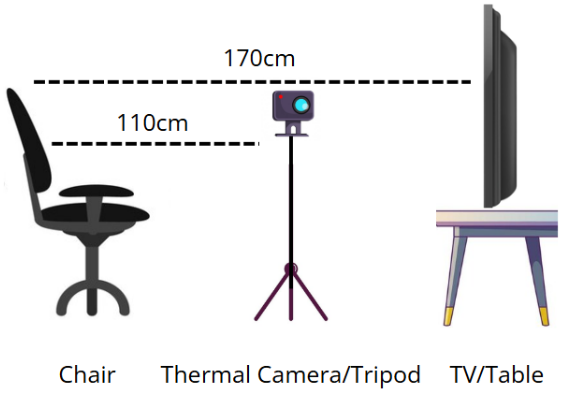

Figure 1: The approximate data collection set up.

To ensure participant discretion and comfort, data collection was conducted in a dedicated and quiet section of the lab with minimal distractions. The 3 × 3 metre area used for recording was positioned away from windows and foot traffic. Blinds were drawn during data collection to reduce visual distractions and control natural light. Only essential personnel were present, and no observers were in the direct line of sight of participants. Participants were seated alone during recordings and instructed to behave naturally, with the assurance that there were no 'correct' emotional responses. They were also informed that they could withdraw at any time without consequence. Room temperature and lighting conditions were kept consistent across all sessions to avoid external influences on emotional expression. The thermal camera setup was non-intrusive, and no visible facial markers or physical contact were used to preserve a neutral and comfortable environment.

2.5. Protocol

Spontaneous and posed facial expressions were collected as part of this study. Before the collection began, each participant was given information about the process and the option to opt out at any time. First, using the TC001 camera recording at 25 frames per second, spontaneous emotions were collected. This involved participants watching 19 short video clips, presented in a random order, on a TV with the camera placed between the TV and the participant. Starting with a sad short video followed by a neutral clip, this pattern continued by first playing a short video that elicited an emotion other than neutral, followed by a neutral video clip, an approach taken by [18] to help alleviate any negative feelings. After each short video clip, the participant reported how that clip made them feel from the six emotions: anger, fear, happiness, neutrality, sadness, surprise, or other. This is a similar approach to that taken by [16]. Once all short videos were watched and emotions were reported, the video recording of the participant was stopped.

After the spontaneous collection was complete, participants were asked if they were comfortable proceeding to the posed collection activity, to which all responded affirmatively. To record the posed emotions, the same setup was used as in the spontaneous collection, with the TC001 recording. However, instead of playing video clips, the participant was asked to mimic the emoticon displayed on the TV. The participants were asked to mimic each posed expression in a random order, three times for three seconds each, reverting to a neutral pose between expressions. After the posed data collection was complete, each participant was debriefed to ensure that they were not affected by any of the materials used to induce the emotions. None of the participants reported feeling any adverse effects, concluding the data collection process.

2.6. Dataset Curation

Separate thermal video recordings were created for each emotion-eliciting clip, allowing individual emotional responses to be evaluated in isolation. After data collection was completed, the recordings were saved in MP4 format and prepared for frame extraction. This separation ensured that only frames corresponding to a specific emotional stimulus were considered during the annotation process. We adhered to the following process when collecting the data:

- Start video clip
- Start thermal recording
- Stop the video clip
- Stop thermal recording
- Collect participants' self-reported emotions

To create the thermal facial expression dataset, each spontaneous and posed video recording was exported into individual frames in JPG format using the OpenCV library in Python [37]. The selection process was performed manually by the authors, as shown in Algorithm 1. Every frame was reviewed and evaluated based on the participant's self-reported emotion following each video clip, as well as visible facial cues in the thermal image, such as changes in the cheek, eye, or mouth regions. Only frames that reflected the intended emotional state were included in the dataset.

This manual annotation process follows similar practices in thermal FER literature. For example, Kopaczka et al. [23] performed manual frame selection and verified annotations through expert agreement, while the KTFE dataset [16] relied on participant self-reports and con-

Algorithm 1 Thermal Frame Annotation Procedure.

```
for each thermal video clip do
    Convert video to image frames
    if participant's reported emotion matches the intended emotion of the clip then
        for each extracted frame do
            if thermal facial cues are visibly aligned with reported emotion then
                Add frame to dataset with corresponding emotion label
            end if
        end for
    end if
end for
```

trolled experimental protocols to guide the inclusion of expression samples. Each selected image was then assigned to a directory corresponding to one of six emotion classes and prepared for preprocessing.

As a result, a total of 61 images for anger, 130 images for happiness, 97 images for neutrality, 92 images for sadness, and 178 images for surprise. Figure 2 shows an example of each expression from the exported frames.

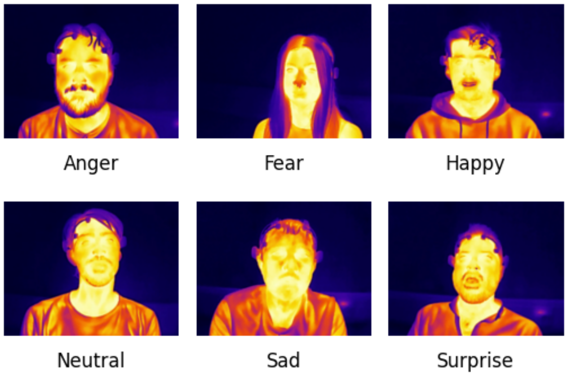

Figure 2: An example image of each emotion in the dataset.

2.7. Preprocessing

Various steps were taken to prepare each image in the dataset for model training. First, a Python script was created in which the images were loaded from the directories previously created during the dataset curation step using the OpenCV library [37]. Next, a pre-trained Haar Cascades classifier was used on each image to detect whether a face was present. As expected, due to the manual diligence in the dataset curation step, a face was successfully found in each image.

Using the coordinates of the face's position identified by the Haar Cascades model, the face was cropped from each image, the remaining background was removed, the image was converted to greyscale, and finally normalised. The image was then resized to 48 × 48, which are common dimensions for many types of image classification tasks, as it is large enough for models to pick up nuances and small enough to be efficient for the training stage and inference. The image channels were then merged to ensure it had three channels, as this format is required by certain models. Finally, the dataset was split into training, validation, and testing sets using a 60/20/20 ratio.

2.8. Convolutional Neural Network

Convolutional Neural Networks (CNNs) are deep learning models primarily used for structured grid data, such as images, using a series of layers that include convolutional layers, pooling layers, and fully connected layers [38]. CNNs employ kernels in the convolutional layers that slide over the input data to extract local features, which are crucial for the model in recognising patterns in images. These fixed patterns can be a limitation of the CNN model, but they save training time as the model doesn't need to learn how to focus.

This process is followed by activation functions, typically Rectified Linear Units (ReLU), which introduce non-linearity into the model, enabling it to learn complex patterns. To reduce the spatial dimensions of feature maps, pooling layers are used, often employing max pooling or average pooling. These layers decrease the computational load and help prevent overfitting. This hierarchical structure enables CNNs to learn increasingly abstract features with each layer, from simple edges in the initial layers to more complex shapes and objects in the deeper layers [39].

CNNs have gained prominence in facial expression recognition due to their ability to learn features from facial images without the need for manual feature extraction, which was a limitation of traditional machine learning methods. The architecture of CNNs allows them to effectively capture both low-level features, such as basic textures and edges, and high-level features, such as facial expressions. Recent advances in deep learning have significantly improved the performance of facial expression recognition models, enabling them to operate effectively in challenging conditions such as variations in lighting, head pose, and angles [40].

To further enhance the performance of CNNs, some approaches leverage transfer learning, where pre-trained CNN models are fine-tuned on facial expression recognition datasets, allowing for better generalisation and performance on new data [41]. In general, CNNs combine their robust feature extraction capabilities with advanced training techniques to achieve high accuracy and reliability, representing a powerful tool for recognising human emotions from facial expressions. In this work, we fine-tune the ResNet50 pre-trained model [13] using the Keras deep learning API [42].

2.9. Vision Transformers

The introduction of the transformer model [43] has significantly influenced natural language processing and has recently advanced into vision and multimodal applications. In recent years, state-of-the-art image classification has been dominated by CNNs until the introduction of vision transformer (ViT) models. Unlike CNNs, which rely on local receptive fields and hierarchical feature extraction, ViTs utilise a self-attention mechanism. This mechanism allows the model to capture global dependencies across the entire image by dividing the image into patches, which are then linearly embedded into a sequence of tokens for the transformer to process [6]. The self-attention mechanism enables the model to assess the importance of different patches, facilitating a better understanding of spatial relationships and contextual information within the image.

ViTs have shown promising performance in the field of facial expression recognition, specifically in challenging scenarios such as varying lighting conditions and occlusions. The authors in [44] present TFE (Transformer Architecture for Occlusion Aware Facial Expression Recognition), demonstrating that transformers can effectively manage occlusions by focusing on relevant facial features while ignoring irrelevant background noise. Additionally, although different types of ViT models exist, Li et al. [45] employs the Mask Vision Transformer (MVT) model, which incorporates a mask generation network to filter out complex backgrounds and occlusions, a technique that further enhances the model's robustness, especially for facial expression recognition in the wild.

Moreover, ViTs address the limitations inherent in CNNs. Traditional CNNs rely on local features and often struggle with fixed-size input requirements [46]. The global receptive field of transformers allows them to learn more comprehensive representations of facial expressions by utilising positional embeddings, enabling the model to capture nuances within the data in unexpected ways. However, this comes at a cost: the training time increases, and larger amounts of training data are required, resulting in high computational resource demands. Research using ViT models for thermal facial expression recognition is not as prevalent as work involving CNNs. In this study, the aim is to address this gap by comparing the performance of a modified ViT model described in [15], herein referred to as the modified ViT, designed for training on smaller datasets utilising the Keras API, and fine-tuning on a pre-trained ViT model developed by researchers from Google [14] using the Hugging Face platform [47].

3. Results

In this section, we describe the feature extraction process and the evaluation metrics chosen to assess the performance of the models. We compare training and testing performance and provide insights into generalisation capabilities by incorporating test data from the KTFE dataset.

3.1. Feature Extraction

To evaluate the performance of the CNN architecture, the feature extraction process utilised the pre-trained ResNet model's convolutional layers as a foundational component. This approach processes input images by leveraging the model's pre-trained weights on the ImageNet dataset [48]. The output of the convolutional layers captures high-level abstract representations of the input images, which are then passed to subsequent layers for further processing. These layers focus on feature refinement and classification based on the extracted features. Some studies have shown that using regions of interest can enhance the performance of CNNs [49], while other work emphasises the importance of using the entire face to capture temperature changes across the facial surface [50].

The pre-trained ViT model developed by Google extracts features following the standard architecture of the model. First, it divides the input image into non-overlapping patches of 16×16 pixels. Patch embeddings are created by flattening each patch before linearly projecting it into a vector. Positional embeddings are added to the patch embeddings to incorporate spatial information. These embeddings serve as input tokens for the Transformer, where the core feature extraction occurs within the Transformer encoder, which comprises several layers. To capture intricate relationships between patches, each layer features multi-head self-attention, a feed-forward neural network (Multi-Layer-Perceptron) for non-linear feature extraction, and layer normalisation to stabilise the network, resulting in a sequence of rich token embeddings.

The modified ViT model also follows the standard architecture of the model in the feature extraction process; however, it implements further customisations. For instance, it tokenises patches using shifted patch tokenisation, whereas the Google model utilises a simpler patch embedding mechanism based on fixed-size non-overlapping patches. Furthermore, the Google model relies on standard multi-head attention, whereas the modified model introduces a trainable temperature parameter to scale the query vectors, providing more flexibility compared to the standard multi-head attention mechanism. Despite these variations, both ViT models utilise the transformative power of attention mechanisms to capture complex relationships for effective feature extraction.

3.2. Evaluation Metrics

To evaluate the effectiveness of the models, the following performance metrics were included: accuracy, precision, recall, and F1 score. Confusion matrices were also utilised to provide insights into accuracy and error patterns across the different emotion categories.

- Accuracy—a straightforward metric that offers an overall sense of model performance by measuring the proportion of correctly classified instances over the total number of instances.
- Precision—a metric that reflects the model's ability to identify positive samples correctly by measuring the ratio of true positive instances to the sum of true positive and false positive instances.
- Recall—a metric that evaluates the model's capability to capture all relevant instances by measuring the ratio of true positive instances to the sum of true positive and false negative instances.
- F1 Score—a metric that balances the importance of precision and recall by measuring the harmonic mean of precision and recall.

3.3. Comparison of Model Training

To train each model, the data was split into 60/20/20 for training/validation/testing. Each model underwent 200 epochs of training; however, early stopping callbacks were implemented to prevent overfitting. These callbacks were based on the evaluation of accuracy and loss from the training and validation data and included a function to reduce the training learning rate when a plateau was detected.

Starting low but gradually increasing over time, the training process for the ResNet model lasted for 125 epochs before stopping early. Figure 3 illustrates that the validation accuracy initially surpassed the training accuracy, suggesting effective generalisation to the validation data.

The ViT model by Google exhibited a rapid learning process, as shown in Figure 4, achieving near-perfect accuracy in just a few epochs. In contrast to the ResNet model, the training accuracy exceeded the validation accuracy early in the training process, although both followed similar trajectories thereafter. This indicates that the model learned quickly and overfitted the training dataset, suggesting that the early stopping patience was insufficient to prevent overfitting. However, the validation accuracy was still respectable, highlighting the model's generalisation capabilities.

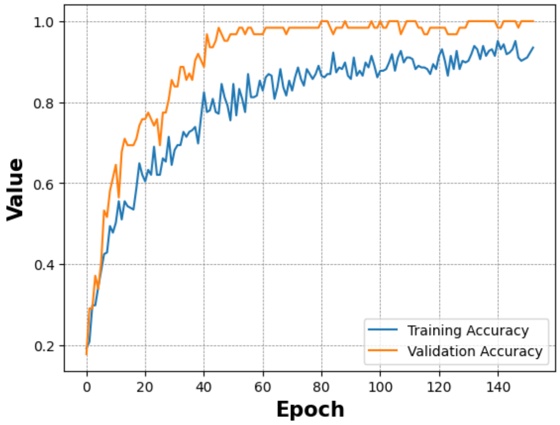

Figure 3: ResNet training and validation accuracy.

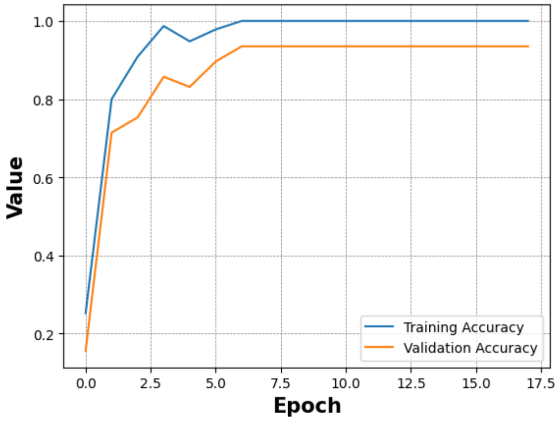

Figure 4: ViT by Google training and validation accuracy.

Figure 5 shows the modified ViT model, which demonstrated a swift learning process, terminating training early at 46 epochs. Despite minor fluctuations, both training and validation accuracies stabilised early and maintained high accuracy, indicating efficient generalisation and adaptation.

In general, each model showed different behaviours in the training process, highlighting their unique characteristics and learning speeds. Notably, the ViT models managed to learn in far fewer epochs than the ResNet model.

3.4. Comparison of Model Performance

After the training process, the models' performance was evaluated using the test data from the training dataset split, with results reported using the metrics described in Section 3.2. Macro averages were used to provide insights across class distributions. The results are summarised in Table 2 while Table 3 shows the results of each model at a more granular level for each class.

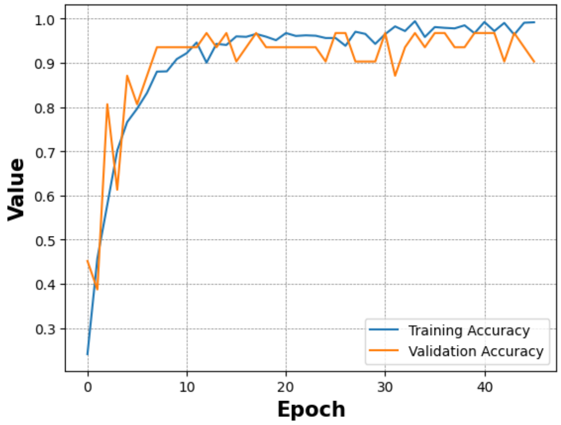

Figure 5: Modified ViT model training and validation accuracy.

Table 2: Each model's performance for Accuracy, Precision, Recall, and F1-Score.

Model	Accuracy	Precision	Recall	F1-Score
ResNet	95	92	96	94
Google ViT	96	96	95	95
Modified ViT	96	97	96	96

Table 3: The performance of each model for each class.

Model	Metric	An	Fe	Ha	Ne	Sa	Su
ResNet	Precision	92	100	100	75	88	100
	Recall	86	100	86	100	100	100
	F1	92	100	93	86	93	100
Google ViT	Precision	100	89	100	89	100	96
	Recall	100	89	100	89	93	100
	F1	100	89	100	89	96	98
Modified ViT	Precision	100	100	085	100	100	100
	Recall	100	100	100	82	93	100
	F1	92	100	92	90	97	100

The ResNet model demonstrated high accuracy at 95. The precision reached 92, indicating a respectable rate of true positive predictions. Recall was slightly higher at 96, suggesting that the model effectively identified the majority of positive instances, resulting in an F1 score of 94.

The model achieved perfect scores when predicting Fear and Surprise; however, it showed some limitations in classifying Neutral and Anger emotions. Although precision and F1 score for the Anger emotion were high, the recall was comparatively lower at 86. In contrast, for the Neutral emotion, precision was relatively low, but both the F1 score and recall were high. Similarly, although precision was high for the Sad emotion, it was the lowest scoring metric with a high F1 score and perfect recall.

The ViT model by Google achieved slightly higher accuracy than the ResNet model, reaching 96. The model also achieved considerably more precision at 96. Finally, with a recall of 95, the model maintained effective detection capabilities, albeit slightly less sensitive than its precision, culminating in an F1 score of 95. Unlike ResNet, this model managed to identify the Anger and Happy emotions with perfect precision, recall and F1 scores. Performance deteriorated when identifying Fear and Neutral classes, although metrics still remained high. For the Sad emotion, the model presented excellent precision of 100 and a respectable recall and F1 score, and when predicting Surprise, the model performed exceptionally well, closely approaching perfect scores.

The modified ViT model excelled and surpassed ResNet and Google ViT in almost all metrics. Although the same accuracy score was achieved as the model by Google, the modified ViT achieved 97 for precision, demonstrating a higher proportion of correct positive predictions. Its recall matched the accuracy score of 96, resulting in an F1 score of 96, marking the modified ViT as the top-performing model in our comparison. Similar to ResNet, this model achieved perfect scores for the Surprise and Fear classes, while detecting Anger was also near-perfect, with an F1 score of 92. The model experienced some challenges with the Happy emotion, where model precision and F1 score were lower despite a perfect recall score. This model achieved the highest precision for Neutral, yet recall slightly lagged, resulting in an F1 score of 90. This was similar to the Sad emotion; however, recall was higher, which resulted in an F1 score of 97.

In summary, while all models exhibited high performance across metrics, ResNet showed strong general performance, particularly for Fear and Surprise classes, while the ViT by Google excelled with Anger and Happy recognition. The modified ViT revealed areas for improvement but demonstrated impressive precision for Surprise and Sad emotions. The modified ViT outperformed the others, particularly excelling in overall precision and F1 score, underscoring its efficacy in thermal emotion classification.

3.5. Model Generalisation

In our study, we evaluated the generalisation capabilities of thermal emotion classification using learning curves—a method that provides insight into a model's performance as the training dataset size gradually increases [51]. We tested our models on different portions of the dataset, ranging from 50% to 100%. Our evaluation focused on two scenarios: using only the KTFE dataset and using a mixed dataset comprising KTFE test samples alongside a 20% test split from our own collected dataset.

For the KTFE dataset, 90 samples were selected as a fixed test set across all generalisation experiments. The remaining 359 images were used for training, with subsets ranging from 50% to 100% of the available training data. The KTFE dataset contained 449 images. In the mixed dataset configuration, an additional 106 test samples from our dataset were included alongside the KTFE test items and these test sets remained unchanged across all training set size comparisons. To ensure reproducibility and consistent evaluation, a fixed random seed (42) was used during preprocessing.

Although some models showed signs of overfitting during training, early stopping was implemented based on validation accuracy and loss, and the best-performing model checkpoint in the validation set was restored for evaluation.

As shown in Figure 6, the ResNet model achieved an initial accuracy of 58% when half of the KTFE dataset was utilised, which increased steadily as more of the training data were incorporated, eventually peaking at 84% with 90% of the dataset—the highest accuracy recorded for the KTFE test data—before dropping slightly to 82% when the complete dataset was utilised. Figure 7 shows, in contrast, testing accuracy on the hybrid dataset remained consistently around the high 80s and 90% mark, starting at 89% when using 50% of the data, suggesting strong model generalisation capabilities.

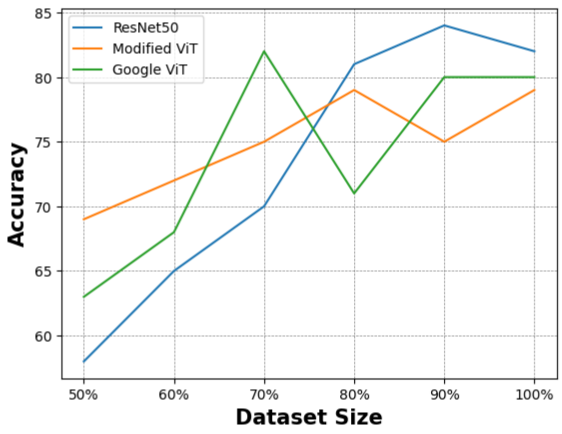

Figure 6: Model accuracy on different sizes of the KTFE dataset.

The Google ViT model's results displayed in Figure 6 fluctuated when testing on the KTFE dataset, starting at 63% accuracy before sharply increasing to 82% when using 70% of the dataset as part of the training data. This was followed by a slight decline in performance, which then stabilized at approximately 80% as larger portions of the dataset were utilized. In the mixed testing data shown in Figure 7, the model started with an impressive 88% accuracy using only 50% of the KTFE dataset in the training data before reaching a high of 94% with 70% of the training data used—the highest accuracy recorded—followed by a slight decrease in performance but remaining above 88%.

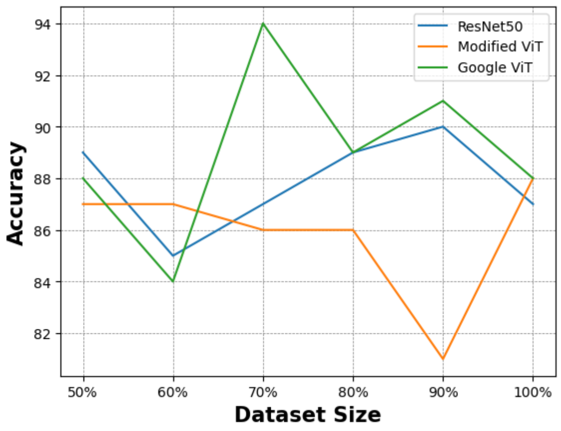

Figure 7: Model accuracy on different sizes of the mixed dataset.

When using the modified ViT model, accuracy on the KTFE test set started at 69% with 50% of the KTFE dataset used in the training data before peaking at 79% when using 80% of the data, then dropping slightly as seen in Figure 6. In the mixed test set shown in Figure 7, the model showed a strong initial accuracy of 87% with 50% of the KTFE training data included, and the accuracy continued to vary across the different dataset sizes while remaining above 80%. The model peaked at 88% when the full dataset was used during the training process.

4. Discussion

The study aimed to evaluate the performance of different deep learning models in classifying emotions using thermal cameras. The performance of ResNet, Google ViT, and a modified ViT was compared, providing some interesting insights. The ResNet model, comprising a CNN architecture, displayed robust general performance, excelling particularly in classifying Fear and Surprise emotions with perfect precision and performing well in test instances involving the KTFE dataset. Meanwhile, both ViT models yielded slightly higher overall accuracies, with the modified ViT achieving the highest precision across all tests. In particular, the Google ViT achieved perfect scores for Anger and Happy emotions and demonstrated exceptional generalisation capabilities when tests included the KTFE dataset.

Although the ResNet architecture has been proven effective for facial expression recognition [52,53], the results in this work suggest that ViT models—especially the modified ViT—are particularly adept at thermal facial expression recognition. This may be attributed to their architecture, which enables more nuanced feature extraction [54]. The rapid learning process, portrayed in the results and overall performance of the ViT models, aligns with previous findings indicating their advantage in image classification tasks [55–57].

Moreover, the use of low-cost cameras and making thermal emotion classification more accessible could have a substantial impact in fields such as affective computing and healthcare. This data could be used to train models and form part of advanced systems that monitor emotional well-being or enhance human-computer interaction by accurately interpreting emotional states in real-time settings. The results show that the generalisation capabilities provide the potential to capture wider variability and incorporate diverse data, which is pivotal for real-world applications where data conditions can significantly vary. As the thermal spectrum is also considered non-invasive data [12], this work contributes meaningfully to developing ethical applications, where privacy and security risks are prevalent.

Nonetheless, this study does contain some constraints and limitations. One significant limitation is the dataset size and diversity. While the models performed well, the dataset does not contain any "in-the-wild" data [58], which might feature variations in head angles, head rotation, and lighting or shadows. In addition, while gender and ethnic diversity were not the main focus of this work, the dataset could be improved and considered more balanced with a better male-to-female ratio and increased ethnic diversity. Further, this work shows that using this dataset to form a hybrid dataset can yield respectable results and offers a potential solution when only low-cost cameras are available, yet data collection remains difficult.

To address these limitations, future work should focus on expanding the dataset, incorporating more diverse test participants, and collecting data under conditions that better reflect real-world scenarios. Exploring hybrid datasets is another promising direction, particularly to determine the optimal ratio of state-of-the-art to low-cost data for maximal model performance on other prominent thermal datasets.

Another avenue worth exploring is the development of hybrid models that combine the strengths of convolutions and transformers, which may offer enhanced performance in specific contexts—especially when both local feature extraction and global context are valuable. Techniques such as transfer learning could also be more deeply leveraged to maintain high performance across diverse environments without extensive retraining. In this study, we applied transfer learning by fine-tuning a pre-trained Google ViT model on our thermal dataset, demonstrating strong generalisation even with limited training data. Additionally, convolutional methods such as region-based cropping or attention-guided feature enhancement could be employed to refine input representations in hybrid architectures.

Lastly, while this work primarily contributes to a novel dataset and comparative model evaluation, the findings have broader implications for real-world system integration. The rapid convergence and inference speed of the ViT models—particularly the modified version trained on smaller datasets—suggest viability for deployment on edge devices or embedded systems. Real-time metrics such as latency, throughput, and energy efficiency could be benchmarked in future implementations, especially in applications like driver monitoring, emotion-aware user interfaces, or mental health support tools [59–61]. Testing these models on platforms such as Raspberry Pi or NVIDIA Jetson would help identify hardware limitations and inform system architecture decisions under real-world constraints.

Although this dataset is relatively small in scale, it offers several advantages that support reliable model evaluation, including moderate participant diversity across age and national background. Along with its structured class distribution and controlled elicitation protocol, these factors enhance its utility as a benchmark for evaluating models in low-cost, real-world emotion recognition scenarios.

5. Conclusions

This work introduces our low-cost thermal facial expression recognition dataset and explores the potential of ViT models as a viable alternative to traditional CNNs for thermal emotion classification, particularly within the context of low-resolution imagery obtained from IoT-based low-cost thermal cameras. Two ViT models were used: one developed by Google and a modified model. Additionally, ResNet was used for comparison within the CNN architecture. Our findings highlight several key insights that contribute to the fields of emotion recognition technology and affective computing.

Compared to ResNet, the ViT models demonstrated a rapid learning capability, with the modified ViT model outperforming both ResNet and the Google model on our newly introduced dataset. Furthermore, our results indicate that diverse training data can significantly enhance model performance, as the models achieved high accuracy using a hybrid dataset, combining the KTFE dataset with our own. The Google ViT model yielded an overall accuracy of 94%.

In conclusion, these findings demonstrate the impact that affordable hardware and state-of-the-art model architectures can have on emotion recognition. This work shows that leveraging low-cost, low-resolution cameras with advanced models such as ViTs can make emotion recognition systems more accessible and scalable across varying applications. The study not only highlights the superiority of ViT models over traditional CNNs but also emphasises the role of diverse datasets in training machine learning models. In addition to its affordability, the dataset's structured composition, manual annotation process, and inclusion of participants from five national backgrounds provide a solid foundation for assessing FER models under controlled and reproducible conditions.

Further exploration and refinement of ViT models and low-cost hardware solutions may potentially revolutionise their application in real-world scenarios. Specifically, future work should investigate lightweight ViT architectures optimised for edge devices, explore hybrid models that integrate CNNs and transformers for improved efficiency, and evaluate deployment on embedded platforms to assess latency, power consumption, and inference speed under practical constraints.

List of Abbreviations

CNN	Convolutional Neural Network
FER	Facial Expression Recognition
IoT	Internet of Things
KTFE	Kotani Thermal Facial Emotion
NVIE	Natural Visible and Infrared Facial Expression
ResNet	Residual Neural Network
RGB	Red, Green, Blue (visible light spectrum)
UWSTF	University of the West of Scotland Thermal Faces
ViT	Vision Transformer

Author Contributions

J.T.B. conceptualized the study, developed the methodology, implemented the software, performed validation and formal analysis, conducted the investigation, curated the data, prepared the original draft of the manuscript, and created the visualizations. M.Z.S. supervised the work, administered the project, and reviewed and edited the manuscript.

Availability of Data and Materials

The dataset described in this work is available to download at https://zenodo.org/records/15830552.

Ethics Committee Approval and Consent to Participate

Written informed consent was obtained from all participants prior to data collection. The study was approved by the ethics board at the University of the West of Scotland under project #18323.

Consent for Publication

Participants were informed that thermal images collected during the study could be used for publication under the Creative Commons Attribution License 4.0. All participants were adults over the age of 18.

Conflicts of Interest

The authors have no conflicts of interest to declare.

Funding

No external funding was received for this research.

Acknowledgments

The authors would like to thank the participants from the University of the West of Scotland who volunteered their time to partake in the data collection.

References

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personal. Soc. Psychol.*, vol. 17, no. 2, p. 124, 1971. [CrossRef]

[2] D. Ghimire, S. Jeong, J. Lee, S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimed. Tools Appl.*, vol. 76, pp. 7803–7821, 2017. [CrossRef]

[3] S. Happy, A. Dantcheva, F. Bremond, "Expression recognition with deep features extracted from holistic and part-based models," *Image Vis. Comput.*, vol. 105, p. 104038, 2021. [CrossRef]

[4] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, 2020. [CrossRef]

[5] B. Jena, G. K. Nayak, S. Saxena, "Convolutional neural network and its pretrained models for image classification and object detection: A survey," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 6, p. e6767, 2022. [CrossRef]

[6] A. Dosovitskiy, "An image is worth 16 × 16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. [CrossRef]

[7] M. Naseer, K. Ranasinghe, S. Khan, F. S. Khan, F. Porikli, "On improving adversarial transferability of vision transformers," in *10th International Conference on Learning Representations, ICLR 2022*, Virtual Conference, Apr. 25–29, 2022. [CrossRef]

[8] M. S. Murshed, C. Murphy, D. Hou, N. Khan, G. Ananthanarayanan, F. Hussain, "Machine learning at the network edge: A survey," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 8, pp. 1–37, 2021. [CrossRef]

[9] D. C. Lepcha, B. Goyal, A. Dogra, V. Goyal, "Image super-resolution: A comprehensive review, recent trends, challenges and applications," *Inf. Fusion*, vol. 91, pp. 230–260, 2023. [CrossRef]

[10] P. Buddharaju, I. T. Pavlidis, P. Tsiamyrtzis, M. Bazakos, "Physiology-based face recognition in the thermal infrared spectrum," *IEEE Trans. Pattern Anal.*

Mach. Intell., vol. 29, no. 4, pp. 613–626, 2007. [CrossRef]

[11] D. Cardone and A. Merla, "New frontiers for applications of thermal infrared imaging devices: Computational psychopshysiology in the neurosciences," *Sensors*, vol. 17, no. 5, p. 1042, 2017. [CrossRef]

[12] R. Gade and T. Moeslund, "Thermal cameras and applications: A survey," *Mach. Vis. Appl.*, vol. 25, pp. 245–262, 2013. [CrossRef]

[13] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 26–Jul. 1, 2016, pp. 770–778. [CrossRef]

[14] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020. [CrossRef]

[15] S. H. Lee, S. Lee, B. C. Song, "Vision transformer for small-size datasets," *arXiv preprint arXiv:2112.13492*, 2021. [CrossRef]

[16] H. Nguyen, N. Tran, H. D. Nguyen, L. Nguyen, K. Kotani, "Ktfev2: Multimodal facial emotion database and its analysis," *IEEE Access*, vol. 11, pp. 17 811–17 822, 2023. [CrossRef]

[17] H. Nguyen, K. Kotani, F. Chen, B. Le, "A thermal facial emotion database and its analysis," in *Image and Video Technology: 6th Pacific-Rim Symposium, PSIVT 2013*, Guanajuato, Mexico, Oct. 28–Nov. 1, 2014, pp. 397–408. [CrossRef]

[18] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimed.*, vol. 12, no. 7, pp. 682–691, 2010. [CrossRef]

[19] R. I. Hammoud, "Guest editorial: Object tracking and classification beyond the visible spectrum," *Int. J. Comput. Vis.*, vol. 71, no. 2, pp. 123–124, 2007. [CrossRef]

[20] K. Panetta, A. Samani, X. Yuan, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. Rao, A. Kaszowska, H. Taylor, "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, pp. 509–520, 2020. [CrossRef]

[21] M. Kowalski and A. Grudzien, "High-resolution thermal face dataset for face and expression recognition," *Metrol. Meas. Syst.*, Vol. 25, no. 2, pp. 403–415. [CrossRef]

[22] A. Wesley, P. Buddharaju, R. Pienta, I. Pavlidis, "A comparative analysis of thermal and visual modalities for automated facial expression recognition," in *International Symposium on Visual Computing*. Berlin, Heidelberg: Springer, 2012, pp. 51–60. [CrossRef]

[23] M. Kopaczka, R. Kolk, D. Merhof, "A fully annotated thermal face database and its application for thermal facial expression recognition," in *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Houston, TX, USA, May 14–17, 2018, pp. 1–6, 2018. [CrossRef]

[24] A. Bhattacharyya, S. Chatterjee, S. Sen, A. Sinitca, D. Kaplun, R. Sarkar, "A deep learning model for classifying human facial expressions from infrared thermal images," *Sci. Rep.*, vol. 11, 2021. [CrossRef]

[25] B. Assiri and M. Hossain, "Face emotion recognition based on infrared thermal imagery by applying machine learning and parallelism," *Math. Biosci. Eng.*, vol. 20, pp. 913–929, 2022. [CrossRef]

[26] S. Manssor and S. Sun, "Tirfacenet: Thermal ir facial recognition," in *12th International Congress on Image and Signal Processing, bioMedical Engineering and Informatics (CISP-bMEI)*, Suzhou, China, Oct. 19–21, 2019, pp. 1–7. [CrossRef]

[27] B. Sathyamoorthy, U. Snehalatha, T. Rajalakshmi, "Facial emotion detection of thermal and digital images based on machine learning techniques," *Biomed. Eng. Appl. Basis Commun.*, vol. 35, p. 2250052, 2022. [CrossRef]

[28] A. Prabhakaran, J. Nair, S. Sarath, "Thermal facial expression recognition using modified resnet152," in *Advances in Computing and Network Communications*, Singapore: Springer, 2021, pp. 389–396. [CrossRef]

[29] K. Kamath, R. Rajendran, Q. Wan, K. Panetta, S. Agaian, "Ternet: A deep learning approach for thermal face emotion recognition," in *Mobile Multimedia/Image Processing, Security, and Applications 2019*, Nuremberg, Germany: SPIE, 2019, p. 10. [CrossRef]

[30] S. Rooj, A. Routray, M. Mandal, "Feature based analysis of thermal images for emotion recognition," *Eng. Appl. Artif. Intell.*, vol. 120, p. 105809, 2023. [CrossRef]

[31] S. Shaees, H. Naeem, M. Arslan, M. Naeem, S. Ali, H. Aldabbas, "Facial emotion recognition using transfer learning," in *2020 International Conference on Computing and Information Technology (ICCIT-1441)*, Tabuk, Saudi Arabia, Sept. 9–10, 2020, pp. 1–5. [CrossRef]

[32] M. K. Uhrig, N. Trautmann, U. Baumgärtner, R.-D. Treede, F. Henrich, W. Hiller, S. Marschall, "Emotion elicitation: A comparison of pictures and films," *Front. Psychol.*, vol. 7, 2016. [CrossRef]

[33] E. A. İyilikci, M. Boğa, E. Yüvrük, Y. Özkılıç, O. İyilikci, S. Amado, "An extended emotion-eliciting film clips set (egefilm): Assessment of emotion ratings for 104 film clips in a turkish sample," *Behav. Res. Methods*, vol. 56, no. 2, pp. 529–562, 2024. [CrossRef]

[34] N. Ding, X. Xu, E. Lewis, "Short instructional videos for the TikTok generation," *J. Educ. Bus.*, vol. 98, no. 4, pp. 175–185, 2023. [CrossRef]

[35] M. K. Uhrig, N. Trautmann, U. Baumgärtner, R.-D. Treede, F. Henrich, W. Hiller, S. Marschall, "Emotion elicitation: A comparison of pictures and films," *Front. Psychol.*, vol. 7, 2016. [CrossRef]

[36] N. Aldunate and R. González-Ibáñez, "An integrated review of emoticons in computer-mediated communication," *Front. Psychol.*, vol. 7, 2017. [CrossRef]

[37] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s J. Softw. Tools Prof. Program.*, vol. 25, no. 11, pp. 120–123, 2000. Available online: https://www.researchgate.net/publication/233950935_The_Opencv_Library (accessed 20 January 2025).

[38] J. Wu, “Introduction to convolutional neural networks,” *Natl. Key Lab Nov. Softw. Technol. Nanjing Univ. China*, vol. 5, no. 23, p. 495, 2017. Available online: https://cs.nju.edu.cn/wujx/paper/CNN.pdf (accessed 20 January 2025).

[39] A. Khan, A. Sohail, U. Zahoora, A. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artif. Intell. Rev.*, vol. 53, pp. 5455–5516, 2020. [CrossRef]

[40] S. Li and W. Deng, “Deep facial expression recognition: A survey,” *IEEE Trans. Affect. Comput.*, vol. 13, pp. 1195–1215, 2022. [CrossRef]

[41] B. Li, “Facial expression recognition via transfer learning,” *EAI Endorsed Trans.-Learn.*, p. 169180, 2018. [CrossRef]

[42] F. Chollet et al., (2015). *Keras* [Online]. Available: https://keras.io (accessed 20 January 2025).

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010. [CrossRef]

[44] J. Gao and Y. Zhao, “Tfe: A transformer architecture for occlusion aware facial expression recognition,” *Front. Neurorobotics*, vol. 15, 2021. [CrossRef]

[45] H. Li, M. Sui, F. Zhao, Z. Zha, F. Wu, “Mvt: Mask vision transformer for facial expression recognition in the wild,” *arXiv preprint arXiv.2106.04520*, 2021. [CrossRef]

[46] H. Hyeonbin, S. Kim, W. Park, J. Seo, K. Ko, H. Yeo, “Vision transformer equipped with neural resizer on facial expression recognition task,” in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore & Virtual, May 22–27, 2022, pp. 2614–2618. Available online: https://researchr.org/publication/icassp-2022 (accessed 20 January 2025).

[47] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al., “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. Available: https://aclanthology.org/2020.emnlp-demos.6/ (accessed 20 January 2025).

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, Jun. 20–25, 2009, pp. 248–255. [CrossRef]

[49] R. Mao, R. Meng, R. Sun, “Facial expression recognition based on deep convolutional neural network,” in *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, Oct. 18–20, 2023, p. 46. [CrossRef]

[50] M. Yamada and Y. Kageyama, “Face detection and analysis of relationship between degree of emotion arousal and facial temperature,” in *International Conference on Industrial Application Engineering 2021*, Kitakyushu, Japan, Mar. 26–29, 2021. [CrossRef]

[51] T. Viering and M. Loog, “The shape of learning curves: A review,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7799–7819, 2023. [CrossRef]

[52] B. Zhang, L. Dong, L. Kong, M. Liu, Y. Zhao, M. Hui, X. Chu, “Prediction of impulsive aggression based on video images,” *Bioengineering*, vol. 10, p. 942, 2023. [CrossRef]

[53] W. Xu and R. Cloutier, “A facial expression recognizer using modified resnet-152,” *Eai Endorsed Trans. Internet Things*, vol. 7, p. e5, 2022. [CrossRef]

[54] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, M. Shah, “Transformers in vision: A survey,” *ACM Comput. Surv.*, vol. 54, no. 10s, Sep. 2022. [CrossRef]

[55] Y. Bai, J. Mei, A. L. Yuille, C. Xie, “Are transformers more robust than cnns?” in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan, Eds., vol. 34. Red Hook, NY, USA: Curran Associates, Inc., 2021, pp. 26 831–26 843 [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/e19347e1c3ca0c0b97de5fb3b690855a-Paper.pdf (accessed 20 January 2025).

[56] M. Usman, T. Zia, S. Tariq, “Analyzing transfer learning of vision transformers for interpreting chest radiography,” *J. Digit. Imaging*, vol. 35, pp. 1445–1462, 2022. [CrossRef]

[57] L. Picek, M. Šulc, Y. Patel, J. Matas, “Plant recognition by ai: Deep neural nets, transformers, and knn in deep embeddings,” *Front. Plant Sci.*, vol. 13, 2022. [CrossRef]

[58] G. Viswanatha Reddy, C. Dharma Savarni, S. Mukherjee, “Facial expression recognition in the wild, by fusion of deep learnt and hand-crafted features,” *Cogn. Syst. Res.*, vol. 62, pp. 23–34, 2020. [CrossRef]

[59] Z. Chen, X. Feng, S. Zhang, “Emotion detection and face recognition of drivers in autonomous vehicles in iot platform,” *Image Vis. Comput.*, vol. 128, p. 104569, 2022. [CrossRef]

[60] S. K. D’mello and J. Kory, “A review and meta-analysis of multimodal affect detection systems,” *Acm Comput. Surv. (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015. [CrossRef]

[61] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, “Emotion recognition in human-computer interaction,” *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001. [CrossRef]