



ViTCN: Hybrid Vision Transformer with Temporal Convolution for Multi-Emotion Recognition

Kamal Zakiieldin¹ · Radwa Khattab¹ · Ehab Ibrahim¹ · Esraa Arafat¹ · Nehal Ahmed² · Elsayed Hemayed³

Received: 31 August 2023 / Accepted: 13 February 2024
© The Author(s) 2024

Abstract

In Talentino, HR-Solution analyzes candidates' profiles and conducts interviews. Artificial intelligence is used to analyze the video interviews and recognize the candidate's expressions during the interview. This paper introduces ViTCN, a combination of Vision Transformer (ViT) and Temporal Convolution Network (TCN), as a novel architecture for detecting and interpreting human emotions and expressions. Human expression recognition contributes widely to the development of human-computer interaction. The machine's understanding of human emotions in the real world will considerably contribute to life in the future. Emotion recognition was identifying the emotions as a single frame (image-based) without considering the sequence of frames. The proposed architecture utilized a series of frames to accurately identify the true emotional expression within a combined sequence of frames over time. The study demonstrates the potential of this method as a viable option for identifying facial expressions during interviews, which could inform hiring decisions. For situations with limited computational resources, the proposed architecture offers a powerful solution for interpreting human facial expressions with a single model and a single GPU. The proposed architecture was validated on the widely used controlled data sets CK+, MMI, and the challenging DAiSEE data set, as well as on the challenging wild data sets DFEW and AFFWild2. The experimental results demonstrated that the proposed method has superior performance to existing methods on DFEW, AFFWild2, MMI, and DAiSEE. It outperformed other sophisticated top-performing solutions with an accuracy of 4.29% in DFEW, 14.41% in AFFWild2, and 7.74% in MMI. It also achieved comparable results on the CK+ data set.

Keywords Emotion-recognition · Computer-vision · Deep-learning · Vision-transformer · Temporal-convolution-network

K. Zakiieldin, E. Ibrahim, N. Ahmed, E. Hemayed, contributed equally to this work.

www.talentinobot.com.

✉ Kamal Zakiieldin
kamal@talentinoai.com

✉ Nehal Ahmed
nehal.khaled@gmail.com; nehal.khaled@acu.edu.eg

Radwa Khattab
radwa.s.khattab@gmail.com

Ehab Ibrahim
ehabibrahim758@gmail.com

Esraa Arafat
esraa.hassan@talentinobot.com

Elsayed Hemayed
ehemayed@zewailcity.edu.eg

¹ Talentino, Maadi, Cairo, Egypt

1 Introduction

The ability to express emotions through facial expressions is crucial for humans to communicate and connect. As technology advances and people increasingly rely on computers for various activities like online learning and shopping, interacting with virtual systems has become an essential aspect of daily life. Understanding and responding to human emotions and mental states is crucial in order to interact with computers similarly to how we interact with people. It is increasingly important to have a natural interaction with technology.

In recent years, there has been a growing interest in emotion recognition (ER), which involves the accurate and automated classification of emotions in images or video sequences [1–3], and [4]. This research topic has gained

² Ahram Canadian University, 6 October, Giza, Egypt

³ CIE, Zewail City of Science and Technology, October Gardens, Giza, Egypt

attention in the fields of psychology, computer vision, and artificial emotional intelligence (AEI).

It assists in recognizing not just the emotional states of humans but also enables the imitation of different emotions in human-machine interactions, which has significant practical uses in real-life situations. Examples of such applications include driver safety monitoring to determine if the driver is distracted or if the driver is paying attention and to predict his actions based on his confusion or frustration levels.

Other applications are human-robot emotional interaction, medical domains to detect signs of depression or pain, or the identification of children with learning or cognitive disabilities involving assessing their level of involvement and linking it to the likelihood of having autism or attention deficit hyperactivity disorder (ADHD). Additionally, it involves recognizing specific elements or situations that capture or irritate the child's attention, as it may be an indication that a child suffers from a certain disease [5, 6], and [7]. In e-learning, facial expressions are utilized to determine which sections of a lecture are perplexing for the majority of students and to gauge the level of engagement among students while watching a video. Finally, in Talentino, to understand the candidates' engagement during their interviews and analyze their behaviors.

Although considerable progress has been made toward enhancing emotion recognition, there are still various challenges in exploring the dynamic emotion variations, and obtaining precise emotional analysis remains challenging in the present time.

Many systems use facial expressions and features to identify human emotions [8, 9], and [10]. Various steps are involved in systems designed to detect human emotions, such as retrieving images, preprocessing them, segmenting the images, extracting features, classifying facial expressions, and conducting training [11].

The unregulated environment presents various difficulties for practical implementation. However, there are more and more social networks and applications being used as data sources. Deep learning networks have also improved the processes of analysis and recognition.

The majority of current efforts [3] and [4] concentrate on using convolutional neural networks (CNN) to extract the feature representation of each frame, but they do not take into account the correlation between the frames in a video sequence. These approaches seek to identify the most significant expression features in each frame and address the problem as an image-based task.

Thus, they are relying mainly on the spatial features in the images. Other recent works [12, 13] have also considered temporal features to enhance recognition accuracy.

There are two main types of techniques used for facial expression recognition (FER): static image-based approaches and dynamic sequence-based approaches. The majority of

static frame-based techniques choose peak (apex) frames from films. Then perform facial emotion detection on these frames using local binary patterns [14], Gabor wavelets [15], and neural features. For instance, Zhao et al. [16] suggest using a sample with peak expression to guide a deep network that is peak-piloted, as it gains knowledge by studying a set of expressions that are not at their highest level. Meng et al. [17] suggest using the attention mechanism to combine multiple distinct frames into a unified representation for a video. These techniques are effective at choosing peak frames, but they do not take into account the changes over time and the relationship between consecutive facial frames.

In contrast to static frame-based techniques, dynamic sequence-based techniques learn spatiotemporal relations using 3D convolutional neural networks (3DCNN) [12] and long-short-term memory (LSTM) [18]. This could mimic long-term dependencies and improve FER performance.

To understand the time-related characteristics of spatial data and increase recognition accuracy, Kim et al. [19] suggest utilizing an LSTM network. Chen et al. [20] suggest a 3D-Inception-ResNet that enhances the representations of learned features. It calculates attention maps based on spatial-temporal and channel-wise factors. Li et al. [13] recently developed clip-aware dynamic facial expression by extracting clip-level features from each clip-based representation and re-weighting them.

The methods' performances are still quite distant from being ideal due to occlusions, different head positions, bad lighting, and other unanticipated challenges in real-world scenarios, even though various ways [12], [18] have been developed for in-the-wild FER. The issue of capturing spatial and temporal discriminative information for in-the-wild FER is difficult. Our grasp of discriminative feature representation and contextual information modeling has significantly deepened as transformer-based approaches to computer vision challenges have recently become increasingly popular [21], [22].

We can summarize our main contributions as follows:

- We proposed a state-of-the-art hybrid architecture, ViT CN, combining the powerful Vision Transformer with a Temporal Convolution Network.
- We processed FER as a sequence of frames in order to recognize the deep expression in the actual context.
- We performed our experiments on numerous standard data sets for comparison, which demonstrated that the proposed architecture outperformed the state-of-the-art methods. It showed that it has obtained the highest results in DFEW[23], AFFWild2[24], MMI[25], and DAiSEE[26] and achieved comparable results on other data sets such as CK+ [27].

- We conducted an ablation study to confirm the effectiveness of every element in the suggested model. All these experiments are done on one single GPU.¹

2 Related Work

The temporal correlations of succeeding frames in a sequence can be useful for face expression identification, even though the majority of the earlier models concentrated on static images. In this section, we present advanced deep networks for FER that take into account the spatial and temporal motion patterns in video frames, as well as the learned features obtained from the temporal structure. The spatio-temporal FER network utilizes both textural and time-related data to capture and represent more nuanced expressions. It uses a set of frames from a temporal window as a single input without knowing the intensity of the expression beforehand.

To identify emotions in a series of video frames, existing approaches frequently use recurrent neural network (RNN) models and their modifications. In several real-world applications, hybrid connections using CNN models have displayed outstanding performance. Deep RNNs, in particular LSTMs, have demonstrated impressive results in capturing the temporal relationships of sequential data.

RNNs are neural networks that have loops. This ability allows them to effectively understand and learn the time-based patterns in sequential data. In order to forecast the current outcome, RNNs can link historical data to the job at hand. The vanishing or exploding gradient issue makes training RNNs difficult. LSTM networks, a type of RNN that has the ability to acquire knowledge of long-term relationships, offer a solution to this issue.

CNNs are a different type of neural network that performs convolutional operations. Instead of the normal linear multiplication operation [28], convolution is another matrix multiplication procedure that generally depends on the weighted sum of neighboring input pixels multiplied with some special kernels. CNN is a convolutional layer that convolves the input into a feature map. In human eyes, convolution is the modeling of the response of a neuron in the visual cortex to a specific stimulus. Each convolutional neuron processes data only for its receptive field.

Memory cells with four neurons each make up the chain-like structure of LSTMs, which were created to interact in extremely unique ways. LSTM models with gated recurrent units (GRU) are one variant of the LSTM architecture. GRU models use less memory since they employ fewer training parameters. LSTM models are more accurate for bigger data sets, but GRU does calculations more quickly. LSTM or GRU networks have been used to attain the most advanced

results to date. The performance of these networks is further enhanced through FER training. An LSTM [29] or GRU [30] network is given a sequence of frames to understand variations in facial expressions and to identify an individual's emotional or mental condition.

For FER, a number of pre-trained models utilizing CNN architectures and similar modifications have been proposed. Self-encoder, CNN, and confidence networks are a few of these networks. Nevertheless, they lack the capacity to record contextual temporal information. Instead, they often show significant potential for automated feature learning. To achieve this, a variety of RNN model variations, such as CNN-LSTM [31, 32] and CNN-GRU [33], have been integrated with CNNs. They enhance their effectiveness in dealing with facial emotion recognition tasks. By reducing the impact of differences and the surrounding environment, these networks may more accurately identify facial expressions by obtaining more detailed information and distinguishing expression information from sequences of facial expressions. The LSTM component of these networks is responsible for learning and recognizing the changing patterns over time, while CNN extracts deep visual information. These networks highlight the significance of reading micro-expressions.

Extensions of the conventional LSTM and GRU architectures, Bidirectional LSTM (BDLSTM) and Bidirectional GRU (BIGRU), respectively, enhance the effectiveness of learning models to optimize the efficiency of FER. The BDLSTM processes the sequence in both directions, utilizing two LSTMs for training. As a result, the network receives more context, which speeds up the learning of an expression's sequence. Consequently, for FER, a CNN is included as a hybrid link at the very end to aid the model in thoroughly processing the variations in facial expressions. The CNN-BDLSTM [34, 35], and CNN-BIGRU [36] hybrid connection models are two examples.

The development of transformer networks in natural language processing (NLP) has attracted the curiosity of many in the field of computer vision. The transformer in NLP was created to represent lengthy sequence inputs. When compared to CNN networks, ViT has produced impressive results by pre-training on large data sets such as ImageNet-1k and then fine-tuning the target data set. ViT divided the picture into patches that allowed it to adapt to the computer vision challenge. The self-attention technique used in the transformer has the ability to collect long-range dependencies between these patches. Chaudhari et al. [21] applied ViT on FER, and Aouayeb et al. [22] applied the ViT structure as well on FER by injecting a Squeeze and Excitation (SE) block prior to the Multiple Layer Perceptron (MLP) heads for the FER job.

Xue and colleagues [37] introduced TransFER, a transformer-based method. Local CNN blocks were made to find various local patches after extracting feature maps using a

¹ Nvidia GPU-RTX 1080 8GB.

backbone CNN. Next, by using a multi-head self-attention-dropping part, a transformer encoder examined the general correlation among these local patches.

A two-stream pyramid cross-fusion transformer network was proposed [38]. In order to address scale sensitivity, intra-class discrepancy, and inter-class similarity issues in FER, it looks into the relationship between landmark features and picture attributes.

The primary difficulties encountered while utilizing CNN for FER are those related to computational complexity, picture quality, lighting fluctuations, high intra-class variations, and strong inter-class similarities brought on by changes in facial appearance. Several studies have considered creating a hybrid system by fusing deep learning approaches to address these problems.

3 Hybrid Model Architecture

Our proposed hybrid model, ViTCN, is composed of two parts: a ViT and a TCN. The ViT is used to extract the important spatial features from the images, while the TCN is used to encode spatiotemporal information extracted from the different video frames, combine the correlated features extracted for each frame, analyze the relationship between them, and classify the accurate expression. Both models are explained in the following subsections. Figure 1 shows our proposed architecture.

3.1 Vision Transformer

ViT [39] architecture is inspired by the basic transformer architecture first used in NLP problems. Its architecture is most similar to the encoder part of the transformers, where the image is split into a set of image patches, which are called visual tokens. These tokens are then embedded into a set of encoded features of a specified dimension. Moreover, the order of the patches is changed according to the positional encoding part of the ViT. This architecture was selected as it is able to identify and capture significant characteristics from the image in a relatively small vector of features that will be fed to the TCN model.

In our proposed architecture, we used the pre-trained ViT model from the PyTorch library. It is trained on the ImageNet-1k data set; however, we replaced the final fully connected convolution layer with one that has the same input/output dimension of 32, which is fed to the TCN model afterward.

3.2 Temporal Convolution Network

For the purpose of action recognition in videos, Lea et al. first proposed TCNs in 2016 [40]. The privilege of TCN architecture comes from its ability to encode spatiotemporal information coming from different frames along a video, which is then passed to a classifier label to classify these features into the corresponding classes. These features could be utilized to detect actions, emotions, or whatever significant information we need to detect. Moreover, TCN can process any sequence length, which enables it to consider more in-depth features.

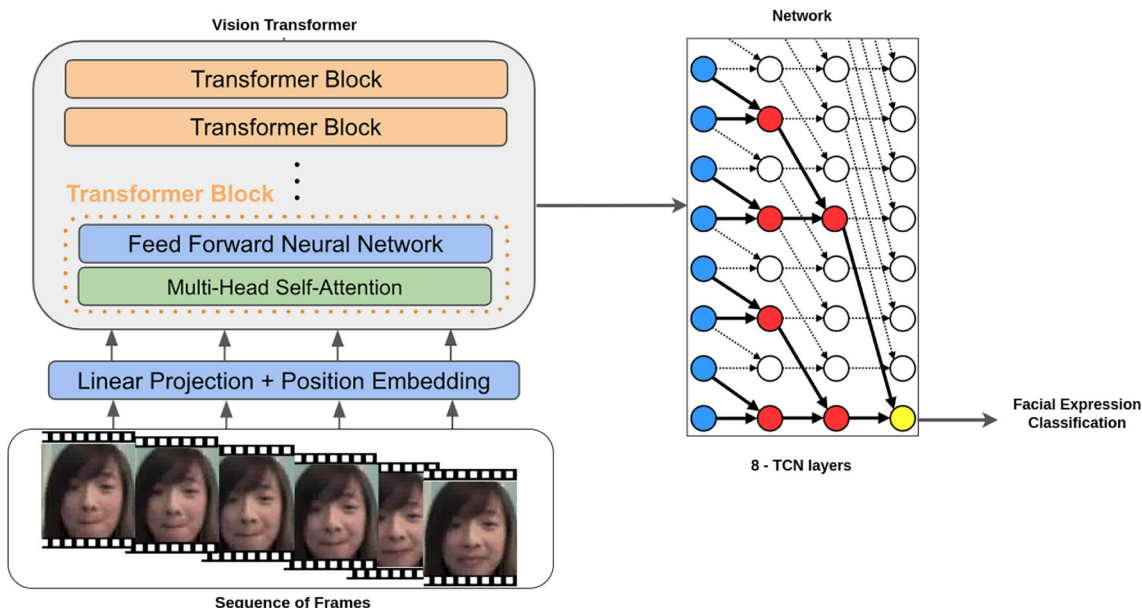


Fig. 1 State-Of-Art ViT with TCN Architecture

Fig. 2 Expressions classes in the CK+ data set

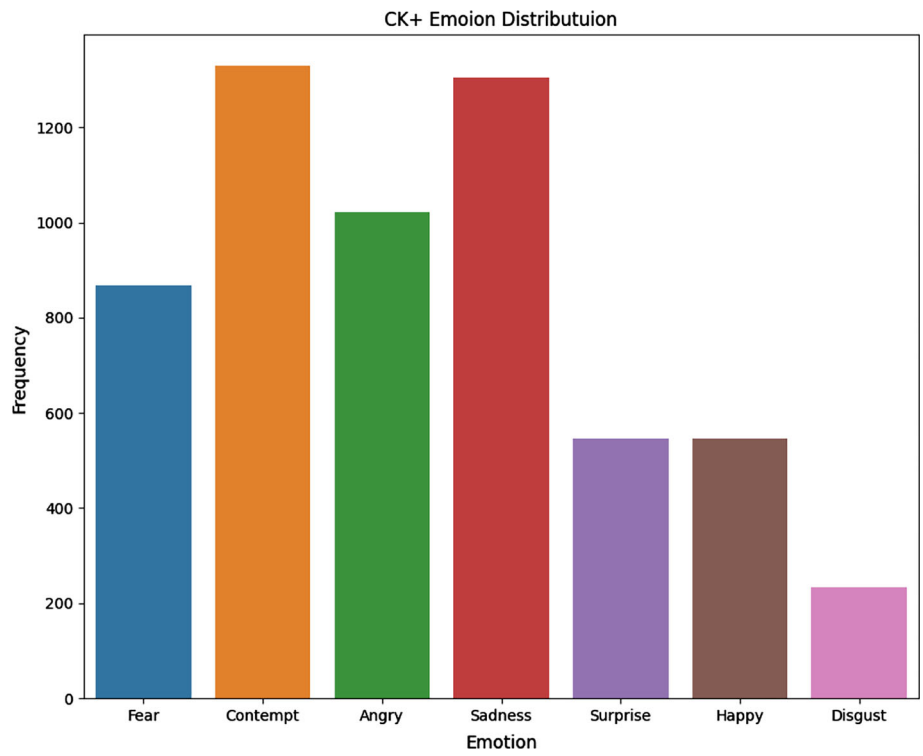


Fig. 3 Expressions classes in the MMI data set

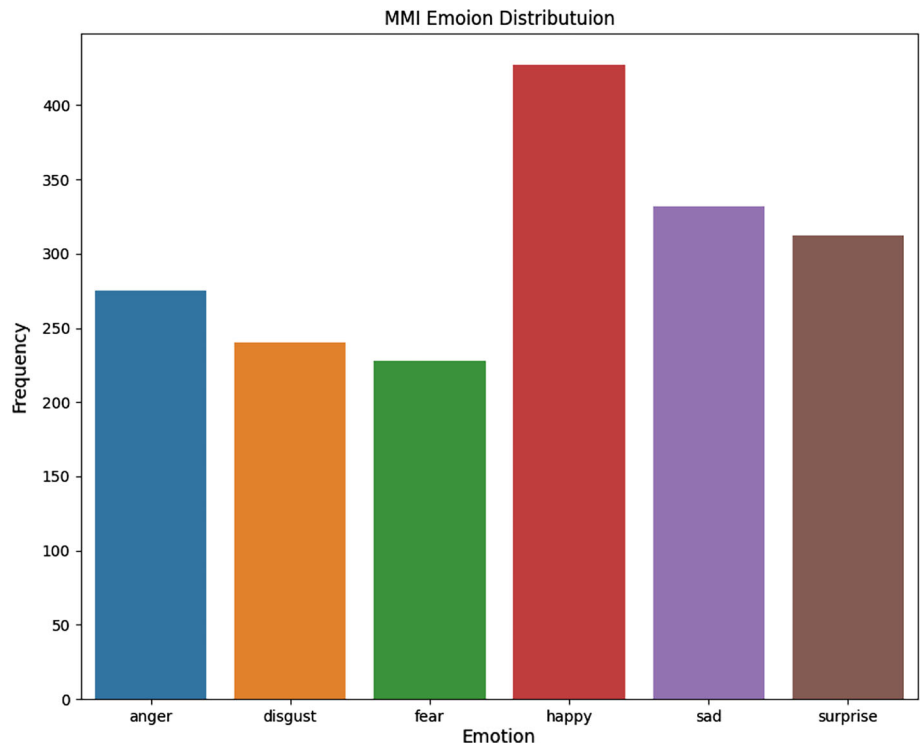




Fig. 4 Samples of CK+ data set

To identify the emotional expression in each frame of the proposed model, we concentrated on the face's changing characteristics. This information is then sent to a fully connected convolution layer, or classifier, which categorizes the input video based on the emotion, whether it exists or not (1 or 0).

3.3 Training Configuration

We have conducted many experiments to select the most optimal hyper-parameters during the proposed architecture training, which will be discussed in the ablation study section. Our model, ViTCN, is composed of two parts: a ViT and a TCN. We used the pre-trained ViT by replacing the final fully connected layer with a new layer with the same dimension to feed the TCN module. Our TCN module consists of 8 TCN layers with 3 kernel sizes. After many trials,

we have chosen the following values as the default: We have used the ADAM [41] optimizer with the default learning rate set to 0.001. The dropout is set to be 10%. We split the data sets into a training set, a validation set, and a testing set with (60% - 20% - 20%), respectively. Our batch size consists of 4 samples, as we have a limitation on our training machine as our single GPU has only 8 GB of RAM. We also train all experiments with 50 epochs; however, we are using early stopping for most of our experiments.

3.4 Loss Calculation

Due to the imbalance of most of the available data sets and benchmarks, as discussed in [42], we needed a modified loss function for better back-propagation and learning of the model. As discussed in the training configuration subsection, we used a modified version of the binary cross entropy (BCE)

Fig. 5 Samples of MMI data set

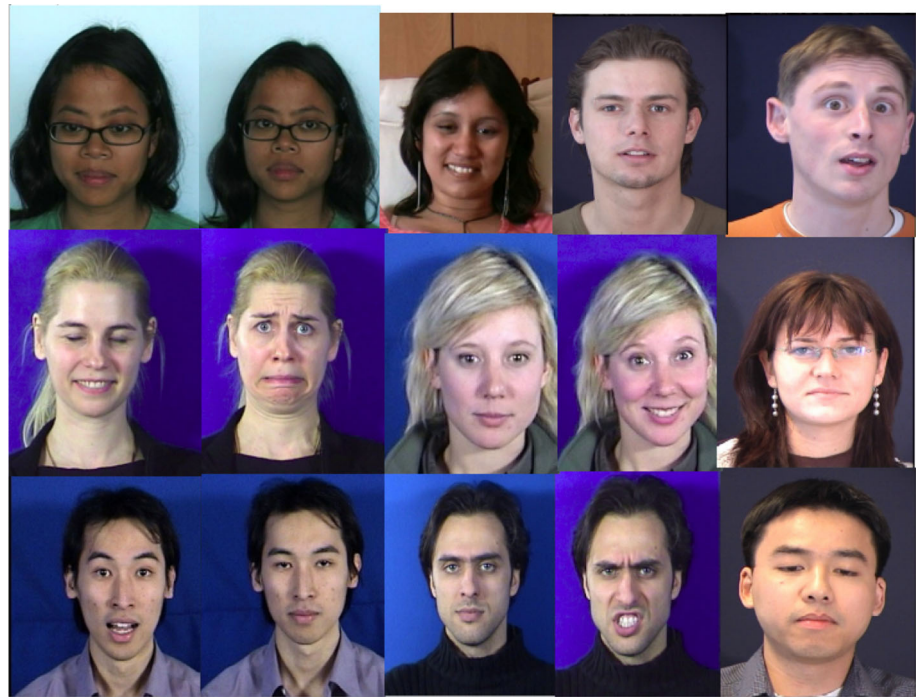


Table 1 DAiSEE Labels File sample

Clip ID	Engaged	Bored	Frustrated	Confused
1100021015.avi	2	2	1	0
1100021039.avi	2	1	2	2
1100021038.avi	2	0	0	0

for calculating the loss. We utilized the weighted loss for each label (0, 1), as explained in equation (1):

$$Loss = \left(1 - \frac{N}{B}\right) \times BCE(real_0, predicted_0) + \frac{N}{B} \times BCE(real_1, predicted_1) \tag{1}$$

where N represents the count of '0' labels within batch B, $real_0$ is the list of '0' labels from batch, and $predicted_0$ is their equivalent output from the model. Same with subscript 1 for label '1'.

However, if N equals '0', we used equation (2):

$$Loss = 0.35 \times BCE(real_1, predicted_1) \tag{2}$$

The weighted loss function improved the results by adjusting the calculated loss based on the label; this gives more attention to learning the unbalanced data by decreasing the loss value for the dominant label, which reduces the overfitting of that label.

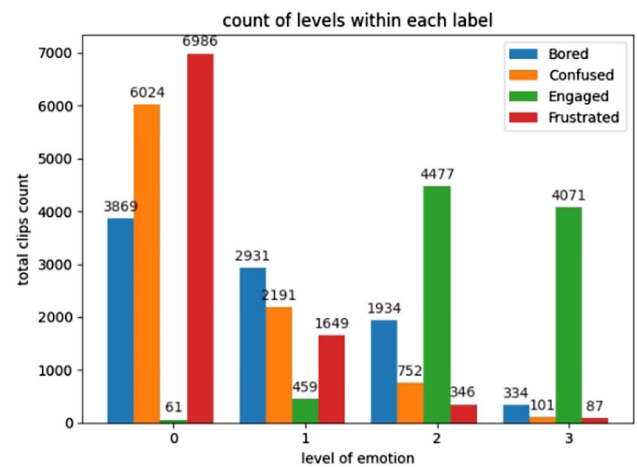


Fig. 6 Original DAiSEE statistics before augmentation on all splits

4 Experiments

In this part, we outlined the summary of the experiments conducted and discussed the obtained outcomes. Firstly, we started by discussing the data sets utilized to train and evaluate the suggested architecture. Then, we addressed the evaluation metrics utilized in evaluating the experimental outcomes and compared them with the latest advancements in the field. In conclusion, we provided an ablation study for different contributions with other details on the proposed solution and an analysis of additional visualization for an in-depth understanding of the application of ViT to the FER task.

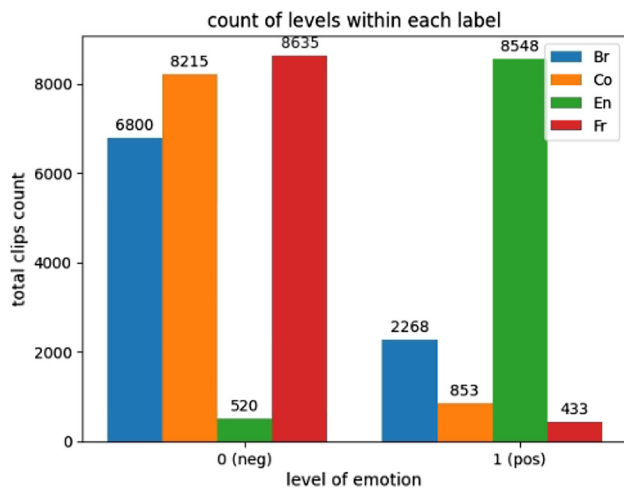


Fig. 7 DAiSEE data set statistics as a binary classification problem

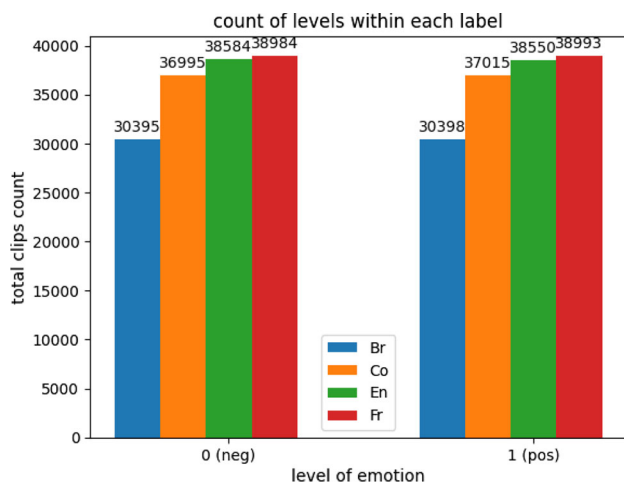


Fig. 8 Final statistics of the data after splitting to two-second clips and augmentation

4.1 Data sets

One of the challenges in ER is finding a suitable data set that suits our application requirements, as most of the available data sets have some challenges, such as non-frontal interview-based views, single images without any sequences, or small data sets with low data distribution. Hence, benchmarking our architecture is limited to a small number of applicable data sets. Frontal view data sets are chosen, with a sequence of images or videos. In the following subsections, we presented each data set and the class distribution, showing some statistics and insights about these data sets.

4.1.1 CK+

The CK+ dataset (also known as Cohn Kanade) is an expanded version of the CK data set [27] and [43]. It consists of 593 sequence videos obtained from 123 individuals; how-

ever, the labeled videos are only 327. The length of image sequences can differ, ranging from 10 frames to 60 frames with frontal views and 30-degree views. The videos were captured at a rate of 30 frames per second (FPS) and had a resolution of either 640x490 or 640x480 pixels. The videos had either an 8-bit grayscale or a 24-bit color value.

There are seven categories for facial expressions, which include anger, disgust, contempt, fear, happiness, surprise, and sadness. The age of the participants in CK+ data collection ranged from 18 to 50 years, with 69% being female and 31% being male. They are from different countries: 81% Euro-American, 13% Afro-American, and 6% other groups. The unequal distribution of expressions in CK+ is evident. Most facial expression classification (FEC) algorithms use the CK+ database, which is widely recognized as the most commonly utilized laboratory-controlled FEC database. Figure 2 shows the total number of videos per emotion. It demonstrates that CK+ offers a diverse range of expressions and that there is no dominant class. Figure 4 shows samples from the CK+ data set.

4.1.2 MMI

The MMI [25] (also known as Maja Pantic, Michel Valstar, and Ioannis Patras) dataset contains videos of the full temporal pattern of facial expressions, where the videos start from natural facial expression, then the peak of emotion, and finally back to the natural facial expression. Prototypical expressions and expressions with a single Facial Action Coding System (FACS) action unit are contained. MMI is composed of about 2900 videos from 75 different subjects. Videos are classified into seven facial expressions: surprise, anger, disgust, fear, sadness, and happy. The distribution of the seven emotions is shown in Figure 3. Figure 5 shows samples from the MMI data set.

4.1.3 DAiSEE

The DAiSEE dataset [26] includes 9068 video clip shots from 112 subjects. Each clip has a duration of 10 seconds and is recorded at a rate of 30 frames per second. It has four labels for each clip, with four degrees for each. These labels are “frustration”, “engagement”, “boredom” and “confusion”. And for each of these emotions, the intensity could range from 0 to 3 (namely, very low, low, high, and very high), representing the intensity level of the user’s emotion. A sample is shown in Table 1.

The DAiSEE data set has odd statistics concerning biasing and labeling, as shown in Fig. 6, which makes the data utilization for efficient training more complicated. Figure 6 shows a huge imbalance regarding the labels or the intensity of emotion within each label. Taking engagement emotion, for instance, it is shown that there are 4071 instances labeled

Fig. 9 Samples of the original DAiSEE data set



Fig. 10 Expressions classes in the AFFWild2 data set

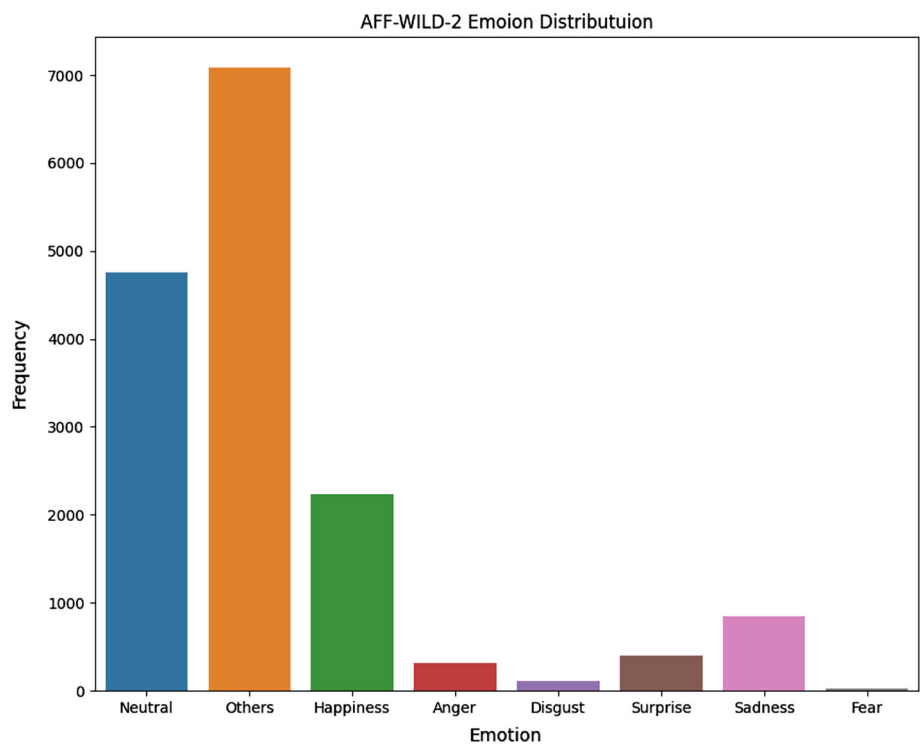


Fig. 11 Expressions classes in the DFEW data set

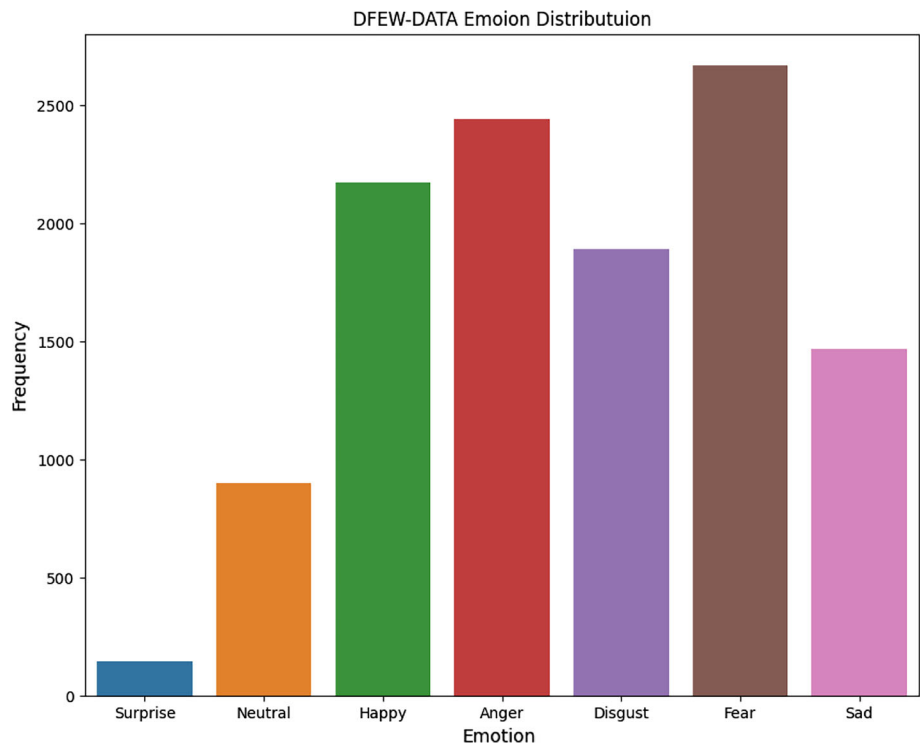


Fig. 12 Samples of the original AFFWild2 data set

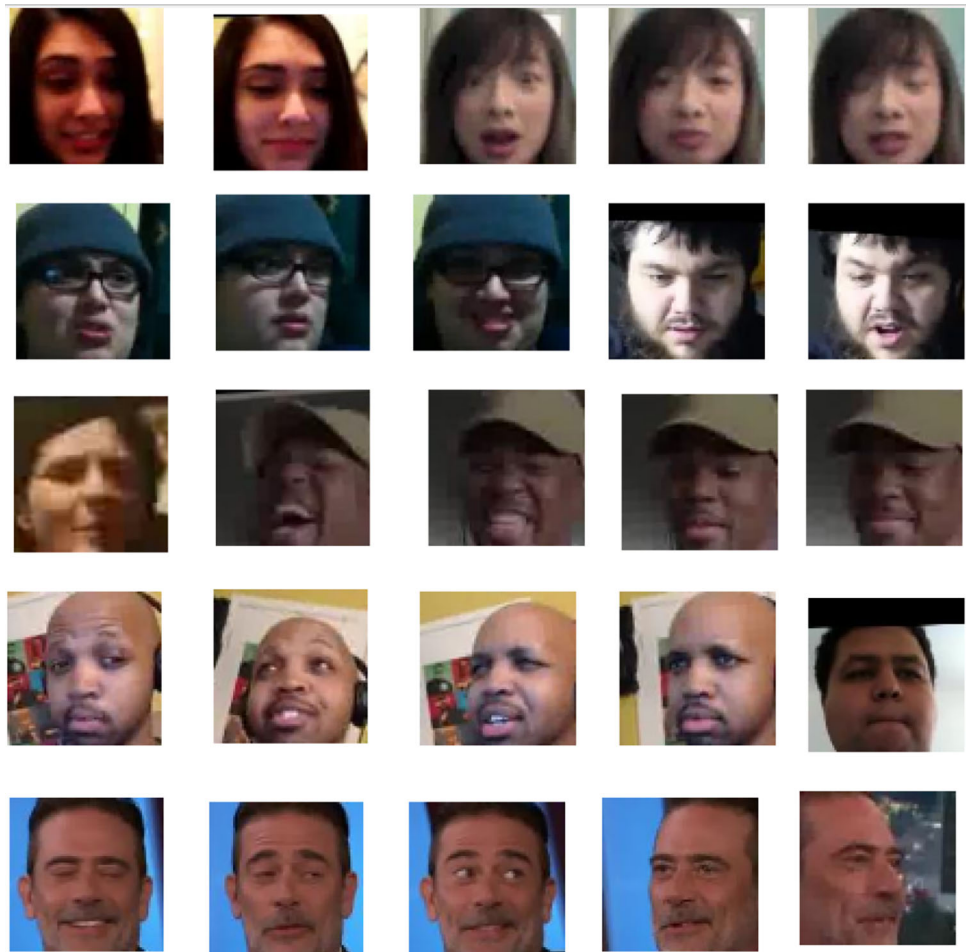


Fig. 13 Samples of the original DFEW data set**Table 2** Data sets summary

Name	Size	Source of data	Labels
CK+	593 sequences, Only 327 labeled	123 different subjects	Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise
MMI	2900 videos	75 different subjects	Anger, Surprise, Disgust, Happy, Sadness, and Fear
DFEW	16372 videos	1500 movies	Surprised, Natural, Happy, Anger, Disgust, Fear and Sad
AFF-Wild2	564 videos, Only 546 labeled	554 different subjects	Happiness, Surprise, Anger, Disgust, Fear, Sadness and the neutral state
DAiSEE	9068 videos	112 different subjects	Frustration, Engagement, Boredom and Confusion

“3”, 4477 instances labeled “2”, 459 instances labeled “1” and only 61 instances labeled “0”. It proved to be a large bias towards levels 2 and 3 compared to levels 0 and 1. We handled the problem as a binary classification problem (level “1” means that emotion exists and “0” means the absence of that emotion). We merged old labels 0 and 1 to new label 0, and labels 2, and 3 to new label 1.

At this point, we would have 8548 clips labeled 1 and only 520 labeled 0. This kind of imbalance makes the deep learning model unable to be trained correctly, and many of the evaluation metrics become unreliable. As a result, it was decided to use intensive data augmentation techniques to balance the data. Figure 7 shows the binary classification statistics. The data still had a significant bias that needed to

be fixed. Instead of focusing on balancing the levels of all emotions at once, we were more concerned with balancing the levels of each emotion because the suggested technique involved distinct models for each emotion.

To prepare the data for the training phase, we first sampled the videos into frames with a sample rate of 5, converting each 10-second video into 250 frames. Then we divided clip frames into two seconds each instead of 10 seconds, which theoretically enlarged the data set size to five times its original size. Each folder has 10 consecutive frames, each representing 2 seconds of the video. Secondly, for each two-second clip, which had a label in the labeling file, augmentation techniques were applied to make the data balanced. For each emotion, we considered the less-presented label from 0 to 1 as

the one to be augmented with different methods with random hyper-parameters. The applied augmentation techniques are adding noise, interpolation, random horizontal flip, random rotation within a small range, random resized cropping, and changing sharpness, saturation, or blurring. After the augmentation step is done, a good, reliable, balanced data set is produced with the statistics shown in Fig. 8. Figure 9 shows samples from the DAiSEE data set.

4.1.4 AFFWild2

AFFWild2 [24] data set consists of 546 videos from 554 subjects, of which 326 are male and 228 are female. These videos have a big variety in terms of nationalities and ages within different environments. The total frames are around 2.8 million, which are classified in various ways. Around 546 videos (2.6 million frames) are labeled according to basic expressions (happiness, surprise, anger, disgust, fear, sadness, and the neutral state), and 541 videos (2.6 million frames) have been categorized based on the FACS. The labeled videos by basic expressions are utilized in this research. The distribution of the seven emotions is shown in Figure 10. Figure 12 shows samples from the AFFWild2 data set.

4.1.5 DFEW

Dynamic Facial Expression in the Wild (DFEW) [23] contains 16372 videos from 1500 movies where every video has various challenging interferences. Examples of challenges are different illumination, occlusions, and crowding. Twelve professional annotators have labeled these videos. Videos are classified into seven labels to describe facial expressions: surprised, natural, happy, angry, disgust, fear, and sad. The distribution of the emotions is shown in Figure 11. Figure 13 shows samples from the DFEW data set.

Table 2 provides a summary of the data sets utilized in training and testing.

Table 3 Comparison of Expression Classification results on CK+ data set image-based

Methods	Acc. (%)	F1-Score	recall
prop+resnet50 [27]	89.56	0.887	0.889
Deep-Emotion [44]	98	–	–
ViT+ SE [22]	99.8	–	–
SE-3DPCANet [45]	93.15	–	–
ResNe t[46]	83.53	0.821	0.812
DeR f [47]	97.3	–	–
ViTCN, cropped faces*	94	0.944	0.94

Table 4 Comparison of Expression classification on CK+ data set sequence-based

Methods	Accuracy (%)
IT-RBM [48], 10 folds	87.16
STM-Explet [49]	94.19
LBVCNN(joint) [50], 10 folds	97.38
MDSTFN [51], 10 folds	97.28
Compact Frame2Seq [52]	98.47
DCPN [53], 10 folds	98.6
ViTCN, peak frames, 10 folds*	95.1

4.2 Frames Transform

Transformations are applied to images in deep learning using frameworks such as PyTorch and are typically known as resizing, normalizing, and converting images to tensors. However, considering a sequence of frames from a video for a model is more complicated than that. We built our class for transforming a sequence of frames, applying the same set of transformations to each image with the equivalent sample rate that was defined, and finally collecting and reshaping them to give one tensor as input to the ViT model.

4.3 Evaluation Criteria

The proposed model is assessed based on two primary factors, namely the F1-score, and accuracy, which are used as the main criteria for evaluation. However, in some experiments, the weighted average recall (WAR) and the unweighted average recall (UAR) are utilized to compare previous work.

The F1-score is the harmonic mean value of the recall (the ability of the classifier to find all the positive samples) and precision (the ability of the classifier not to label as positive a sample that is negative). The F1-score reaches its best value at 1 and its worst score at 0. The F1-score equation (3) is defined as:

$$F1 = \frac{2 \times \mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}} \quad (3)$$

The F1-score for emotions is calculated by considering the prediction made for each frame, where an emotion category is identified in every frame.

Accuracy (abbreviated as Acc.) is a measure of how well the test samples are predicted, expressed as the proportion of correctly predicted samples. The highest possible accuracy score is 1, indicating perfect predictions, while the lowest score is 0, indicating no correct predictions. The formula for calculating accuracy is as follows:

$$Acc = \frac{\text{Number of Correctly Predicted Samples}}{\text{Total Number of samples}} \quad (4)$$

UAR is defined as the mean accuracy of each class divided by the total number of classes, regardless of samples per class distribution, which makes it a better metric to optimize when the sample-class ratio is imbalanced.

WAR also called overall accuracy, is the ratio of accurately classified samples to the total number of samples, which is related to the number of samples in each category.

4.4 Implementation details: Training and Testing Settings

Our model is trained using the PyTorch platform, utilizing a single NVIDIA-GTX 1080Ti 8GB GPU card. By default, we train the model with a batch size of 4; however, in DFEW and CK+, we use 2 as our batch size to fit in our small GPU memory. We utilize the ADAM optimizer to optimize our proposed model, starting with a learning rate of 0.001.

Our backbone is the pre-trained ViT `vit_base_patch16_224`. While training, we split each video clip into 10 frames; however, some data sets have fewer frames per clip discussed in the experiment.

4.5 Comparison with state-of-the-art

In the following subsections, the outcomes of the experiments that were conducted to evaluate the proposed architecture are shown and compared to the previous work.

4.5.1 CK+ experiments

CK+ experiments are usually conducted as image-based experiments; hence, we have compared the obtained results in two different modes: the image-based mode and the sequence-based mode. First, we benchmark over the image-based mode. Table 3 shows a comparison of the achieved results on single image processing. In order to run a single image on the TCN network, the network was fed with two vector representations: the first is the feature vector from ViT, and the second is the original image.

The proposed model has achieved acceptable results compared to others, and ViT+SE [22] has obtained almost 99.8% accuracy. Although [22] are using ViT as their backbone, they still have the best result up to date because of the SE block they have combined with the transformer, where two fully connected layers are used with a single operation of point-wise multiplication. We think that integrating this mechanism with TCN might show better results across all benchmarks, as the SE block can optimize the TCN architecture with a channel-wise attention module; however, we consider this point to be future work.

On the other side, to have a fair comparison, the main proposed contribution is tested in sequence-based mode as well. Training a sequence of frames in CK+ was hard due to

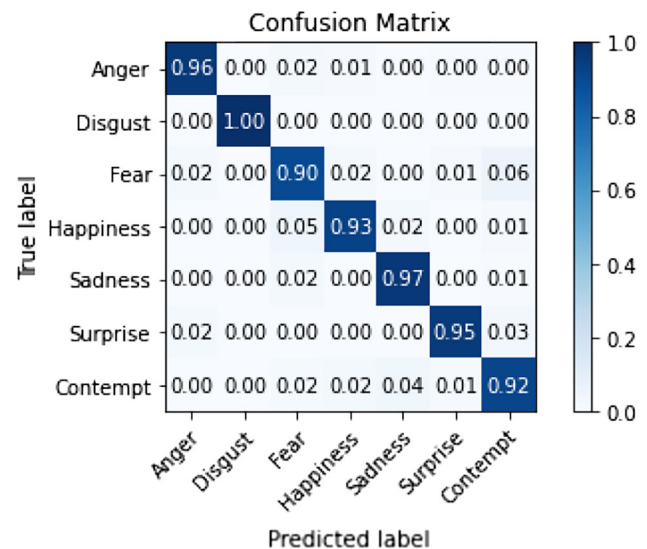


Fig. 14 The confusion Matrix for the CK+ data set

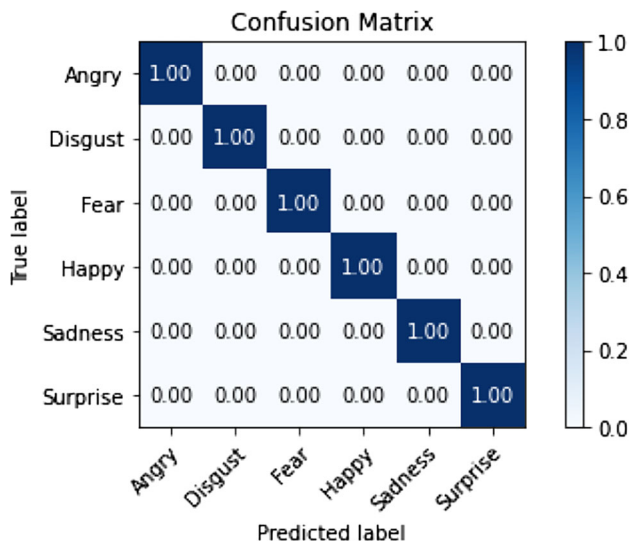
the absence of the peak expressions in the first few frames, hence, it is chosen to train over the peak frames starting from the 7th frame.

A comparison of the achieved results on the sequence of images is shown in table 4. The proposed architecture operates better than IT-RBM [48] and STM-Explet [49] by 7.94% and 0.91% respectively. DCPN [53] has obtained the highest performance due to its high training capabilities. First, its architecture consists of three cascaded inception deep neural networks, the first network is pre-trained over the ImageNet data set with augmented images from the CK+ data set, and subsequently, it is fine-tuned on the CK+ data set. This gives the network prior information about the data set, and it becomes familiar with the data during the pre-training step. The first network makes predictions about the emotion and transfers this information to the second network. The second network chooses two frames from the sequence; the highest prediction score is considered a peak frame, and the lowest prediction score is considered a weak frame. This assisted them in achieving a significant contrast between the highest (peak) and lowest (non-peak) frames, and then these two selected frames were used in the third network to define the emotion for the whole sequence. Due to computation limits, we could not fine-tune ImageNet. Although DCPN [53] achieved the highest accuracy using a complex architecture, the proposed architecture accomplished comparable results using a simple architecture.

Figure 14 shows the confusion matrix for the CK+ data set. It is shown that most of the classes are relatively easy to distinguish except for contempt vs. sadness and fear vs. contempt.

Table 5 Comparison of Expression classification on MMI data set

Methods	Accuracy (%)
STM-Explet [49]	75.12
IDFERM [54]	81.13
IT-RBM [48]	82.21
GCN [55]	85.89
MDSTFN [51]	91.46
ViTCN*	99.2

**Fig. 15** The confusion Matrix for the MMI data set

4.5.2 MMI experiments

Table 5 reports the comparison of the suggested model with other advanced video-based methods on the MMI data set. The proposed model demonstrates superior performance by achieving an accuracy of 99.2% surpassing the accuracy of the previous top-performing models MDSTFN [51] by 7.74% and STM-Explet [49], IDFERM [54], IT-RBM [48], GCN [55] by 24.08%, 18.07%, 16.99%, and 13.31% respectively.

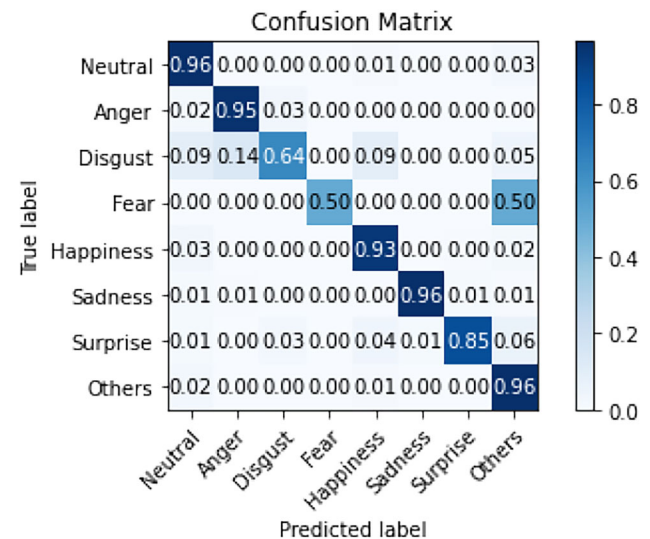
The confusion matrix is shown in Fig. 15, and it is shown that the proposed model perfectly distinguishes all classes.

4.5.3 AFFWild2 experiments

We report results by accuracy and F1-score in Table 6. The accuracy comparison reveals that the proposed model outperforms both the baseline [56] and the most sophisticated approaches, like TSAV [57] and NeteaseFuxi [58], by 34.5%, 21.6%, and 14.41% respectively. It outperforms them by 0.7, 0.452, and 0.087, according to the F1-score. The outcomes demonstrate the efficiency of the suggested model in classifying facial expressions.

Table 6 Comparison of Seven Basic Expression classification on AFFWild2 test set

Methods	Accuracy (%)	F1-Score
Baseline [56]	60.5	0.15
TNT [57]	73.4	0.398
FaceBehaviorNet [59]	–	0.53
TSAV [57]	70	0.4
NISL2020 [60]	68	0.27
NeteaseFuxi [58]	80.59	0.763
NISL2022 [61]	–	0.548
ViTCN*	95	0.85

**Fig. 16** The confusion Matrix for the AFFWild2 data set**Table 7** Comparison of Expression classification on DFEW data set

Methods	Accuracy (%)	UAR (%)	WAR (%)
3D Resnet18 [23]	56.51	44.73	54.98
CEFLNet [63]	65.35	–	–
EST [64]	65.85	–	–
STT [62]	–	54.58	66.65
ViTCN*	70.14	56	71

Figure 16 shows the confusion matrix of the AffWild2 data sets. It is shown that “Neutral”, “Anger”, “Happiness”, and “Sadness” are relatively easy to distinguish. However, “Disgust”, “Surprise” and “Fear” are mostly confused with other emotions, as these classes have few examples.

4.5.4 DFEW experiments

Experiments on the in-the-wild DFEW data set demonstrate that the suggested model offers a successful approach to identifying and understanding changing facial expressions.

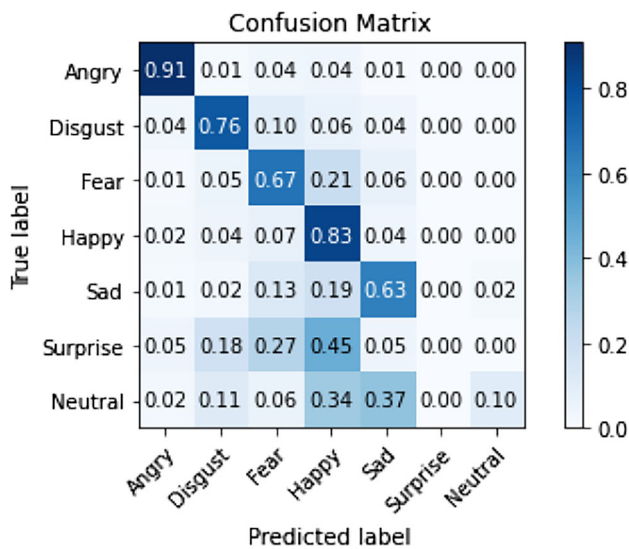


Fig. 17 The confusion Matrix for the DEFW data set

Table 8 Comparison of Expression classification on DAiSEE data set multi-classes, FS refers to Full Screen

Methods	Accuracy (%)
MTCNN [66]	58.84
ResnetX50LSTM-FS*	79
ResnetX50TCN-FS*	82.9
ViTCN-FS*	83.42

We evaluate our suggested model by comparing it with current achieved results on DFEW with respect to accuracy, the UAR, and the WAR. The comparison outcomes are presented in Table 7. The proposed model obtains the best results utilizing the three reported metrics, and we have outperformed all results up to date.

The previous leading method, known as STT [62], achieved a UAR of 54.85% and a WAR of 66.65%. However, the proposed model surpasses STT [62] by 1.42% in UAR and 4.35% in WAR. Furthermore, the proposed model outperforms 3D Resnet18 [23] by 11.27% and 16.02% in UAR and WAR. It is also shown that the proposed model

Table 9 Engagement class classification

Method	Accuracy (%)	F1 Score
DERN [67]	60	–
Weighted Loss, ResNet+TCN [65]	53.7	0.33
ResNet+LSTM [65]	61.15	--
ResNet+TCN [65]	63.9	–
BoS [68]	66.58	–
Ordinal TCN [71]	67.37	0.47
ViTCN-CF, without augmentation, Weighted Loss	94.8	0.973

Table 10 Confusion class classification

Methods	Accuracy (%)	F1 Score
ViTCN-FS	76.7	0.864
ViTCN-CF	81.77	0.8976

Table 11 Frustration class classification

Methods	Accuracy (%)	F1 Score
ViTCN-FS	86.48	0.926
ViTCN-CF	91.3	0.954

achieves superior accuracy outcomes in comparison to alternative approaches.

Figure 17 shows the confusion matrix of the DEFW data set; it shows that our architecture can distinguish the “Anger”, “Happy”, and “Disgust” classes. However, the “fear”, and “sad” classes are a bit confused with the “happy” class. Nonetheless, the “neutral”, and “surprise” classes are still not up to par (as they have a limited number of examples) and could be improved in future research.

However, the “fear” and “sad” categories are somewhat unclear when compared to the “happy” category. Nonetheless, the “neutral” and “surprise” categories are still not up to par, as they have a limited number of examples and could be improved in future research.

4.5.5 DAiSEE experiments

The DAiSEE data set is a bit challenging; as we explained earlier in Table 1, each video could contain more than one label. First, we trained our network on a multi-classification problem over the whole data set; Table 8 shows the results we have obtained with multi-class classification. We have tried different approaches instead of ViTCN, inspired by [65], such as combining ResNet-Backbone with TCN layers and ResNet-Backbone with LSTM layers.

Many research papers, such as [65, 67–71], and others, have investigated working with the engagement class. This encourages us to fine-tune our architecture to do a binary

Table 12 Boredom class classification

Methods	Accuracy (%)	F1 Score
ViTCN-FS	75.5	0.858
ViTCN-CF	78.8	0.881

classification for each class. Hence, using binary classification for each class separately leads us to have four different models. In this section, we explain the experiments for each class. Tables 9, 10, 11 and 12 investigate the obtained results over the engagement, confusion, frustration, and boredom expression classes, respectively.

Table 9 shows the obtained results compared with the reported results in [65, 67, 68], and [71]. It was noticed that the proposed architecture, ViTCN, has higher results. However, we have noticed that our architecture has overfitted the dominant class. The overfitting problem was discussed in the ablation study.

There are no previously reported results for the other three classes in the literature as a binary classification problem; hence, we show only our obtained results for those classes.

Table 10 shows the classification of the confusion emotion. It is shown that the proposed architecture has achieved promising results utilizing the default configuration of cropping the faces (CF) in each video sequence. The obtained results outperformed other experiments we have performed over the full screen (FS) of the video sequence.

Table 11 shows the classification of the frustration emotion. It is shown that the proposed architecture has achieved promising results utilizing the default configuration with CF and using an augmentation ratio of 35% in each video sequence. The obtained results outperformed other experiments we have performed over FS of the video sequence.

Despite the imbalance in the boredom class, the classification presented in Table 12 has yielded satisfactory results when employing the default configuration by CF in every video sequence, along with an augmentation ratio of 35% to maintain data set balance.

4.5.6 Discussion

By successfully integrating ViT with TCN in our ViTCN architecture, we unlock substantial performance gains for ViT in FER tasks. This hybrid approach outperforms existing advanced models while being trained under restrictive conditions (single GPU), demonstrating its potential for real-world applications. The proposed architecture achieved accuracy improvements exceeding 8% in MMI, 14% in AFFWild2, and 4% in DFEW. We delve deeper into the influence of training-phase choices like ViT freezing, input type (full frame vs. cropped faces), and loss function (normal vs.

Table 13 Ablation study in terms of accuracy with and without TCN block

Model	MMI	CK+	DFEW
ViT	98.3%	95.2%	49.6%
ViTCN	99.2%	95.1%	70%

Table 14 Ablation study in terms of F1-Score with and without TCN block

Model	MMI	CK+	DFEW
ViT	97.7%	93.6%	41.7%
ViTCN	99.1%	94.45%	55.6%

Table 15 Engagement class on full-screen images across different augmentation (Aug.) ratios

Our Methods	Aug. (%)	Accuracy (%)	F1 Score
ViTCN-FS	0	95	0.974
ViTCN-FS	15	94.8	0.973
ViTCN-FS	35	94	0.969
ViTCN-FS	100	75.2	0.8557

weighted) on the model's effectiveness in the following section.

4.6 Ablation Study

In this section, we perform ablation experiments to assess the influence of each element of our model, specifically the ViT architecture, TCN block, and data augmentation techniques.

First, the CK+, MMI, and DFEW data sets are utilized to conduct the experiments. Second, we assess our hyperparameter tuning on the DAiSEE data set.

4.6.1 ViT Study

First, we assess the performance of the ViT architecture, the added TCN block, and the use of Hugging Face (ViT_base Patch 16) as a pre-training model trained on ImageNet-21k. Tables 13 and 14 show the accuracy and the F1-score, respectively. It was noticed that adding the TCN block outperforms the ViT basic architecture on MMI and DFEW data sets. In the CK+ data set, although the accuracy is slightly decreased with only 0.1%, the F1-score has seen a 0.85% improvement.

In the MMI data set, the accuracy has seen a 0.9% improvement, and the F1-score has seen a 1.4% improvement as well. In the DFEW data set, adding the TCN block has enhanced the accuracy by 20.4%, and the F1-score by 13.9%.

Table 16 Engagement class on cropped faces across different augmentation (Aug.) ratios

Our Methods	Aug. (%)	Accuracy (%)	F1 Score
ViTCN-CF	0	95	0.974
ViTCN-CF	15	95	0.974
ViTCN-CF	35	93	0.963
ViTCN-CF	100	84.2	0.871

4.6.2 Data Processing Study on DAiSEE data set

In Tables 15 and 16, different experiments are conducted, and we first discuss the augmentation ratio over the data set. We generate more frames from the original DAiSEE frames by increasing the number of frames using the augmentation techniques discussed earlier. We have augmented the data set to rebalance the data by increasing the weak class. 100% augmentation indicates that our augmented data set is balanced. We have noticed that rotating the images usually leads to overfitting across any rotated image, as most of the rotated images are generated with our data augmentation technique; hence, we have eliminated the rotation procedure from the augmentation techniques that were utilized in all the conducted experiments.

Increasing the augmentation level on full-screen images (Table 15) decreases the accuracy and F1-score. It is noticed that using augmentation with a 35% balancing ratio is sufficient, as 15% is a bit low, and more than 50% leads to overfitting problems. Also, the model became activated with augmentation. Furthermore, we observed that training over the FS of the video sequence decreases the accuracy, and the model misclassifies more data and becomes more activated with outliers and noises. On the other side, using CF led the model to neglect unrelated noises and focus only on facial expressions. Video frames were cropped using multiple techniques, such as Multi-task Cascaded Convolutional Neural Networks (MTCNN) [72] and dLib [73].

In Table 17, we have investigated freezing the ViT during fine-tuning our ViTCN, and we noticed that freezing the parameters has reduced the obtained results.

Hence, our best setup is using the CF without freezing the ViT parameters, using a sufficient augmentation ratio of 35%.

Table 17 Studying the effect of freezing and non-freezing ViT with Engagement class across different proposed models, Aug. stands for the augmentation ratio applied

Methods	Aug. (%)	Accuracy (%)	F1 Score
ViTCN-CF, Weighted Loss, freeze	15	81.8	0.898
ViTCN-CF, Weighted Loss	15	94.4	0.971
ViTCN-CF, Weighted Loss, freeze	35	84.7	0.916
ViTCN-CF, Weighted Loss	35	86	0.92

Table 18 Studying the effect of using different augmentation ratios (Aug.) with engagement class

Methods	Aug. (%)	Acc. (%)	F1 Score
ViTCN-FS, Weighted Loss	35	89.7	0.945
ViTCN-CF, Weighted Loss	0	94.8	0.973
ViTCN-CF, Weighted Loss	15	94.4	0.971
ViTCN-CF, Weighted Loss	35	86	0.92
ViTCN-CF, Weighted Loss	65	94.4	0.971
ViTCN-CF, Weighted Loss	100	92.6	0.961

5 Conclusion and Future Work

In this work, we introduced the ViTCN, a hybrid architecture that combines the learned spatial features of the ViT with the temporal features of the TCN extracted from the different video frames and correlates the features extracted for each frame. It shows remarkable success in enhancing the performance of ViT on FER tasks. The performance of the proposed hybrid architecture was evaluated on controlled data sets like CK+ and MMI, as well as on wild data sets like DFEW and AFFWild2. It was shown that the suggested architecture outperforms other sophisticated solutions when utilizing a single model trained on a single GPU, notably on the MMI, DFEW, and AFFWild2 data sets. Our architecture has outperformed other sophisticated solutions with an accuracy of more than 8% in MMI, 14% in AFF-Wild2, and 4% in DFEW. Also, it produces competitive results on the CK+ data set. Nevertheless, due to the imbalanced classes in the DAiSEE data set, we examined the effects of augmentation methods and ratios. We discussed the implications of both freezing and non-freezing ViT during the training phase.

We aim to develop a facial expression detection and recognition system that utilizes a single GPU to minimize computational power while maintaining accuracy. Our goal is to achieve superior performance on a majority of benchmark data sets or match the performance of existing methods while minimizing computational resources across a series of frames.

In our future research, we plan to expand the ViTCN architecture to tackle a more challenging task, such as identifying micro-expressions. Additionally, we intend to improve the ViTCN architecture by incorporating an attention mechanism

called SE block [22] into the TCN architecture. Further augmenting our computational resources will help us employ 16 TCN layers with larger kernel sizes, which will boost the module's capacity and potentially yield superior performance on the CK+ data set.

Acknowledgements Thanks are extended to the ITIDA center for their support of this work in cooperation with the TIEC center in the collaborative funding program (Grant No. CFP220).

Author Contributions Architecture Conceptualization is prepared by KZ, NA, EH; Data curation, collection, and processing by KZ, EI, EA; All formal analysis is made by KZ, RK; Research and data investigation are done by NA, EH, EA; The methodology is designed by RK, KZ, EH; Installation and resources are supplied by KZ; Software is developed by KZ, EI, RK; Technical Experiments are prepared by KZ, EI, RK; With the supervision of NA, EH; and validation by NA, EH; The original draft is written by KZ, NA; and the review and editing of the paper are written and reviewed by NA, EH, KZ. All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). This research is partially funded in cooperation with ITIDA, the TIEC Center.

Availability of data and material The data sets used and/or analyzed during the current study are available from the corresponding author upon reasonable request. The data sets are also available at the following links: DAiSEE: <https://datasets.activeloop.ai/docs/ml/datasets/daisee-dataset/> MMI: <https://mmifacedb.eu/> CK+: <https://github.com/spenceryee/CS229/tree/master/> AFFWild2: <https://ibug.doc.ic.ac.uk/resources/aff-wild2/> DFEW: <https://dfew-dataset.github.io/download.html>.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. zbey, N. O., Topal, C.: Expression recognition with appearance-based features of facial landmarks. *Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2018, pp. 1–4. (2018)
2. Liu, M., R. W., S. Shan, Chen, X.: Learning expressionlets via universal manifold model for dynamic facial expression recognition. *IEEE Transactions on Image Processing*, 2016. (2016)
3. Monkaresi, H., R. A. C., N. Bosch, D'Mello, S. K.: Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing* **8**, 15–28 (2016)
4. Zhang, K., Y. D., Y. Huang, Wang, L.: Facial expression recognition based on deep evolutionary spatial-temporal networks. *IEEE Transactions on Image Processing*, 2017. (2017)
5. Kayadibi, I., U. E., Güraksin, G. E., Özmen Süzme, N.: An eye state recognition system using transfer learning: Alexnet-based deep convolutional neural network. *International Journal of Computational Intelligence Systems*. (2022)
6. Kayadibi, I., Güraksin., G. E.: An early retinal disease diagnosis system using oct images via cnn-based stacking ensemble learning. *International Journal for Multiscale Computational Engineering* **21**, 1–25 (2023)
7. Lecciso, F., Levante, A.: Emotional expression in children with asd: A pre-study on a two-group pre-post-test design comparing robot-based and computer-based training. *Front Psychol*. 2021;12:678052. (2021)
8. Khan, G., U. G., Siddiqi, A., Waqar, S.: Geometric positions and optical flow based emotion detection using mlp and reduced dimensions. *IET Image Process* 13:634–643 634–643 (2019)
9. Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., Zareapoor, M.: Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters* **115**, 101–106 (2018). *Multimodal Fusion for Pattern Recognition*
10. Fan, Y., D. L., Lu, X., Liu, Y.: Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*. 445–450 (2018)
11. Marian Stewart, B., F. I., Gwen, L., Javier, M.: Real time face detection and facial expression recognition: development and applications to human-computer interaction. *Computer vision and pattern recognition workshop*, 2003. *CVPRW'03*. **5** (2003)
12. Ayral, T., S. B., Pedersoli, M., Granger, E.: Temporal stochastic softmax for 3d cnns: An application in facial expression recognition. *IEEE/CVF Winter Conference on Applications of Computer Vision*. 3029–3038. 3029–3038 (2021)
13. Liu, Y., et al.: Clip-aware expressive feature learning for video-based facial expression recognition. *Information Sciences* **598**, 182–195 (2022)
14. Huang, M., Z. W. & Ying., Z.: A new method for facial expression recognition based on sparse representation plus lbp. *International Congress on Image and Signal Processing*, 1750–1754. **4** (2010)
15. Ho Lee, S., W. J. B., Ro., Y. M.: Collaborative expression representation using peak expression and intra-class variation face images for practical subject-independent emotion recognition in videos. *Pattern Recognition* **54** 52–67 (2016)
16. Xiangyun, Z., et al.: Peak-piloted deep network for facial expression recognition. *Computer Vision - ECCV 2016*, 425–442 (2016)
17. Meng, D., K. W., Peng, X., Qiao, Y.: Frame attention networks for facial expression recognition in videos. *IEEE International Conference on Image Processing*. IEEE, 3866–3870. (2019)
18. Vielzeuf, V., S. P. & Jurie, F.: Temporal multimodal fusion for video emotion classification in the wild. *19th ACM International Conference on Multimodal Interaction*. 569–576 (2017)
19. Hoe Kim, D., J. J., Baddar, W. J., Ro, Y. M.: Multiobjective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing* **10**, 2 223–236 (2017)
20. Chen, W., M. L., Zhang, D., Lee., D.-J.: Stcam: Spatialtemporal and channel attention module for dynamic facial expression recognition. *IEEE Transactions on Affective Computing* (2020)
21. Chaudhari, A., A. K., Bhatt, C., Mazzeo, P. L.: Vitfer: facial emotion recognition with vision transformers. *Applied System Innovation*. **5**, 80 (2022)

22. Aouayeb, M., Hamidouche, W., Soladie, C., Kpalma, K., Seguiet, R.: Learning vision transformer with squeeze and excitation for facial expression recognition. arXiv preprint arXiv (2021)
23. Jiang, X.: et al. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. ACM Multimedia, 2020. (2020)
24. Kollias, D., Zafeiriou, S.: Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770, 2018. (2018)
25. Pantic, M., R. R., Valstar, M., Maat, L.: Web-based database for facial expression analysis. IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005; p. 5. (2005)
26. Abhay, G., A. S., Richik, J., Vineeth, B.: Daisee: Dataset for affective states. E-Learning Environments. (2016)
27. Gupta, S., Tekchandani, R.K.: Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications **82**, 11365–11394 (2023)
28. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**, 2278–2324 (1998)
29. Yu, Z., Q. L., Liu, G., Deng, J.: Spatio-temporal convolutional features with nested lstm for facial expression recognition. Neurocomputing **317** (2018) 50–57. 50–57 (2018)
30. Huan, R.-H., et al.: Video multimodal emotion recognition based on bi-gru and attention fusion. Multimed. Tools Appl. **80**, 8213–8240 (2021)
31. Hung, B., Tien, L.: Facial expression recognition with cnn-lstm. Research in Intelligent and Computing in Engineering, Springer 549–560 (2021)
32. Abedi, W. M. S., A. T. S. Nadher, I.: Modified cnnlstm for pain facial expressions recognition. **29**, 304–312 (2020)
33. Vu, M. T., M. B.-A., Marchand, S.: Multitask multi-database emotion recognition. IEEE/CVF International Conference on Computer Vision 3637–3644 (2021)
34. Liu, Z.-X., Zhang, D.-G., Luo, G.-Z., Lian, M., Liu, B.: A new method of emotional analysis based on cnn-bilstm hybrid neural network. Clust. Comput. **23** (4) 2901–2913. (2020)
35. Du, P., X. L., Gao, Y.: Dynamic music emotion recognition based on cnn-bilstm. IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC) 1372–1376 (2020)
36. Yan, W., Zhou, L., Qian, Z., Xiao, L., Zhu, H.: Sentiment analysis of student texts using the cnn-bigru-at model. Scientific Programming 1058–9244 (2021)
37. Xue, F., Q. W., Guo, G.: Transfer learning relation-aware facial expression representations with transformers. IEEE/CVF International Conference on Computer Vision (ICCV) 3601–3610 (2021)
38. Zheng, C., M. M., Chen, C.: Poster: A pyramid cross-fusion transformer network for facial expression recognition. arXiv preprint arXiv (2022)
39. Alexey, D.: et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv (2020)
40. Colin, L., Rene, V., Austin, R., D, H. G. Temporal convolutional networks: A unified approach to action segmentation. Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, Proceedings, Part III 14 47–54 (2016)
41. Kingma, D. P., Ba., J. Adam: A method for stochastic optimization. CoRR, abs/1412.6980. (2015)
42. Denis, D., Wolfgang, M., Alexey, K.: Deep learning-based engagement recognition in highly imbalanced data. Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23 166–178 (2021)
43. Patrick, L.: et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. computer society conference on computer vision and pattern recognition-workshops 94–101 (2010)
44. Shervin, M., Mehdi, M., Amirali, A.: Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors (Basel, Switzerland) **21**, 166–178 (2021)
45. Chang, L., Chenglin, W., Yiting, Q.: A video sequence face expression recognition method based on squeeze-and-excitation and 3dPCA network. Sensors, 1424–8220. **23** (2023)
46. Sugianto, N., D. T., Tydd, B.: Deep residual learning for analyzing customer satisfaction using video surveillance. (2018)
47. Yang, H., U. C., Yin, L.: Facial expression recognition by de-expression residue learning. IEEE Conference on Computer Vision and Pattern Recognition 2168–2177 (2018)
48. Wang, S., Zheng, Z., Yin, S., Yang, J., Ji., Q.: A novel dynamic model capturing spatial and temporal patterns for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
49. Liu, M., R. W., Shan, S., Chen, X.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. IEEE Conference on Computer Vision and Pattern Recognition, pages 1749–1756, 2014. 1749–1756 (2014)
50. Kumawat, S., M. V. & Raman, S.: LbvCNN: Local binary volume convolutional neural network for facial expression recognition from image sequences. IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0. (2019)
51. Sun, N., Li, Q., Huan, R., Liu, J., Han, G.: Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recognition Letters. (2019)
52. Kuo, C.-M., S.-H. L., Sarkis, M.: A compact deep learning model for robust facial expression recognition. IEEE Conference on Computer Vision and Pattern Recognition Workshops 2121–2129 (2018)
53. Yu, Z., Q. L., Liu, G.: Deeper cascaded peak-piloted network for weak expression recognition. The Visual Computer (2017)
54. Liu, X., P. J., Kumar, B. V., You, J.: Hard negative generation for identity-disentangled facial expression recognition. Pattern Recognition **88**, 1–12 (2018)
55. Liu, D., H. Z. Zhou, P.: Video-based facial expression recognition using graph convolutional networks. Proc. 25th Int. Conf. Pattern Recognit. (ICPR) **25**, 607–614 (2021)
56. Kollias, D., E. H., Schulc, A., Zafeiriou, S.: Analysing affective behavior in the first abaw 2020 competition. arXiv preprint arXiv. (2020)
57. Kuhnke, F., L. R. Ostermann, J.: Two-stream aural-visual affect analysis in the wild. arXiv preprint arXiv. (2020)
58. Zhang, W.: et al. Prior aided streaming network for multi-task affective recognition at the 2nd abaw2 competition. arXiv preprint arXiv. (2021)
59. Kollias, D., V. S. & Zafeiriou, S.: Face behavior la carte: Expressions, affect and action units in a single network. arXiv preprint arXiv. (2019)
60. Deng, D., Z. C., Fau, B. E. S.: Facial expressions, valence, and arousal: A multi-task solution. (2020)
61. Deng, D.: Multiple emotion descriptors estimation at the abaw3 challenge. arXiv preprint arXiv. (2022)
62. Ma, F., B. S. Li, S.: Spatio-temporal transformer for dynamic facial expression recognition in the wild. arXiv preprint arXiv. (2022)
63. Liu, Y.: et al. Clip-aware expressive feature learning for video-based facial expression recognition. Information Sciences (2022)
64. Liu, Y.: et al. Expression snippet transformer for robust video-based facial expression recognition. arXiv preprint arXiv (2021)
65. Ali, A., Shehroz, K.: Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. 18th Conference on Robots and Vision (CRV) 151–157 (2021)

66. Liao, J., Y. L. Pan, J.: Deep facial spatiotemporal network for engagement prediction in online learning. *Appl. Intell.* 51 6609–6621 (2021)
67. Huang, T., Mei, Y., Zhang, H., Liu, S., Yang, H.: Finegrained engagement recognition in online learning environment. *IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*. 338–341 (2019)
68. Ali, A., D. B. J., Thomas, C., Shehroz, K.: Bag of states: A non-sequential approach to video-based engagement measurement. *arXiv preprint arXiv.* (2023)
69. Verma, M., N. T., Nakashima, Y., Nagahara, H.: Multi-label disengagement and behavior prediction in online learning. *International Conference on Artificial Intelligence in Education*. Springer, Cham. (2022)
70. Xusheng, A., V. S. S. Li, C.: Class-attention video transformer for engagement intensity prediction. *arXiv preprint arXiv* (2022)
71. Ali, A., Shehroz, K.: Affect-driven ordinal engagement measurement from video. *arXiv preprint arXiv.* (2021)
72. Zhang, K., Z. L., Zhang, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10) 1499–1503 (2016)
73. An improved faster rcnn approach: X. Sun, P. W. & Hoi, S. C. Face detection using deep learning. *Neurocomputing* **299**, 42–50 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.