PLOS ONE

RESEARCH ARTICLE

VT-3DCapsNet: Visual tempos 3D-Capsule network for video-based facial expression recognition

Zhuan Li[1], Jin Liu[1]*, Hengyang Wang[1], Xiliang Zhang[1], Zhongdai Wu[2]*, Bing Han[2]

1 College of Information Engineering, Shanghai Maritime University, Shanghai, China, **2** National Engineering Research Center of Ship Transportation Control Systems, Shanghai Ship and Shipping Research Institute, Shanghai, China

* jinliu@shmtu.edu.cn (JL); wu.zhongdai@coscoshipping.com (ZW)

Citation: Li Z, Liu J, Wang H, Zhang X, Wu Z, Han B (2024) VT-3DCapsNet: Visual tempos 3D-Capsule network for video-based facial expression recognition. PLoS ONE 19(8): e0307446. https://doi.org/10.1371/journal.pone.0307446

Editor: Qionghao Huang, Zhejiang Normal University, CHINA

Received: May 12, 2023

Accepted: July 5, 2024

Published: August 23, 2024

Data Availability Statement: Previously reported datasets were used to support this study and are available at http://www.jeffcohn.net/Resources/ and https://cs.anu.edu.au/few/AFEW.html. Our code and models are available at https://github.com/shmtu-lz/vt-3dCapsNet.git.

Funding: The work was supported by the National Key Technologies Research and Development Program of China under Grant 2021YFC2801001, in part by the National Social Science Foundation of China under Grant 20&ZD130. The funders had no role in study design, data collection and analysis,

Abstract

Facial expression recognition(FER) is a hot topic in computer vision, especially as deep learning based methods are gaining traction in this field. However, traditional convolutional neural networks (CNN) ignore the relative position relationship of key facial features (mouth, eyebrows, eyes, etc.) due to changes of facial expressions in real-world environments such as rotation, displacement or partial occlusion. In addition, most of the works in the literature do not take visual tempos into account when recognizing facial expressions that possess higher similarities. To address these issues, we propose a visual tempos 3D-CapsNet framework(VT-3DCapsNet). First, we propose 3D-CapsNet model for emotion recognition, in which we introduced improved 3D-ResNet architecture that integrated with AU-perceived attention module to enhance the ability of feature representation of capsule network, through expressing deeper hierarchical spatiotemporal features and extracting latent information (position, size, orientation) in key facial areas. Furthermore, we propose the temporal pyramid network(TPN)-based expression recognition module(TPN-ERM), which can learn high-level facial motion features from video frames to model differences in visual tempos, further improving the recognition accuracy of 3D-CapsNet. Extensive experiments are conducted on extended Kohn-Kanada (CK+) database and Acted Facial Expression in Wild (AFEW) database. The results demonstrate competitive performance of our approach compared with other state-of-the-art methods.

1 Introduction

Facial expression recognition has been widely used in the fields of mental health, virtual reality, synthetic animation, intelligent monitoring, and intelligent robots. A FER system [1] is mainly composed of three stages: face detection, facial feature extraction, and expression recognition. Generally, Face detection detects whether there is a face in the given image area based on unique facial characteristics and then locates the facial coordinate and segments the face from the image. Facial feature extraction is only performed on the facial area, which is the most

decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

important stage of FER. After the facial expression features are obtained, the facial expressions are classified through using the neural network based method. The robustness and completeness of the extracted features will decisively influence the final recognition result.

With the development of convolutional neural network in many fields [2–5], substantial breakthroughs have been made in facial expression recognition. Zhang et al. [6]. proposed a Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN) to analyze the facial expression information of temporal sequences, which employed a multi-signal convolutional neural network (MSCNN) to extract spatial features from still frames to complement the still appearance information. Fan et al. [7] proposed MRE-CNN framework for FER, which aimed to enhance the learning power of CNN models by capturing both the global and local features. Li et al. [8] put forward a new DLP-CNN method that enhanced the discriminative power of deep features by preserving the locality closeness while maximizing the inter-class scatter. However, these methods for FER only focus on local or global feature representation, and they neglect relationship (relative position relationship, scale, feature direction, etc.) among local facial features [9]. There are several important studies as follows:

The proposal of the capsule network [10] can well tackle the above problems. The capsule network uses neuron vectors to learn the pose information for targets and uses dynamic routing mechanism to transfer the capsule information between layers. This network has achieved the higher accuracy rate and strong robustness in digital recognition. However, the traditional capsule network only uses one layer of convolution for spatial feature extraction, which limits its performance [11]. Therefore, in the 3D-CapsNet model proposed in this paper, the 3D convolutional architecture is used in the feature extraction stage to extract spatiotemporal features, and then extraxted features are further encoded by dynamic routing mechanism. In addition, the visual tempos of facial expressions [12] can be employed to improve the recognition effect. Specifically, different facial expressions usually hold different visual tempos. However, In some cases there are larger similarities in visual appearance of different expressions(e.g. Fear and Surprise) and the key to distinguish them is their visual tempos.

To this end, we utilize improved deep 3-dimensional convolutional neural network model (3D-ResNet) to replace the traditional convolutional layer of capsule networks, and we named the improved capsule network that applyed improved 3D-ResNet architecture 3D-CapsNet. Specifically, the improved 3D-ResNet architecture employs 3D convolution kernels to learn spatiotemporal features in the residual block part. Meanwhile, we incorporate AU-perceived attention module into the 3D residual block to better focus on the key facial area. Then a non-local attention block is used to capture long-range dependencies on spatial and temporal dimension. Next we exploit dynamic routing algorithms of CapsNet to calculate the weight allocation between different capsules to capture the hierarchical structure of features and better handle complex feature relationships.

Given that 3D-CapsNet focuses on improving the ability of feature representation, the ability of modeling the dynamic visual tempos of facial expressions is insufficient. Therefore, we propose TPN-based expression recognition module (TPN-ERM) to further improve our 3D-CapsNet model and we named the improved 3D-CapsNet architecture that integrated with TPN-ERM VT-3DCapsNet. Overall, our main contributions can be summarized as follows:

1. First, we propose 3D-CapsNet model for emotion recognition, in which we introduced improved 3D-ResNet architecture that integrated with AU-perceived attention module to enhance the ability of feature representation of capsule network through extracting deeper hierarchical spatiotemporal features and extracting latent information (position, size, orientation) in key facial areas.

2. Second, we propose the temporal pyramid network(TPN)-based expression recognition module(TPN-ERM), which can learn high-level facial motion features from video frames to model differences in visual tempos, further improving the recognition accuracy of 3D-CapsNet.
3. Finally, extensive experiments are conducted on the CK+ and AFEW datasets, the results demonstrate that our method significantly outperforms other state-of-the-art methods.

The remaining of this article is organized as follows. The related work is presented in Section 2, while the 3D-CapsNet model and TPN-based expression recognition module are discussed in Section 3. Experimental results and discussion are shown in section 4, and conclusion is drawn in Section 5.

2 Related work

2.1 Traditional hand-crafted feature extraction for FER

Most of the traditional facial expression feature extraction algorithms used artificially designed features or shallow learning methods. Local Binary Pattern (LBP) is one of the finest manual feature extraction techniques. Niu et al. [13] offered an approach based on a combination of LBP features and an enhanced ORB, which effectively solved the problem of overlapping and redundant feature points in the feature extraction process and produced decent results on multiple experimental controlled datasets. Xiang et al. [14] proposes a novel illumination insensitive feature descriptor by integrating the center-symmetric local binary pattern (CS-LBP) into a common feature description framework. However, these approaches, due to their low semantic intensity, fail to perform accurate recognition in more complex real-world scenarios. To tackle this problem, Liao et al. [15] introduced RCL-Net that is based on ResNet-CBAM residual attention branch and the local binary feature (LBP) extraction branch (RCL-Net), which aimed to emphasize the local detail feature information of facial expressions and extract texture feature information. However, these manual feature extraction approaches have a high workload, lengthy stages and significant restrictions in terms of practical applications, and these methods depend heavily on researcher's experience and their performance is strongly affected by image quality variations.

2.2 Deep learning based methods for FER

Recently, the methods based on deep learning have been widely used for FER and achieve state-of-the-art performance. But most CNN-based models fail to learn long-range inductive biases between different facial regions in most neural layers, which limits the performance of the models. To address this problem, Huang et al. [16] introduced a novel FER framework with two attention mechanisms for CNN-based models. In particular, a visual transformer attention mechanism is used to learn high-level semantic representation. Wu et al. [17] proposed a novel cross-hierarchy contrast (CHC) framework FER-CHC to utilize these crucial features in improving the performance of CNN-based models for FER through employing a contrastive learning mechanism. Specifically, the CHC captures common and differential features from different facial expressions with a cross-hierarchy contrast mechanism, which can regularize the feature learning of the backbone network and enhance global representations of facial expressions. Recent methods based on Vision Transformer (ViT) [16, 18] have been introduced to improve the performance of CNN-based models. However, ViT-based approaches are vulnerable to facial regions unrelated to expressions and may learn redundant correlation representations due to their self-attention mechanism. To address these issues, Fan et al. [19] proposed a novel graph-based model called Face2Nodes, which can flexibly learn the

graph representations of facial expressions without requiring additional auxiliary facial information such as landmarks. In addition, Zhou et al. [20] realized that existing models failed to capture the crucial complementary gains between face and context information in video clips. And they presented a novel cross-attention and hybrid feature weighting network to achieve accurate emotion recognition from large-scale video clips through fully exploiting the complementary information between face and context features.

Kensho et al. [21] proposed a 3D CNNs based on ResNets toward a better action representation and described the training procedure of 3D ResNets in details. Teng et al. [22] proposed a new network called Typical Facial Expression Network (TFEN) to extract temporal features and the spatial structure of facial expressions in an integrated manner, which uses two deep two-dimensional (2D) convolutional neural networks (CNNs) to extract facial and expression features from input video. A facial feature decoupler decouples facial features from expression features to minimize the influence from inter-subject face variations. These networks combine with a 3D CNN and form a spatial-temporal learning network to jointly explore the spatial-temporal features in a video. But 3D convolutions usually employ structures with fixed temporal depth that decreases the potential to extract discriminative representations. Based on these, Melo et al. [23] proposed a novel deep learning architecture called the Maximization and Differentiation Network (MDN) to effectively represent facial expression variations that are relevant for depression assessment. Khanna et al. [24] combines two commonly used deep 3-dimensional Convolutional Neural Networks (3D CNN) models with slight modifications. Initially, the 3D ResNet model extracted feature vectors from video frame sequences, then these feature vectors are fed to the 3D DenseNet model's blocks, which are then used to classify the predicted emotion.

The convolution neural network does not consider the spatial and interlayer features of the target. The use of dynamic routing in capsule networks [10] can well address the above problems. However, the traditional capsule neural network does not extract features sufficiently before the dynamic routing between the capsules. To achieve better feature representation, Shu et al. [25] presented RES-CapsNet to investigate the recognition of micro-expression, which proposed an improved capsule network that used Res2Net as the backbone to extract multi-level and multi-scale characteristics. low-quality 3D face recognition(FR) with missing facial features still suffers from insufficient discriminative feature extraction for visible face regions. Zhao et al. [26] proposed a dual-stream multi-scale fusion network (DSNet) for low-quality 3D FR. Which introduced a capsule network as the second stream to enhance the expression of 3D facial spatial position information, thereby further improving the performance of low-quality 3D FR with missing facial features. Ye et al. [27] designed SCapsNet for FER, which used a shallow small convolution kernel to reduce the network parameters of the capsule network and optimize the quantity and quality of capsule of the capsule network to improve the calculation efficiency.

Feature pyramid is a widely-used method to learn multi-scale feature representation for detecting objects of various scales. FPN [28] exploited the inherent multi-scale pyramidal hierarchy of deep convolutional networks to construct feature pyramids with marginal extra cost. Liu et al. [29] proposed PANet which added a bottom-up path to enhance the entire feature hierarchy with accurate localization signals in lower layers. However, FPN and PANet cannot handle the information about temporal series well, and they are designed in complex and multi branch manner. On the other hands, visual tempo characterizes the dynamics and the temporal scale of an action. Therefore, modeling such visual tempos of different actions will facilitate their recognition. Yang et al. [12] proposed a temporal pyramid network (TPN) for modeling the visual tempo. The extraction of features and fusion of features form feature hierarchy structure for the backbone so that they can capture action instances at various tempos.

Yang et al. [30] proposed a CNN-based framework called Pyramidal Spatio-Temporal Network (PSTNet), which employed Spatial encoding for spatial representation of external factors, while prior pyramid enhances feature dependence of spatial scale distances and temporal spans, then post pyramid is proposed to fuse the heterogeneous spatio-temporal features of multiple scales. To enrich temporal information of the inputs, a Multiple Frame Rate Module (MFM) is proposed to mix different frame rates at a fine-grained pixel-wise level. Chen et al. [31] introduced a Multiple Frame Rate Module (MFM) which mixed different frame rates at a fine-grained pixel-wise level to enrich temporal information of the inputs. But these methods remain computationally expensive when dealing with the dynamic visual tempos of action instances at the input frame level. In this work, we improve the ways of feature fusion and introduce adaptive weighted fusion to better learn parameters in the feature fusion stage to tackle this problem.

To sum up, the performance of FER system can still be greatly improved. In the existing works, failing to make full use of spatiotemporal features limits the performance of the model, and visual tempos of facial expressions are rarely used in FER. As a result, we propose a fine-grained method that uses 3D-CapsNet to fully utilize and encode the feature presentation in the input video. We also propose TPN-ERM, which employs visual tempos of different facial expressions to further improve the performance of 3D-CapsNet.

3 Our methods

An overview of the proposed approach is depicted in Fig 1. The input of the 3D-CapsNet is limited to a small number of contiguous video frames owing to the increased trainable parameters as the size of input window (the time dimension of the convolution) increases. On the other hand, many facial actions of expressions are extended in various frames, it is necessary to encode facial motion information in the 3D-CapsNet model. Toward this end, we propose to use improved TPN-ERM to calculate facial motion features representing expressions from a considerable number of video frames and use these features as auxiliary outputs to regularize 3D-CapsNet model as shown in Fig 1. Particularly, we generate a feature vector encoding long-term facial motion information beyond the information contained in the input frames to the 3D-CapsNet for each trained expression. 3D-CapsNet is made to learn a feature vector close to this feature, which is accomplished via connecting many auxiliary output units to the

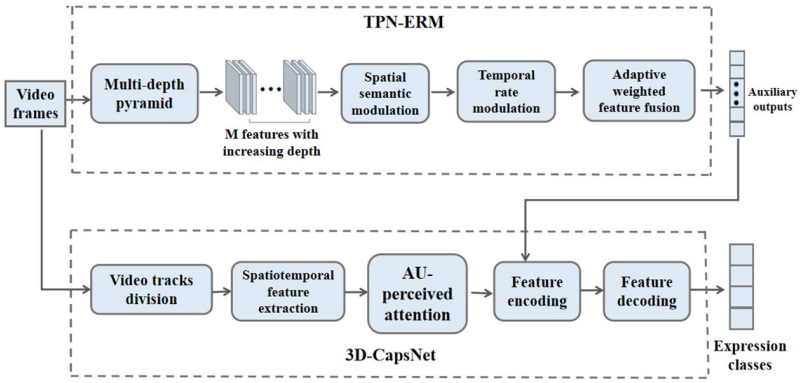

Fig 1. An illustration of our visual tempos 3D-capsule network framework(VT-3DCapsNet). 3D-CapsNet is made to learn a feature vector close to high-level motion feature via connecting auxiliary outputs to the PrimaryCaps layer for feature encoding.

https://doi.org/10.1371/journal.pone.0307446.g001

PrimaryCaps layer in 3D-CapsNet. This will enable 3D-CapsNet to learn high-level expression features.

In the following, we describe our proposed 3D-CapsNet model which is used to better learn spatiotemporal feature. Then, we describe TPN-based expression recognition module (TPN-ERM) that models the variances in visual tempos to further optimize the performance of 3D-CapsNet model.

3.1 The 3D-CapsNet model

The overall architecture of 3D-CapsNet is shown in Fig 2. The improved 3D-ResNet extracts deep spatiotemporal features from video tracks (which can be regarded as a video clip). And the encoding and decoding abilities of the dynamic routing mechanism in the capsule network can obtain features represented by vectors, thus improving the accuracy of facial expression recognition. Each module of 3D-CapsNet will be described in detail in the following sections.

3.1.1 Spatiotemporal features extraction. The whole system could be divided into two important parts: extracting spatiotemporal features from video tracks through improved 3D ResNet, integrating spatiotemporal features by the non-local blocks, and learning the features in key facial areas through the AU perception enhancement module. The video clip is first divided into continuous non-overlapping small segments, and each small segment contains N frames. Assuming that each fragment is represented as:

$$c_k = \{x_t \mid x_t \in R^{H \times W}\}_{t=1}^{N} \tag{1}$$

where N is the length, H and W are the height and width of the image respectively. Specifically, our network structure is shown in Fig 3, in which (a) is the overall improved 3D ResNet structure, (b) is the residual block structure of the 3D-RseNet we used, and (c) shows the bottleneck block after average pooling to speed up training and improve performance.

We adopt 3D convolution kernels of 3D ResNet-50 architecture to extract spatiotemporal features. However, 3D-ConvNet is hard to optimize because of the large number of parameters. To tackle this problem, we added an extra time dimension to all 2D ResNet-50 convolution filters. For example, a 2D $k \times k$ kernel can be exaggerated to a 3D $t \times k \times k$ kernel that spans t frames. Specifically, we use a 1D convolution layer which is purely to learn temporal sequence features and a 2D convolution layer to learn the spatial features in the residual block. As shown in the blue dashed box in Fig 3(b). Meanwhile, we have also incorporated AU-perceived enhancement module into the 3D residual block to better focus on the key parts of the

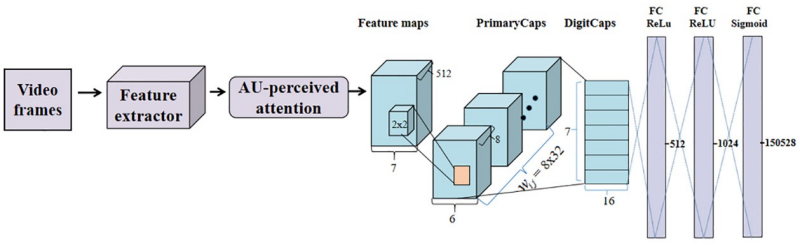

Fig 2. The overall architecture of 3D-CapsNet. (1) The improved 3D-ResNet servers as a feature extractor to better learn spatiotemporal features, and the AU-perceived attention mechanism is introduced to perceive the specific AU changes related to facial expression and focus on effective feature information of key facial areas. Thus enhancing model's ability of feature representation. (2) The capsule network module encodes the enhanced feature mapping through dynamic routing mechanism, decodes it through three fully connected layers, and implements the final expression classification through the squeeze function.

https://doi.org/10.1371/journal.pone.0307446.g002

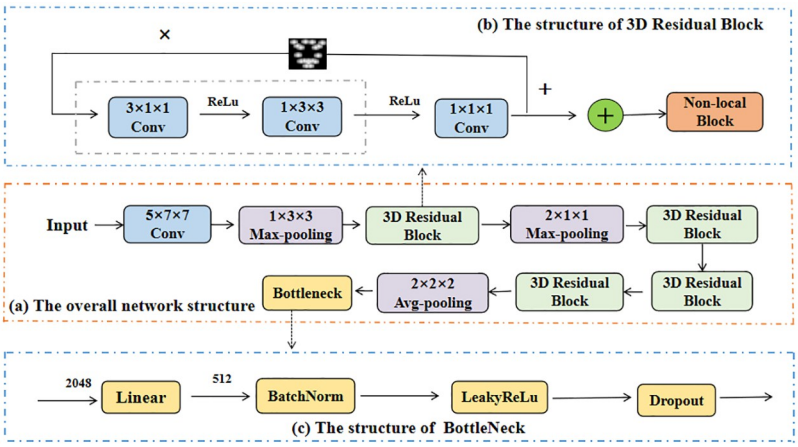

Fig 3. The overall structure of improved 3D-ResNet.

https://doi.org/10.1371/journal.pone.0307446.g003

facial area. Finally, our network also contains a non-local attention block [32] to capture long-range dependencies on spatial and temporal dimension.

3.1.2 The AU-perceived attention module. Since the face in video is an image containing structured targets(facial organs), the contribution of facial active areas will be more prominent. Facial expression recognition needs to pay more attention to the facial key areas. We refer to method proposed by Wei Li [33] and show specific implementation algorithm of attention mapping in Table 1.

First, 68 landmarks based on the key area of the face are obtained, and then the AU center is gained by moving the zoom distance or using the existing landmarks. The rules for defining AU centers also refer to Wei Li [33]. Finally, we build the attention map for AUs based on key facial landmarks.

After that, we adjust the attention map to the 100 x 100 pixels to ensure that the attention map of each image is uniform. For each AU center, 7 pixels near the center are regarded as AU area, so the size of each AU area is 15 x 15 pixels. Higher weight is assigned to the closer points to the AU center. The relationship follows the equation:

$$w_{\alpha} = 1 - 0.095 d_m, \quad d_m = |x_1 - x_2| + |y_1 - y_2| \tag{2}$$

where d_m is the Manhattan distance to the AU center. Suppose the coordinates of the two

Table 1. The algorithm of attention map generation.

Algorithm 1 Attention map generation algorithm.
Input: The path of face image(path); Key area of the face(keyArea);
Output: Coordinates of the bounding box(bbc); the attention map(am).
1: image ← Image.open(path)
2: img_array ← np.array(image)
3: landmark_point ← ObtainLandmark(keyArea, img_array)
4: au_point ← ComputeAuCenter(landmark_point, img_array)
5: dm ← ManhattanDistance(au_point)
6: wa ← 1 – 0.095 * dm
7: bbc, am ← GetBoundingBox(wa, au_point, img_array)
8: return bbc, am

https://doi.org/10.1371/journal.pone.0307446.t001

points are (x_1, y_1) and (x_2, y_2) respectively. The Manhattan distance represents the sum of the absolute axis distances of the two points on the standard coordinate system. We use a relatively large local area to cover the key areas of AU so that the generation method of attention mapping can have a certain degree of fault tolerance to landmark shift.

However, applying the attention layer directly will discard a lot of information which are not included in the attention layer. We use the skipping layer of the residual network to combine the attention enhancement area with other areas so as to address the problem of losing information in non-critical areas. Fig 4 shows the skipping layer connection in residual network and attention mapping in our model. The skipping layer structure combines the lower level spatial features in the convolution layer with the higher-level semantic features to form richer feature representation of the input image.

As shown in Fig 3, the specific operation is to embed the generated attention map in the convolution operation of the first 3D residual block, and set the attention map as a parallel stream separately. After relevant feature map is generated through the foregoing maximum pooling, we will multiply the feature map by the attention map. After the convolution in the 3D residual block is completed, the parallel attention feature generation map and the convolution result will be added element by element for fusion to input the pooling layer of next stage.

3.1.3 Feature encoding and decoding. The capsule network (CapsNet) uses neuron vectors instead of traditional neuron nodes. These neuron vectors can characterize in an image, such as the position, size, and orientation of an object. Therefore, the features of the image can be comprehensively learned. We use the length of the output vector to express the probability of the existence of the entity, and the direction of the vector represents the instantiation

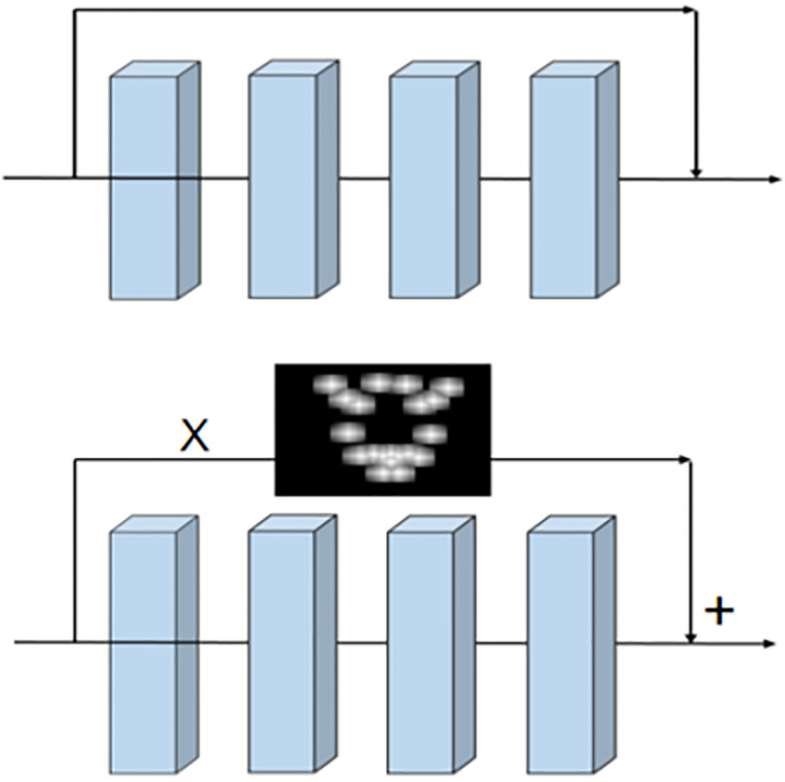

Fig 4. Skipping layer connection in ResNet and application of attention map in our model.

https://doi.org/10.1371/journal.pone.0307446.g004

parameter. A nonlinear squeeze function called squash is used to compress the capsule vector length to a value between 0 and 1. The longer the length is, the higher probability of the entity appearing in the input. To prevent the loss function from failing to converge when the length of the vector is 0, a minimal value $\varepsilon(10^{-7})$ is added to $\|b\|$ to ensure that training proceeds normally, as shown in Eq 4. The formula of the "squashing" function is as follows:

$$v_j = \frac{\|b_j\|^2}{1+\|b_j\|^2} + \frac{b_j}{\|b_j\|} \tag{3}$$

$$\begin{aligned} \|b\| &= \sqrt{\sum_i b_i^2 + \varepsilon}, \\ b_j &= \sum_i c_{ij} w_{ij} \alpha_i, \\ \hat{\alpha}_{j|i} &= w_{ij} \alpha_i, \\ c_{ij} &= \frac{exp(x_{ij})}{\sum_k exp(x_{ik})} \end{aligned} \tag{4}$$

here v_j is the output vector of the output layer, and b_j is the input vector of the output layer, which is a weighted sum of the prediction vectors $\hat{\alpha}_{j|i}$. These prediction vectors are obtained by multiplying the output α_i of the fully connected capsule layer and a pose matrix w_{ik}, as shown in Eq 4. c_{ij} is the coupling coefficient determined by the iterative dynamic routing process. This variable is used to measure the consistency between the fully connected layer capsules and the output layer capsules. The initial value of x_{ij} is set to 0, thereby ensuring that the prior probabilities of the information transmitted by the low-level capsules to the high-level capsules are equal. The coupling coefficients are iteratively determined from the initial values, and the dynamic routing process is shown in Fig 5.

The capsule network architecture and specific network parameters are shown in Fig 2. The 512 x 7 x 7 feature map obtained through the AU attention-constrained dual enhancement network are sent to the PrimaryCaps layer, and a 2 x 2 convolution kernel with a step length of 1 is used to obtain a 256 x 6 x 6 feature map, which is adjusted to 32 8-dimensional feature vectors, the feature map size is 6 x 6. Between the PrimaryCaps layer and the DigitCaps layer, each capsule receives input from all capsules in the previous layer, and the network executes a dynamic routing consensus algorithm. The output of the DigitCaps layer is a 16 x 7

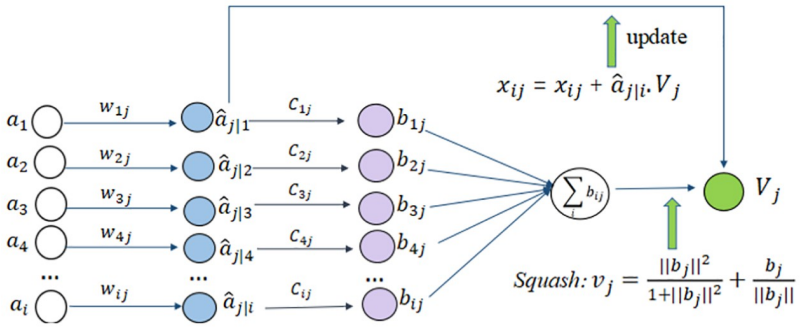

Fig 5. Dynamic routing process.

https://doi.org/10.1371/journal.pone.0307446.g005

dimensional vector, where 16 represents the vector dimension and 7 represents the number of expression category.

For the capsules in PrimaryCaps layer, the activated capsules encode position information. After passing through the DigitCaps layer, the activated capsule transfers the spatial position information of the image to the predicted probability output of the vector. The transition from position information encoding to probability encoding indicates that as the level of the capsules increases, the dimensionality of the capsules should increase. From the perspective of the proposed architecture, the dimension of capsules in the PrimaryCaps layer is an 8-dimensional vector, while the dimension of capsule in the DigitCaps layer is 16 dimensions. The high-level capsules possess more degrees of freedom and can represent more complex entities.

We have three fully connected layers for image reconstruction after the DigitCaps layer, which is the decoding stage of capsule network. The real label is used as the reconstruction target in the training process. The loss function of image reconstruction is constructed by calculating the Euclidean distance between the output pixels of the Sigmoid layer and the pixels of the original image. The coupling coefficient c_{ij} is updated through dynamic routing, while the update operation of other convolution parameters and the weight matrix in the capsule is completed through the loss function. Our loss function contains margin loss L_c and reconstruction loss L_r, and we add a scaling factor ρ to reconstruction the loss to make the margin loss dominate the training process and not be dominated by the reconstruction loss L_r. The formula of each loss function is as follows.

$$L_c = T_c max(0, m^+ - \|v_c\|)^2 + \lambda(1 - T_c) max(0, \|v_c\| - m^-)^2 \tag{5}$$

$$L_r = (x_r - x)^2, \quad L = L_c + 6 * 10^{-4} L_r \tag{6}$$

where c represents the classification category, and T_c represents the indicator function of the classification. If there is a facial expression of the category c, then $T_c = 1$. m^+ represents the upper limit and m^- represents the lower limit. x and x_r denote original image and reconstructed image respectively.

3.2 TPN-based expression recognition module (TPN-ERM)

Generally, various facial expressions usually hold different visual tempos. However, there are always higher similarities in some expressions(Fear and Surprise), and the key to distinguish them is their visual tempos. To this end, we propose to use TPN-based expression recognition module (TPN-ERM) to model the variances in visual tempos of different facial expressions precisely to further optimize the performance of expressions recognition system.

3.2.1 Collection of hierarchical features. TPN is built upon a set of M hierarchical features that have increasing temporal receptive fields from bottom to top, and we employ multi-depth pyramid to collect these features from a backbone network, which is defined as:

$$F = \{F_1, F_2, \ldots, F_M\} \tag{7}$$

Where F_i represents i-th feature, F is a set of M hierarchical features, and the size of each feature F_i is:

$$sizes = \{C_1 \times T_1 \times W_1 \times H_1, \ldots, C_M \times T_M \times W_M \times H_M\} \tag{8}$$

Where H, W, and C represent height, width and numbers of channels of each frame,

repectively. T is the number of frames. The size usually satisfies the following formula:

$$\{C_{i1} \geq C_{i2}, W_{i1} \geq W_{i2}, H_{i1} \geq H_{i2}; i_{i1} \leq i_{i2}\} \tag{9}$$

3.2.2 Spatial and temporal semantic modulation. Multi-depth pyramids can well integrate features of different scales, but there is the problem of unaligned spatial semantics. To solve this problem, spatial semantic modulation is applied to TPN. For each feature (top-level features are not included), a series of convolutions with a specific stride are applied, so that the spatial shape and receptive field of these features match the top-level feature. Therefore, the overall loss function of backbone network of TPN is:

$$L_{total} = L_{CE,o} + \sum_{i=1}^{M-1} \lambda_i L_{CE,i} \tag{10}$$

where $L_{CE,o}$ is the cross entropy loss, $L_{CE,i}$ is the loss of the i-th auxiliary classifier. $\{\lambda_i\}$ is the balance factor. After spatial semantic modulation, features have aligned shapes and consistent semantics in spatial dimensions.

Meanwhile, we also use temporal rate modulation to calibrate in time dimension. In order to improve the flexibility of TPN, we use a set of hyper-parameters $\{\alpha_i\}_{i=1}^{M}$ for temporal rate modulation. Specifically, α means that the parameter subnet will be used after performing spatial semantic modulation, and the updated feature will be temporarily down-sampled at the i-level using α_i as a factor. The use of such hyper-parameters allows us to better control the differences of features on the temporal scale, so that we can perform feature aggregation more effectively. In the following section, we will call F_i whose size is $C_i \times T_i \times W_i \times H_i$ the i-feature after performing spatial semantic modulation and temporal rate modulation.

3.2.3 Adaptive weighted fusion. After the hierarchical features are collected and preprocessed, the next step is to aggregate these features effectively. Assuming that the aggregated feature at i-th level are F'_i, there are three basic ways:

$$IsolationFlow : F'_i = F_i \tag{11}$$

$$Bottom - upFlow : F'_i = F'_i \oplus g(F_i, T_i/T_{i-1}) \tag{12}$$

$$Top - downFlow : F'_i = F'_i \oplus g(F_i, T_i/T_{i+1}) \tag{13}$$

Where $\oplus$ means element-wise addition. $g(F, \delta)$ is applied along the temporal dimension, where F represents the feature.

Besides the top-down flow and bottom-up flow, we could also combine them to achieve two other ways, namely Cascade Flow and Parallel Flow. While applying a bottom-up flow after a top-down flow will form the cascade flow, applying them simultaneously will lead to the parallel flow.

However, TPN does not employ the correlation of features at different levels. Specifically, when these features are aggregated, each of them made a different contribution to the aggregated features. Therefore, based on the FPN [28] and PANet [29], we introduce an adaptive weighted fusion method which can learn parameters to improve the ways of feature aggregation. The top-down and bottom-up flow are integrated and added behind each convolution layer, which can enhance the information fusion. The specific structure is shown in the Fig 6.

Regarding the computational process of feature fusion, the previous pyramid attention network used global self-attention mechanism, but it dose not employ the contribution of input

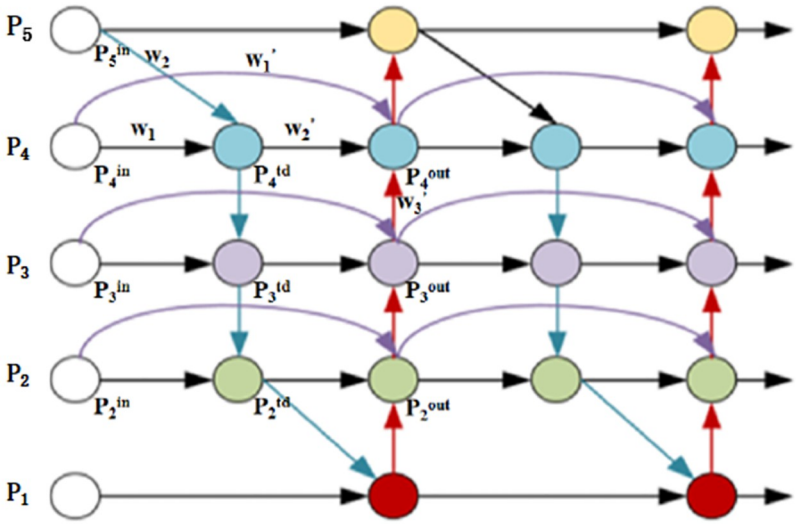

Fig 6. Adaptive weighted fusion. Similar to PANet, but it deletes the nodes with only one in-degree in PANet to eliminate some redundant calculations. Secondly, we added some new connections called cross-node connections shown as the three purple connection curves. The first feature node of the same layer (the white node) in these connections is connected to the output node without top-down feature fusion, while the output nodes participate in bottom-up feature fusion. Finally, inspired by the idea of recursive networks, we integrate top-down with bottom-up flow and add some cross-node connections to enhance fusion of features.

https://doi.org/10.1371/journal.pone.0307446.g006

features to the output features is different. Therefore, we apply a weight to each input feature so that the network can learn the importance of each input feature. Two weighted fusion methods are considered:

$$Softmax - based - fusion : O = \sum_i \frac{e^{w_i}}{\sum_j e^{w_j}} * I_i \tag{14}$$

where w_i is a learnable weight that can be updated during network training, and I_i represents the input weight of i-th layer. The softmax function normalizes the weights to a probability with a range of 0 to 1, which indicates the importance of each input feature. But it will result in more hardware consumption, to minimize additional computational costs, we further propose an efficient fusion method.

$$Fast - standardized - fusion : O = \sum_i \frac{w_i}{\varepsilon \sum_j e^{w_j}} * I_i \tag{15}$$

Applying the ReLU function behind each w_i can ensure $w_i > 0$. Fast standardized fusion also makes the value of each weight range between 0 and 1, but it is more efficient than the softmax-based fusion. We take the fourth layer P4 in Fig 6 as an example to describe our adaptive weighted fusion method that uses fast standardized fusion:

$$P_4^{td} = Conv\left(\frac{w_1 * P_4^{in} + w_2 * Resize(P_5^{in})}{w_1 + w_2 + \varepsilon}\right) \tag{16}$$

$$P_4^{out} = Conv\left(\frac{w_1' * P_4^{in} + w_2' * P_4^{td} + w_3' * Resize(P_3^{out})}{w_1' + w_2' + w_3' + \varepsilon}\right) \tag{17}$$

among them P_4^{td} is the intermediate feature in the top-down path, and P_4^{out} is the output feature in the bottom-up path. All other features are constructed in the similar way.

4 Experiments

In this section, we first describe two public datasets, and then illustrate the specific experimental details and evaluation metrics. The experiment results demonstrate the effectiveness of our 3D-CapsNet and TPN-ERM. Moreover, we compare our integrated model to current effective models on public datasets. Finally, the ablation studies show the effectiveness of each module in our proposed model.

4.1 Datasets

4.1.1 CK+. The Extended Cohn-Kanade (CK+) database [34] collects facial expressions of 123 subjects by videos, and a total of 593 facial expression video sequences. The subjects are young people between ages of 18 and 30. The data sequence vary in duration from 10 to 60 frames. Seven basic emotion classes (anger, contempt, disgust, fear, happiness, sadness, and surprise) are marked in these videos, and the annotation work is based on the Facial Action Coding System (FACS). Since CK+ is not clearly divided into the training set, validation set and test set, the algorithms evaluated on this database are not same. Each image sequence changes from the onset (the neutral frame) to the peak (the expressive frame). Moreover, the X-Y coordinates of 68 facial landmark points were given for each image in the database. The landmark points of key frames within each video sequence were manually labeled, while the remaining frames were automatically aligned using the Active Appearance Model(AAM) fitting algorithm [35].

4.1.2 AFEW. The AFEW database [36] has been used as the official database in the EmotiW since 2013. The AFEW database contains facial expressions collected from different TV and film works which are believed to be closed to real world conditions. The database is comprised of training set, validation set and test set. There are 578 video clips in the training set. The validation and test sets have 383 video clips and 407 video clips, respectively. The video clips are marked with seven expression labels: anger, disgust, fear, happiness, sadness, surprise and neutral. Besides, this database provides original video clips and aligned face sequences. Different from the CK+ database, facial expressions in AFEW are more natural and spontaneous. The variations in illumination, pose and background in image sequences expand the complexity of facial expression analysis.

4.2 Experimental settings

4.2.1 Implementation details. We used tensorflow and pytorch frameworks to conduct experiments, and trained the model with a total of 500 iterations. We adopted the ADAM algorithm as the model trainer in the training process, which can utilize hyper-parameters to greatly accelerate the speed of network convergence. In addition, our training parameters on different data sets are slightly different. On the CK+ and AFEW data sets, we set the trainer parameter such as β_1, β_2 and ε as 0.9, 0.999 and 10–8. The batch size of each epoch is 16. The value of scaling factor ρ is 0.0006 to reconstruction loss in the feature decoding stage.

We used improved TPN for expression auxiliary recognition, which can be integrated into our 3D-CapsNet. We use the inflated 3D-ResNet as the backbone network to ensure good performance on various datasets, and the original ResNet serves as the 2D backbone for contrast. M-level TPN network processes the i-th level features through a series of convolutions with a stride of M_i in spatial semantic modulation and the feature dimension is fixed to 1024. The time rate modulation is realized through the convolution layer and the maximum pooling

layer. Finally, the five kinds of information flow mentioned in section 3 aggregate features separately to make a comparison. All aggregated features in TPN will be rescaled by max pooling operation, and their cascade will be fed into the fully connected layer for final prediction.

4.2.2 Evaluation metrics. Multiple performance and evaluation criteria are used to evaluate the performance of proposed model. Following prior work, we adapt Accuracy, Precision (P), Recall(R) and F1 score. The formulas are as follows.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{19}$$

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + TN} \tag{20}$$

where TP represents the number of samples correctly predicted as positive class, FP represents the number of samples incorrectly predicted as positive class, TN is the number of samples correctly predicted as negative class, and FN is the number of samples incorrectly predicted as negative class. The same is true for multiple classifications, as long as all other categories that do not belong to the current category are considered as negative cases. Higher values denote better performance for all metrics.

4.3 Results of 3D-CapsNet

First, we drew the confusion matrix for emotion prediction on the CK+ and AFEW data sets, and the results are shown in Fig 7. Through the confusion matrix, we can find that the highest recognition accuracy rate on CK+ is happy, which can reach 70.54%, followed by anger, and the lowest recognition rate is disgust, which only reached 23.08%, because most of disgust expressions are recognized as angry and sad. The expression with higher recognition accuracy rate on AFEW is also happy and angry, reaching 50.54% and 47.3% respectively. The lowest recognition rate is surprise, whose recognition accuracy rate is 27.99%, and quite a few surprise expressions are recognized as happy. To sum up, we can see that the recognition effect of facial expressions with obvious characteristics like happy and angry is far greater than that of contempt and other expressions whose facial features are not obvious, which can also illustrate the necessity of paying attention to the key areas of facial expression during the recognition process.

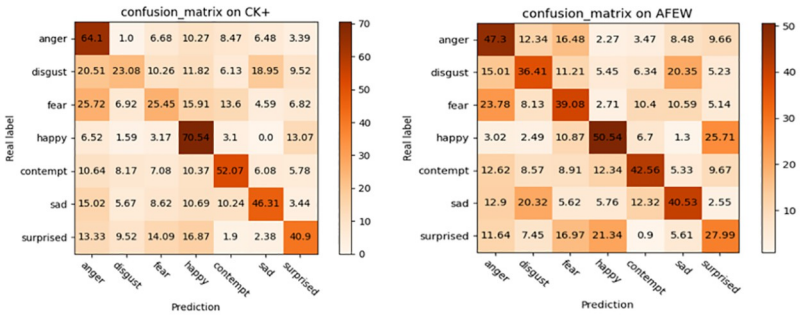

Fig 7. The confusion matrix on CK+(left) and AFEW(right).

https://doi.org/10.1371/journal.pone.0307446.g007

Table 2. The performance (%) comparison of 3D-CapsNet with state-of-the-art methods on CK+ and AFEW.

CK+			AFEW		
Methods	**Data type**	**Acc(%)**	**Methods**	**Data type**	**Acc(%)**
LOMo [37]	Landmark	95.10	SSE-HoloNet [38]	Video	46.47
ST-GCN [39]	Landmark	93.64	FAN [40]	Video	51.18
DGNN [41]	Landmark	96.02	AGCN [42]	Landmark	24.21
CTSLSTM [43]	Landmark	93.90	CAER-Net [44]	Video	51.68
AGCN [42]	Landmark	94.18	uGMM-IMK [45]	Video	49.50
DAM [46]	Video	95.88	uGMM-SVK [45]	Video	47.50
HCIA [47]	Video	96	PST-BLN w/MCD [48]	Landmark	33.33
ST-BLN w/MCD [48]	Landmark	95.47	PST-BLN wo/MCD [48]	Landmark	30.15
ST-BLN wo/MCD [48]	Landmark	93.19	IFERCV [49]	Video	51.62
EFE [50]	Video	94.44	MIC-MI [51]	Video	52.62
EFE+EMR+Encoder(FC) [50]	Video	95.63	MLCNNs+3DCNN [52]	Video	49.3
3D-CapsNet	Video	96.26	3D-CapsNet	Video	52.63

https://doi.org/10.1371/journal.pone.0307446.t002

We compare our model with other effective methods. These methods mainly enhance their learning ability by capturing global and local features on the corresponding dataset. Considering that the face possesses specific structure, we use dynamic routing between capsules to obtain the relationship between AUs. The capsule network encodes spatial information when calculating the possibility of existence of an object. So it is very suitable for FER. In addition, our network can focus on the facial activity area enhanced by the attention map. In general, the benefits of our model are attributed to two enhancement modules, namely 3D-CNN with AU-perceived attention mechanism and the CapsNet with multiple convolution layers. The performance of the proposed method is compared with both video-based and landmark-based state-of-the-art methods on AFEW and CK+ datasets in Table 2. The visualization representation of the experimental results is shown in Fig 8. Some methods such as [45, 50] use videos or image sequences as the main data stream for training the model while some methods such as [37, 48] utilize local facial landmark data to highlight the most important parts of the facial images and improve the performance. the accuracy of our proposed method on CK+ and AFEW reached 96.26% and 52.63%, respectively, which is better than most methods.

Then we further analyzed time complexity, space complexity and recognition accuracy of proposed 3D-CapsNet module and other state-ofthe-art modules in terms of the number

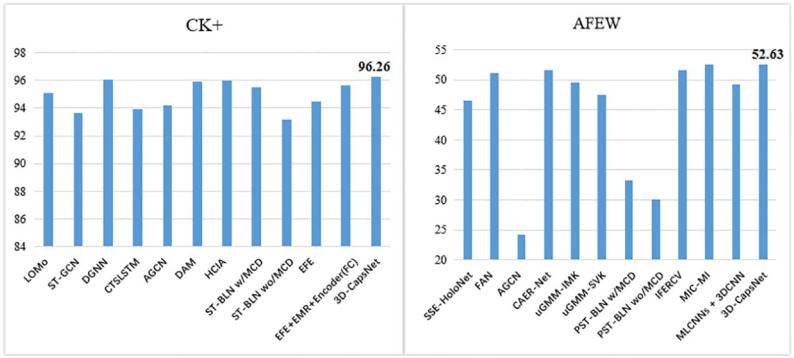

Fig 8. The visualization representation of the experimental results.

https://doi.org/10.1371/journal.pone.0307446.g008

Table 3. The FLOPs, Params and accuracy comparison of 3D-CapsNet with state-of-the-art methods.

			Accuracy(%)	
Methods	**FLOPs/G**	**Params/M**	**CK+**	**AFEW**
SSA [53]	-	23.52	93.56	51.33
SA-Net [54]	-	23.58	92.12	49.80
FcaNet [55]	-	26.04	91.50	48.75
CapsNet [56]	325.37	5.37	76.12	32.30
Ours	445.38	6.40	96.26	52.63

https://doi.org/10.1371/journal.pone.0307446.t003

floating point of operations(FLOPs), the number of model parameters (Params) and accuracy (Acc) in Table 3. It is obvious from Table 3 that our method outperforms others on recognition accuracy. FcaNet utilized a frequency attention method that rethinked channel attention using frequency analysis. Spectral and Spatial Attention (SSA) module integrates spectral semantics with spatial locations to address the problems of interaction between spatial action units and the inadequacy of semantic information about spectral expressions. SA-Net separated channels into two equal portions for channel and spatial attention to address the problem of high computational overhead when fusing spatial attention and channel attention together. Compared to these methods, our model uses fewer parameter quantities and achieves better recognition effect. But the parameter quantity and FLOPs of our model is slightly higher than that of CapsNet. This is because the iteration of dynamic routing in the capsule network results in longer computation time and improved 3D-ResNet introduced extra parameters. But experiment results in Table 3 have shown that the performance of our model is far better than traditional capsule networks and other three models although the importation of extra parameters.

Through training and experiments with the above methods in section 3, the loss functions of 3D convolution network and our 3D-CapsNet are shown in Fig 9. As can be seen from the figures, the loss of our 3D-CapsNet model is lower than that of single 3D convolution network, and the difference between the predicted value and the true value of the facial expression recognition result gets smaller, which means our model can better optimize the previous facial expression classification model and have better recognition performance.

4.4 Impact of TPN-ERM on 3D-CapsNet

In this section, we show the visualization results after applying the improved TPN and compare 3D-CapsNet+TPN-ERM(VT-3DCapsNet) with 3D-CapsNet and other state-of-the-art methods to analyze the impact of TPN-ERM on 3D-CapsNet.

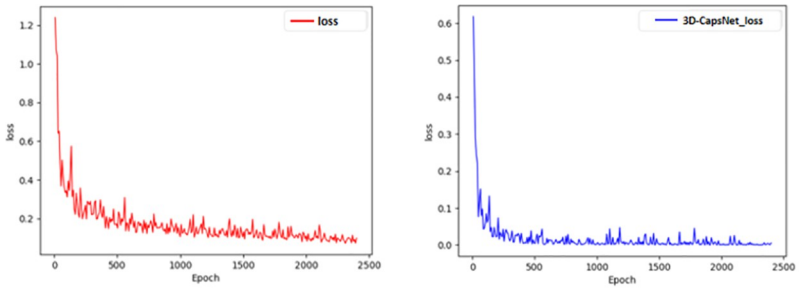

Fig 9. The loss functions of 3D convolution network and 3D-CapsNet.

https://doi.org/10.1371/journal.pone.0307446.g009

Table 4. The performance (%) comparison of 3D-CapsNet with state-of-the-art methods on CK+ and AFEW.

CK+			AFEW		
Methods	**Data type**	**Acc(%)**	**Methods**	**Data type**	**Acc(%)**
(N+M)-tuplet [57]	Landmark	93.90	SSE-HoloNet [38]	Landmark	28.17
DAM [46]	Video	95.88	E-ConvLSTM [58]	Video	45.29
NSVT [59]	Video	96.5	DGNN [41]	Landmark	32.64
CUDL [60]	Video	96.6	C3D-GRU [61]	Video	49.87
PST-BLN wo/MCD [48]	Landmark	93.10	MRAN [62]	Video	49.01
PST-BLN w/MCD [48]	Landmark	93.34	IFERCV+Adv [49]	Video	52.01
Proposed Method w/o Attention [63]	Video	93.84	Emotion-BEEU [64]	Video	52.49
EFE+EMR+Encoder [50]	Video	97.17	IFERCV+T^t [49]	Video	51.86
EFE+EMR+EPMG [50]	Video	98.06	ST-BLN w/MCD [48]	Landmark	36.11
DRL [65]	Video	89.8	ST-BLN wo/MCD [48]	Landmark	34.13
			CEFLNet [66]	Video	53.98
VT-3DCapsNet	Video	96.5	VT-3DCapsNet	Video	55.01

https://doi.org/10.1371/journal.pone.0307446.t004

We compared the proposed integrated model with other facial expression recognition methods. The results of the comparison experiments are shown in Table 4. The visualization representation of the experimental results is shown in Fig 10.

It can be known from the in Table 4 that our method has achieved better and more accurate results on the CK+ and AFEW data sets. Especially, our method can achieve the highest accuracy rate of 98.57% on the CK+, which improves the accuracy by 8.77% compared with DRL in Table 4. The effect on the AFEW data set is also better than most other methods, which can reach an accuracy of 55.01%.

Table 5 shows the precision(P), recall(R), and F1 obtained by our model for recognizing various expressions. From the above table, it can be seen that proposed integrated model(VT-3DCapsNet)has the highest recall rate in terms of happy and surprise, which can achieve 99.68% and 99.34% on the CK+ dataset. The lower recall rates are sad and fear, which only reached 92.82% and 93.74% on CK+, respectively. From the perspective of precision, the precision of happy is the highest, reaching 100% on CK+, while the precision of fear is low, which only achieved 92% on CK+ and 30% on AFEW. F1 represents the model's ability to recognize various facial expressions. The most prominent F1 values in Table 5 are happy and surprise, which reached 99.8% and 99.1% on CK+, and they also reached optimal values of 66% and

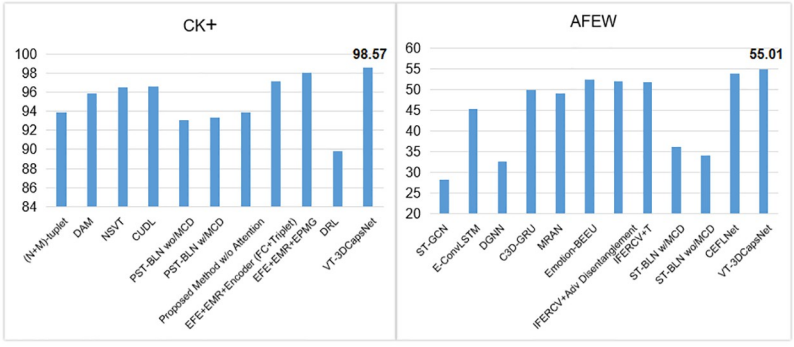

Fig 10. The visualization representation of the experimental results.

https://doi.org/10.1371/journal.pone.0307446.g010

Table 5. The precision, recall, and F1 obtained by our model for recognizing various expressions.

	AFEW			CK+		
Label	**P(%)**	**R(%)**	**F1(%)**	**P(%)**	**R(%)**	**F1(%)**
anger	37.13	41.92	39.15	95.17	94.32	95.82
disgust	49.43	44.76	47.07	94.57	96.14	95.17
fear	30.00	31.33	31.85	92.00	93.74	93.22
happy	63.21	70.16	66.00	100	99.68	99.80
contempt	46.72	56.25	50.17	97.43	95.61	96.68
sad	33.18	39.34	36.27	95.81	92.82	94.41
surprise	56.29	99.34	71.00	99.70	99.34	99.10

https://doi.org/10.1371/journal.pone.0307446.t005

71% on AFEW. The experiment results indicate that our model has the strongest recognition ability for happy and surprise.

4.5 Ablation experiment

Both 3D-CapsNet for feature representations learning and TPN-ERM for optimizing the performance of expressions recognition system gain improvements on FER. We conducted a quantitative evaluation of these two models in order to better understand our method. For a more detailed analysis of the FER results, we also explored how different network components affect the performance of 3D-CapsNet and TPN-ERM.

4.5.1 Evaluation of 3D-CapsNet on AFEW and CK+. In order to prove the effectiveness of the 3D-CapsNet with AU-perceived attention mechanism, we conducted ablation experiments on the AFEW dataset. We compare our model with VGG16 used as basic backbone network for feature extraction. The experimental results are shown in Table 6. It can be seen from the Table 6 that using 3D convolution is better than using VGG16 in processing video tasks. When 3D convolution with AU-perceived attention is used for feature extraction, the model improves the accuracy by 2.1% compared with using VGG16 singly and is also better than the accuracy of the 3D convolution without the attention mechanism. When adding the capsule network, our 3D-CapsNet improves the accuracy by 7% compared to the VGG16 and is slightly better than the model that uses 3D convolution and attention mechanism but does not integrate with the capsule network.

After that, we used different models as contrasts on CK+ to evaluate the influence of the components of 3D-CapsNet and the results are shown in the following Table 7. It can be seen from the table that when the capsule network does not contain multi-layer convolution and attention mechanism is also not used, the recognition accuracy is 76.12% and is relatively low. VGG16 has a deep convolution structure and does not have the dynamic routing mechanism of the capsule network and attention mechanism, which achieves an accuracy of 78.14%. The

Table 6. The ablation study result on AFEW.

3D convolution	capsule neural network	AU perceptual attention	Acc
×	×	×	45.6
√	×	×	46.8
√	×	√	47.7
×	√	×	32.3
√	√	×	48.2
√	√	√	**52.6**

https://doi.org/10.1371/journal.pone.0307446.t006

Table 7. The ablation study result on CK+.

Method	capsule dynamic routing	deep convolution	attention mechanism	Acc
VGG16	×	√	×	78.14
CapsNet	√	×	×	76.12
AVGGNet	×	√	√	79.29
RCCnet	√	√	×	81.12
Ours	√	√	√	**96.26**

https://doi.org/10.1371/journal.pone.0307446.t007

AVGGNet network contains both deep convolution and attention mechanism and further improve the accuracy to 79.29%. RCCnet contains deep convolution and capsule, and the accuracy rate is increased to 81.12%. Our model contains the attention mechanism of capsule, deep convolution and attention with AU perception and can achieve the highest accuracy rate of 96.26%, thus proving the effectiveness of our method.

To illustrate the superiority of our 3D-CapsNet more intuitively, we provide the accuracy curves of our method and other methods on the CK+ data set. As can be seen from the Fig 11, the accuracy of our model in the training process has always been higher than other traditional methods.

4.5.2 Evaluation of TPN-ERM on AFEW and CK+. To prove the effectiveness and flexibility of the TPN-ERM we used, we conduct ablation experiments on the CK+ and AFEW datasets. The experimental results are shown in the following Fig 12.

Among them, the purple histogram is the result of only using the 3D-CapsNet model, and the yellow histogram is the result of 3D-CapsNet integrated with improved TPN. It can be seen from the figure that the using the improved TPN can improve the recognition accuracy of the model, which improves the accuracy of 2.31% and 2.38% on the CK+ dataset and AFEW dataset, respectively. These prove that using TPN module is feasible and effective for improving the accuracy of facial expression recognition.

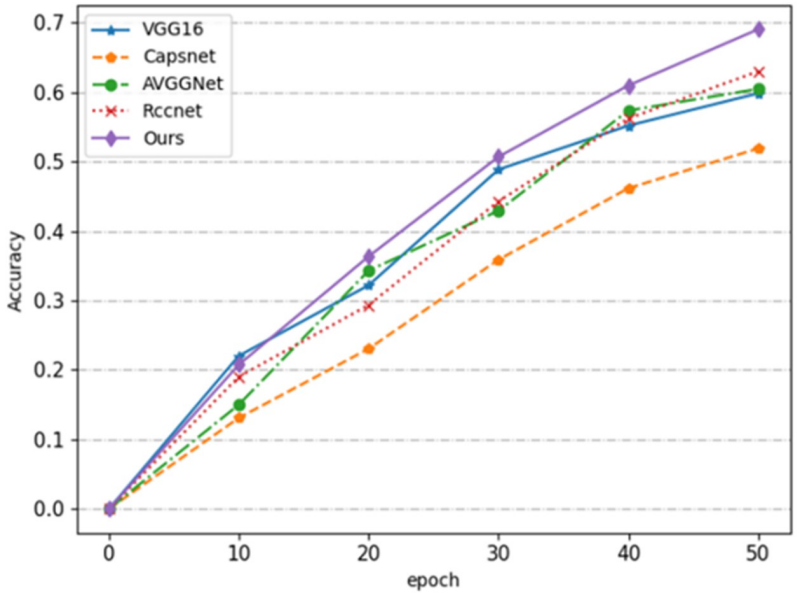

Fig 11. Comparison of the training process of each model.

https://doi.org/10.1371/journal.pone.0307446.g011

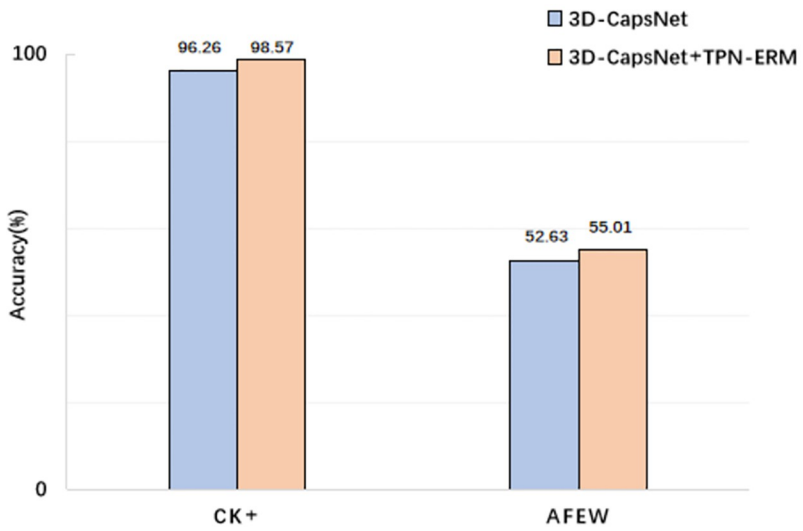

Fig 12. The result of using and not using improved TPN on CK+ and AFEW.

https://doi.org/10.1371/journal.pone.0307446.g012

There are four methods of information fusion in the original TPN structure, but we use a novel feature fusion method, namely adaptive weighted fusion. In order to prove that our improvement to TPN are effective, we give the detailed graphs of the accuracy of the first 50 rounds of training on the two data sets in Fig 13.

As can be seen from the Fig 14, our improved TPN with adaptive weighted fusion is more accurate than other information flow fusion methods in the first 50 rounds of training, which proves that using TPN with adaptive weighted fusion is effective.

There are two main calculation methods in our adaptive weighted fusion, namely softmax-based fusion and fast standardized fusion. But the additional softmax will cause more hardware consumption. Therefore, to prove that our fast standardized fusion can process faster, we select a piece of video on the CK+ and compare the processing speed of two calculation methods.

The processing speed of faster standardized fusion is significantly faster than that of softmax-based fusion in Fig 15. When the average velocity of processing video is equal, the fast standardized fusion is 2 seconds faster than softmax-based fusion and obtains a better level.

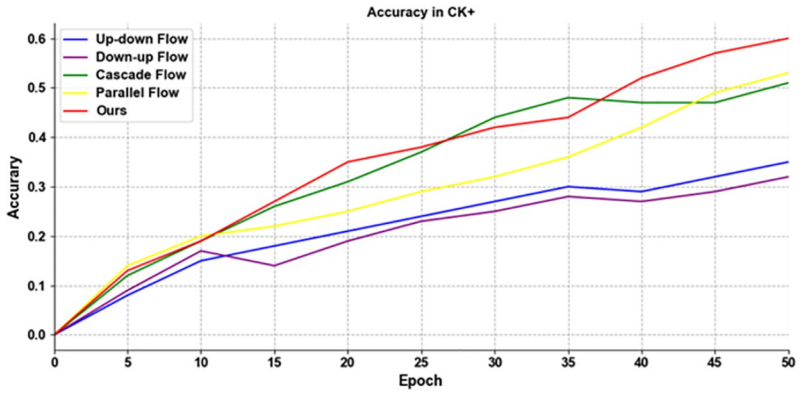

Fig 13. The results of using different information flow methods for TPN training on CK+.

https://doi.org/10.1371/journal.pone.0307446.g013

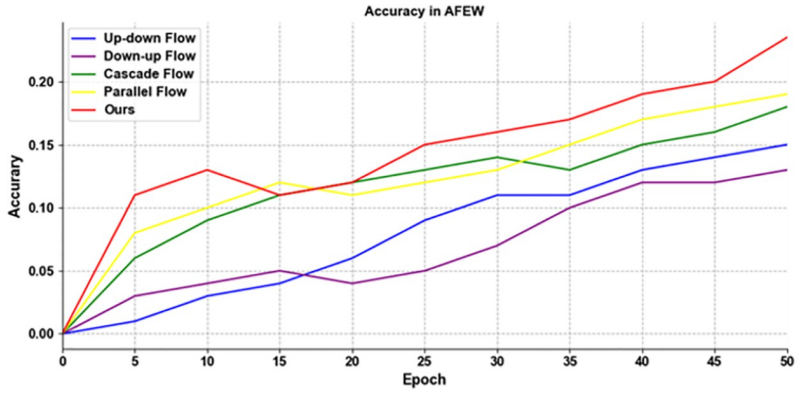

Fig 14. The results of using different information flow methods for TPN training on AFEW.

https://doi.org/10.1371/journal.pone.0307446.g014

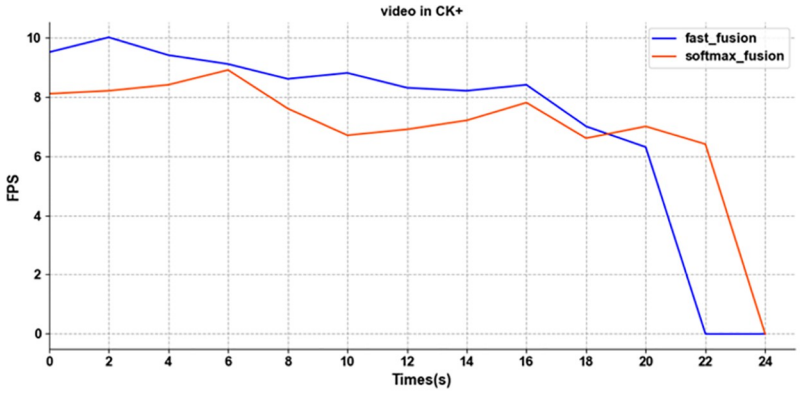

Fig 15. The processing speed result of softmax-based fusion and fast standardized fusion.

https://doi.org/10.1371/journal.pone.0307446.g015

4.6 Discussion

In the experimental section of 3D-CapsNet, we aimed to prove that the improved capsule network that applyed improved 3D-ResNet architecture achieved the goal of expressing deeper hierarchical spatiotemporal features and handle complex feature relationships. Furthermore, the confusion matrices on the CK+ and AFEW datasets, when contrasting expressions with obvious features against those with less obvious features, demonstrate a significant impact of facial key regions on the recognition effectiveness. Comparative evaluations with other state-of-the-art models further underscored the outstanding performance of our proposed model in recognition accuracy.

TPN-ERM calculates facial motion features representing expressions from a considerable number of video frames and uses these features as auxiliary outputs to further enhance the performance of 3D-CapsNet. To analysis of TPN-ERM's impact on 3D-CapsNet, we conducted relevant experiments for performance comparison and complexity analysis, and the results revealed that revealed that TPN-ERM excelled in restore the detailed information of the original images, capturing variations in video, and significantly improving recognition performance of 3D-CapsNet.

To gain a comprehensive understanding of proposed model, we conducted a series of ablation experiments to show the influence of different model components on recognition performance. Specifically, we combined different components of 3D-CapsNet for ablation experiments and validated significant impact of 3D convolution, AU-perceived attention module and capsule neural network on recognition accuracy. Furthermore, we conducted other ablation experiments to prove the impact of TPN-ERM on 3D-CapsNet, and the results established the feasibility and effectiveness of TPN-ERM, which can significantly elevate recognition accuracy of 3D-CapsNet.

Although our model has demonstrated excellent performance in facial expression recognition, there are still some limitations we need to overcome in the future:

1. The proposed 3D-CapsNet model introduces additional parameters compared to the original capsule network, which increases training time and additional hardware memory costs.
2. There are a large number of "ambiguous phenomenon" in more complex real-world scenarios such as low image resolution, occlusion and ambiguous expressions, which can result in low recognition accuracy. So our model can be further improved and discussed in this respect.

5 Conclusion

In this paper, we propose a visual tempos 3D-CapsNet to better learn spatiotemporal feature. We also propose a TPN-based expression recognition module (TPN-ERM) that models the variances in visual tempos of facial expressions actions precisely to further optimize the performance of 3D-CapsNet. Extensive experiments demonstrate that 3D-CapsNet outperforms most state-of-the-art models in terms of the accuracy after adding improved 3D-ResNet architecture that integrated with AU-perceived attention module. It also proves that the feature representation of 3D-CapsNet are more informative after integrating with the TPN-ERM. In the future, we will consider optimizing the routing algorithm of CapsNet to reduce the network parameters. Furthermore, the datasets more complex real-world scenarios(low image resolution, occlusion and ambiguous expressions) we used in the experiments are not sufficient, so the performance of expression recognition in complex scenes may not be satisfactory. The next step is to test the performance of our model on relevant datasets for further improvement.

Author Contributions

Conceptualization: Zhuan Li, Jin Liu, Hengyang Wang.

Data curation: Zhuan Li.

Formal analysis: Zhuan Li.

Funding acquisition: Jin Liu.

Investigation: Zhuan Li, Hengyang Wang.

Methodology: Zhuan Li, Jin Liu, Hengyang Wang.

Project administration: Zhuan Li, Jin Liu, Hengyang Wang.

Resources: Zhuan Li.

Software: Zhuan Li.

Supervision: Zhuan Li, Jin Liu.

Validation: Zhuan Li.

Visualization: Zhuan Li.

Writing – original draft: Zhuan Li, Xiliang Zhang, Zhongdai Wu, Bing Han.

Writing – review & editing: Zhuan Li, Jin Liu.

References

1. Shan L, Deng W. Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing. 2018; PP(99).
2. Yang Y, Xu C, Dong F, Wang X. A new multi-scale convolutional model based on multiple attention for image classification. Applied Sciences. 2019; 10(1):101. https://doi.org/10.3390/app10010101
3. Liu J, Yang Y, Lv S, Wang J, Chen H. Attention-based BiGRU-CNN for Chinese question classification. Journal of Ambient Intelligence and Humanized Computing. 2019; p. 1–12.
4. Shang S, Liu J, Yang Y. Multi-layer transformer aggregation encoder for answer generation. IEEE Access. 2020; 8:90410–90419. https://doi.org/10.1109/ACCESS.2020.2993875
5. Han S, Liu J, Zhang J, Gong P, Zhang X, He H. Lightweight dense video captioning with cross-modal attention and knowledge-enhanced unbiased scene graph. Complex & Intelligent Systems. 2023; 9 (5):4995–5012. https://doi.org/10.1007/s40747-023-00998-5 PMID: 36855683
6. Zhang K, Huang Y, Du Y, Wang L. Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Transactions on Image Processing. 2017; 26(9):4193–4203. https://doi.org/10.1109/TIP.2017.2689999 PMID: 28371777
7. Fan Y, Lam JC, Li VO. Multi-region ensemble convolutional neural network for facial expression recognition. In: International Conference on Artificial Neural Networks. Springer; 2018. p. 84–94.
8. Li S, Deng W, Du J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2852–2861.
9. Liu J, Yang Y, He H. Multi-level semantic representation enhancement network for relationship extraction. Neurocomputing. 2020; 403:282–293. https://doi.org/10.1016/j.neucom.2020.04.056
10. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. Advances in neural information processing systems. 2017; 30.
11. Chang S, Liu J. Multi-lane capsule network for classifying images with complex background. IEEE Access. 2020; 8:79876–79886. https://doi.org/10.1109/ACCESS.2020.2990700
12. Yang C, Xu Y, Shi J, Dai B, Zhou B. Temporal pyramid network for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 591–600.
13. Niu B, Gao Z, Guo B. Facial expression recognition with LBP and ORB features. Computational Intelligence and Neuroscience. 2021; 2021:1–10. https://doi.org/10.1155/2021/8828245 PMID: 33505453
14. Xiang Z, Yang R, Deng C, Teng M, She M, Teng D. An illumination insensitive descriptor combining the CSLBP features for street view images in augmented reality: experimental studies. ISPRS International Journal of Geo-Information. 2020; 9(6):362. https://doi.org/10.3390/ijgi9060362
15. Liao J, Lin Y, Ma T, He S, Liu X, He G. Facial expression recognition methods in the wild based on fusion feature of attention mechanism and LBP. Sensors. 2023; 23(9):4204. https://doi.org/10.3390/s23094204 PMID: 37177408
16. Huang Q, Huang C, Wang X, Jiang F. Facial expression recognition with grid-wise attention and visual transformer. Information Sciences. 2021; 580:35–54. https://doi.org/10.1016/j.ins.2021.08.043
17. Wu X, He J, Huang Q, Huang C, Zhu J, Huang X, et al. FER-CHC: Facial expression recognition with cross-hierarchy contrast. Applied Soft Computing. 2023; 145:110530. https://doi.org/10.1016/j.asoc.2023.110530
18. Zakieldin K, Khattab R, Ibrahim E, Arafat E, Ahmed N, Hemayed E. ViTCN: Hybrid Vision Transformer with Temporal Convolution for Multi-Emotion Recognition. International Journal of Computational Intelligence Systems. 2024; 17(1):64. https://doi.org/10.1007/s44196-024-00436-5
19. Jiang F, Huang Q, Mei X, Guan Q, Tu Y, Luo W, et al. Face2nodes: learning facial expression representations with relation-aware dynamic graph convolution networks. Information Sciences. 2023; 649:119640. https://doi.org/10.1016/j.ins.2023.119640

20. Zhou S, Wu X, Jiang F, Huang Q, Huang C. Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks. International Journal of Environmental Research and Public Health. 2023; 20(2):1400. https://doi.org/10.3390/ijerph20021400 PMID: 36674161
21. Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE international conference on computer vision workshops; 2017. p. 3154–3160.
22. Teng J, Zhang D, Zou W, Li M, Lee DJ. Typical facial expression network using a facial feature decoupler and spatial-temporal learning. IEEE Transactions on Affective Computing. 2021;.
23. de Melo WC, Granger E, Lopez MB. MDN: A deep maximization-differentiation network for spatio-temporal depression detection. IEEE transactions on affective computing. 2021;.
24. Khanna D, Jindal N, Rana PS, Singh H. Enhanced spatio-temporal 3D CNN for facial expression classification in videos. Multimedia Tools and Applications. 2023; p. 1–18.
25. Shu X, Li J, Shi L, Huang S. RES-CapsNet: an improved capsule network for micro-expression recognition. Multimedia Systems. 2023; 29(3):1593–1601. https://doi.org/10.1007/s00530-023-01068-z
26. Zhao P, Ming Y, Hu N, Lyu B, Zhou J. DSNet: Dual-stream multi-scale fusion network for low-quality 3D face recognition. AIP Advances. 2023; 13(8). https://doi.org/10.1063/5.0153077
27. Ye M, Liu G. Facial expression recognition method based on shallow small convolution kernel capsule network. Journal of Circuits, Systems and Computers. 2021; 30(10):2150177. https://doi.org/10.1142/S0218126621501772
28. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2117–2125.
29. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 8759–8768.
30. Yang E, Liu S, Liu Y, Fang K. PSTNet: Crowd Flow Prediction by Pyramidal Spatio-Temporal Network. IEICE TRANSACTIONS on Information and Systems. 2021; 104(10):1780–1783. https://doi.org/10.1587/transinf.2020EDL8111
31. Chen Y, Ge H, Liu Y, Cai X, Sun L. Agpn: Action granularity pyramid network for video action recognition. IEEE Transactions on Circuits and Systems for Video Technology. 2023;.
32. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7794–7803.
33. Li W, Abtahi F, Zhu Z, Yin L. EAC-Net: A Region-based Deep Enhancing and Cropping Approach for Facial Action Unit Detection. IEEE transactions on pattern analysis and machine intelligence. 2018; 40 (11):2583–2596. https://doi.org/10.1109/TPAMI.2018.2791608 PMID: 29994168
34. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops; 2010. p. 94–101.
35. Tzimiropoulos G, Pantic M. Optimization problems for fast aam fitting in-the-wild. In: Proceedings of the IEEE international conference on computer vision; 2013. p. 593–600.
36. Dhall A, Goecke R, Lucey S, Gedeon T. Acted facial expressions in the wild database. Australian National University, Canberra, Australia, Technical Report TR-CS-11. 2011; 2:1.
37. Sikka K, Sharma G, Bartlett M. Lomo: Latent ordinal model for facial analysis in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 5580–5589.
38. Hu P, Cai D, Wang S, Yao A, Chen Y. Learning supervised scoring ensemble for emotion recognition in the wild. In: Proceedings of the 19th ACM international conference on multimodal interaction; 2017. p. 553–560.
39. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence; 2018.
40. Meng D, Peng X, Wang K, Qiao Y. Frame attention networks for facial expression recognition in videos. In: 2019 IEEE international conference on image processing (ICIP). IEEE; 2019. p. 3866–3870.
41. Shi L, Zhang Y, Cheng J, Lu H. Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 7912–7921.
42. Shi L, Zhang Y, Cheng J, Lu H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 12026–12035.

43. Hu M, Wang H, Wang X, Yang J, Wang R. Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks. Journal of Visual Communication and Image Representation. 2019; 59:176–185. https://doi.org/10.1016/j.jvcir.2018.12.039

44. Lee J, Kim S, Kim S, Park J, Sohn K. Context-aware emotion recognition networks. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 10143–10152.

45. Perveen N, Roy D, M Chalavadi K. Facial expression recognition in videos using dynamic kernels. IEEE Transactions on Image Processing. 2020; 29:8316–8325. https://doi.org/10.1109/TIP.2020.3011846 PMID: 32746249

46. Xie S, Hu H, Wu Y. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. Pattern recognition. 2019; 92:177–191. https://doi.org/10.1016/j.patcog.2019.03.019

47. Chowdary MK, Nguyen TN, Hemanth DJ. Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications. 2021; p. 1–18.

48. Heidari N, Iosifidis A. Progressive spatio-temporal bilinear network with Monte Carlo dropout for landmark-based facial expression recognition with uncertainty estimation. In: 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP). IEEE; 2021. p. 1–6.

49. Liu X, Jin L, Han X, Lu J, You J, Kong L. Identity-aware facial expression recognition in compressed video. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE; 2021. p. 7508–7514.

50. Zhang J, Yu H. Improving the facial expression recognition and its interpretability via generating expression pattern-map. Pattern Recognition. 2022; 129:108737. https://doi.org/10.1016/j.patcog.2022.108737

51. Liu X, Jin L, Han X, You J. Mutual information regularized identity-aware facial expression recognition in compressed video. Pattern Recognition. 2021; 119:108105. https://doi.org/10.1016/j.patcog.2021.108105

52. Nguyen HD, Kim SH, Lee GS, Yang HJ, Na IS, Kim SH. Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks. IEEE Transactions on Affective Computing. 2019; 13(1):226–237. https://doi.org/10.1109/TAFFC.2019.2946540

53. Gao H, Wu M, Chen Z, Li Y, Wang X, An S, et al. SSA-ICL: Multi-domain adaptive attention with intra-dataset continual learning for Facial expression recognition. Neural Networks. 2023; 158:228–238. https://doi.org/10.1016/j.neunet.2022.11.025 PMID: 36473290

54. Zhang QL, Yang YB. Sa-net: Shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2021. p. 2235–2239.

55. Qin Z, Zhang P, Wu F, Li X. Fcanet: Frequency channel attention networks. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 783–792.

56. Madhu G, Govardhan A, Srinivas BS, Sahoo KS, Jhanjhi N, Vardhan K, et al. Imperative dynamic routing between capsules network for malaria classification. CMC-Computers Materials & Continua. 2021; 68(1):903–919. https://doi.org/10.32604/cmc.2021.016114

57. Liu X, Ge Y, Yang C, Jia P. Adaptive metric learning with deep neural networks for video-based facial expression recognition. Journal of Electronic Imaging. 2018; 27(1):013022–013022. https://doi.org/10.1117/1.JEI.27.1.013022

58. Miyoshi R, Nagata N, Hashimoto M. Facial-expression recognition from video using enhanced convolutional lstm. In: 2019 Digital Image Computing: Techniques and Applications (DICTA). IEEE; 2019. p. 1–6.

59. Shehu HA, Browne W, Eisenbarth H. Emotion categorization from video-frame images using a novel sequential voting technique. In: Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15. Springer; 2020. p. 618–632.

60. Muhammad G, Hossain MS. Emotion recognition for cognitive edge computing using deep learning. IEEE Internet of Things Journal. 2021; 8(23):16894–16901. https://doi.org/10.1109/JIOT.2021.3058587

61. Lee MK, Choi DY, Kim DH, Song BC. Visual scene-aware hybrid neural network architecture for video-based facial expression recognition. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE; 2019. p. 1–8.

62. Lee J, Kim S, Kim S, Sohn K. Multi-modal recurrent attention networks for facial expression recognition. IEEE Transactions on Image Processing. 2020; 29:6977–6991. https://doi.org/10.3390/s20185184 PMID: 32932939

63. Zhu X, Ye S, Zhao L, Dai Z. Hybrid attention cascade network for facial expression recognition. Sensors. 2021; 21(6):2003. https://doi.org/10.3390/s21062003 PMID: 33809038

64. Kumar V, Rao S, Yu L. Noisy student training using body language dataset improves facial expression recognition. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer; 2020. p. 756–773.

65. Mishra S, Joshi B, Paudyal R, Chaulagain D, Shakya S. Deep residual learning for facial emotion recognition. In: Mobile Computing and Sustainable Informatics: Proceedings of ICMCSI 2021. Springer; 2022. p. 301–313.

66. Liu Y, Feng C, Yuan X, Zhou L, Wang W, Qin J, et al. Clip-aware expressive feature learning for video-based facial expression recognition. Information Sciences. 2022; 598:182–195. https://doi.org/10.1016/j.ins.2022.03.062