BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **24**, No 1
Sofia • 2024 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2024-0010

Efficient DenseNet Model with Fusion of Channel and Spatial Attention for Facial Expression Recognition

Duong Thang Long

Faculty of Information Technology, Hanoi Open University, Vietnam
E-mail: duongthanglong@gmail.com

***Abstract*:** *Facial Expression Recognition* (*FER*) *is a fundamental component of human communication with numerous potential applications. Convolutional neural networks, particularly those employing advanced architectures like Densely connected Networks* (*DenseNets*), *have demonstrated remarkable success in FER. Additionally, attention mechanisms have been harnessed to enhance feature extraction by focusing on critical image regions. This can induce more efficient models for image classification. This study introduces an efficient DenseNet model that utilizes a fusion of channel and spatial attention for FER, which capitalizes on the respective strengths to enhance feature extraction while also reducing model complexity in terms of parameters. The model is evaluated across five popular datasets: JAFFE, CK+, OuluCASIA, KDEF, and RAF-DB. The results indicate an accuracy of at least 99.94% for four lab-controlled datasets, which surpasses the accuracy of all other compared methods. Furthermore, the model demonstrates an accuracy of 83.18% with training from scratch on the real-world RAF-DB dataset.*

***Keywords*:** *Convolutional neural networks, Dense connected network architectures, Channel and spatial attention mechanisms, Facial expression recognition.*

1. Introduction

Facial Expression Recognition (FER) presents significant challenges in computer vision, with applications in various real-life scenarios. It plays a crucial role in interpersonal communication by enabling others to understand a person's emotions and intentions, making it an essential element in human interaction. Its applications are diverse, including human-computer interaction, image captioning, video transcription, and social communication.

Paul Ekman and Wallace Friesen initially identified six fundamental human facial expressions, which include Happiness (Ha), Sadness (Sa), Surprise (Su), Disgust (Di), Anger (An), and Fear (Fe), as mentioned in [1, 2]. These facial expressions are believed to be universally recognized across various nationalities, ethnicities, and religions. Additionally, some authors [3, 4] have proposed including

Contempt (Co) and Neutral (Ne) as basic facial expressions. These six basic facial expressions are illustrated in Fig. 1, sourced from the JAFFE [5] dataset.

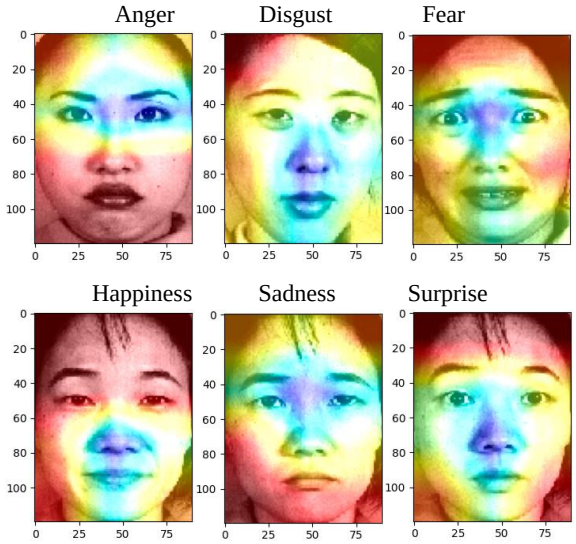


Fig. 1. Example images of six basic facial expressions from the JAFFE dataset. The highlighted areas on these images are generated by the gradient-based localization of the proposed model, which they indicate as important for FER

Intuitively, facial expressions are closely tied to the extent of deformation observed in facial landmarks and muscles, particularly in critical regions like the eyes, nose, and mouth. Consequently, these areas play a pivotal role in addressing the FER challenge [6]. For instance, as depicted in Fig. 1, the eye region consistently appears in most facial expressions but gains exceptional prominence in expressions of anger, fear, surprise, and sadness. Therefore, specific facial movements such as raised eyebrows, locked eyebrows, and movements in the corners of the mouth are regarded as fundamental components of expression changes. These factors can significantly influence the performance and efficiency of FER systems, especially those relying on computer vision techniques. Hence, when developing models for facial expression recognition, meticulous attention to these distinctive features becomes imperative.

DenseNet, introduced by H u a n g et al. [4], has gained prominence as a breakthrough architecture. Its key innovation lies in the dense connectivity scheme, illustrated in Fig. 2, where each layer connects densely to all others within a Dense Block (DB). This design promotes feature reuse, optimizes gradient flow, and reduces parameters. Consequently, it effectively addresses challenges like the vanishing gradient problem, enabling the training of deep networks. DenseNet excels in tasks with limited data, forming robust feature hierarchies that capture details ranging from edges to complex objects in image recognition. In FER, where capturing spatial and semantic information is crucial, DenseNet's ability to preserve fine-grained details and extract meaningful features becomes invaluable. This is especially valuable in

scenarios with limited data or computational resources, making DenseNet a cornerstone in various CNN models across research [1, 10, 16, 17].

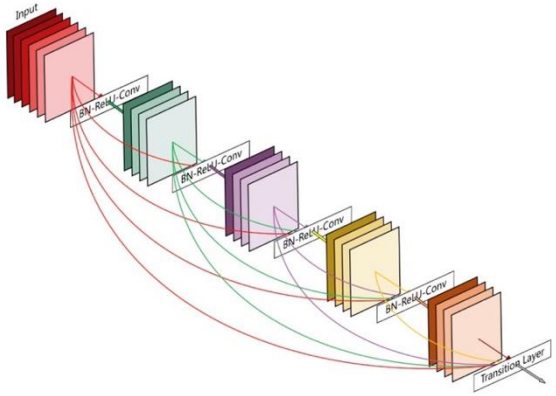


Fig. 2. Illustration of a dense block in DenseNet architecture [7]

On the other hand, attention mechanisms [8] have gained prominence in recent years for their ability to explicitly establish dependencies among features. They excel at enhancing feature representations generated by standard convolutional layers. In FER, this can help the network focus on critical regions of the face, such as the eye area, nose area, and mouth area, where crucial facial expressions are often manifested.

These insights motivate our approach of integrating attention mechanisms with densely connected convolutional layers to improve the performance of Convolutional Neural Networks (CNNs) for FER. This helps reduce the number of parameters in the network and enhances its generalization ability, allowing it to capture both spatial and semantic information in the images. The proposed model is assessed on the performance of well-known FER datasets, encompassing various poses and lighting conditions.

The remainder of this paper is organized as follows: Section 2 shows a brief review of the literature related to CNNs for FER, DenseNet, and attention mechanisms. In Section 3, the approach in detail is explained and illustrated. Section 4 describes our experiments and results. Finally, Section 5 includes the conclusion of this research and a discussion on future work.

2. Related works

In recent years, significant efforts have been dedicated to enhancing FER performance in both controlled laboratory settings and real-world applications. Researchers have predominantly focused on harnessing modern Convolutional Neural Networks (CNNs) with various architectures, including well-established ones such as VGGNet [9], ResNet [10], InceptionNet [11], SENet [12], and DenseNet [7]. CNNs exhibit the remarkable ability to extract intricate features from images, consistently achieving state-of-the-art results on benchmark datasets [3, 13].

DenseNet is an efficient architecture for FER in recent articles. Its dense connectivity between convolutional layers allows it to learn more complex features than traditional CNNs. B h a t t i et al. [9] employed the DenseNet architecture for extracting deep features from facial images. They evaluated their model's performance in FER, specifically utilizing DenseNet201, with the highest accuracy of 96.8% on the JAFFE dataset, surpassing the second-highest accuracy of 96% accuracy of other methods [15]. Furthermore, the authors in [1] proposed a lightweight model using dense blocks. They assessed the performance of their model on three benchmark datasets: JAFFE, CK+, and OuluCASIA. The experimental results revealed that their model achieved state-of-the-art performance on all datasets, with an average accuracy of 99.08%, 99.90%, and 100%, respectively.

In the integration of attention mechanisms into CNNs, authors have utilized channel or spatial attention models, or combinations of them. In [23-25], researchers introduced end-to-end CNN models equipped with an attention mechanism for FER. In [23], the attention module assigns higher weights to critical features, directing the network's focus toward these crucial elements for expression recognition. On the other hand, [24] proposed a co-attentive multi-task CNN for two tasks: facial landmark detection and facial expression recognition. In [25], the authors used an attention module through a spatial transformer network to focus on important facial regions. These models were evaluated on benchmark datasets, including JAFFE, CK+, OuluCASIA, NCUFE, FER2013, and SFEW2, achieving high accuracies. The best performance reported in [23] is 98.68% on CK+, while [24] achieved a maximum accuracy of 96.71% on CK+, and [25] has the highest accuracy of 98.0% on CK+. The CK+ dataset is collected in a controlled lab environment, making it relatively easier for FER models compared to the others.

In [16], the authors used state-of-the-art models, including VGG-19, GoogLeNet, and ResNet-152, as backbone networks for feature extraction. They employed a fusion attention process, combining both channel attention and spatial attention, to create the model. It analyses the link between different channels and assigns weights to them, generating spatial attention weights in the spatial domain to suppress or raise the relevance of certain regions. On the other hand, in [18], an end-to-end DenseNet model integrated with spatial attention is proposed for FER. This architecture consists of two key components: a densely connected module and a spatial attention module, each with its specific function. The densely connected module generates a deep representation of the entire image, while the spatial attention module identifies and highlights regions of local and global levels. The combination of these two components allows for the extraction of emotion-relevant features, resulting in accurate classification. Extensive experiments conducted on publicly available facial expression databases demonstrate the reliability of this proposed method. It achieves accuracies of 76.95% on RAF-DB and 95.71% on CK+. Additionally, the model in [18] has a relatively small size, comprising approximately 1.2 million parameters.

In this review, the exploration of combining DenseNet architecture with the fusion of channel and spatial attention to harness its substantial potential has not been

undertaken. This issue will be discussed in detail in the next section, addressing a key aspect of our proposed method.

3. Proposed method

In this section, the proposed model is presented, which integrates a fusion of attention mechanisms with an end-to-end DenseNet architecture. This model is referred to as FCSDNet (Fusion of Channel and Spatial Attention in Densely Connected Convolutional Layers Network), and it is tailored for the FER task. The FCSDNet model encompasses two fundamental stages:

(a) Feature extraction from images, with a focus on crucial regions using attention mechanisms.

(b) Classification of the extracted features into their respective labels for FER.

The quality of recognition and the computational complexity of models are often influenced by factors such as the number of filters and the depth of networks. Researchers frequently adjust these factors based on specific application requirements to achieve high recognition accuracy while maintaining acceptable computational complexity. Thus, the FCSDNet model is designed with a moderate number of layers and an appropriate number of filters in each convolutional layer. This design choice ensures compatibility with computational resources and broadens the model's applicability.

3.1. Fusion of channel and spatial attention module

Channel attention modules can learn to focus on the most important channels in a feature map, while spatial attention modules learn to emphasize critical spatial locations within a feature map. Combining these two types of attention modules can lead to improved performance in the FER task. This combination is achieved by establishing a linear collaborative relationship between spatial and channel attention modules, as they are sensitive to different network depths and tasks.

Given an intermediate feature map $F^i \in R^{H\times W\times C}$ as input, where H, W, and C denote spatial height, width, and the number of channels. For the Channel Attention Module (CAM), global information aggregation is performed through average-pooling (P_{ga}) and max-pooling (P_{gm}) operations, yielding descriptors $P_{\mathrm{ga}}(F^i)$ and $P_{\mathrm{gm}}(F^i)$ in the dimension of $1\times 1\times C$. These descriptors are then processed by a scale network ($\mathbb{S}_C = \{\mathbb{S}_C^{1,r},\ \mathbb{S}_C^2\}$, which denotes two fully connection layers, the first layer has a ratio (r) for compacting features) to create a channel attention map optimized for the FER task. This scale network performs the same as a Squeeze-and-Excitation network block [12], utilizing a set of learnable parameters (weights matrix $\mathbb{W}_C$ and biases vector $\mathbb{b}_C$). The output of the scale network is computed as $\mathbb{S}_C(x) = \mathbb{W}_C^2 \cdot \delta(\mathbb{W}_C^{1,r}\cdot x + \mathbb{b}_C^{1,r}) + \mathbb{b}_C^2$, where $\mathbb{W}_C^2,\ \mathbb{b}_C^2 \in \mathbb{S}_C^2$ and $\mathbb{W}_C^{1,r},\ \mathbb{b}_C^{1,r} \in \mathbb{S}_C^{1,r}$ are the network parameters, $x \in \{P_{\mathrm{ga}}(F^i),\ P_{\mathrm{gm}}(F^i)\}$ is the input to the network, δ is the activation function (in this study, it is the rectified linear unit (ReLU) function). The output dimensions of P_{ga}, P_{gm}, and $\mathbb{S}_C$ are $[1\times 1\times C]$. The output feature vectors are combined using element-wise summation $\oplus$ and a sigmoid-gated mechanism σ.

The final channel attention output is obtained by element-wise multiplication with the input feature map. The CAM is formulated as follows:

(1) $$\mathrm{CAM}(F^i) = F^i \otimes \sigma\left(\mathbb{S}_C\left(P_{\mathrm{ga}}(F^i)\right) \oplus \mathbb{S}_C\left(P_{\mathrm{gm}}(F^i)\right)\right).$$

For the Spatial Attention Module (SAM), it is first produced by applying a convolutional operation ($\mathbb{C}_s$) to the concatenation ([;]) of channel-based average-pooling (P_{ca}) and max-pooling (P_{cm}) on the input to obtain a spatial attention map. The concatenation is done by stacking the channels of the input. The output dimensions of P_{ca}, P_{cm}, and $\mathbb{C}_S$ are $[W \times H \times 1]$. Then, the final output is obtained by element-wise multiplication of the input by the spatial attention map. The SAM is formulated as follows:

(2) $$\mathrm{SAM}(F^i) = F^i \otimes \mathbb{C}_S([P_{\mathrm{ca}}(F^i); P_{\mathrm{cm}}(F^i)]).$$

A fusion of channel and spatial attention ($\mathcal{F}_{cs}$) can involve maximum, addition, multiplication, or concatenation [8]. In this study, a normalized weight function for addition on channel and spatial attention is employed, as shown in the next equation. This weight (w^f) is a hyperparameter and can be chosen heuristically based on specific tasks:

(3) $$\mathcal{F}_{CS}(F^i) = w^f.\mathrm{CAM}(F^i) \oplus (1 - w^f).\mathrm{SAM}(F^i).$$

The process of this fusion is illustrated in Fig. 3.

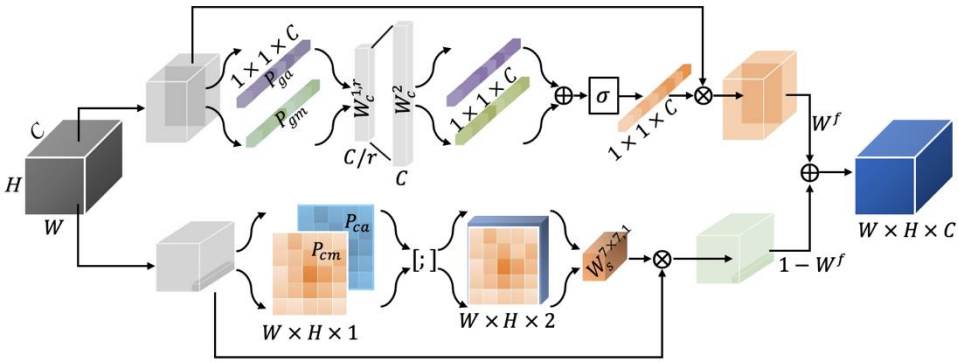


Fig. 3. Fusion of channel and spatial attention in the proposed FCSDNet model

3.2. Designing the FCSDNet model

The FCSDNet model utilizes the architecture of densely connected convolutional layers for the feature extraction stage. It incorporates initial convolutional layers to better preserve the details of low-level features from the original image. For the main body, dense blocks are employed to capture high-level and abstract features. These layers are integrated with a fusion of attention mechanisms to enhance the richness and effectiveness of the features by focusing on crucial regions of images. Therefore, the FCSDNet model has two phases of extracting features: the first involves raw and low-level feature extraction by traditional convolutional layers, and the second refines features, producing increasingly high-level features through densely connected blocks with integrated attention modules.

The proposed model consists of a series of Dense Blocks (DBs), illustrated in Fig. 4. Each DB contains multiple pairs of ConVolutional (CV) layers, each of which includes immediately preceding Batch Normalize (BN) and ReLU activation, referred to as components ($\mathcal{C}$). In each $\mathcal{C}$, the first CV has a kernel size of 1×1 with stride 1, denoted by $\mathrm{CV}^1(1 \times 1{:}\,1)$, and the second one is $\mathrm{CV}^2(3 \times 3{:}\,1)$. Given a desired number of output channels for DBs, denoted as m, the number of output channels of the first and the second CV is set to m and $m/4$, respectively. The processing of the component with input feature maps F' can be formulated as $\mathcal{C}(F') = \mathcal{F}^2(\mathcal{F}^1(F'))$, where $\mathcal{F}^k(x) = \mathrm{CV}^k(\mathrm{ReLU}(\mathrm{BN}(x)))$, $k \in \{1, 2\}$. Since F' is of size $W' \times H' \times C'$, then $\mathcal{C}(F')$ has an output size of $W' \times H' \times m/4$. At the end of each component, a concatenation ([;]) of all previous feature maps is performed, instead of just the feature maps from the immediately preceding layer in traditional CNNs. This means that each convolutional layer has access to all previous feature maps, which allows it to learn more complex features. To simplify the computational complexity of the network, Transition Layers (TLs) are applied after each DB to reduce the number of channels. The TLs consists of a $1 \times 1{:}\,1$ convolution operation with ReLU activation to refine the information in the feature maps, and a 2×2 average pooling with strides of 2, denoted by $P_{\mathrm{a}}(2 \times 2{:}\,2)$, is responsible for reducing the spatial size of the feature maps.

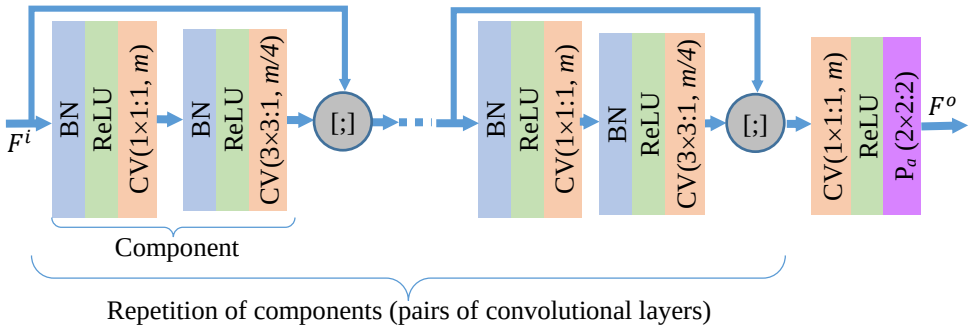


Fig. 4. Dense blocks used in the proposed FCSDNet model

The processing of DBs can be formulated as below:

(4) $F^{\mathrm{o}} = \mathrm{TL}\left([;]\left(\ldots,[;](\mathcal{C}(F'),F')\right)\right)\left(f^{[;]}\left(\cdots;f^{[;]}\left(f^{p_2}\left(f^{p_1}\left(F^i\right)\right);F^i\right);F^i\right)\right)$.

The output feature map of the DBs is denoted by F^{o}. Since the input size of the DBs is $W^i \times H^i \times C^i$, the output size is $\frac{W^i}{2} \times \frac{H^i}{2} \times m$. The frequency notation[;] corresponds to the number of components (pairs of convolutional layers) used in the DBs. $\mathrm{DB}(m, p)$ is also used to denote the dense block with the desired number of output channels as m, and p represents the number of repetition pairs of convolutional layers.

Overall, the proposed FCSDNet model consists of three DBs preceded by an initial convolutional layer, as shown in Fig. 5. Following each DB is the integration of a fusion of channel and spatial attention module, as detailed in the section above. This attention module operates on the output feature maps of the DB, emphasizing

important feature regions for the Facial Expression Recognition (FER) task. These DBs are configured with different numbers of output channels, specifically 32, 64, and 128, and they consist of pairs of convolutional layers with one, two, and four pairs, respectively. This repetition of pairs of convolutional layers with varying filters enables the network to capture hierarchical features. Typically, lower layers focus on extracting low-level features like edges and textures, while deeper layers learn more abstract and complex representations. By concatenating features from all pairs of convolutional layers, the FCSDNet model creates a feature hierarchy that spans from simple to intricate details. This hierarchical representation is highly valuable for the FER task, as it allows the model to recognize facial expressions based on both basic facial features and more complex patterns.

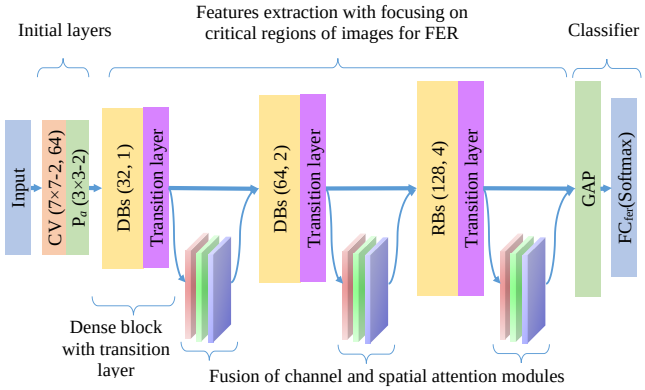


Fig. 5. Overall structure of the proposed FCSDNet model

Table 1. Parameters of the FCSDNet model

Layers /blocks	Operations (kernel size – strides, filters)	#Parameters (thousand)
Input	-	-
Initial layers	C(7×7 – 2, 64) P_a(3×3 – 2)	9.47 -
1st dense block	[C(1×1 – 1, 128) →C(3×3 – 1, 32)]	8.32 36.89
Transition	C(1×1 – 1, 128) →P_a(2×2 – 2)	12.41 -
Attention module	D(8)×2, C(7×7 – 1, 1)	4.33
2nd dense block	[C(1×1 – 1, 256) →C(3×3 – 1, 64)]×2	82.42 295.04
Transition	C(1×1 – 1, 256) →P_a(2×2 – 2)	65.79 -
Attention module	D(8)×2, C(7×7 – 1, 1)	16.77
3rd dense block	[C(1×1 – 1, 512) →C(3×3 – 1, 128)]×4	919.55 2359.80
Transition	C(1×1 – 1, 512) →P_a(2×2 – 2)	393.72 -
Attention module	D(8)×2, C(7×7 – 1, 1)	66.21
Aggregation	Global average pooling	-
Classifier (FC layer)	Softmax	3.59
Total	-	4.27M

The FCSDNet model has a total depth of 18 convolutional layers, divided into three DBs with transition layers and initial layers, which serve as feature extractors.

Additionally, three fusion modules of channel and spatial attention are incorporated into the model to emphasize crucial regions of images by weighting feature maps. Despite having a relatively moderate number of layers, this model contains only about 4.27 million parameters, making it less complex compared to other modern CNN models for image classification problems. This parameter efficiency is achieved by using small kernel sizes for the convolutional operations within the DBs. The detailed parameters of the model with a specific input size in height×width×channels of 100×100×3 are shown in Table 1, where the symbol "×" indicates repetition in DBs; *C* represents a convolutional layer with specified kernel size, strides, and the number of filters; P_{a} indicates an average pooling layer with a specified window size and strides; and "→" signifies the forward connection between two layers.

3.3. Model classifying and training loss

The feature maps extracted by the DBs, which incorporate the fusion of channel and spatial attention modules, are further processed through a Global Average Pooling (GAP) layer, as in Fig. 5. This layer reduces the spatial dimensions of the feature maps while preserving crucial information. It operates by calculating the average value of each feature map across all spatial locations. The results are then consolidated into a vector, where each element corresponds to the average activation of a specific feature map. This vector (denoted by f^*) serves as a compact representation of the essential features derived from the input image.

Following the GAP, a softmax function is applied in conjunction with a fully connected layer for classification. This step generates a probability distribution across different expression categories. The classification process can be summarized as follows:

(5) $$y^* = \mathrm{argmax}_{k=1,\ldots,N^{\mathrm{c}}}\{\hat{y}_k\}, \quad \hat{y} = \mathrm{softmax}(W_{\mathrm{o}} \times f^* + b_{\mathrm{o}}),$$

where y^* is the output class of the model classification; $\hat{y}_k \in \hat{y}$ indicates the probability distribution of the prediction of *k*-th category (or class), $k = 1, \ldots, N^{\mathrm{c}}$; N^{c} is the number of categories; W_{o} and b_{o} are the weight matrix and bias of the final fully connected layer (or output layer), respectively; softmax(.) stands for the normalized exponential function.

Cross-entropy loss is a commonly used loss function for multi-class classification tasks, such as FER. It is a nonlinear function that measures the difference between two probability distributions. In the context of FER, cross-entropy loss measures the difference between the predicted probability distribution of facial expressions and the actual probability distribution of facial expressions.

During training, the FCSDNet model is optimized to minimize cross-entropy loss. This means that the model learns to predict the correct facial expression with the highest probability possible. The employed loss function is defined as follows:

(6) $$\mathcal{L}(y, \hat{y}) = -\frac{1}{N^{\mathrm{s}}} \sum_{i=1}^{N^{\mathrm{s}}} y^{(i)} \log \hat{y}^{(i)},$$

where N^{s} is the number of samples in a dataset; $y^{(i)}$ and $\hat{y}^{(i)}$ represent the ground-truth distribution and the predicted distribution of the *i*-th sample.

The non-linear, exponential behavior of cross-entropy loss helps the model to converge to optimal performance. This is because cross-entropy loss penalizes the

model more heavily for incorrect predictions than for correct predictions. This helps to ensure that the model learns to predict the correct facial expression even when the input image is noisy or ambiguous.

4. Experimental results

In this section, the datasets and running parameters to be employed for training the FCSDNet model will be initially described. Subsequently, the results and discussion will be presented to assess the performance of the model.

4.1. Dataset and experimental setup

For experimental running, five datasets were used, namely, JAFFE [5], CK+ (Extended Cohn-Kanade) [26], OuluCASIA [27], KDEF [28], and RAF_DB [29].

The JAFFE dataset consists of 213 images captured from 10 Japanese women, showcasing six basic emotions along with a "neutral" emotion. On the other hand, the CK+ dataset comprises 981 images collected from 118 individuals, each displaying six basic emotions in addition to the "contempt" emotion. The OuluCASIA dataset features 1440 images from 80 individuals, offering a range of six basic facial expressions under varying illumination and head poses, all in color. Both the JAFFE and CK+ datasets are presented in grayscale. The KDEF dataset contains 4900 images capturing six basic emotions and the "neutral" expression, all from five different angles. It includes images of 70 individuals (35 females and 35 males) aged between 20 and 30 years. Notably, all images were taken without any occlusions such as mustaches, earrings, or eyeglasses.

Lastly, the RAF_DB (Real-world Affective Faces database) is a unique facial expression dataset, collected from the internet, with a total of 29,672 real-world facial images. This dataset includes six basic emotions along with the "neutral" expression, and it is divided into a training set, consisting of 12,271 face-aligned images, and a testing set, which contains 3068 images. The total images and distributed images in classes of these datasets are also shown in Table 2.

Table 2. Distributed images in classes of datasets

Dataset	Number of classes	Total images	Distributed images in classes		
			Min	Max	Average
JAFFE	7	213	29	32	30.4
CK+	7	981	54	249	140.1
OuluCASIA	6	1440	240	240	240.0
KDEF	7	4900	700	700	700.0
RAF_DB	7	29,672	355	5957	2191.3

This experiment utilized a 5-fold cross-validation scenario for the first four datasets (JAFFE, CK+, OuluCASIA, and KDEF). Therefore, the images of each dataset are randomly divided into 5 equal-sized folds for five runs. In each run, one fold is selected for testing (D^{te}), and the remaining folds are used for training the model, with 80% used for training set (D^{tr}) and 20% for model evaluation (D^{va}) to

select the best one. The final results were reported as the mean and standard deviation of the five runs. In the case of the RAF_DB dataset, which involves separated training images and testing images (D^{te}), 80% of the training images are used for training set (D^{tr}), and the remaining 20% are allocated for validation set (D^{va}) to select the best model. Therefore, only one run was performed for this dataset due to a fixed testing set.

To enhance the performance of the model and prevent overfitting, the training data is augmented by using 2D image transform operations as in [30] like rotation, scaling, noise addition, and translation. This increases the diversity of the training data, improving the model's ability to generalize and perform well under different conditions.

The FCSDNet model was trained from scratch using Adam as the optimizer with an initial learning rate (lr) of 10^{-3}. The training process is terminated when it reaches the maximum epochs or the model's performance on the validation set has not improved for a certain number of epochs. Then, the best model is chosen to get the results on the test set. Augmentation parameters, as outlined in Table 3, are randomly chosen within specified ranges. The extent of augmentation for each image is contingent upon the dataset's volume; it is set at three for datasets containing over 20,000 images, five for datasets with images numbering between 20,000 and 5000, and ten for datasets with fewer than 5000 images.

Table 3. Parameters for augmentation and training model

No	Parameters	Range value
	***A*. Data augmentation**	
1	Variance of Gaussian noise addition	[0, 0.05]
2	Rotation relative to the original image (degree, negative is counter-clockwise)	$[-18^o, 18^o]$
3	Translation relative to the original image (percentage, negative is left translation for width or up translation for height)	$[-10\%, 10\%]$
4	Scaling relative to the size of the image (negative is downscaling, for both width and height)	$[-10\%, 10\%]$
5	Horizontal flipping image	Yes/No
	***B*. Training model**	
6	Initial learning rate (lr)	10^{-3}
7	Batch size	128
8	Max epochs	150

The experiments were run on a computer system equipped with TPU and 32 Gb RAM. The FCSDNet model is developed by using the Python programming language on the TensorFlow platform, which is a widely used deep learning framework known for its powerful features in image processing and CNN modeling. (**https://github.com/duongthanglong/duongthanglong/blob/main/y23attentiondensenet4fer.py**)

4.2. Results and discussion

The training results of the FCSDNet model are an average from the number of runs, depicted in Fig. 6, for four datasets, excluding CK+ due to its small size with well-distinguished images, and ease in achieving high accuracy during training. Each subfigure exhibits the accuracies of the training data (blue line) and validation data (yellow line) for a specific dataset. Within the JAFFE dataset, where the number of images is limited, and facial expressions across classes are notably similar, fluctuations in validation accuracy during training are observed. Conversely, for the other datasets, improvements in validation accuracy align with those in training accuracy. Notably, the model rapidly improved its accuracy within the first 20 epochs, reaching its peak performance around the midway point of the total epochs. This trend persisted, indicating that the model continually learned and enhanced its performance.

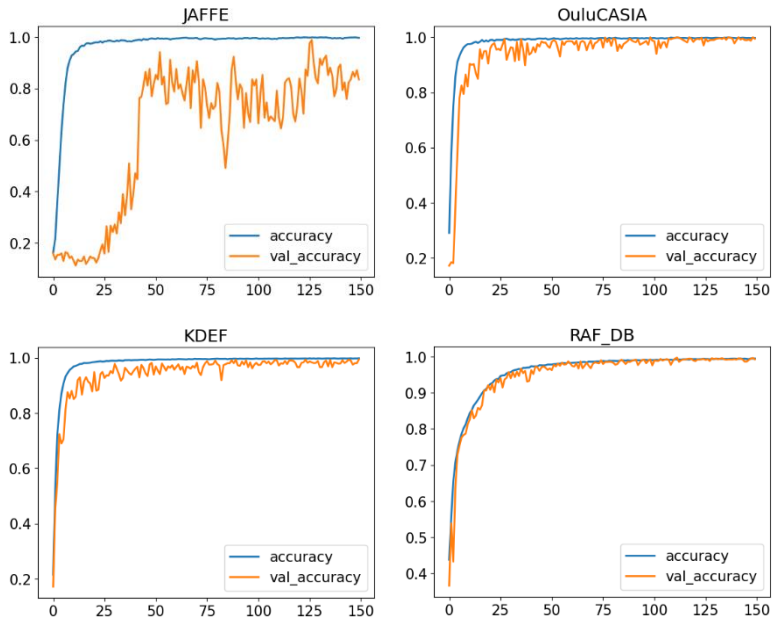


Fig. 6. The training progress with accuracies of training and validation data

The testing data accuracies for the FCSDNet model are presented in detail in Table 4. In the table, the symbol "-" indicates that no running, and the appended "R#" indicates the corresponding experiment run. The last column represents the average accuracy across all runs for each dataset. The first three datasets, including JAFFE, CK+, and OuluCASIA, achieved the highest testing accuracy of 100%. The model also performed exceptionally well on the KDEF dataset, with an accuracy of 99.94%. It has approximately the highest accuracy. These datasets achieved such high accuracies because they are based on controlled laboratory conditions, resulting in consistent image capture conditions. In contrast, the RAF-DB dataset consists of real-world data captured under diverse and uncontrolled conditions, leading to the lowest

testing accuracies (83.18%) compared to the other datasets. Only one run was conducted for this dataset since it has two separate sets for training and testing data.

Table 4. Accuracies of experiment running in details

Dataset	R#1	R#2	R#3	R#4	R#5	Average
JAFFE	100	100	100	100	100	100
CK+	100	100	100	100	100	100
Oulu-CASIA	100	100	100	100	100	100
KDEF	99.80	100	100	100	99.90	99.94
RAF-DB	83.18	-	-	-	-	83.18

Fig. 7 presents some misclassified images from the testing data. The first row displays three misclassified images from the KDEF dataset, while the subsequent rows show such images from the RAF-DB dataset. Each image's title indicates the target emotion and the predicted emotion, separated by the symbol ">". These misclassified images also prove to be challenging for intuitive recognition.

For details of Fig. 7. In the first row, the images have faces oriented towards the left or right, where only one side's features are extracted, making it difficult for FER. On the other hand, the images in the RAF-DB dataset exhibit various challenging conditions. For example, in the second row, the first two images have distorted faces, making it nearly impossible to discern facial expressions. The last image is marred by white streaks and significant noise, further complicating the recognition of facial expressions. In the last row, the first image is partially obscured by the lower part of the face, and the middle image is blurred, presenting additional challenges for observation and identification. The last image features are relatively clear facial characteristics, but the determination of facial expressions remains problematic due to their unclear visibility. These examples underscore the greater difficulty of the FER task when dealing with real-world, uncontrolled images compared to lab-controlled ones.

In the FCSDNet model, the operations of the DBs with a fusion of channel and spatial attention play a crucial role in feature extraction for FER. To visually represent the DBs' operations, gradient-based localization is employed. This method highlights the areas in the images where the model concentrates or shows interest when extracting features, often referred to as a heat map of activated neurons on the image. In Fig. 8, heat maps of the model are presented for various facial expressions in some images of the KDEF (a) and RAF-DB (b) datasets. These images clearly illustrate that the heat maps predominantly focus on areas crucial for representing facial expressions, such as the nose, mouth, and eyes. This intuitive demonstration emphasizes that the FCSDNet model prioritizes important image regions for extracting descriptive features for FER. Conversely, when these image areas are not considered, it becomes challenging to correctly identify the intended facial expression.

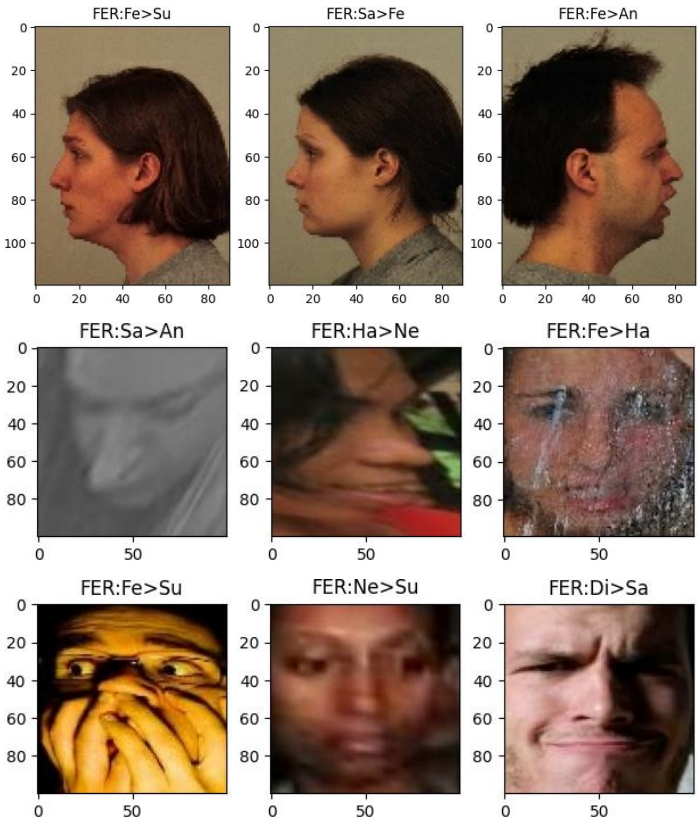


Fig. 7. Misclassified images on testing data

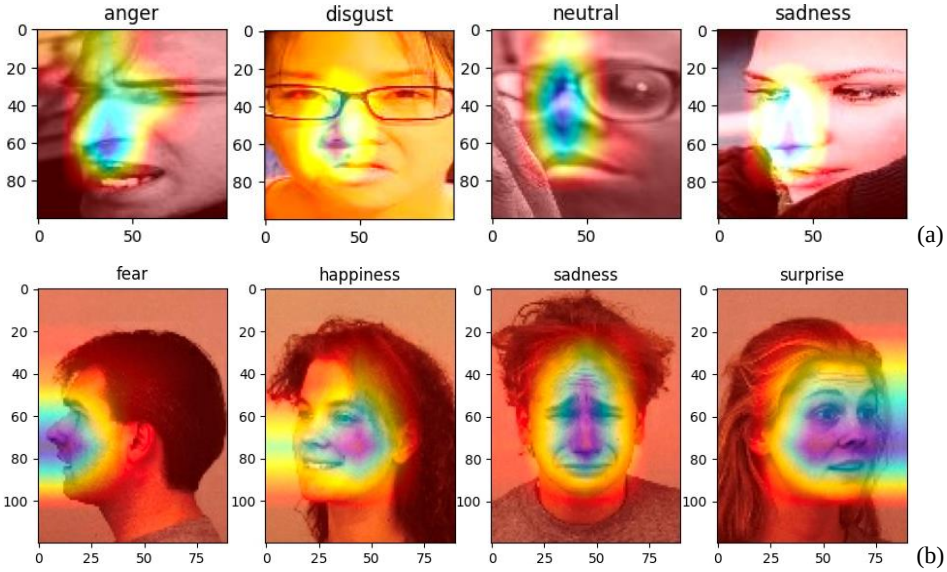
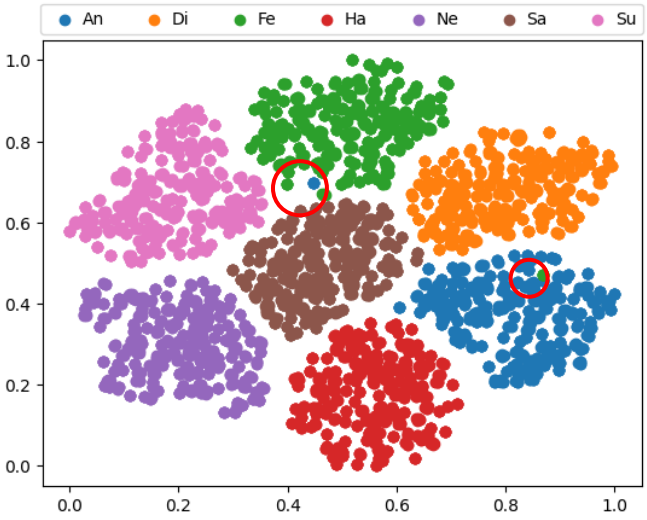
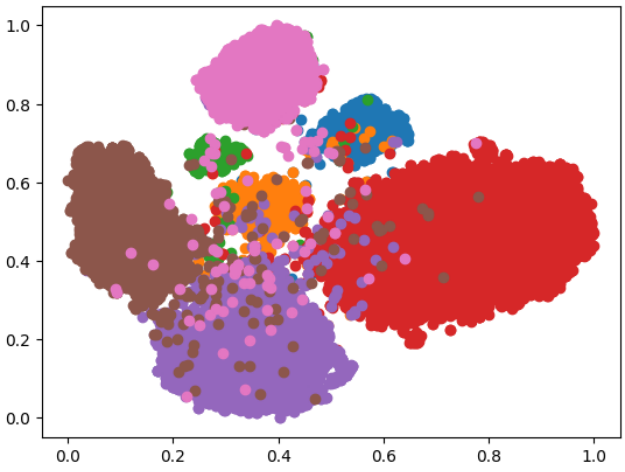


Fig. 8. Heatmap of the FCSDNet model for FER on some images

t-SNE is employed, following the approach in [24], for a feature visualization study on the FCSDNet model's learned features. The results are presented in Fig. 9, with the findings for the KDEF (a) dataset shown in the upper part and those for RAF-DB (b) in the lower part of Fig. 9. The labels of facial expressions are shortened to the first two letters to fit the figure. This visualization reveals that the proposed model forms compact feature clusters, with notably few outliers, as highlighted by the red circles, particularly for the KDEF dataset. This observation provides evidence that the FCSDNet model is efficient in learning highly distinctive features. It serves to reduce intra-class feature variations while enhancing inter-class feature distinctions.



(a)

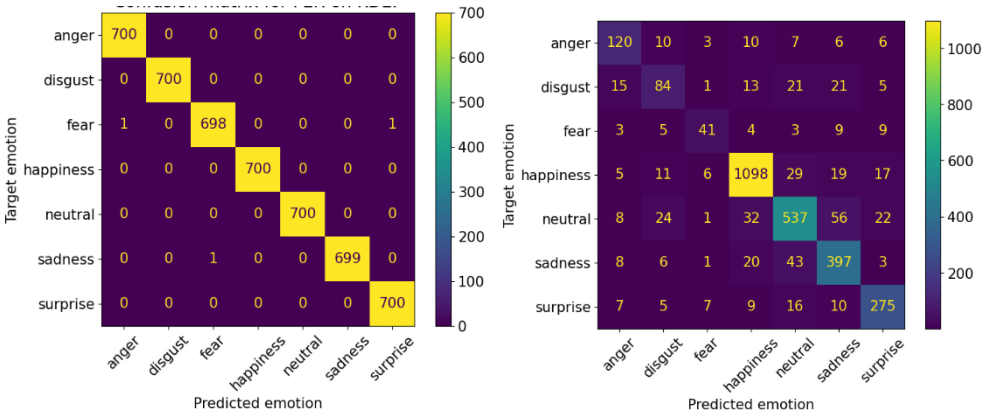


(b)

Fig. 9. t-SNE visualization of the FCSDNet model on KDEF (a) and RAF-DB (b)

To illustrate overall of the FCSDNet model's performance, we have constructed confusion matrices based on results from all runs using two datasets: KDEF and RAF-DB, as shown in Fig. 10. Each row in these matrices represents a target emotion label in the dataset, while each column corresponds to a predicted facial expression label generated by the model.

In a typical run of the trained model, we apply it to recognize all images within the testing dataset, aggregating the results across all runs. Looking at the confusion matrix in subfigure (a) for the KDEF dataset, it's evident that the model achieved perfect classification on the emotion label of "happiness", with no misclassifications for both predicted and being predicted by the model. This suggests a clear distinction between these facial expressions from others. Similarly, the "anger" and "neutral" emotion labels were all correctly predicted. However, each of these labels had one case of misclassification, being predicted from "fear" and "disgust", respectively. In subfigure (b) for the RAF-DB dataset, the highest number of misclassifications is 56, where "neutral" is predicted to be "sadness". Every cell in this matrix has a nonzero value, and three cases involved only one misclassification, where "disgust", "neutral", or "sadness" was predicted to be "fear". For instance, the first image in the middle row of Fig. 7 is one of eight misclassified images from the cell in the second-to-last row and the first column in Fig. 8. Once again, this matrix highlights the challenges associated with FER when dealing with real world, life-wild images.



(a) (b)

Fig. 10. Confusion matrices on KDEF (a) and RAF-DB (b) datasets

Table 5 provides a comparison of the experimental results of the FCSDNet model with those of other studies. All methods in the comparison used CNN-based models and conducted experiments in various data scenarios, as indicated in parentheses next to the method name along with the model size in the number of parameters ("M" representing million, "F" stands for fold). The symbol "*" presents the method that has utilized pre-trained models on large-scale datasets, and the symbol "-" indicates cases where no experimental data were reported. The best accuracies are highlighted in bold. The FCSDNet model achieved a perfect accuracy of 100% in three cases: JAFFE, CK+, and OuluCASIA. Another perfect accuracy

of 100% was also achieved in the OuluCASIA dataset [1]. Despite its relatively low complexity, with the fourth-lowest number of model parameters, the FCSDNet model outperformed the other models in all cases for the JAFFE, CK+, OuluCASIA, and KDEF datasets. For the KDEF dataset, the proposed model achieved the highest accuracy of 99.94%, surpassing [17] by 0.04% and outperforming [31] by 12.17%. The proposed model achieved the fourth-highest accuracy in five cases for the RAF-DB dataset. However, it's worth noting that the three models with better performance above used pre-trained models on very large-scale datasets for conducting their experiments, which makes these cases not fair for comparison.

Table 5. Comparison of accuracies on testing data

Method (model-size, num-fold)	JAFFE	CK+	Oulu CASIA	KDEF	RAF-DB
Wu et al. [3]* (-, -)	92.90	99.75	-	-	**90.06**
Devaram and Cesta [11] (1.6M, 5F)	80.09	84.27	-	99.90	-
Yu and Xu [18]* (11M, 10F)	-	98.33	87.32	-	85.22
Kollias, Sharmanska and Zafeiriou [27]* (-, -)	-	-	-	-	78.00
Farzaneh and Qi [28]* (11M, -)	-	-	-	-	87.78
Zhou, Liang and Shi [26] (0.06M, -)	-	-	-	87.71	-
Ming et al. [29] (39M, 10F)	-	99.50	89.60	-	-
Long, Tung and Dung [1] (2.4M, 5F)	99.08	99.90	**100**	-	-
Long [2] (23.5M, 5F)	96.20	99.68	98.47	-	-
Proposed FCSDNet (4.27M, 5F)	**100**	**100**	**100**	**99.94**	83.18

5. Conclusion

In this paper, a novel CNN-based model is introduced for FER known as FCSDNet. This model leverages state-of-the-art architectures, specifically dense connected networks and a fusion of channel and spatial attention, to improve performance while maintaining a moderate model complexity. FCSDNet consists of three dense blocks with attention mechanisms to emphasize features crucial for FER. It comprises a total of 18 convolutional layers designed for feature extraction. Despite its medium number of layers, FCSDNet has a relatively low parameter count, approximately 4.27 million parameters, making it less complex compared to contemporary CNN models employed in image classification tasks. This reduction in parameters is achieved by using compact kernel sizes in the convolutional operations within the dense blocks.

In the experiments, five popular datasets are employed, following a 5-fold cross-validation scenario except for the RAF-DB dataset, which is organized differently. FCSDNet exhibits impressive accuracy, achieving results ranging from 99.94% to 100% in FER across four lab-controlled datasets. However, it demonstrates a lower accuracy of 83.18% on the real-world RAF-DB dataset, although it is higher when compared fairly to other methods. It's important to note that these results are obtained under certain computing limitations, resulting in relatively short training times and the use of moderately sized datasets. When the model is trained at a deeper level with larger datasets, it can anticipate even higher performance.

Future work should improve accuracy and efficiency, especially for real-world datasets characterized by diverse and uncontrolled conditions. One potential avenue for enhancement is the incorporation of well-established network architectures.

Additionally, broadening the model's scope to encompass tasks like face identification and head-pose estimation holds promise for further research and development.

Acknowledgments: The authors would like to express gratitude to colleagues for their valuable assistance in this research. This work also received partial support through a grant from Hanoi Open University, Vietnam.

References

1. Long, D. T., T. T. Tung, T. T. Dung. A Facial Expression Recognition Model Using Lightweight Dense-Connectivity Neural Networks for Monitoring Online Learning Activities. – International Journal of Modern Education and Computer Science, Vol. **6**, 2022, pp. 53-64.
2. Long, D. T. A Facial Expressions Recognition Method Using Residual Network Architecture for Online Learning Evaluation. – Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. **25**, 2021, No 6, pp. 953-962. DOI: https://doi.org/10.20965/jaciii.2021.p0953.
3. Wu, X., J. He, Q. Huang, C. Huang, J. Zhu, X. Huang, H. Fujita. FER-CHC: Facial Expression Recognition with Cross-Hierarchy Contrast. – Applied Soft Computing, Vol. **145**, 2023, pp. 1-12. DOI: https://doi.org/10.1016/j.asoc.2023.110530.
4. Huang, G., Z. Liu, L. V. D. Maaten, K. Q. Weinberger. Densely Connected Convolutional Networks. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), 2017, pp. 2261-2269. DOI: https://doi.org/10.1109/CVPR.2017.243.
5. Guo, M., T. Xu, J. Liu, Z. Liu, P. Jiang, T. Mu, S. Zhang, R. Martin, M. Cheng, S. Hu. Attention Mechanisms in Computer Vision: A Survey. – Computational Visual Media, Vol. **8**, 2022, No 3, pp. 331-368. DOI: https://doi.org/10.1007/s41095-022-0271-y.
6. Alom, M., T. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. Nasrin, M. Hasan, B. Essen, A. Awwal, V. Asari. A State-of-the-Art Survey on Deep Learning Theory and Architectures. – Electronics, Vol. **8**, 2019, No 292, pp. 1-67.
7. Deng, W., S. Li. Deep Facial Expression Recognition: A Survey. – IEEE Transactions on Affective Computing, Vol. **13**, 2022, pp. 1195-1215.
8. Nan, Y., J. Ju, Q. Hua, H. Zhang, B. Wang. A-MobileNet: An Approach of Facial Expression Recognition. – Alexandria Engineering Journal, Vol. **61**, 2022, pp. 4435-4444. DOI: https://doi.org/10.1016/j.aej.2021.09.066.
9. Bhatti, Y., A. Jamil, N. Nida, M. Yousaf, S. Viriri, S. Velastin. Facial Expression Recognition of Instructor Using Deep Features and Extreme Learning Machine. – In: Computational Intelligence and Neuroscience. Vol. **2021**. 2021, pp. 1-17. DOI: https://doi.org/10.1155/2021/5570870.
10. Cao, Y. An Expression Recognition Model Based on Channel and Spatial Attention Fusion. – Journal of Physics: Conference Series, 2022, pp. 1-6. DOI: 10.1088/1742-6596/2363/1/012016.
11. Devaram, R. R., A. Cesta. LEMON: A Lightweight Facial Emotion Recognition System for Assistive Robotics Based on Dilated Residual Convolutional Neural Networks. – Sensors, Vol. **22**, 2022, No 3366, pp. 1-20.
12. Gan, C., J. Xiao, Z. Wang, Z. Zhang, Q. Zhu. Facial Expression Recognition Using Densely Connected Convolutional Neural Network and Hierarchical Spatial Attention. – Image and Vision Computing, Vol. **117**, 2022, No 104342, pp. 1-9. DOI: https://doi.org/10.1016/j.imavis.2021.104342.
13. Lai, S., C. Chen, J. Li. Efficient Recognition of Facial Expression with Lightweight Octave Convolutional Neural Network. – Journal of Imaging Science and Technology, Vol. **66**, 2022, No 4, pp. 040402-1-040402-9.
14. Zhu, Q., Q. Mao, H. Jia, O. Noi, J. Tu. Convolutional Relation Network for Facial Expression Recognition in the Wild with Few-Shot Learning. – Expert Systems with Applications, Vol. **189**, 2022, No 116046, pp. 1-9. DOI: https://doi.org/10.1016/j.eswa.2021.116046.

15. C h e n, X., X. Z h e n g, K. S u n, W. L i u, Y. Z h a n g. Self-Supervised Vision Transformer-Based Few-Shot Learning for Facial Expression Recognition. – Information Sciences, Vol. **634**, 2023, pp. 206-226. DOI: https://doi.org/10.1016/j.ins.2023.03.105.
16. X i a o, J., C. G a n, Q. Z h u, Y. Z h u, G. L i u. CFNet: Facial Expression Recognition via Constraint Fusion under Multi-Task Joint Learning Network. – Applied Soft Computing, Vol. **141**, 2023, No 110312, pp. 1-12. DOI: https://doi.org/10.1016/j.asoc.2023.110312.
17. L i, J., K. J i n, D. Z h o u, N. K u b o t a, Z. J u. Attention Mechanism-Based CNN for Facial Expression Recognition. – Neurocomputing, Vol. **411**, 2020, pp. 340-350. DOI: https://doi.org/10.1016/j.neucom.2020.06.014.
18. Y u, W,. H. X u. Co-Attentive Multi-Task Convolutional Neural Network for Facial Expression Recognition. – Pattern Recognition, Vol. **123**, 2022, No 108401, pp. 1-11. DOI: https://doi.org/10.1016/j.patcog.2021.108401.
19. M i n a e e, S., M. M i n a e i, A. A b d o l r a s h i d i. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. – Sensors, Vol. **21**, No 3046, 2021. DOI: https:// doi.org/10.3390/s21093046.
20. L y o n s, M., S. A k a m a t s u, M. K a m a c h i, J. G y o b a. Coding Facial Expressions with Gabor Wavelets. – In: Proc of 3rd IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200-205. DOI: https://doi.org/10.48550/arXiv.2009.05938.
21. L u c e y, P., J. F. C o h n, T. K a n a d e, J. S a r a g i h, Z. A m b a d a r. The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. – In: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops, 2010, pp. 94-101. DOI: https://doi.org/10.1109/CVPRW.2010.5543262.
22. Z h a o, G., X. H u a n g, M. T a i n i, S. Z. L i, M. P i e t i k ä i n e n. Facial Expression Recognition from Near-Infrared Videos. – Image and Vision Computing, Vol. **29**, 2011, pp. 607-619.
23. E l l e n, G., D. R. R u d i, L. L e m k e, V. B r u n o. The Karolinska Directed Emotional Faces: A Validation Study. – Cognition & Emotion, Vol. **22**, 2008, No 6, pp. 1094-1118.
24. L i, S., W. D e n g, J. D u. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. – In: Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), 2017, pp. 2584-2593.
25. L o n g, D. T. Efficient Multi-Task CNN for Face and Facial Expression Recognition Using Residual and Dense Architectures for Application in Monitoring Online Learning. – International Journal of Fuzzy Logic and Intelligent Systems, Vol. **23**, 2023, No 3, pp. 229-243. DOI: http://doi.org/10.5391/IJFIS.2023.23.3.229.
26. Z h o u, N., R. L i a n g, W. S h i. A Lightweight Convolutional Neural Network for Real-Time Facial Expression Detection. – IEEE Access, Vol. **9**, 2021, pp. 5573-5584. DOI: 10.1109/ACCESS.2020.3046715.
27. K o l l i a s, D., V. S h a r m a n s k a, S. Z a f e i r i o u. Distribution Matching for Heterogeneous Multi-Task Learning: A Large-Scale Face Study. – IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. DOI: https://doi.org/10.48550/arXiv.2105.03790.
28. F a r z a n e h, A. H., X. Q i. Facial Expression Recognition in the Wild via Deep Attentive Center Loss. – In: Proc of IEEE Winter Conference on Applications of Computer Vision (WACV'21), 2021, pp. 2401-2410. DOI: 10.1109/WACV48630.2021.00.
29. M i n g, Z., J. X i a, M. L u q m a n, J.-C. Bu r i e, K. Z h a o. Dynamic Multi-Task Learning for Face Recognition with Facial Expression. – In: Proc. of Lightweight Face Recognition Challenge Workshop during the 2019 International Conference on Computer Vision (ICCV'19), 2019. DOI: https://doi.org/10.48550/arXiv.1911.03281.

Received: 06.11.2023; Second Version: 27.12.2023; Accepted: 22.01.2024