EEG-based Emotion Recognition via Transformer Neural Architecture Search

Chang Li, Zhongzhen Zhang, Xiaodong Zhang, Guoning Huang, Yu Liu, and Xun Chen

***Abstract*—Emotion recognition based on Electroencephalogram (EEG) plays an increasingly important role in the field of brain-computer interfaces (BCI). Recently, deep learning (DL) has been widely applied to EEG decoding owning to its excellent capabilities in automatic feature extraction. Transformer holds great superiority in processing time series signals due to its long-term dependencies extraction ability. However, most existing Transformer architectures are designed manually by human experts, which is a time-consuming and resource-intensive process. In this paper, we propose an automatic Transformer neural architectures search (TNAS) framework based on multi-objective evolution algorithm (MOEA) for EEG-based emotion recognition. The proposed TNAS conducts MOEA strategy that considers both accuracy and model size to discover the optimal model from well-trained supernet for the emotion recognition. We conducted extensive experiments to evaluate the performance of the proposed TNAS on the DEAP and DREAMER datasets. The experimental results showed that the proposed TNAS outperforms the state-of-the-art methods.**

***Index Terms*—Electroencephalogram (EEG), deep learning (DL), emotion recognition, Transformer neural architecture search (TNAS).**

I. Introduction

EMOTION is a comprehensive manifestation of human psychological and physical state, which largely affects our consciousness, behavior, and interpersonal communication. With the rapid development of interactive technologies, affective computing (AC) has shown great potential in the field of human-computer interaction (HCI) [1]–[4]. Whether the system can efficiently understand the emotion states of users largely affects the interactive experience and interaction ability. Besides, emotion recognition has also been applied to the healthcare and has become an emerging method of machine-aided diagnosis for emotional disorders, (*e.g.*, depression, autism, and attention deficit hyperactivity) [5]. Therefore, it is significant to explore efficient emotion recognition methods to meet the demand of practical applications.

Emotion recognition methods could be roughly divided into non-physiological signals based and physiological signals based methods. The non-physiological signals mainly include facial expression, speech, body gesture, and eye blinking, *etc.* The physiological signals contain electrocardiogram (ECG), electrooculogram (EOG), electromyogram (EMG) and electroencephalogram (EEG), *etc.* However, non-physiological signals are easily affected by subjective consciousness, resulting in inconsistent external performance and internal emotional state. In addition, it is tough to obtain the emotional state of some disabled persons with facial or limb problems through external behaviors [6]. On the contrary, physiological signals are intrinsic manifestations independent of subjective control and can reflect the actual emotional state. It is worth noting that EEG is the most preferred source among the physiological signals due to its high time resolution [7]. Many psychophysiological studies have pointed out that emotion is related to the cerebral cortex which clustered electrical activities of the neural cells [8]. In addition, compared with invasive measurement methods (the sensor must be placed under the scalp, *e.g.*, electrocorticography (ECoG) and depth electrode), EEG-based devices are more wearable and portable, which make real-time emotion recognition (*e.g.*, sleep scoring, fatigue driving, and workload calculation) possible. Therefore, EEG-based emotion recognition has been widely utilized in the brain-computer interface (BCI) and has attracted increasing attention in many interdisciplinary fields, ranging from neuroscience to computer engineering [9].

To better describe the emotional states, dimensional model is adopted in emotion recognition. Dimensional theory believes that emotion is a highly correlated continuous variable. The emotional states are described as the coordinate points in space using several basic dimensions with continuous values (such as arousal, valence, and dominance), each of which is a measure of a certain aspect of emotion. Russell [10] proposed the bipolar dimensional model including valence and arousal dimensions as is illustrated in Fig. 1. The valence dimension ranges from negative to positive to indicate the pleasantness of the emotion. The arousal dimension represents the excitement of the emotion ranging from low to high. Similarly, the three-

This work is supported by the National Key Research and Development Program of China (Grant 2019YFA0706203), the National Natural Science Foundation of China (Grants 61922075, 41901350, 62176081 and 32150017), the Chongqing Municipal Health Commission Medical Research Project (Grant 2022WSJK094), the USTC Research Funds of the Double First-Class Initiative (Grant KY2100000123), the Provincial Natural Science Foundation of Anhui (Grant 2008085QF285), and the Fundamental Research Funds for the Central Universities (Grant JZ2021HGTB0078). (Corresponding author: Xun Chen.)

C. Li, Z. Zhang, and Y. Liu are with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China, and are also with Anhui Province Key Laboratory of Measuring Theory and Precision Instrument, School of Instrument Science and Optoelectronics Engineering, Hefei University of Technology, Hefei 230009, Anhui, China. E-mail: changli@hfut.edu.cn, zzzhang@mail.hfut.edu.cn, yuliu@hfut.edu.cn.

X. Zhang is with the Chongqing Key Laboratory of Human Embryo Engineering, Chongqing 400010, China. E-mail: zhangxd207@163.com.

G. Huang is with the Reproductive and Genetic Institute, Chongqing Health Center for Women and Children, Chongqing 400010, China. E-mail: gnhuang217@sina.com.

X. Chen is is with the Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, 230001, China and USTC IAT-Huami Joint Laboratory for Brain-Machine Intelligence, Institute of Advanced Technology, University of Science and Technology of China, Hefei 230088, China. E-mail: xunchen@ustc.edu.cn.

dimensional model further explains the emotional state by adding the dominance dimension, which refers to the degree of subjective control [11]. In this paper, we adopt the dimensional model for emotion recognition.

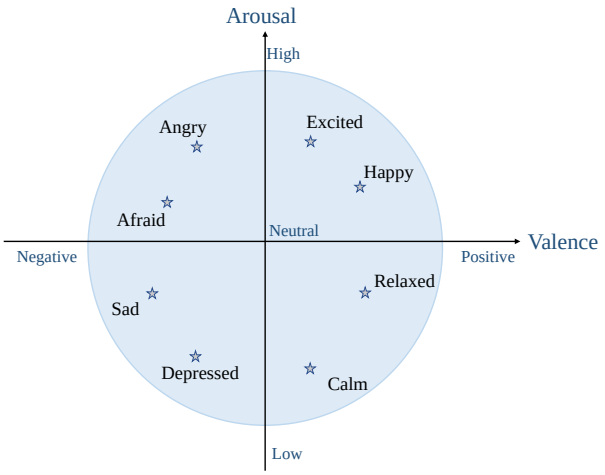

Fig. 1: Valence-Arousal dimensional model.

The rest of this paper is organized as follows: Section II briefly presents the related work in emotion recognition. Section III introduces the proposed method on the emotion recognition tasks. Section IV reports the experimental results evaluated on two commonly used emotion datasets, *e.g.*, DEAP and DREAMER datasets. Finally, the conclusions are carried out in Section V.

II. Related Work

In this section, we comprehensively present the related research on EEG-based emotion recognition and introduce the development of NAS and Transformer.

A. EEG-based emotion recognition

In recent years, EEG-based emotion recognition has attracted widespread attention due to its excellent characteristics, *e.g.*, portability, simplicity, and high time resolution. Traditionally, researchers usually first extract features manually from raw EEG signal and then put them into specific classifiers for emotion recognition. For example, Li *et al.* selected the optimal band to extract frequency features from EEG signal. Common spatial patterns (CSP) and linear support vector machine (SVM) were utilized to classify the emotions [12]. Jie *et al.* chose the notable EEG channels to calculate the sample entropy and adopted SVM for classification [13]. However, manual feature design and extraction largely depend on specific expertise, which is a time-consuming and complicated process. Therefore, efficient feature extraction methods for emotion recognition have attracted the hot attention of many researchers.

Deep learning (DL) technique has achieved outstanding performance in the field of computer vision (CV), speech recognition, and natural language processing (NLP) due to its automatic feature extraction and feature selection abilities. Therefore, many researchers have applied such an efficient algorithm to the EEG decoding to capture discriminative high-level features and underlying dependencies. In addition, the end-to-end DL framework combines both feature extraction and classification without complicated preprocessing, which has demonstrated its advantages over traditional methods. Cimtay *et al.* exploited pretrained state-of-the-art CNN from raw EEG signal, which removed manual feature extraction to overcome the risk of eliminating hidden features of the raw signal [14]. Cui *et al.* proposed an end-to-end regional-asymmetric CNN to capture regional information among physically adjacent channels and the discriminative information between the left and right hemispheres of the brain [15]. Alhagry *et al.* adopted long-short term memory (LSTM) to learn features from EEG signal, and utilized dense layer for classification [16]. Although the above EEG emotion recognition methods based on CNN or RNN have achieved promising results, there still remain a couple of limitations. For example, long-range correlation is an important feature of time series such as EEG, however, CNNs usually adopt the small size kernels to extract deeper features at the cost of limiting the receptive field. Such operations may lose part of time series information and it is hard to explore global internal features from long-distance signals [17]. On the other hand, the sequence models, *e.g.*, RNN and LSTM, sequentially process the EEG signal but cannot be parallelized, which increases the convergence time [18]. Compared with the above-mentioned methods, the self-attention based Transformer has shown its superiority in processing long-distance time series signal [19]. Some researchers have used Transformer for EEG decoding. Liu *et al.* directly utilized Transformer to capture spatiospectral features for EEG-based emotion recognition [20]. Wang *et al.* proposed a squeezed joint-dimension-aware Transformer to focus on the space, frequency, and time information and has proved its effectiveness [21].

B. NAS

DL has made major breakthroughs and progress in many challenging fields such as machine translation, image classification, speech recognition, and emotion recognition. However, such progress largely relies on the design of the network architecture, which is crucial to its final performance. To develop more superior and complex network architectures that can solve challenging tasks, many researchers have spared no effort to try new solutions in this field. As the complexity of the network architecture grows, the cost of manually designing also increases.

Currently, most employed neural architectures are manually designed by experts, which is a resource-consuming and error-prone process. In addition, it is hard to jump out of the fixed thinking paradigm from inherent knowledge to develop excellent architectures. To reduce the onerous design cost, novel technology has emerged to allow the algorithm to automatically design the network architectures with less human intervention, *i.e.*, neural architecture search (NAS).

Theoretically speaking, NAS can be cast as a search problem over a group of choices that determine the different components of the network architecture. Zoph *et al.* first utilized

reinforcement learning (RL) algorithm to reach outstanding classification accuracy on the image classification tasks, which demonstrated the feasibility of automatic architecture design [22]. Shortly thereafter, Real *et al.* also achieved similar result by adopting the evolution algorithm (EA), which once again verified the feasibility of NAS [23]. Therefore, many researchers were committed to further developing and utilizing this automatic technology. NAS has also quickly been applied to various fields, such as image classification [24], object detection [25], semantic segmentation [26], and multi-objective optimization [27]. As the hottest topic of automated machine learning (AutoML), NAS has become one of the fastest-growing methods in this field.

In this paper, we propose an automatic Transformer neural architecture search named TNAS. The proposed TNAS first builds a super network that covers all candidates in the design space, and trains it to convergence on the training set, then searches through multi-objective evolution algorithm (MOEA) to obtain the optimal network architecture for EEG-based emotion recognition.

The main contributions of this paper can be summarized as follows:

1. We proposed an end-to-end TNAS framework based on MOEA to automatically search the optimal Transformer architecture to solve EEG-based emotion recognition tasks.
2. To comprehensively balance accuracy and resource efficiency, we proposed MOEA to consider both classification performance and model size in the search process.
3. We conducted extensive comparison experiments on two widely used datasets, *e.g.*, DEAP and DREAMER datasets. The experimental results showed that the proposed TNAS achieved state-of-the-art classification accuracy, which demonstrated its effectiveness.

III. Methodology

In this section, we introduce the framework of the proposed TNAS for EEG-based emotion recognition and explain the construction of the vanilla Transformer. Then, the principle of MOEA is expanded in detail. Finally, we detail the construction of the proposed TNAS.

A. Framework of TNAS

The proposed TNAS-based emotion recognition is an end-to-end system, which can capture global interaction from raw EEG signal for emotion classification tasks. The pipeline of the system is illustrated as Fig. 2. We first preprocess the EEG data and divide it into three parts for training, validating, and testing, then we train the constructed supernet on the training set to convergence. The multi-objective evolution algorithm (MOEA) is applied to search the optimal architecture from well-trained supernet, expecting to strike balance between outstanding performance and limited source consumption. Finally, the optimal architecture is evaluated on the test data, and the average classification accuracy is regarded as the final performance of the system. The details will be expanded in the following sections.

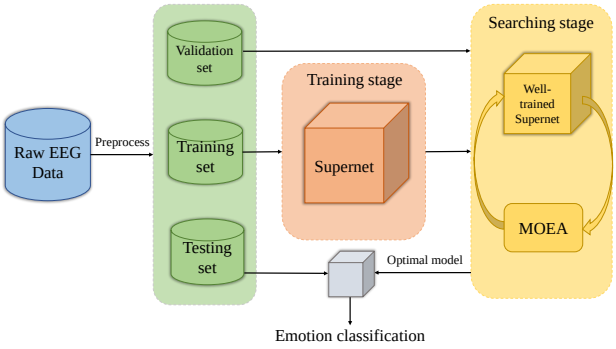

Fig. 2: The pipeline of the proposed method.

B. Transformer

Transformer was originally applied to NLP tasks and achieved significant results. Inspired by the achievements of Transformer in the field of NLP, Dosovitskiy *et al.* first modified this technique and applied it to CV tasks, named Vision Transformer (ViT) [28]. Notably, NLP tasks usually deal with $1D$ data, while CV problems correspond to $2D$ data. EEG is multi-channel time series signal, and $2D$ EEG data is utilized in this paper. Therefore, we employ ViT architecture for emotion recognition. Fig. 3 illustrates the overview of ViT, which is mainly composed of projection head, multiple stacked encoders, and classifier. The encoders contain alternating layers of MSA and MLP (1 hidden layer with Gaussian Error Linear Unit (GELU) nonlinearity) blocks. In addition, layer normalization (LN) and residual connections are applied before and after each block, respectively.

To facilitate the $2D$ input processing, ViT first divides the input $x \in \mathbb{R}^{H\times W\times C}$ into a sequence of flatten $2D$ patches of size $x_p \in \mathbb{R}^{N\times(P_1\cdot P_2\cdot C)}$, where (H, W) is the solution of the input and C is the number of channels. (P_1, P_2) is the fixed size of patches. $N = \frac{HW}{P_1P_2}$ is the number of patches, which also denotes the actual length of the input sequence. Then a trainable projection is adopted to map the x_p into a d dimension vector, *i.e.*, patch embedding. In addition, a learnable class embedding x_{class} is laid at the head of the patch embedding sequence to present the whole input. Standard $1D$ position embedding $P \in \mathbb{R}^{(N+1)\times d}$ is also added to the patch embedding to retain positional information. The final embedding input to the encoder z_0 is illustrated as follows:

$$z_0 = \left[x_{class}, x_p^1E, x_p^2E, \ldots, x_p^NE\right] + P, \quad E \in \mathbb{R}^{(P_1\cdot P_2\cdot C)\times d}, P \in \mathbb{R}^{(N+1)\times d}. \tag{1}$$

The core mechanism of ViT is multi-head self-attention (MSA), and we will detail the principle in the following.

Multi-head self attention: self attention holds unique advantages for extracting global information. In the standard self attention, the input embedding sequence $z \in \mathbb{R}^{N\times d}$ will be first transformed to the matrix $Q \in \mathbb{R}^{N\times d_h}$, $K \in \mathbb{R}^{N\times d_h}$ and $V \in \mathbb{R}^{N\times d_h}$ through a linear projection, where N is the length of the embedding sequence, d_h is the Q-K-V dimension, and D is the embedding dimension. Self attention computes the

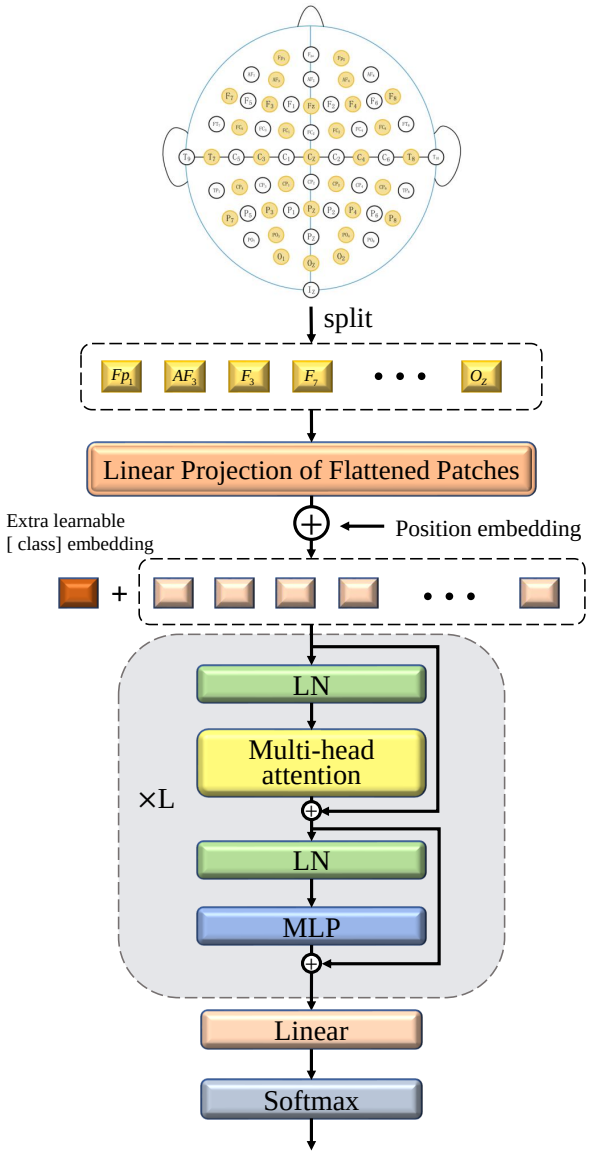

Fig. 3: The overview of vanilla ViT. The input is split and flattened into fixed-size patches, and embedded through the Linear projection. Then we feed the embedding vector into L encoders. An extra learnable [class] embedding and position embedding are added to present the classification and position information of the input. Where LN means layernorm operation.

weighted sum of all values for each element in the sequence as the output:

$$\text{Attention}\left(Q,K,V\right)=\text{Softmax}\left(\frac{QK^{T}}{\sqrt{d_h}}\right)V, \tag{2}$$

where $\sqrt{d_h}$ is the scaling factor. Multi-head self attention is based on single-head self attention, adopting h projections to transform embedding sequence to obtain multiple sets of Q, K and V. Finally, MSA performs self attention in parallel and projects their concatenated outputs (Eq. 3), which can jointly focus on the information from different representation subspaces at different positions. Fig. 4 illustrates the construction of MSA.

$$\text{MSA}\left(Q,K,V\right)=\text{Concat}\left(head_1,\ldots,head_h\right)W_o, \tag{3}$$

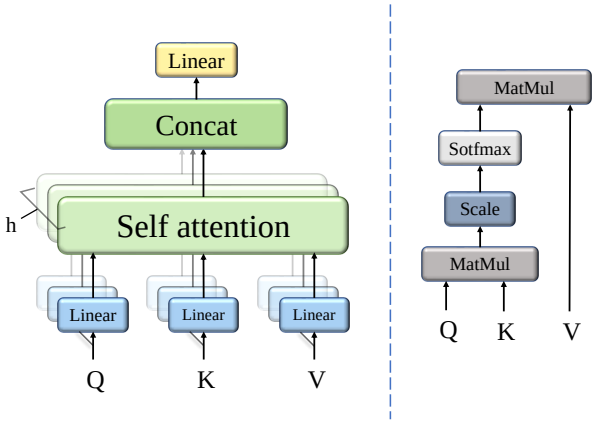

Fig. 4: The construction of MSA. Left: MSA based on several parallel self attention. Right: standard self attention.

where $head_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$, $W_i^Q \in \mathbb{R}^{d_h \times d_h^{'}}$, $W_i^K \in \mathbb{R}^{d_h \times d_h^{'}}$, $W_i^V \in \mathbb{R}^{d_h \times d_h^{'}}$ and $W_o \in \mathbb{R}^{hd_h^{'} \times d_h}$. Therefore, the output $z_l^{'}$ of MSA block and z_l of MLP block illustrated in Fig. 3 can be formulated as follows:

$$z_l^{'}=\text{MSA}\left(\text{LN}\left(z_{l-1}\right)\right)+z_{l-1}, l=1,\ldots,L, \tag{4}$$

$$z_l=\text{MLP}\left(\text{LN}\left(z_l^{'}\right)\right)+z_l^{'}, l=1,\ldots,L, \tag{5}$$

where the L denotes the number of the layers, LN means layernorm operation.

C. MOEA-basewd TNAS

The proposed MOEA-based TNAS mainly solves two challenges in searing ViT structure: 1) How to effectively combine the key hyperparameters of the model, *e.g.*, embedding dimension, MLP ratio (the ratio of hidden dimension to the embedding dimension in MLP), number of heads in MSA and the model depth. 2) How to efficiently search various ViT architectures suitable for different resource-constrainted application scenarios (the end-to-end system cannot support large models).

To address the above challenges, we will detail the principles of the proposed method from two aspects: the weight sharing mechanism and MOEA strategy.

Weight sharing of TNAS: weight sharing equipped TNAS can search ViT structure in a fast and less resource-consuming fashion. Considering the unique structure of ViT: the homogeneous building blocks such as MSA modules with different number of heads, and MLP with different hidden dimensions *etc.* We proposed a special weight sharing mechanism for TNAS, in which the homogeneous blocks are structurally compatible to achieve mutual weight sharing. The proposed TNAS first constructs a supernet that contains all subnets in the search space. Supernet stacks the maximum number of encoder blocks with the largest key hyperparameters. The searched optimal subnets directly inherit the weights from the well-trained supernet, which can achieve better performance than the comparison methods. Fig. 5 presents the overview of the constructed supernet. The center of the weight sharing is to

enable different blocks to share their weight of common parts at the same layer. Suppose a subnet in the search space $\alpha \in \mathcal{S}$ with l layers, the structure and its weight can be presented as follows:

$$\begin{cases} \alpha = \left(\alpha^{(1)}, \dots \alpha^{(i)}, \dots \alpha^{(l)}\right) \\ w = \left(w^{(1)}, \dots w^{(i)}, \dots w^{(l)}\right), \end{cases} \tag{6}$$

where $\alpha^{(i)}$ and $w^{(i)}$ mean the selected block and its weight in the i-th layer, respectively. There are multiple choices of blocks in each layer, $\alpha^{(i)}$ and $w^{(i)}$ are sampled from a set of n block candidates in the search space as follows:

$$\begin{cases} \alpha^{(i)} \in \left\{b_1^{(i)}, \dots b_j^{(i)}, \dots b_n^{(i)}\right\} \\ w^{(i)} \in \left\{w_1^{(i)}, \dots w_j^{(i)}, \dots w_n^{(i)}\right\}, \end{cases} \tag{7}$$

where $b_j^{(i)}$ and $w_j^{(i)}$ are the candidate blocks and their corresponding weight in the search space. Due to the advantages of the homogeneous blocks structure, we only need the weights of the largest block among n candidate blocks in each layer. The smaller block can directly utilize the weight from the largest one. Thus, for any two blocks in the same layer: $w_j^{(i)} \subseteq w_k^{(i)}$ or $w_k^{(i)} \subseteq w_j^{(i)}$. Supernet is composed of the largest blocks in each layer. Therefore, the subnets can inherit the weights from the well-trained supernet.

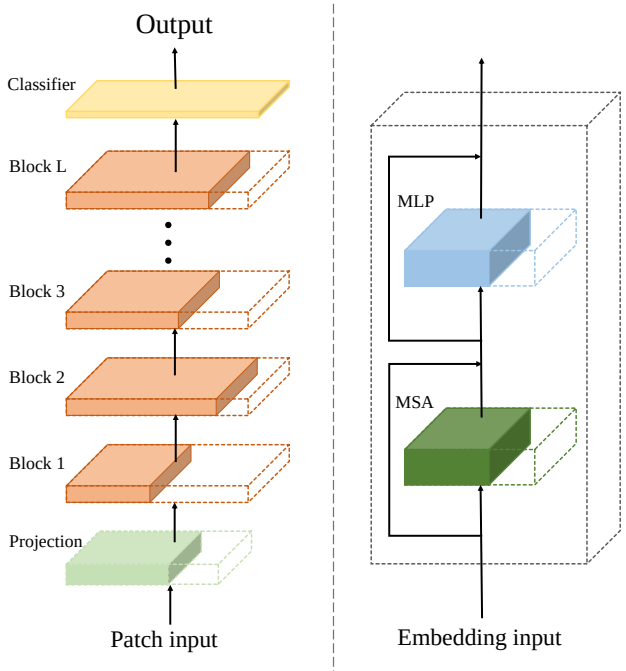

Fig. 5: Left: the construction of Supernet with weight sharing strategy. Supernet consists of L encoder blocks, while subnet inherits the weight from the solid parts of the supernet. The blocks in dashed means weight of supernet but not inherited currently (the layers of subnet may be less than L, thus subnet could flexibly inherit the weight from any layer). Right: the weight sharing of MLP and MSA in encoder blocks. We search for the best block with optimal embedding dimension, number of heads, and MLP ratio for each layer in subnet. In addition, the optimal depth is also searched.

MOEA: to strike balance between accuracy and the model size simultaneously, we utilize the Pareto theory [29] to solve the multi-objective optimization problems. Pareto front means a set of optimal resolutions, which represents the trade-off between the criteria and allows a link with the decision variables. Therefore, we adopt weighted sum approach to assign different weights to each objective function as follows:

$$score = \max_{\alpha \in \mathcal{S}} \left(\lambda_1 val\_acc_\alpha + \frac{\lambda_2}{modelsize_\alpha}\right), \tag{8}$$

where val_acc_α denotes the accuracy of subnet α on the validation dataset. Then MOEA is adopted to search the optimal submodel in the search space, as is illustrated in Fig. 6. We first random sample a certain number of subnets as the original population, and calculate the score as Eq. 8 for each of them. Then we employ crossover and mutation on the superior candidates with high scores for iterations. Finally, we select the candidate with the highest score as the optimal model.

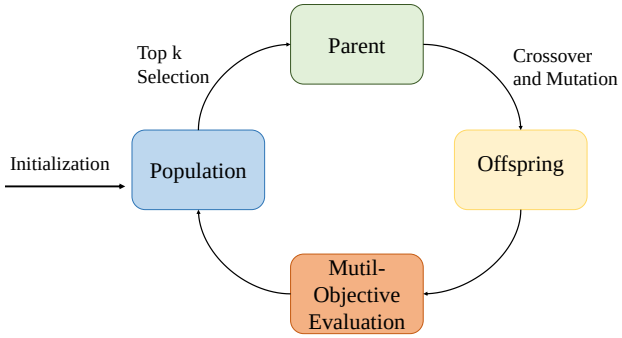

Fig. 6: The detail of multi-objective evolution algorithm (MOEA). Note that we first initial to generate the original population.

D. Search Pipeline

To search the optimal subnet α^*, we create a large search space, which contains diverse ViT structures with different model complexities. In addition, the structures in the search space cover the combination of key hyperparameters mentioned above. Unlike vanilla Transformer that all layers share an identical structure, we utilize different structures of building blocks in each layer. The search space $\mathcal{S}$ is encoded into a supernet as $\mathcal{N}(\mathcal{S}, W)$, where W is the weight of supernet which is shared across all candidate subnets $\alpha \in \mathcal{S}$. There are two stages to search the optimal model α^*. For the first stage, the weight W of the supernet is optimized as follows:

$$W_{\mathcal{S}} = \arg\min_{W} \mathcal{L}_{\text{train}}(\mathcal{N}(\mathcal{S}, W)), \tag{9}$$

where $\mathcal{L}_{\text{train}}$ means loss function in the training process. We adopt random sample to select subnets α from supernet $\mathcal{N}(\mathcal{S}, W)$ to train the parameter W with less resource cost. The detail of the training process is illustrated in Algorithm 1.

The second stage is to search the optimal subnet α^* via MOEA on the well-trained supernet with $W_{\mathcal{S}}$:

$$\alpha^* = \arg\max_{\alpha \in \mathcal{S}} \text{score}(\mathcal{N}(\alpha, w)), \tag{10}$$

where the subnet α is sampled with MOEA, which inherits weight w from well-trained $W_{\mathcal{A}}$. The score computes both the

Algorithm 1: Supernet Training with Weight Sharing

Input: Training epochs N,search space $\mathcal{S}$, supernet $\mathcal{N}$, initial supernet weights $W_{\mathcal{S}}$, train dataset D_{train} , loss function $\mathcal{L}_{\text{train}}$, epoch N;
Output: Well-trained supernet;

for i = *1 to N* **do**
  for *data, labels in* D_{train} **do**
    Random sample one Transformer architecture $\alpha = \left(\alpha^{(1)}, \ldots \alpha^{(i)}, \ldots \alpha^{(l)}\right)$ from the space $\mathcal{S}$;
    Obtain the corresponding weights $w = \left(w^{(1)}, \ldots w^{(i)}, \ldots w^{(1)}\right)$ from $W_{\mathcal{S}}$, where l is the maximum depth;
    Compute the gradients ∇w based on $\mathcal{L}_{train}$, data, and labels;
    Update the corresponding part of w in $W_{\mathcal{S}}$ while freeze the rest part of the superet $\mathcal{N}$;
  end
end

accuracy and model size of the sampled subnet α on validation dataset (Eq. 8).

In the search process, we first generate the original population P through random select algorithm, then we calculate the score function and take the top K subnets. The top K subnets are taken as parents for crossover and mutation to generate the next generation. Crossover generates new subnets by randomly selecting two top K candidates to cross. For mutation, the selected candidate mutates its depth and blocks to generate new subnets. The detailed process is illustrated in Algorithm 2.

Algorithm 2: Multi-Objective Evolution Search

Input: Search space $\mathcal{S}$, supernet $\mathcal{N}$, supernet weights $W_{\mathcal{S}}$, population size P, number of generation iteration $\mathcal{T}$, validation dataset D_{val} ;
Output: The optimal subnet α^*;

$G_{(0)}$= Random sample P subnets $\{\alpha_1, \alpha_2, \cdots \alpha_P\}$ from $\mathcal{S}$;
while *search step* $t \in (0, \mathcal{T})$ **do**
  while $\alpha_i \in G_{(t)}$ **do**
    Obtain the corresponding weight W_{α_i} from the weights $W_{\mathcal{S}}$ of supernet;
    Obtain the $score$ of the subnet $\mathcal{N}\left(\alpha_i, W_{\alpha_i}\right)$ on D_{val} ;
  end
  G_{topK} = the top K candidates by $score$ order;
  $G_{\text{crossover}}$ = Crossover $\left(G_{\text{topK}}, S\right)$;
  G_{mutation} = Mutation $\left(G_{\text{topK}}, S\right)$;
  $G_{(t+1)}$ = $G_{\text{crossover}} \cup G_{\text{mutation}}$;
end

IV. Experiment and Analysis

In this section, we first introduce two widely used datasets, *e.g.*, DEAP and DREAMER. Then we detail the implementation in the experiment. Finally, we conducted extensive experiments to compare our method with state-of-the-art methods on both datasets. Note that the proposed method focuses on the user-dependent emotion classification tasks.

A. Data Materials

DEAP and DREAMER have been widely used datasets in the field of EEG-based emotion recognition. Table I shows the data format, and the details of above datasets are as follows:

TABLE I
Data Introduction

Dataset	Array	Array shape	Array content
DEAP	Data	$40 \times 32 \times 8064$	videos $\times$ channels $\times$ data
	Labels	40×3	videos $\times$ label
DREAMER	Data	$18 \times 14 \times 25472$ (M)	videos $\times$ channels $\times$ data
	Labels	18×3	videos $\times$ label

1) DEAP: the DEAP dataset is a multimodal dataset that records the EEG (32 electrodes placed according to the international 10-20 system) and peripheral physiological signals of 32 healthy subjects when stimulated by 60s videos in 40 trials. Each participant performs a self-assessment to rate the level of arousal, valence, liking, dominance, and familiarity from 1 to 9. Each video trial data consists of 60s stimulated signals and 3s unstimulated signals (baseline signal). In our experiment, we adopt the blind source separation technique to extract EEG signal and set the threshold to 5 to divide trials into two classes (high and low) according to the levels of arousal, valence, and dominance, respectively.

2) DREAMER: the DREAMER dataset records EEG signal with 14 channels and ECG signals with 2 channels from 23 subjects (14 male and 9 female) when they are stimulated by 18 videos ranging from 65s to 193s (M = 199s). The participants rate the levels of arousal, valence, and dominance from 1 to 5 to present their emotional states. In our experiment, we only extract EEG signal, and set the threshold to 3 according to the levels of valence, arousal, and dominance, respectively.

Considering both DEAP and DREAMER record baseline signal for each subject recorded without stimulation, and it has been proved that baseline signal removal is beneficial for emotion recognition [30]. Therefore, we calculate the average baseline signal and subtract the average baseline signals from the stimulated trial signals for each subject on both datasets. For DEAP dataset, each subject contains 2400 (40 trails $\times$ 60s) EEG samples with 32 channels, and we downsample EEG samples to 128Hz to obtain the final samples $\mathbf{X} \in \mathbb{R}^{2400 \times 128 \times 32}$. For DREAMER dataset, the EEG samples of each subject is 3728 with 14 channels, finally we obtain the samples $\mathbf{X} \in \mathbb{R}^{3728 \times 128 \times 14}$.

B. Implementation Details

We adopt 10-fold cross-validation to evaluate the performance of the proposed TNAS and the comparison methods,

i.e., we calculate the 10-fold cross-validation accuracy of each subject, and take the average of the 10-fold accuracy from the whole subjects as the final result.

The key hyperparameters of the model, *e.g.*, embedding dimension, MLP ratio, the number of MSA heads, and model depth are ranging as follows:

$$\begin{cases} \text{embedding dimension} \in (168, 192, 216, 240) \\ \text{number of head} \in (2, 4, 6) \\ \text{MLP ratio} \in (3, 3.5, 4) \\ \text{depth} \in (6, 7, 8, 9, 10) \end{cases}, \tag{11}$$

As we build different layers with different blocks, the final search space $\mathcal{S}$ reaches 1.4×10^{10}. We train the W of supernet with cross-entropy loss and Adam optimization in cosine learning rate scheduler. The supernet is trained for 400 epochs to convergence. The λ_1, λ_2 in score function (Eq. 8) are both set to 1. In MOEA search process, we set the original population P = 100, mutation number = 50, crossover number = 50, top K = 50, generation iteration $\mathcal{T}$ = 50. In addition, we split EEG signal into patches according to the channels, thus the patch size of both datasets is set to (128×1).

TABLE II
THE STRUCTURES AND PARAMETERS OF SUPERNET AND THE REPRESENTATIVE SEARCHED OPTIMAL SUBNET

	Hyperparameters	Subnet	Supernet
Embedding dimension	-	168	240
Encoder block 1	MLP ratio	4.0	4.0
	Head number	2	6
Encoder block 2	MLP ratio	3.5	4.0
	Head number	4	6
Encoder block 3	MLP ratio	3.5	4.0
	Head number	2	6
Encoder block 4	MLP ratio	4.0	4.0
	Head number	4	6
Encoder block 5	MLP ratio	3.0	4.0
	Head number	4	6
Encoder block 6	MLP ratio	3.5	4.0
	Head number	4	6
Encoder block 7	MLP ratio	3.5	4.0
	Head number	4	6
Encoder block 8	MLP ratio	-	4.0
	Head number	-	6
Encoder block 9	MLP ratio	-	4.0
	Head number	-	6
Encoder block 10	MLP ratio	-	4.0
	Head number	-	6
Parameters	-	2.50M	6.98M

C. Result and Analysis

The proposed MOEA-based TNAS takes about 1 hour on NVIDIA RTX 2080 GPU to discover the optimal subnet for a single classification task. Fig. 7 and Table II show the representative searched optimal blocks of 1-th subject on arousal dimension in DEAP dataset. The model size of optimal subnet searched by MOEA strategy reaches only 2.50M, which is less than supetnet's 6.98M.

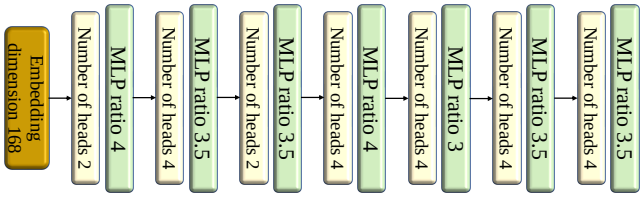

Fig. 7: The blocks of the optimal model for arousal dimension classification task of 1-th subject in DEAP dataset.

To evaluate the performance of our method, we conducted extensive comparison experiments on DEAP and DREAMER datasets. Note that all comparison experiments are conducted under the same conditions to ensure fairness. We compare our method with some state-of-the-art DL methods, *e.g.*, dynamical graph convolution neural networks (DGCNN) [31], convolution recurrent attention model (CRAM) [32], continuous convolution neural network (Cont-CNN) [33], gcForest [34] and NAS [35]. Song *et al.* adopted a DGCNN, which utilized a graph to model the multichannel EEG features to dynamically learn the intrinsic relationship between different channels. Zhang *et al.* presented a CRAM to encode the high-level representation of EEG signal and explored the temporal dynamics and discriminative temporal periods by recurrent attention mechanism. Yang *et al.* proposed Cont-CNN with $3D$ representation of EEG segments, which combined the features of EEG signal from different frequency bands while preserving spatial information among channels. Cheng *et al.* adopted gcForest while constructing $2D$ frame sequences with spatial position relationship across channels, which can explore the spatial and temporal information of EEG signal. Similar to our work, Li *et al.* utilized an LSTM controller with reinforcement learning to discover the optimal CNN architecture, which has demonstrated the effectiveness of automatic architecture search in emotion recognition. Besides, we also compared some traditional methods, *e.g.*, decision tree (DT), SVM, and MLP [33].

Table III shows the average accuracy and standard deviation of TNAS and the comparison methods on the DEAP dataset. Compared with three traditional methods (DT, SVM, and MLP), TNAS achieves the best performance and demonstrates its automatic feature extraction capability in EEG signal. Besides, the proposed TNAS also outperforms five compared DL methods (DGCNN, CRAM, Cont-CNN, gcForest, and NAS). Compared with CRAM, the average recognition accuracy on arousal, valence, and dominance dimension increases by 14.75%, 13.25%, and 27.04%, respectively. In contrast with Cont-CNN, TNAS improves the average classification accuracy by 8.42%, 9.23%, and 8.42% on three dimensions. Compared with NAS which is similar to our work, the average accuracy increases by 0.72%, 0.94%, and 0.85%, on arousal, valence, and dominance, respectively. Fig. 8 details the classification accuracy of 32 subjects in the DEAP dataset on arousal, valence, and dominance, respectively.

Table IV reports the average accuracy and standard deviation of TNAS and compared methods on the DREAMER dataset. Overall, the proposed TNAS reaches the highest average accuracy in the emotion classification task. For CRAM, the improved average accuracy on three dimensions reaches 16.66%, 17.93%, and 18.19%, respectively. Compared with Cont-CNN, the average accuracy increases by 12.11%,

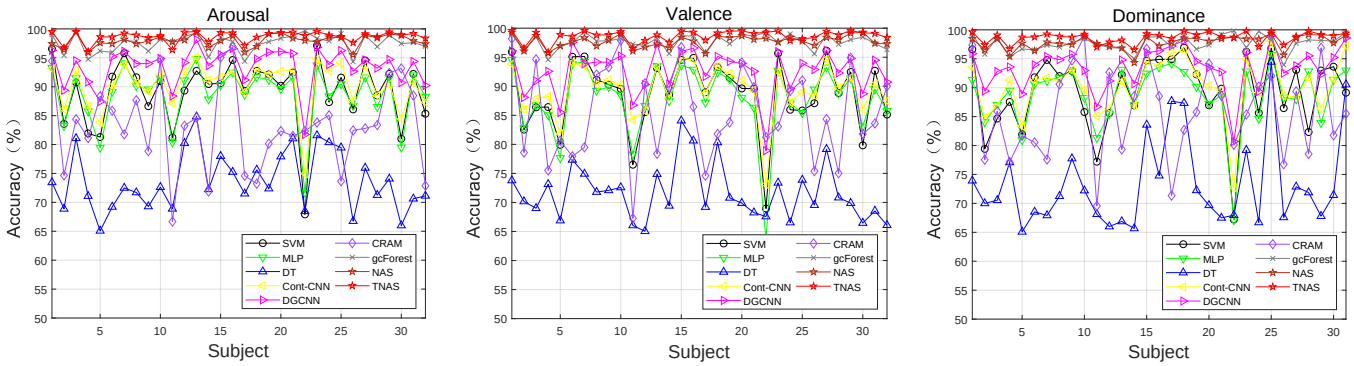

Fig. 8: Average accuracy (%) on each subject of different methods on arousal, valence, and dominance classification tasks on DEAP database, respectively.

TABLE III
AVERAGE ACCURACY AND STANDARD DEVIATION (%) OF DIFFERENT METHODS ON DEAP DATASET

Methods	Arousal	Valence	Dominance
DT	73.70 ± 5.09	71.63 ± 4.71	73.36 ± 7.83
MLP	88.68 ± 5.04	87.82 ± 6.15	88.75 ± 5.53
SVM	89.07 ± 5.89	88.65 ± 6.18	89.14 ± 6.70
DGCNN	93.50 ± 3.35	92.55 ± 3.93	93.50 ± 3.75
CRAM	83.19 ± 7.49	85.34 ± 7.94	71.63 ± 4.71
Cont-CNN	90.24 ± 4.08	89.45 ± 4.42	90.25 ± 4.95
gcForest	97.53 ± 1.52	97.69 ± 1.22	97.62 ± 1.39
NAS	97.94 ± 1.04	97.74 ± 1.02	97.82 ± 1.20
TNAS	**98.66 ± 0.94**	**98.68 ± 0.98**	**98.67 ± 0.95**

11.87%, and 11.85% on arousal, valence, and dominance dimension. Similarly, the proposed TNAS also outperforms similar work NAS, with the increased average accuracy of 0.33%, 0.12%, and 0.29% on arousal, valence, and dominance dimension, respectively. Fig. 9 illustrates the detailed comparison results of 23 subjects in DREAMER on arousal, valence, and dominance dimension.

We also conducted the experiment to validate the effectiveness of the proposed MOEA. Table V shows the average accuracy and standard deviation of searched optimal subnets and supernet on DEAP and DREAMER datasets. Note that the optimal subnets searched by MOEA strategy can achieve better performance. MOEA strategy can find the optimal subnets according to the characteristics of different classification tasks. In addition, we guess the reason for the poor performance of supernet may be overfitting due to redundant parameters.

TABLE IV
AVERAGE ACCURACY AND STANDARD DEVIATION (%) OF COMPARISON METHODS ON DREAMER DATASET

Methods	Arousal	Valence	Dominance
DT	75.74 ± 6.44	75.53 ± 6.71	76.76 ± 5.70
MLP	83.71 ± 5.39	83.64 ± 5.97	83.90 ± 5.32
SVM	87.03 ± 4.88	87.14 ± 6.21	87.18 ± 4.87
DGCNN	88.93 ± 3.93	89.59 ± 5.13	88.63 ± 5.13
CRAM	80.29 ± 9.64	78.48 ± 7.40	78.71 ± 7.61
Cont-CNN	84.84 ± 4.86	84.54 ± 5.01	85.05 ± 4.96
gcForest	90.41 ± 5.33	89.03 ± 5.56	89.89 ± 6.19
NAS	96.62 ± 3.52	96.29 ± 3.82	96.61 ± 4.04
TNAS	**96.95 ± 3.35**	**96.41 ± 3.61**	**96.90 ± 3.49**

Extensive experimental results have shown that the proposed TNAS outperforms the state-of-the-art methods. In addition, the proposed method holds not only the highest average accuracy but also the smallest deviation, which proves its effectiveness and robustness. Fig. 10 shows the average weight matrix of heads in MSA, which indicates the attention distribution of input embeddings. The proposed TASN assigns different attention weight to EEG channels according to the correlation between EEG channels to better solve EEG-based emotion recognition tasks. We also report the related studies on EEG-based emotion recognition with DEAP and DREAMER datasets in Tables VI, VII. From the report results we can notice that the weight sharing equipped TNAS reaches the highest accuracy on both datasets.

TABLE V
AVERAGE ACCURACY (%) OF SUPERNET AND THE SEARCHED OPTIMAL SUBNETS ON DEAP AND DREAMER DATASETS (MEAN ± STD. DEV.)

Dataset	Models	Valence	Arousal	Dominance
DEAP	Supernet	94.88 ± 2.82	93.39 ± 3.14	93.47 ± 3.21
	Optimal subnets	**98.68 ± 0.98**	**98.66 ± 0.94**	**98.67 ± 0.95**
DREAMER	Supernet	94.62 ± 4.39	95.47 ± 4.01	94.99 ± 4.53
	Optimal subnets	**96.41 ± 3.61**	**96.95 ± 3.35**	**96.90 ± 3.49**

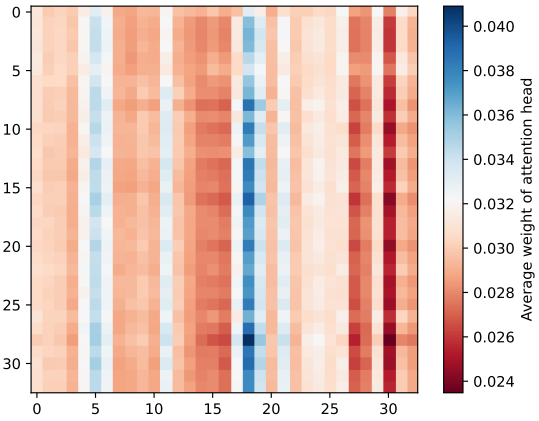

Fig. 10: An example of average weight matrix of self-attention heads in the MSA. The ith row of each attention weight matrix corresponds to the attention distribution of the ith embedding over the entire embedding sequence.

Note that the weight sharing strategy adopted in this study benefits from the homogeneity of the building blocks. The basic components of Transformers are quite different from CNNs: only using MSA and MLP blocks instead of con-

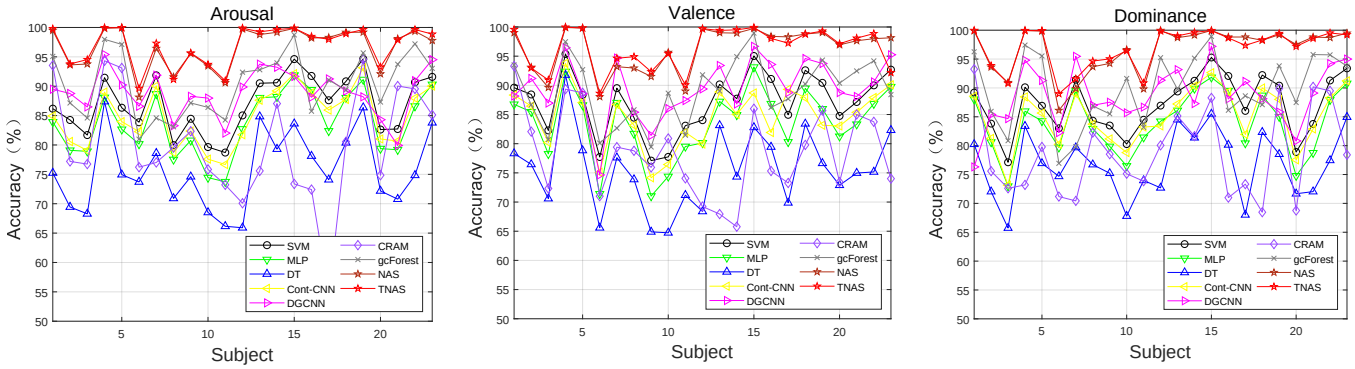

Fig. 9: Average accuracy (%) on each subject of different methods on arousal, valence, and dominance classification tasks on DREAMER database, respectively.

TABLE VI
DETAILS OF SEVERAL REPORTED STUDIES ON DEAP DATASET

Studies	Input	Classifier	Evaluation Method	Accuracy (%)		
				V	A	D
Huang *et al.* [36]	spatial+temporal features	BiDCNN	10-fold cross validation	94.38	94.72	-
Islam *et al.* [37]	PCC	CNN	10-fold cross validation	78.22	74.92	-
Ozdemir *et al.* [38]	spectral+spatial features	DCNN	10-fold cross validation	90.62	86.13	88.48
An *et al.* [39]	2D EEG features	DCNN and LSTM	10-fold cross validation	87.84	87.69	-
Liu *et al.* [40]	Raw EEG signal	Capsule network	10-fold cross validation	97.97	98.31	98.32
Tao *et al.* [41]	Raw EEG signal	ACRNN	10-fold cross validation	93.72	93.38	-
Li *et al.* [42]	Raw EEG signal	Capsule network	10-fold cross validation	97.41	97.25	98.35
Proposed method	Raw EEG signal	TNAS	10-fold cross validation	**98.68**	**98.66**	**98.67**

A, V, and D mean arousal, valence, and dominance dimension, respectively.

TABLE VII
DETAILS OF SEVERAL REPORTED STUDIES ON DREAMER DATASET

Studies	Input	Classifier	Evaluation Method	Accuracy (%)		
				V	A	D
Sharma *et al.* [43]	SM-SSA features	GCBN	10-fold cross validation	93.30	92.06	92.38
Pandey *et al.* [44]	Raw EEG signal	CNN	10-fold cross validation	75.93	81.48	-
Topic *et al.* [45]	HOLO-FM features	CNN and SVM	10-fold cross validation	88.20	90.43	-
Li *et al.* [46]	DE features	ASTG and LSTM	10-fold cross validation	96.27	96.44	-
Liu *et al.* [40]	Raw EEG signal	Capsule network	10-fold cross validation	94.59	95.26	95.13
Li *et al.* [42]	Raw EEG signal	Capsule network	10-fold cross validation	95.54	94.96	95.52
Proposed method	Raw EEG signal	TNAS	10-fold cross validation	**96.41**	**96.95**	**96.90**

A, V, and D mean arousal, valence, and dominance dimension, respectively.

volution operations. The corresponding units of the subnets are independent of other hidden units in the training process with weight sharing. Recent works have indicated that too many heads in MSA are redundant [47], and we also noticed that when we increased the number of heads in MSA but found no benefit for performance. In addition, the advantage of Transformer is that they can be flexibly extended to high parameter complexity. Although such significant features allows training large models, it usually requires high training and inference cost, which is unrealistic for real-time and resource-constrainted applications. Lightweight and portable EEG devices have become an important part in the field of BCI. This study is dedicated to design efficient and less resource-consuming Transformer architectures for EEG-based emotion recognition.

V. CONCLUSION

In this paper, we propose an end-to-end DL method TNAS to automatically search the optimal architecture for EEG-based emotion recognition tasks. The proposed TNAS adopts the weight sharing strategy to train the supernet and searches the optimal model with MOEA within 1 GPU-hour for a single classification task. The proposed TNAS achieves the average accuracy of 98.66%, 98.68%, and 98.67% for arousal, valence, and dominance dimension classification tasks on the DEAP dataset, respectively. On the DREAMER database, the proposed TNAS achieves the average accuracy of 96.95%, 96.41%, and 96.90% on arousal, valence, and dominance dimension classification tasks, respectively. The results demonstrate that the proposed TNAS reaches the state-of-the-art performance compared with baseline methods, which indicates the effectiveness of TNAS. In our future work, we will explore user-independent problems in EEG-based emotion recognition. Moreover, most NAS systems only focus on solving specific tasks on some fixed datasets, and a high-quality NAS system should have the capability of lifelong learning. Therefore, how to embed lifelong learning into NAS system to make the searched network competent for different tasks is worth studying.

REFERENCES

[1] Z. Gao, Y. Li, Y. Yang, N. Dong, X. Yang, and C. Grebogi, "A coincidence-filtering-based approach for cnns in eeg-based recognition," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 11, pp. 7159–7167, 2020.

[2] J. Jin, Z. Wang, R. Xu, C. Liu, X. Wang, and A. Cichocki, "Robust similarity measurement based on a novel time filter for ssveps detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[3] J. Jin, R. Xiao, I. Daly, Y. Miao, X. Wang, and A. Cichocki, "Internal feature selection method of csp based on l1-norm and dempster–shafer theory," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4814–4825, 2021.

[4] Y. Yu, Y. Liu, E. Yin, J. Jiang, Z. Zhou, and D. Hu, "An asynchronous hybrid spelling approach based on eeg–eog signals for chinese character input," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1292–1302, 2019.

[5] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[6] M. M. Rahman, A. K. Sarkar, M. A. Hossain, M. S. Hossain, M. R. Islam, M. B. Hossain, J. M. Quinn, and M. A. Moni, "Recognition of human emotions using eeg signals: A review," *Computers in Biology and Medicine*, vol. 136, p. 104696, 2021.

[7] J. Rizkallah, "Characterization of neocortical networks from high-resolution eeg: application to disorders of consciousness," Ph.D. dissertation, Rennes 1, 2019.

[8] D. Mathersul, L. M. Williams, P. J. Hopkinson, and A. H. Kemp, "Investigating models of affect: relationships among eeg alpha asymmetry, depression, and anxiety." *Emotion*, vol. 8, no. 4, p. 560, 2008.

[9] A. Al-Nafjan, M. Hosny, Y. Al-Ohali, and A. Al-Wabil, "Review and classification of emotion recognition based on eeg brain-computer interface system research: a systematic review," *Applied Sciences*, vol. 7, no. 12, p. 1239, 2017.

[10] J. A. Russell, "Affective space is bipolar." *Journal of personality and social psychology*, vol. 37, no. 3, p. 345, 1979.

[11] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[12] M. Li and B.-L. Lu, "Emotion classification based on gamma-band eeg," in *2009 Annual International Conference of the IEEE Engineering in medicine and biology society*. IEEE, 2009, pp. 1223–1226.

[13] X. Jie, R. Cao, and L. Li, "Emotion recognition based on the sample entropy of eeg," *Bio-medical materials and engineering*, vol. 24, no. 1, pp. 1185–1192, 2014.

[14] Y. Cimtay and E. Ekmekcioglu, "Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset eeg emotion recognition," *Sensors*, vol. 20, no. 7, p. 2034, 2020.

[15] H. Cui, A. Liu, X. Zhang, X. Chen, K. Wang, and X. Chen, "Eeg-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network," *Knowledge-Based Systems*, vol. 205, p. 106243, 2020.

[16] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on eeg using lstm recurrent neural network," *Emotion*, vol. 8, no. 10, pp. 355–358, 2017.

[17] Y. Song, X. Jia, L. Yang, and L. Xie, "Transformer-based spatial-temporal feature learning for eeg decoding," *arXiv preprint arXiv:2106.11170*, 2021.

[18] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems*, 2017.

[20] J. Liu, L. Zhang, H. Wu, and H. Zhao, "Transformers for eeg emotion recognition," 2021.

[21] Z. Wang, Z. Zhou, H. Shen, Q. Xu, and K. Huang, "Jdat: Joint-dimension-aware transformer with strong flexibility for eeg emotion recognition," 2021.

[22] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[23] E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. Tan, Q. V. Le, and A. Kurakin, "Large-scale evolution of image classifiers," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2902–2911.

[24] Y. Weng, T. Zhou, L. Liu, and C. Xia, "Automatic convolutional neural architecture search for image classification under different scenes," *IEEE Access*, vol. 7, pp. 38 495–38 506, 2019.

[25] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7036–7045.

[26] V. Nekrasov, H. Chen, C. Shen, and I. Reid, "Fast neural architecture search of compact semantic segmentation models via auxiliary cells," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9126–9135.

[27] Y. Xiong, R. Mehta, and V. Singh, "Resource constrained neural network architecture search: Will a submodularity assumption help?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1901–1910.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] A. Gaspar-Cunha and J. A. Covas, "Robustness in multi-objective optimization using evolutionary algorithms," *Computational Optimization and Applications*, vol. 39, no. 1, pp. 75–96, 2008.

[30] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, "Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[31] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2020.

[32] D. Zhang, L. Yao, K. Chen, and J. Monaghan, "A convolutional recurrent attention model for subject-independent eeg signal analysis," *IEEE signal processing letters*, vol. 26, no. 5, pp. 715–719, 2019.

[33] Y. Yang, Q. Wu, Y. Fu, and X. Chen, "Continuous convolutional neural network with 3d input for eeg-based emotion recognition," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 433–443.

[34] J. Cheng, M. Chen, C. Li, Y. Liu, R. Song, A. Liu, and X. Chen, "Emotion recognition from multi-channel eeg via deep forest," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 453–464, 2021.

[35] C. Li, Z. Zhang, R. Song, J. Cheng, Y. Liu, and X. Chen, "Eeg-based emotion recognition via neural architecture search," *IEEE Transactions on Affective Computing*, 2021.

[36] D. Huang, S. Chen, C. Liu, L. Zheng, Z. Tian, and D. Jiang, "Differences first in asymmetric brain: A bi-hemisphere discrepancy convolutional neural network for eeg emotion recognition," *Neurocomputing*, vol. 448, pp. 140–151, 2021.

[37] M. R. Islam, M. M. Islam, M. M. Rahman, C. Mondal, S. K. Singha, M. Ahmad, A. Awal, M. S. Islam, and M. A. Moni, "Eeg channel correlation based model for emotion recognition," *Computers in Biology and Medicine*, vol. 136, p. 104757, 2021.

[38] M. A. Ozdemir, M. Degirmenci, E. Izci, and A. Akan, "Eeg-based emotion recognition with deep convolutional neural networks," *Biomedical Engineering/Biomedizinische Technik*, vol. 66, no. 1, pp. 43–57, 2021.

[39] Y. An, N. Xu, and Z. Qu, "Leveraging spatial-temporal convolutional features for eeg-based emotion recognition," *Biomedical Signal Processing and Control*, vol. 69, p. 102743, 2021.

[40] Y. Liu, Y. Ding, C. Li, J. Cheng, R. Song, F. Wan, and X. Chen, "Multi-channel eeg-based emotion recognition via a multi-level features guided capsule network," *Computers in Biology and Medicine*, vol. 123, p. 103927, 2020.

[41] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen, "Eeg-based emotion recognition via channel-wise attention and self attention," *IEEE Transactions on Affective Computing*, 2020.

[42] C. Li, B. Wang, S. Zhang, Y. Liu, R. Song, J. Cheng, and X. Chen, "Emotion recognition from eeg based on multi-task learning with capsule network and attention mechanism," *Computers in Biology and Medicine*, p. 105303, 2022.

[43] L. D. Sharma and A. Bhattacharyya, "A computerized approach for automatic human emotion recognition using sliding mode singular spectrum analysis," *IEEE Sensors Journal*, vol. 21, no. 23, pp. 26 931–26 940, 2021.

[44] P. Pandey and K. Seeja, "A one-dimensional cnn model for subject independent emotion recognition using eeg signals," in *International Conference on Innovative Computing and Communications*. Springer, 2022, pp. 509–515.

[45] A. Topic and M. Russo, "Emotion recognition based on eeg feature maps through deep learning network," *Engineering Science and Technology, an International Journal*, 2021.

[46] X. Li, W. Zheng, Y. Zong, H. Chang, and C. Lu, "Attention-based spatio-temporal graphic lstm for eeg emotion recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[47] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" *arXiv preprint arXiv:1905.10650*, 2019.