TSception:A Deep Learning Framework for Emotion Detection Using EEG

Yi Ding[1], Neethu Robinson[1], Qiuhao Zeng[1], Duo Chen[1], Aung Aung Phyo Wai[1], Tih-Shih Lee[2,3], Cuntai Guan[1*]
[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Neuroscience and Behavioral Disorders Program,
Duke University - National University of Singapore Medical School, Singapore, Singapore
[3]Singapore General Hospital, Singapore, Singapore
[1]{ding.yi, nrobinson, qiuhao.zeng, chenduo, apwaung, ctguan}@ntu.edu.sg,
[2, 3]tihshih.lee@duke-nus.edu.sg

***Abstract*—In this paper, we propose a deep learning framework, TSception, for emotion detection from electroencephalogram (EEG). TSception consists of temporal and spatial convolutional layers, which learn discriminative representations in the time and channel domains simultaneously. The temporal learner consists of multi-scale 1D convolutional kernels whose lengths are related to the sampling rate of the EEG signal, which learns multiple temporal and frequency representations. The spatial learner takes advantage of the asymmetry property of emotion responses at the frontal brain area to learn the discriminative representations from the left and right hemispheres of the brain. In our study, a system is designed to study the emotional arousal in an immersive virtual reality (VR) environment. EEG data were collected from 18 healthy subjects using this system to evaluate the performance of the proposed deep learning network for the classification of low and high emotional arousal states. The proposed method is compared with SVM, EEGNet, and LSTM. TSception achieves a high classification accuracy of 86.03%, which outperforms the prior methods significantly (p<0.05).**

***Index Terms*—Deep learning, convolutional neural network, electroencephalography, emotional arousal, virtual reality**

I. Introduction

Emotions are fundamental in the daily life of human beings. Emotions can be mapped into the Valence, Arousal, and Dominance (VAD) dimensions [1]. Among three dimensions, Emotional Arousal (EA) detection plays an important role in the diagnosis and therapy of psychological disabilities, such as anxiety disorder [2] [3], autistic spectrum disorders (ASD) [4]. Research [5] on emotion-focused therapy (EFT) demonstrated that the emotional arousal is critical to psychotherapeutic success. However the detection of the emotional arousal is still a challenging task for the human-machine interaction system.

Brain computer interface (BCI) system enables the computer to perceive the arousal mental state of the human, using machine learning and signal processing technology [6]. Electroencephalography (EEG) is collected by several electrodes located on the surface of the human head, reflecting the potential neural activity directly. Compared to other emotional signals, such as facial expression and natural language, EEG contains more comprehensive information regarding human mental state and objective evaluation. Emotional arousal detection by EEG contains three main parts: pre-processing, feature extraction and classifier training. The artifact and noise (e.g. eyes blinks, 60 Hz noise) will be removed during the pre-processing stage. Power Spectral Density (PSD) of different frequency bands, Differential Entropy (DE), Event-Related De/Synchronizations (ERD/ERS), Event-Related Potentials (ERP), etc. are commonly extracted as features. A set of features is then selected to train a classifier.

A lot of research work has been conducted to solve the EEG emotional state classification problem [1] [7]. Atkinson *et al*. [8] proposed an efficient feature selection method to improve the SVM classifier performance on emotional arousal detection, with the accuracy being 73.14%. Zheng *et al* [9] investigated stable patterns of electroencephalogram (EEG) over time for emotion recognition, using Discriminative Graph Regularized Extreme Learning Machine with DE features. Li *et al.* [10] constructed emotion-related brain networks with phase-locking value and adopted a multiple feature fusion approach for emotion recognition. Recently, deep learning methods have shown promising classification performance in BCI, such as motor imagery classification [11] [12] [13] [14] [15], emotion recognition [16] [17] [18] [19], and mental-task classification [20] [21] [22]. Yang *et. al.* [16] designed a hierarchical network structure with sub-network nodes to classify three emotional states. Li *et al.* [17] proposed a Hierarchical Convolutional Neural Networks (HCNN) to extract the spatial information of the EEG electrodes by mapping the EEG signal into a 2D location map. Li *et. al.* [18] applied 18 kinds of linear and non-linear features to study the cross-subject emotion recognition problems, achieving 59.06% and 83.33% on two public datasets. Although many machine learning methods have been proposed for emotional arousal detection, most of them highly rely on the hand-extracted features. Vernon J Lawhern *et al* [14] proposed EEGNet, an end-to-end deep learning framework that can extract the hidden temporal and spatial patterns from the raw EEG data.

Inspired by the Inception block of GoogleNet [23], we

* Cuntai Guan is the Corresponding Author.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

proposed TSception, a deep learning framework for EEG signal classification. It uses temporal and spatial learners to learn more discriminative representations for EEG signals in time and space domains simultaneously. There are two types of convolutional learners in TSception: temporal learner and spatial learner. The temporal learner has multi-scale convolutional kernels, learning more discriminative multiple temporal and frequency representations. Psychophysiological evidence [24] indicates that the left and right halves of the human frontal brain areas differentially associate with particular emotions and affective traits. The spatial learner takes advantage of the frontal area of brain emotional asymmetry, using hemisphere kernels to learn the proper representation of the information from the right and left brain.

Studies [25] [3] [26] have shown the VR can induce targeted emotion effectively. In order to study the emotional arousal in the immersive VR environment and evaluate the proposed algorithm, we designed a VR-BCI system and collected EEG data from 18 healthy subjects in the VR environment.

The major contribution of this work can be summarised as:

- Designed a new deep learning framework, which uses temporal/spatial learners to learn time/space discriminative EEG representations of low and high emotional arousal states. The convolutional kernels in temporal learner have multi-scale lengths related to the sampling rate, learning multiple temporal and frequency representations parallelly. The spatial learner considers the frontal area of brain emotional asymmetry to learn global-local spatial representations of EEG signals.
- Designed and implemented the experiment and system to collect emotional arousal EEG data in the VR environment.
- Finally, the proposed method is compared with SVM using relative power and DE as features, EEGNet, LSTM, together with two simplified self-comparison versions, namely, Tception and Sception.

II. Related Work

A. GoogleNet and Inception Block

GoogleNet, also known as Inception-V1, won the 2014-ILSVRC competition [23]. In GoogleNet, small blocks are used instead of conventional convolutional layers. The inception block was introduced in the GoogleNet architecture. In inception blocks, split, transform and merge operations are achieved to improve the learning of varying represents for the same object in different images. The main idea of the Inception block is to extract spatial patterns using multi-scale convolutional kernels (1x1, 3x3, 5x5) on each layer.

B. Convolutional Neural Network for EEG Data

Different from images, the EEG data can be treated as 2D time series, whose dimensions are channels (EEG electrodes) and time respectively. The channels in this paper are the EEG electrodes instead of RGB dimensions in images or the input/output channels for convolutional layers. Because the electrodes are located at different areas on the surface of the human's head, the channel dimension contains spatial information of EEG; the time dimension is full of temporal information instead. In order to train a classifier, the EEG signal will be split into shorter time segments by a sliding window with a certain overlap along the time dimension. Each segment will be one input sample for the classifier.

Recently the convolutional neural networks have shown promising results in BCI [27] [28] [14]. J. Li *et al*. [17] constructed the EEG data into a sparse 2D map according to the relative location of the electrodes for each time point. Then (N, N) sized convolutional kennels were applied to do convolution. It can capture the local spatial pattern by sharing the kernels step by step but will lose the global spatial information since N is usually smaller than the length of the input. R. T. Schirrmeister *et. al.* [11] designed a deep ConvNet, which has 1D temporal convolutional kernels and global spatial convolutional kernels to extract temporal and global spatial information from EEG signal. N. Robinson *et. al.* [27] presented a deep learning-driven EEG-BCI system to perform decoding of hand motor imagery using deep convolution neural network architecture. Fahimi *et. al.* [20] use a deep convolutional neural network to build an inter-subject transfer learning framework for attentive mental state detection. Recently, Vernon J Lawhern *et al.* [14] proposed EEGNet, which contains the depth-wise convolution kernel of size $(N, 1)$. It can extract global spatial dependency by letting the N equal to the length of the channel dimension. In EEGNet, there are also 1D temporal kernels with a single size in each layer to learn temporal information.

III. Methodology

A. General Structure of Proposed Network - TSception

Temporal Spatial Inception (TSception) can be divided into 3 main parts: temporal learner, spatial learner and classifier. Inspired by Inception block, TSception uses multi-scale convolution kernels in temporal/spatial learners to learn time/space diverse representations simultaneously. Fig. 1 shows the general structure of TSception. The input of TSception is the raw EEG signal, making it an end-to-end classification structure. The features are learnt by the temporal and spatial learners automatically. The input is fed into the temporal learner first followed by spatial learner. Finally, the learned feature vector will be passed through 2 fully connected layer to map it to the corresponding label.

B. Temporal Learner

The temporal learner consists of multi-scale 1D temporal kernels (T kernels) whose lengths are in different ratios of EEG signal sampling rate f_S. The ratio coefficients are defined as $\alpha^i \in \mathbb{R}$, where the i is the level of the temporal learner. If the temporal learner has L levels, then i varies from 1 to L, and the temporal learner will have L types of temporal kernels. Hence S_T^i, the size of T kernels in i th level, can be defined as:

$$S_T^i = \left(1, \alpha^i \cdot f_S\right) \tag{1}$$

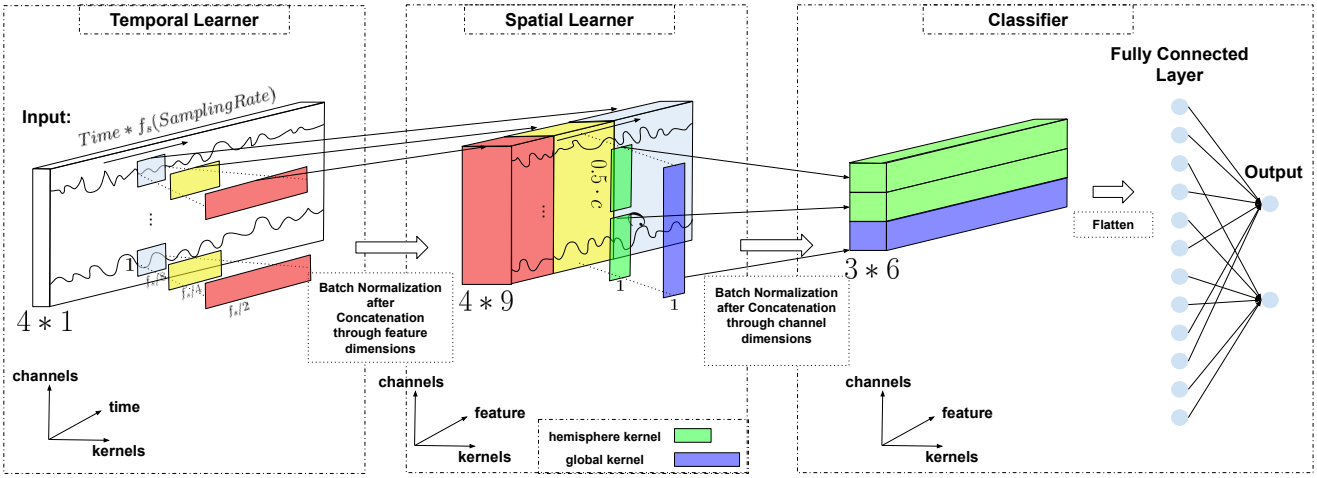

Fig. 1. The structure of TSception. The convolution results correspond to the kernels with the same color. TSception can be divided into 3 main parts: temporal learner, spatial learner and classifier. The input is fed into the temporal learner first followed by spatial learner. Finally, the feature vector will be passed through 2 fully connected layer to map it to the corresponding label. The dimension of the input EEG segment is (4 x 1 x 1024) since it has 4 channels, and 1024 data points per channel. There are 9 kernels for each type of temporal kernels in temporal learner, and 6 kernels for each type of spatial kernels in spatial learner. The multi-scale temporal convolutional kernels will operate convolution on the input data parallelly. For each convolution operation, $ReLU(\cdot)$ and average pooling are applied to the feature. The output of each level temporal kernel are concatenated along feature dimension, after which batch normalization is applied. In spatial learner, the global kernel and hemisphere kernel are used to extract spatial information. The output of the two spatial kernels will be concatenated along the channel dimension after $ReLU(\cdot)$ and average pooling. The flattened feature map will be fed into a fully connected layer. After the dropout layer and softmax activation function the classification result will be generated.

From the frequency perspective, in EEGNet, the length of the T kernel is set at half the sampling rate allows for capturing frequency information at 2 Hz and above [14]. Emotional states are more related to Alpha (8-15 Hz), Beta (15-32 Hz) and Gamma (>32 Hz) [1], in this work, we expand the temporal perception ranges, letting $L = 3, i = 1$ to 3 and $\alpha = 0.5$, the ratio coefficients are [0.5, 0.25, 0.125], which can further capture frequency at 4 Hz to above and 8 Hz to above. We hold the hypothesis that by using the multi-scale temporal kernels, the temporal learner can learn multi-frequency representations related to emotional state. From the time perspective, multi-scale T kernels can capture long short-term temporal pattern, providing more diverse representations. The lower level T kernel has a larger ratio coefficient, which gives longer convolutional kernel length and vice versa. The long kernel can learn long term temporal and low-frequency diverse representations. The short kernel extracts short term temporal and high-frequency representations instead. Let X be the raw EEG input segments array. $X = \left[x^0, x^1, ..., x^n\right], x^n \in \mathbb{R}^{C \times L}$, where n is the number of EEG segments, C is the number of channels, L is the length of the segments. The multi-scale temporal kernels will operate convolution on the input data parallelly. EEG signal has a low signal-noise ratio, using average pooling can reduce the effect of the noise as well as the feature dimension. After activated by $ReLU(\cdot)$, the feature map is further down-sampled by average pooling. Let z^i_{conv} be the output of the i th level temporal kernel, $z^i_{conv} \in \mathbb{R}^{B \times T \times C \times F^i}$, where the B is the number of samples in each mini-batch, T is the number of each level's T kernel, C is the number of channels, F^i is the length of the feature after i th level convolution operation. z^i_{conv} is defined as:

$$z^i_{conv} = AvgPool(ReLU(Conv1D(X, S^i_T))) \tag{2}$$

where the S^i_T is the T kernel size, X is the input raw EEG segments array, $Conv1D(\cdot)$ is the 1D convolution operation with the kernel size being S^i_T, step being (1,1).

The output of each level's T kernel will be concatenated. In order to reduce the internal covariate shift problems in neural networks, batch normalization [29] is added. Hence the final output of the temporal learner Z_T, $Z_T \in \mathbb{R}^{B \times T \times C \times \sum F^i}$ is defined as:

$$Z_T = f_{bn}([z^0_{conv}, ..., z^i_{conv}]) \tag{3}$$

where the f_{bn} is the batch normalization operation, $[\cdot]$ stands for concatenation operation along the feature (F) dimension.

C. Spatial Learner

The spatial learner has multi-scale 1D convolutional kernels whose sizes are related to the location of the EEG channels. There are three types of spatial kernels: global kernel, hemisphere kernel and local kernel. In order to apply three types kernels, the sequence of channels in the input EEG segments should be carefully arranged. The order of the channels should be $[channel_{left}, channel_{right}]$, where the $channel_{left}$ are the channels located at left hemisphere, the $channel_{right}$ are the ones on the right hemisphere. Let the input of spatial learner be $X = [x_0, x_1, ..., x_n], x_n \in \mathbb{R}^{C \times F}$, where n is the number of EEG segments, C is the number of channels, F is the length of the feature for each channel.

For the global kernel, it is the same as the ones in EEGNet [14], whose size is $(C, 1)$, where C is the number of channels. Since the length of the kernel is the same as the channel

dimension of the input EEG segment, it can get the global spatial relation pattern.

Inspired by EEGNet, we further combine the frontal area of brain emotional asymmetry [30] into the kernel design. The hemisphere kernel is used to extract the relation pattern between left and right hemispheres by sharing the convolutional kernels. The size of the hemisphere kernel is $(0.5 \cdot C, 1)$, and the step is $(0.5 \cdot C, 1)$, where C is the total number of channels. The hemisphere kernel is shared by two hemispheres without overlapping so that the asymmetry pattern can be extracted.

Further design of the local kernel has the same principle as the above kernels. We can define the sub-area of the brain surface according to functions, and the length of the local kernel would be the number of channels located in the sub-areas. In this work, only the global and hemisphere kernel are used. The local kernel will be considered in future research as a possible alternative.

D. Optimization of TSception

To optimize the network parameters, we adopt the back propagation method to iteratively update the network parameters until the desired criterion is achieved. Cross-entropy cost is used as the objective function. $L1$ Regularization term is added to keep the weights small, making the model simpler and avoiding over-fitting [31]. The final loss function is expressed as:

$$L_{\varepsilon}(y, \hat{y}) = L_{Cross-entropy}(y, \hat{y}) + \lambda \sum_{i=1}^{n} |\theta_i| \qquad (4)$$

where the y is ground truth label and the $\hat{y}$ is the predicted label. λ is the $L1$ regulation coefficient, θ_i is the i th weight of the model.

To overcome the over-fitting problem, we adopt early stopping in the training process. The stopping criterion is set as the validation accuracy stops increasing for certain epochs.

IV. Experiment

A. Data Acquisition

In order to study the emotional arousal in an immersive VR environment and evaluate the proposed algorithm, we collected EEG data from 18 healthy subjects (9 Males/9 Females, between the ages of 23-49) using a VR-BCI system. HTC VIVE pro is used as the VR device. Four channels (TP9, AF7, AF8, TP10) EEG data are collected using MUSE EEG headband [32] [33]. The sampling rate is 256 Hz. The experiments are conducted in an isolated room with soft illumination to avoid external disturbance. Subjects were seated in a comfortable armchair and instructed to avoid undesired movements. The experiment description and the tasks need to be achieved are described to the subject before the experiment. A demo session is added to let the subject be familiar with the system. After the experiment, a survey form will be given to the subject to get the feedback and their emotional state during the experiment.

The system is developed using Unity 3D development platform. There are 2 types of stimuli: low arousal and high arousal.

Algorithm 1: Training Procedure for TSception

Input: Raw EEG data X, ground truth label Y, model $TSception(\cdot)$, number of temporal kernels N_T, number of spatial kernels N_S, early stopping patient p

```
Initialization;
p_stop = 0;
acc_max = 0;
while p_stop < p do
    if acc_validation > acc_max then
        acc_max = acc_validation;
        p_stop = 0;
    else
        p_stop += 1;
    end
    Ŷ = TSception(X, N_T, N_S);
    loss = L_ε(Y, Ŷ);
    back_propagation(loss, optimizer = Adam);
end
Save model;
```

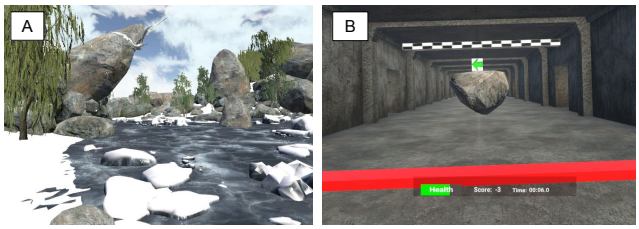

Fig. 2. Low arousal stimuli (A) and high arousal stimuli (B). In low arousal stimuli, there is a white bird flies low above the frozen lake in snowing weather, presented in a first-person perspective. For high arousal stimuli, a stone avoiding game is designed to induce high arousal to subject.

1) Low arousal stimuli: For this stimulus, as shown in Fig. 2 (A), there is a white bird flies low above the frozen lake in snowing weather, presented in a first-person perspective. The bird flies slow and elegantly, with the soft music played in the background. The design of this stimulus follows Bilgin *et. al.* [26], which indicates that the nature-based and low illumination environment can induce low emotional arousal effectively.

2) High arousal stimuli: Studies [34] [35] show that using complex input (two hands), making subject stressful and increasing difficulty properly can induce higher level arousal to the player. For this stimulus, a stone avoiding game is designed. As seen in Fig. 2 (B), there is a stone coming to the first-person perspective fast and randomly, with the audio effect of moving stone. There will be an arrow appearing when the stone reaches a certain distance from the subject. To avoid the stone, the subject needs to press the left handheld controller button if the arrow points left, and the right handheld controller button if the arrow points right. If the subject presses the correct button, the stone will disappear, the player's score

increases by 1. If the wrong button is pressed, the subject will see the stone hits the screen, and the score will be decreased by 1. In order to maintain the arousal level of subjects, adaptive difficulties mechanism [35] is adopted in the game. The more scores the subject gets, the faster the stone moves, and vice versa.

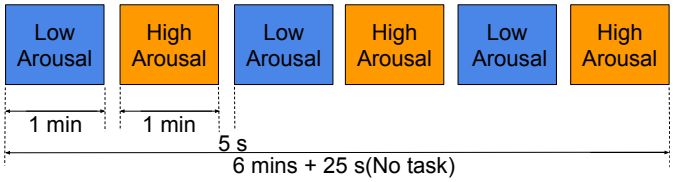

Fig. 3. Protocol of the experiment. There are 2 types of stimuli: low arousal and high arousal. For one experiment session, each stimulus lasts for 1 minute, between which there is 5 seconds' relaxing break. Every subject participated in 3 sessions in total.

The experiment protocol is shown in Fig. 3. Every subject participated in 3 sessions in total. There are 2 trials in each session. For one experiment session, each stimulus lasts for 1 minute, between which there is 5 seconds' relaxing break. The TABLE I summarizes the experiment information.

TABLE I
EXPERIMENT INFORMATION OF DATA ACQUISITION

Factor	Value
Stimuli	VR scenes and game-playing
Number of subjects	18
Number of males	9
Number of females	9
Range of age	23-49
Rating type	Emotional Arousal
Rating value	High/Low
Channels	TP9, AF7, AF8, TP10
Sampling rate	256 Hz
Duration of each subject	6 mins (3 mins high arousal, 3 mins low arousal)

This study, including the data acquisition, was approved by Institutional Review Board of Nanyang Technological University (NTU), Singapore [IRB-2018-12-011].

B. Signal Processing

For the collected data, a band-pass filter from 0.3 Hz to 45 Hz is applied to remove low and high-frequency noise. The electrooculography (EOG) is removed by using MNE open-source python software [36]. The processed data will be used for deep learning methods directly. For SVM, feature extraction is needed.

The EEG signal is band-pass filtered into multiple frequency bands, using zero-phase Chebyshev Type II filters [37]. A total of 9 band-pass filters are used, namely, 4-8 Hz, 8-12 Hz,..., 36-40 Hz. Then the relative power (RP) and DE in 9 frequency bands are used as the features of the SVM input. The RP in i th frequency band can be calculated by:

$$RP_i = \frac{\sum x_i^2}{\sum_j^n \sum x_i^2} \tag{5}$$

where x_i is the data in i th frequency band, n is the number of the frequency bands.

The DE is calculated as [17]:

$$DE = \frac{1}{2} log 2\pi e \sigma^2 \tag{6}$$

where the e is Euler's constant and σ is the standard deviation of x_i.

C. Parameter Setting

The PyTorch library [38] is used to implement the proposed model, the source code can be found via $website$ [1]

Parameters of TSception are selected empirically. There are 3 levels' temporal kernels whose corresponding ratio coefficients are $\alpha = [0.5, 0.25, 0.125]$. For each level, there are 9 convolutional kernels. For the spatial learner, the global and hemisphere kernels are used with 6 convolutional kernels in each type. The hidden node is set as 128 in the first fully connected layers.

For the training process, Adam optimizer is adopted, with the learning rate being 0.001. The size of the mini-batch is 128. The dropout rate and early stopping patient are 0.3 and 4 respectively. The L1 regulation coefficient λ is 1e-06.

D. Experiment Setting

The subject-dependent experiments are conducted. Since there are 3 sessions for each subject, a "Leave One Session Out" cross-validation strategy is applied. The average accuracy of all subjects and standard deviation are reported as the evaluation criterion. In order to get enough data segments to train the deep learning model better, the raw EEG data is split into 4 seconds' segments by a sliding window, whose moving step is 100 ms (25 data points). Hence 574 samples will be generated per stimuli. For each subject, there are 3 sessions, each session contains 2 stimuli. For leave one session out cross-validation, one session is used as the testing set, another 2 sessions are used as the training set. Among the training set, 80% of the training set is used for the training process, and the rest 20% is used as the validation set for early stopping. In one cross-validation step of each subject, the dimension of training set is (1836 x 1 x 4 x 1024), the one for validation set is (460 x 1 x 4 x 1024) and the dimension of testing set is (1148 x 1 x 4 x 1024). The raw data segments will be fed into deep learning methods directly. However for SVM, The manually extracted feature array is used.

V. RESULT AND ANALYSIS

The proposed model is compared with EEGNet, LSTM (Using 3 1D CNN layers as feature extractor followed by a 4 layers LSTM) [39], SVM using RP as features, SVM using DE as features respectively. For better evaluation of the proposed model, two simplified versions, namely Tception and Sception is added in the comparison. As the names show, Tception is achieved by removing spatial learner of TSception and Sception is the one without temporal learner.

[1] https://github.com/deepBrains/TSception

TABLE II
COMPARISON AGAINST CLASSIFICATION/STANDARD DEVIATION(%) ON SUBJECT-DEPENDENT EXPERIMENT WITH SVM(RP), SVM(DE), EEGNET, LSTM, TCEPTION, SCEPTION, TSCEPTION

Method	***ACC***	***STD***
SVM(RP)	80.73 **	8.51
SVM(DE)	82.23 *	9.07
EEGNet	79.96 *	11.47
LSTM	80.81 **	8.69
Tception	83.9	7.42
Sception	77.39	11.47
TSception	**86.03**	8.99

p-value between the method and TSception: * indicating ($p < 0.05$), ** indicating ($p < 0.01$).

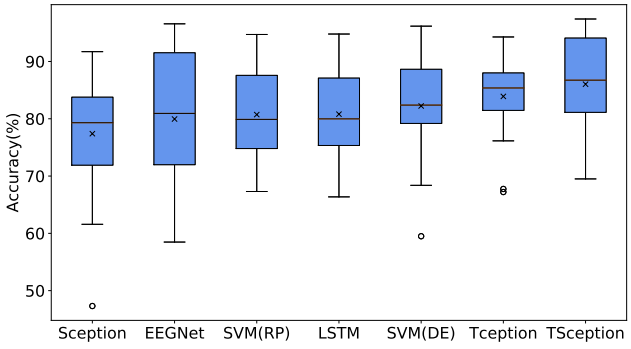

Fig. 4. Results of different methods for all subjects. The X-axis is the method, the Y-axis is the classification accuracy(%). 'x' is the mean of the accuracy for all the subjects. TSception gives the highest accuracy being 86.03%, followed by Tception (83.9%) and SVM using DE (82.23%). The acc of LSTM and SVM using RP are close to each other, being 80.81% and 80.73%. The EEGNet is better than Sception with the accuracy being 79.96% and 77.39% respectively.

TABLE II and Fig. 4 show the classification results for leave one session out cross-validation with SVM(RP), SVM(DE), EEGNet, LSTM, Tception, Sception, TSception.

As shown in the table, DE feature gives higher classification accuracy (82.23%) than RP feature (80.73%) for SVM classifier. For deep learning models, TSception gives the highest accuracy being 86.03%, followed by Tception (83.9%) and LSTM (80.81%). The EEGNet is better than Sception with the accuracy being 79.96% and 77.39% respectively. But both EEGNet and Sception give lower accuracy than other deep learning models, indicating more useful patterns are in temporal information than spatial one. EEGNet only has single size temporal kernels and the Sception only extracts the spatial pattern by 1D spatial kernels. Both of them can't extract the dynamic temporal information effectively, even have lower accuracy than SVM with DE feature in 9 frequency bands.

To better understand which part of the TSception contributes more to the classification results, a self-comparison among the TSception and its modified versions is conducted. The detailed structure parameters for TSception and its two simplified versions are shown in TABLE III.

As TABLE III shows, the TSception has less trainable parameters than another 2 models. Compared with Tception and Sception, the final feature vector is much shorter in TSception since it extracts both temporal and spatial patterns in a sequence operation. Hence it dramatically reduces the number of trainable parameters in fully connected layers. Since it extracts the pattern in temporal and spatial information, the classification accuracy is even higher than the other two, which have much more trainable parameters. As for the accuracy of proposed methods, TSception has the highest ACC among the three proposed methods, with 8.34% improvement over Sception and 2.4% improvement over Tception. From the results, the temporal learner contributes more than the spatial learner. Although the Sception gives the lowest accuracy among all the compared methods, the combination of temporal and spatial learner still gives the highest accuracy. There are two possible reasons: 1) the cross information among temporal and spatial helps to improve accuracy; 2) the sequential structure of using two types learners can decrease the parameters of the model significantly, which can overcome over-fitting problems better.

TABLE III
SELF-COMPARISON FOR TSCEPTION AND TWO SIMPLIFIED VERSION, TCEPTION AND SCEPTION

Method	Trainable Parameters	Number of temporal kernels	Number of spatial kernels	ACC(%)
Tception	822,671	9	- / -	83.9
Sception	147,902	- / -	6	77.39
TSception	**53,483**	9	6	**86.03**

VI. CONCLUSION

In this paper, we proposed TSception, a deep learning framework for EEG emotion classification. It uses multi-scale temporal and spatial convolutional kernels in temporal and spatial learners to learn more discriminative representations in the time and space domain simultaneously. The temporal learner extracts multi-frequency and multi-temporal pattern. The spatial learner takes advantage of the frontal area of brain emotional asymmetry, using hemisphere kernels to extract the information from the right and left hemispheres.

We collect EEG data from 18 healthy subjects in a VR-BCI system to study the emotional arousal in the immersive VR environment and evaluate the proposed algorithm. Compared with the state of art methods in BCI, such as SVM (RP), SVM (DE), EEGNet, LSTM together with two simple variants of TSception, TSception achieves the highest classification accuracy, being 86.03%. The proposed model can be applied in EEG signal classification generally due to its general structure. The code of TSception is also made open-access. Exploration for the potential ability of TSception will be included in the future work.

REFERENCES

[1] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, July 2019.

[2] M. Morena, K. D. Leitl, H. A. Vecchiarelli, J. M. Gray, P. Campolongo, and M. N. Hill, "Emotional arousal state influences the ability of amygdalar endocannabinoid signaling to modulate anxiety," *Neuropharmacology*, vol. 111, pp. 59 – 69, 2016.

[3] X. B. Lin, T.-S. Lee, Y. B. Cheung, J. Ling, S. H. Poon, L. Lim, H. H. Zhang, Z. Y. Chin, C. C. Wang, R. Krishnan, and C. Guan, "Exposure therapy with personalized real-time arousal detection and feedback to alleviate social anxiety symptoms in an analogue adult sample: Pilot proof-of-concept randomized controlled trial," *JMIR Ment Health*, vol. 6, no. 6, p. e13869, Jun 2019.

[4] A. Tseng, Z. Wang, Y. Huo, S. Goh, J. A. Russell, and B. S. Peterson, "Differences in neural activity when processing emotional arousal and valence in autism spectrum disorders," *Human Brain Mapping*, vol. 37, no. 2, pp. 443–461, 2016.

[5] R. D. Lane, L. Ryan, L. Nadel, and L. Greenberg, "Memory reconsolidation, emotional arousal, and the process of change in psychotherapy: New insights from brain science," *Behavioral and Brain Sciences*, vol. 38, p. e1, 2015.

[6] S. K. Ehrlich, K. R. Agres, C. Guan, and G. Cheng, "A closed-loop, music-based brain-computer interface for emotion mediation," *PLOS ONE*, vol. 14, no. 3, pp. 1–24, 03 2019.

[7] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *Journal of Neural Engineering*, vol. 16, no. 3, p. 031001, apr 2019.

[8] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Systems with Applications*, vol. 47, pp. 35 – 41, 2016.

[9] W. Zheng, J. Zhu, and B. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 417–429, July 2019.

[10] P. Li, H. Liu, Y. Si, C. Li, F. Li, X. Zhu, X. Huang, Y. Zeng, D. Yao, Y. Zhang, and P. Xu, "EEG based emotion recognition by combining functional connectivity network and local activations," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2869–2881, Oct 2019.

[11] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[12] O. Kwon, M. Lee, C. Guan, and S. Lee, "Subject-independent brain-computer interfaces based on deep convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2019.

[13] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *Journal of Neural Engineering*, vol. 14, no. 1, p. 016003, nov 2016.

[14] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, Jul 2018.

[15] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain-computer interface using convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5619–5629, Nov 2018.

[16] Y. Yang, Q. M. J. Wu, W. Zheng, and B. Lu, "EEG-based emotion recognition using hierarchical network with subnetwork nodes," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 2, pp. 408–419, June 2018.

[17] J. Li, Z. Zhang, and H. He, "Hierarchical convolutional neural networks for EEG-based emotion recognition," *Cognitive Computation*, vol. 10, no. 2, pp. 368–380, Apr 2018.

[18] X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou, and B. Hu, "Exploring EEG features in cross-subject emotion recognition," *Frontiers in Neuroscience*, vol. 12, p. 162, 2018.

[19] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[20] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan, "Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI," *Journal of Neural Engineering*, vol. 16, no. 2, p. 026007, Jan 2019.

[21] Z. Jiao, X. Gao, Y. Wang, J. Li, and H. Xu, "Deep convolutional neural networks for mental load classification based on EEG data," *Pattern Recognition*, vol. 76, pp. 582 – 595, 2018.

[22] Z. Gao, X. Wang, Y. Yang, C. Mu, Q. Cai, W. Dang, and S. Zuo, "EEG-based spatio-temporal convolutional neural network for driver fatigue evaluation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2755–2763, 2019.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[24] J. J. B. Allen, P. M. Keune, M. Schönenberg, and R. Nusslock, "Frontal EEG alpha asymmetry and emotion: From neural underpinnings and methodological considerations to psychopathology and social cognition," *Psychophysiology*, vol. 55, no. 1, p. e13028, 2018.

[25] A. Felnhofer, O. D. Kothgassner, M. Schmidt, A.-K. Heinzle, L. Beutl, H. Hlavacs, and I. Kryspin-Exner, "Is virtual reality emotionally arousing? investigating five emotion inducing virtual park scenarios," *International Journal of Human-Computer Studies*, vol. 82, pp. 48 – 56, 2015.

[26] P. Bilgin, K. Agres, N. Robinson, A. A. P. Wai, and C. Guan, "A comparative study of mental states in 2D and 3D virtual environments using EEG," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Oct 2019, pp. 2833–2838.

[27] N. Robinson, S. Lee, and C. Guan, "EEG representation in deep convolutional neural networks for classification of motor imagery," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Oct 2019, pp. 1322–1326.

[28] S. Sakhavi and C. Guan, "Convolutional neural network-based transfer learning and knowledge distillation using multi-subject data in motor imagery BCI," in *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*, May 2017, pp. 588–591.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv 1502.03167*, 2015.

[30] A. B. Craig, "Forebrain emotional asymmetry: a neuroanatomical basis?" *Trends in Cognitive Sciences*, vol. 9, no. 12, pp. 566 – 571, 2005.

[31] E. Tartaglione, S. Lepsø y, A. Fiandrotti, and G. Francini, "Learning sparse neural networks via sensitivity-driven regularization," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 3878–3888.

[32] J. Amores, R. Richer, N. Zhao, P. Maes, and B. M. Eskofier, "Promoting relaxation using virtual reality, olfactory interfaces and wearable EEG," in *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, March 2018, pp. 98–101.

[33] G. Wiechert, M. Triff, Z. Liu, Z. Yin, S. Zhao, Z. Zhong, R. Zhaou, and P. Lingras, "Identifying users and activities with cognitive signal processing from a wearable headband," in *2016 IEEE 15th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, Aug 2016, pp. 129–136.

[34] M. Lankes, W. Hochleitner, C. Hochleitner, and N. Lehner, "Control vs. complexity in games: Comparing arousal in 2D game prototypes," in *Proceedings of the 4th International Conference on Fun and Games*, ser. FnG '12. Association for Computing Machinery, 2012, p. 101–104.

[35] H. Qin, P.-L. P. Rau, and G. Salvendy, "Effects of different scenarios of game difficulty on player immersion," *Interacting with Computers*, vol. 22, no. 3, pp. 230–239, 12 2009.

[36] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen, "MNE software for processing MEG and EEG data," *NeuroImage*, vol. 86, pp. 446 – 460, 2014.

[37] Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, June 2008, pp. 2390–2397.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[39] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312 – 323, 2019.