TFF-Former: Temporal-Frequency Fusion Transformer for Zero-training Decoding of Two BCI Tasks

Xujin Li
Research Center for Brain-inspired Intelligence & National Laboratory of Pattern Recognition, CASIA
University of Chinese Academy of Sciences
Beijing, China
lixujin2021@ia.ac.cn

Wei Wei
Research Center for Brain-inspired Intelligence & National Laboratory of Pattern Recognition, CASIA
Beijing, China
weiwei2018@ia.ac.cn

Shuang Qiu
Research Center for Brain-inspired Intelligence & National Laboratory of Pattern Recognition, CASIA
University of Chinese Academy of Sciences
Beijing, China
shuang.qiu@ia.ac.cn

Huiguang He*
Research Center for Brain-inspired Intelligence & National Laboratory of Pattern Recognition, CASIA
University of Chinese Academy of Sciences
Beijing, China
huiguang.he@ia.ac.cn

ABSTRACT

Brain-computer interface (BCI) systems provide a direct connection between the human brain and external devices. Visual evoked BCI systems including Event-related Potential (ERP) and Steady-state Visual Evoked Potential (SSVEP) have attracted extensive attention because of their strong brain responses and wide applications. Previous studies have made some breakthroughs in within-subject decoding algorithms for specific tasks. However, there are two challenges in current decoding algorithms in BCI systems. Firstly, current decoding algorithms cannot accurately classify EEG signals without the data of the new subject, but the calibration procedure is time-consuming. Secondly, algorithms are tailored to extract features for one specific task, which limits their applications across tasks. In this study, we proposed a Temporal-Frequency Fusion Transformer (TFF-Former) for zero-training decoding across two BCI tasks. EEG data were organized into temporal-spatial and frequency-spatial forms, which can be considered as two views. In the TFF-Former framework, two symmetrical Transformer streams were designed to extract view-specific features. The cross-view module based on the cross-attention mechanism was proposed to guide each stream to strengthen common representations of features across EEG views. Additionally, an attention-based fusion module was built to fuse the representations from the two views effectively. The mean mask mechanism was applied to adaptively decrease redundant EEG tokens aggregation for the integration of common representations. We validated our method on the self-collected RSVP dataset and benchmark SSVEP dataset. Experimental results demonstrated that our TFF-Former model achieved competitive performance compared with models in each of the above paradigms. It can further promote the application of visual evoked EEG-based BCI system.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing**;

KEYWORDS

Brain-computer Interface (BCI), Zero-training, Transformer, Event-related Potential (ERP), Steady-state Visual Evoked Potential (SSVEP)

ACM Reference Format:
Xujin Li, Wei Wei, Shuang Qiu, and Huiguang He. 2022. TFF-Former: Temporal-Frequency Fusion Transformer for Zero-training Decoding of Two BCI Tasks. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548269

1 INTRODUCTION

The Brain-computer interface (BCI) system collects and decodes brain neural activity information to build a direct information interaction pathway between the brain and the external machine, which has shown a variety of potential applications for communication, control, and rehabilitation [1]. Electroencephalogram (EEG) -based BCIs have attracted extensive attention because of their non-invasive technology, high temporal resolution and low equipment cost [2].

Visual evoked BCI systems are widely used because of their strong brain responses and wide applications. Representative visual evoked EEG-based BCI paradigms include Event-related Potential (ERP) and Steady-state Visual Evoked Potential (SSVEP). ERP is a special type of evoked potential generated by multiple stimuli with particular psychological meaning. Rapid Serial Visual Presentation (RSVP) paradigms induce specific event-related potential (ERP) components which can be used to recognize sonar images to detect some objects [3] and implement speller [4] [5], image

*Dr. Huiguang He is the corresponding author and he is also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3548269

retrieval system [2], image classification system [6] [7] and abnormal pattern recognition system [3]. SSVEPs are neural oscillations from the parietal and occipital regions of the brain that are evoked from flickering visual stimuli and have shown good performance in various applications, such as spellers [8], [9], device control [10], and games [11] [12].

Previous studies have proposed several classic and state-of-the-art methods with calibration, which required training data were obtained from the same subject on the same day as test data. These methods can be divided into conventional methods and deep learning methods. In terms of ERP decoding methods, the conventional methods include the Hierarchical Discriminant Component Analysis (HDCA) [6] and Minimum Distance to Riemannian Mean (MDRM) [13] in RSVP task. With the great breakthrough of Convolutional Neural Network (CNN) in image classification tasks [14], CNN have been proven to be useful in decoding EEG signals of RSVP tasks [15] such as MCNN [16], EEGNet [17], and OCLNN [18] in RSVP task. However, the above methods require training data obtained from the same subject on the same day as test data and a large amount of calibration data is required to improve decoding performance [19], which is a time-consuming procedure and limits the application of BCI. We call this kind of training methods with calibration the within-subject method and methods without calibration the zero-training method. Recently, Lee proposed a CNN-based model training on a large amount of data from other subjects [20], which showed no statistical difference compared to within-subject methods. Therefore, this demonstrated the potential of zero-training approach in ERP signals decoding. In SSVEP task, the original methods including Canonical Correlation Analysis (CCA) [21] and Filter Bank CCA (FBCCA) [22] belong to zero-training methods. Many within-subject methods had also been proposed such as multi-way CCA [23], multi-set CCA [24], Task-Related Component Analysis (TRCA), ensemble TRCA (eTRCA) [25] Compact-CNN [26], and Conv-CA [27]. Although they can achieve higher classification accuracies in SSVEP task, they still have the disadvantage of requiring individual calibration data.

Although these previous studies can improve the decoding performance to a certain extent, they still have the following bottlenecks: 1). There are great differences in EEG data among different subjects, which is a major difficulty in achieving high performance in zero-training. 2). The design of model structure is tailored to extract features for one specific task, so that the models will ignore the features related to other paradigms and cause the extracted features to be not robust, which limits their applications across tasks. Inspired by the multimodal pre-training transformers, which can achieve competitive performance on downstream tasks even without fine-tuning. One of the main reasons is that transformers can be trained on large dataset to learn common representations of data. We aimed to build a novel framework based on transformer which can learn robust common features between subjects (i.e. task-related features) from the data of many different subjects and can be applied to zero-training tasks in different paradigms. In addition, the multi-view learning of temporal-frequency fusion was also applied to use multidomain information of EEG signals, which can realize the complementary information of temporal and frequency views to further improve the accuracy and thus improve the performance in the zero-training task. Transformer is an attention-based structure whose self-attention has been shown to effectively learn global interactions [28] [29], which can adjust on the local content while modeling global relationships. Meanwhile, Multimodal Transformer [30] can achieve interactions between multimodal sequences across latently adaptive streams from one modality to another, which can be used to achieve the feature fusion between temporal information and frequency information. Therefore, Transformer has the potential to be used to decode EEG signals.

In this paper, we built a Temporal-Frequency Fusion Transformer (TFF-Former) to realize zero-training decoding for both RSVP and SSVEP tasks. Data are organized into two views including raw EEG signals as temporal view and frequency-spectrum data after Fast Fourier transformation as frequency view. The model is composed of two symmetrical Transformer streams which are designed to extract view-specific features. Firstly, we fed temporal and frequency data slices into the model, each of them represented the information of brain activity over a period of time or a frequency band. Secondly, we used encoder layers sharing parameters between the two streams to extract view-specific features and projected them to the same feature space, which can not only reduce parameters, but also implement subsequent view interaction in the same space to maximize the extraction of common features. To fuse the representations from two views effectively, we also proposed attention-based cross-view module and fusion module. Finally, the decision module was used to predict labels. We make the following contributions in this work:

(1) We proposed a zero-training EEG decoding model Temporal-Frequency Fusion Transformer (TFF-Former) which is appropriate for both RSVP and SSVEP tasks. To our knowledge, this is the first EEG temporal-frequency fusion framework based on transformer.

(2) The TFF-Former learn to extract subject-invariance and task-related features by training with labeled data of many existing subjects. We also proposed the cross-view module to guide each stream to strengthen common representations across views and the mean masked attention mechanism to increase the proportion of relevant tokens in aggregation and ignore irrelevant information.

(3) We conducted a lot of experiments to verify the zero-training performance of TFF-Former, which has a significant improvement over the compared methods in two datasets. On the self-collected RSVP dataset [31], the performance of TFF-Former is equivalent to that of the calibration methods with four-block data. On the SSVEP public dataset [32], the performance of TFF-Former is better than other zero-training methods. The code has been released at: https://github.com/lixujin1999/TFF-Former.

2 RELATED WORK

We briefly review previous studies which are divided into RSVP task studies, SSVEP task studies and the studies of decoding physiological signals that introduce Transformer models.

RSVP Task In 2006, Gerson, A.D et al. proposed HDCA, which introduced linear discrimination to reveal target images' differences [6]. Barachan et al. (2014) proposed the MDRM [13] method to classify the covariance matrices transformed from original EEG data according to the minimum distance to mean. In 2018, Lawhern Vernon J et al. proposed EEGNet [17] with depthwise separable convolution [33] which can simplify the structure and reduce the

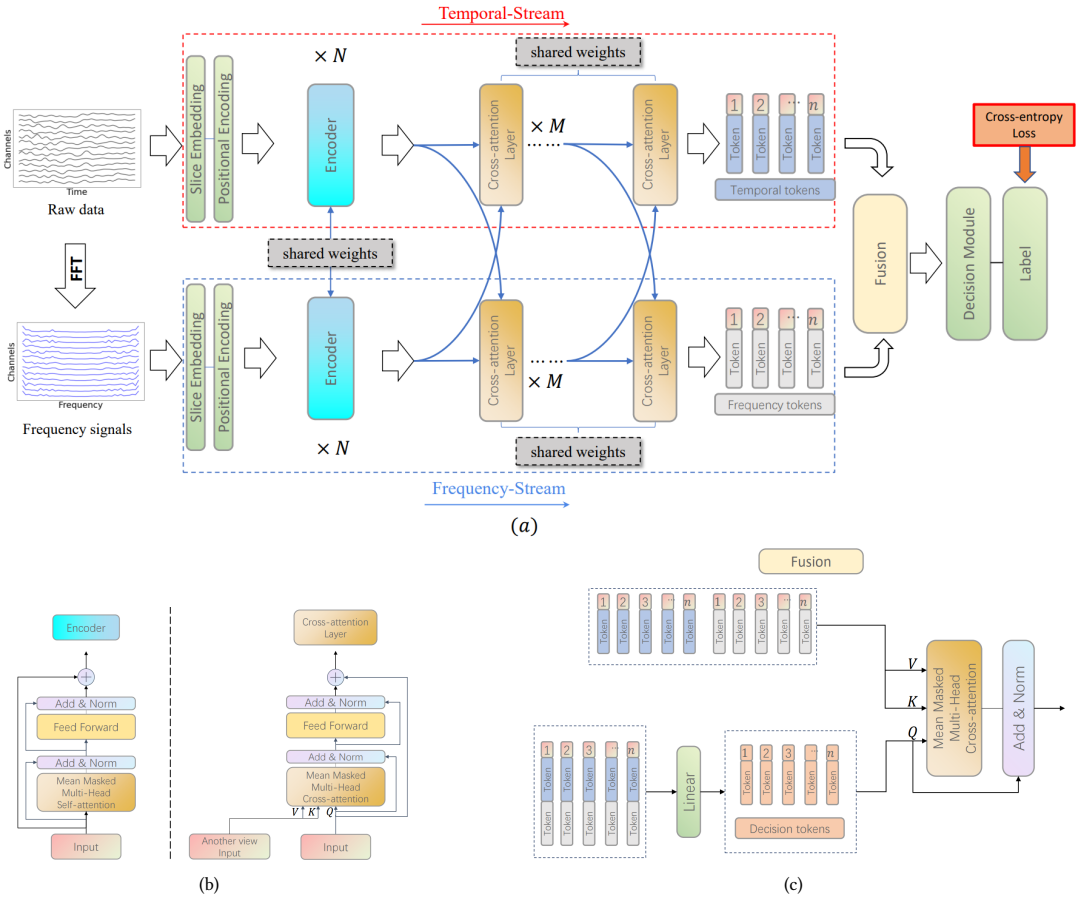


Figure 1: The structure of the proposed TFF-Former model. (a).TFF-Former is composed of two symmetrical Transformer streams including a feature extractor and a cross-view module followed by a fusion module and a decision module. (b).The structure of the fusion module. (c).The structure of encoder layer and cross-attention layer.

number of parameters. Shan et al. proposed One Convolution Layer Neural Network (OCLNN) [18] in P300 speller which achieved high generalization performance. In 2020, Lee et al. proposed a zero-training model [20] based on EEGNet for P300 BCI speller which showed no statistical difference when compared to the within-subject methods. In 2021, Zang et al. proposed PLNet which used the phase-locked characteristic to extract spatiotemporal features for classification [34]. Meanwhile, a phase preservation neural network (PPNN) was also proposed to learn phase information which also improved the classification performance in RSVP task [35].

SSVEP Task In 2006, CCA [21] was developed and adopted in SSVEP classification tasks, which can efficiently identify the targets in the time domain. In 2015, Chen et al. built FBCCA [22] to decompose signals into several sub-band components and fuse the classification results from all sub-bands, which obtained higher classification accuracies than CCA. Compared with conventional zero-training methods, involving individual calibration data can achieve higher classification accuracies in SSVEP tasks such as TRCA and eTRCA [25]. In 2018, Waytowich et al. proposed Compact CNN to extract temporal and spatial feature separately [26] which outperformed CCA on a 12-class SSVEP classification task.

Transformer-based Method In recent two years, Transformers have been used to classify physiological signals. In emotion recognition, Emotion Transformer Fusion (ETF) model has been proposed for emotion recognition, which proved the capability of Transformer based architecture on multimodal emotion recognition with EEG and eye movement signals [36]. In 2022, Wang et al. also proposed a transformer-based model to hierarchically learn the discriminative spatial information from a brain-region-level [37] for the same task. In order to classify the sleep stage in awake or asleep, a network based on Temporal Convolutional Network and Transformer using only HR signals [38] was proposed. However, there have been no studies which apply the Transformers to the ERP and SSVEP classification in zero-training. Our proposed TFF-Former model is a zero-training framework using multi-view fusion Transformer for visual evoked signals decoding.

3 METHOD

In this study, we proposed a Temporal-Frequency Fusion Transformer (TFF-Former) model for zero-training decoding of RSVP and SSVEP tasks. Figure. 1.(a) demonstrates the architecture of TFF-Former.

Our method is a two-stream symmetrical Transformer architecture including a temporal stream and a frequency stream. The raw

EEG signals ($S_t \in \mathbb{R}^{C\times T}$) and the corresponding frequency-spectral signals ($S_f \in \mathbb{R}^{C\times T}$) which can be considered as two views were fed temporal stream and frequency stream respectively. Each stream had a feature extractor and a cross-view module. The feature extractor included a slice embedding layer and N successive Transfomer encoder layers in each stream sharing parameters between the two streams. Then the extracted temporal and frequency features were sent to a cross-view module which consists of M successive cross-attention layers sharing parameters in each stream to improve generalization performance with insufficiently large EEG datasets. Meanwhile, an attention-based fusion module was also proposed to achieve the interaction and complementary fusion between EEG temporal and frequency features. In addition, in order to reduce redundant aggregation for the integration of common representations, we employed the mean mask operation in each multi-head attention mechanism in TFF-Former. Finally, the fusion features obtained by the fusion module were sent to a decision module to obtain the probability of different classes. The TFF-Former model was trained end-to-end using cross-entropy loss.

3.1 Multi-Head Mean Masked Attention

The mean masked attention mechanism (See Figure. 2.[39]) consists of a linear projection layer and a mean masked attention layer. The linear projection layer maps input sequences $\boldsymbol{X} \in \mathbb{R}^{n_x\times d_{model}}$, $\boldsymbol{Y} \in \mathbb{R}^{n_y\times d_{model}}$ to three different sequential vectors (query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$), which are generated as:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{Q}}, \qquad \boldsymbol{K} = \boldsymbol{Y}\boldsymbol{W}_{\boldsymbol{K}}, \qquad \boldsymbol{V} = \boldsymbol{Y}\boldsymbol{W}_{\boldsymbol{V}}$$

where n_x, n_y and d_{model} are the length and dimension of the input sequences of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively and $\boldsymbol{W}_{\boldsymbol{Q}} \in \mathbb{R}^{d_{model}\times d_k}, \boldsymbol{W}_{\boldsymbol{K}} \in \mathbb{R}^{d_{model}\times d_k}, \boldsymbol{W}_{\boldsymbol{V}} \in \mathbb{R}^{d_{model}\times d_v}$ are linear matrices, d_k is dimension of query ($\boldsymbol{Q}$) and key ($\boldsymbol{K}$), and d_v is dimension of the value ($\boldsymbol{V}$). The mechanism is called the self-attention mechanism when the query, key and value are all projected from $\boldsymbol{X}$.

The mean masked attention layer computes attention weights using the query and the corresponding key, then assigns them to the value after mean masked to update the output vector. We can formulate the process into a function as:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax\left[MeanMask\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\right]\boldsymbol{V}$$

where the attention weights are generated by a dot-product operation between the query and the key. Comparing each element of attention matrix with the mean value of the row where the element is located, we set the element which is smaller than the mean value as negative infinity. In this way, we mask tokens that are uncorrelated to the query tokens to increase the weights of relevant tokens and adaptively decrease redundant aggregation.

To tackle the issue that the modeling capability of a single-head attention block is coarse, we use multi-head mean masked attention mechanism that linearly projects the input into multiple feature subspaces with dimension $d_k = d_{model}/h$ where h denotes the number of heads and processes them by several independent mean masked attention heads parallelly. The resulting vectors are concatenated and mapped to the final output with dimensions d_{model}.

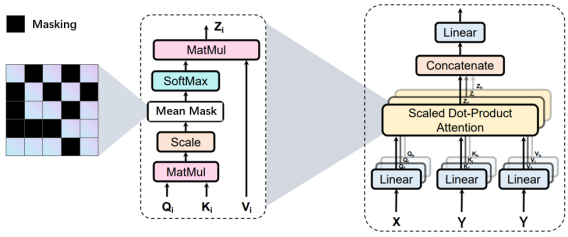


Figure 2: The structure of the mean masked attention layer: Scaled Dot-Product Attention and Multi-Head Mean Masked Attention Mechanism.

3.2 Feature Extractor

In order to extract view-specific features before view interaction, we designed the feature extractor which can be grouped into two parts which are the slice embedding layer and encoder layers. To handle 2D EEG signals, we sliced and reshaped the input signals $\boldsymbol{S_t}, \boldsymbol{S_f} \in \mathbb{R}^{C\times T}$ into a sequence of flattened 2D slices $\boldsymbol{x_s}, \boldsymbol{x_f} \in \mathbb{R}^{n\times(C\times t)}$ each represented brain activities in a period of time or a frequency band respectively, where C denotes the number of channels, T denotes the number of sampling points, t is the slice length and $n = [T/t]$ is the number of slices, which also served as the input sequence length for the encoder layers. The TFFformer used constant latent vector size d_{model} through all of its layers, so we flattened the slices and map to d_{model} dimensions with a trainable linear projection:

$$\boldsymbol{X} = [\boldsymbol{x}^1; \boldsymbol{x}^2, \cdots; \boldsymbol{x}^n]^T\boldsymbol{E} + \boldsymbol{E}_{pos}$$

$$\boldsymbol{E} \in \mathbb{R}^{(C\times t)\times d_{model}}, \boldsymbol{E}_{pos} \in \mathbb{R}^{n\times d_{model}}$$

where the output of these EEG slice embeddings were refered as EEG tokens. Learnable positional parameters ($\boldsymbol{E}_{pos}$) were added to the slice embeddings to retain positional information. The resulting sequence of embedding vectors serves as the input to the encoder layers. In our model, we adopt the slice length (t) as 4.

An encoder layer is composed of two sub-layers (See Fig. 1.(b)). A multi-head mean masked self-attention (MHMSA) layer aggregates the relationship within input tokens. A position-wise Feed-Forward Network (FFN) layer which is an MLP using one hidden layer with $h \times d_{model}$ hidden dimensions extracts feature representation. All of the sub-layers employ a residual connection and a Layer Normalization to enhance the scalability of Transformer. Moreover, we added a skip connection between the start and end of each encoder layer. Because the original temporal and frequency tokens belong to different feature spaces, there are semantic differences. Therefore, we shared the encoder parameters between the two streams to decrease the differences between views features, which maked the view interaction more reasonable and feasible in the subsequent cross-view module.

3.3 View Interaction

To realize view interaction and extract common representations of view features, we employed the cross-view module which consists of M successive cross-attention layers which replaced the first MHMSA layer in encoder layer with multi-head mean masked cross-attention (MHMCA) layer and the rest of the structure is identical (See Figure 1.(b)). In MHMCA layer, two views mutually guided

interaction to each other according to the relationship between their tokens to realize common feature extraction. In order to adapt to the fact that the EEG dataset does not have massive samples, M cross-view attention layers in the same stream sharing parameters to improve the generalization performance of the model and we also added a skip connection between the start and the end of M cross-view attention layers to strengthen feature propagation and feature reuse.

After cross-view module, we proposed an attention-based fusion module (See Figure 1.(c)) after cross-attention layers to achieve complementary feature fusion by fusing the temporal and frequency tokens interacted across views into the high-order decision tokens. Firstly, We token-wise concatenated the two view tokens together and linear projected them into the d_{model} dimension as the original decision fusion tokens. Then we set them as query and set temporal and frequency tokens both as key and value to feed a MHMCA layer to obtain the decision tokens. The process can be formulated as follows:

$$\boldsymbol{X}_{dec} = Add\&LN[Attention([\boldsymbol{X}_t;\boldsymbol{X}_f]\boldsymbol{W}, [\boldsymbol{X}_t^T;\boldsymbol{X}_f^T]^T, [\boldsymbol{X}_t^T;\boldsymbol{X}_f^T]^T)]$$

where $Add\&LN$ denotes residual connection and layernormalization, $\boldsymbol{W} \in \mathbb{R}^{2d_{model}\times d_{model}}$ is a learnable parameter matrix and $\boldsymbol{X}_{dec} \in \mathbb{R}^{n\times d_{model}}$, $\boldsymbol{X}_t \in \mathbb{R}^{n\times d_{model}}$,$\boldsymbol{X}_f \in \mathbb{R}^{n\times d_{model}}$ denotes the decision tokens, temporal tokens and frequency tokens respectively.

3.4 Decision Module

At the end of the model, we designed a decision module which consists of a convolution layer using 16 convolution kernels to aggregate all tokens and a linear layer with softmax activation function to classify the features. The convolution kernel size is $(16, d_{model}/8)$ with stride $(16, d_{model}/8)$ and zero padding. All the activation functions in the model adopt ReLU. The loss function used cross entropy loss:

$$L_{cross\_entropy} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{r=1}^{R} y_{n,r}\log\hat{y}_{n,r}$$

where R denotes the number of classes. y indicates the real label and $\hat{y}$ is the value predicted by the model.

4 EXPERIMENTS

4.1 Datasets

Our proposed TFF-Former model and the compared methods are validated on two EEG datasets: 1) self-collected RSVP dataset [31]. 2) the benchmark dataset [32] for SSVEP tasks.

RSVP Dataset: The experiment included 31 participants (19 males and 12 females; aged 24.9 ± 2.8, 28 right-handed). The visual stimuli for our experiment included 1,400 images (500×500 pixels) from the scene and object database [40] published by MIT CSAIL. These images were divided into target images with pedestrians and non-target images without pedestrians. Images were randomly presented at a frequency of 10 Hz, where the probability of the target image appearance was 4%. Each experimental session had 10 blocks, and each block contained 1400 images, divided into 14 sequences.

Benchmark Dataset: This SSVEP dataset has 35 subjects (17 females, aged 17-34 years), including 40 targets. The 40 targets were coded using the JFPM method. The frequencies range from 8 Hz to 15.8 Hz with an interval of 0.2 Hz, and the phase difference between two adjacent targets was 0.5 π. For each subject, the experiment included 6 blocks, and each block contained 40 trials corresponding to all targets indicated once in random order. In the public dataset, the trial length of 6 s includes 0.5 s before stimulus onset, 5 s for stimulation, and 0.5 s after stimulus offset.

4.2 Data Preprocessing

In the preprocessing stage, the RSVP dataset were down-sampled to 250 Hz. After that, a linear phase 3-order ButterWorth filter with a bandpass between 0.5 and 15 Hz is used to filter the signal to remove slow drift and high-frequency noise and prevent delay distortions. Then the preprocessed data of each block were segmented into EEG trials each containing 1 second EEG data. For each trial, data was normalized to zero mean and variance one. The subsequent analysis and classification of EEG were based on these segmented EEG trials (samples). According to our experimental paradigm, each subject had 10 (blocks) ×1400 (trials) EEG samples per session, where 560 are target samples and the rest are non-target samples. The SSVEP recordings were passed through a Chebyshev Type I band-pass filter with the range of 8 Hz to 90 Hz. We applied a notch filter at 50 Hz to remove the common powerline noise and also normalized each trial to zero mean and variance one. We used data from 0.64 to 2.64 seconds in each trial which contains 500 sampling points.

We employed the Fast Fourier transform (FFT) on each channel of raw data ($\boldsymbol{S_t} \in \mathbb{R}^{C\times T}$) and organized these frequency-spectrum signals as the frequency view input ($\boldsymbol{S_f} \in \mathbb{R}^{C\times T}$). The calculation was implemented based on the Python package numpy.

4.3 Experimental Setup

We conducted zero-training experiment in a Leave-One-Subject-Out (LOSO) way. Each subject will be as the test set alone and the rest as the training set. Especially in RSVP tasks, to overcome the influence made by the extreme imbalance of two classes, we adopt resampling. Down-sampling the non-target class to the same number as the target class. This operation is limited to the training set.

For convenience, we set $d_v = d_k$ in the mean masked attention mechanism. Therefore, there are four hyperparameters in our model, which are the embedding dimension (d_{model}), the number of heads (h) in multi-head attention mechanism, the number of successive encoder layers (N) and the number of cross-attention layers (M). Since we do not have massive data to feed a huge model, we appropriately reduced the size of the model. The parameters in RSVP dataset are set as follows:

$$d_{model} = 128 \qquad h = 4 \qquad N = 1 \qquad M = 2$$

In the benchmark dataset we use the same position encoding as [39] and set M to 1. We use PyTorch framework. The overall network is trained by minimizing the cross-entropy loss function. Adam optimizer is adopted for model optimization and the learning rate is 0.0005 in RSVP task and 0.001 in SSVEP task with a 20% decrease every 40 epochs. The L2 regularization is adopted, and the weight decay coefficient is 0.01. In SSVEP task we also used label smoothing regularization with $\epsilon = 0.005$. The batch size is set to 64 and the maximum number of training epochs is set to 100.

Table 1: CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON RSVP DATASET (mean ± standard deviation)

Method	BA(%)↑	TPR(%) ↑	FPR(%)↓	Accuracy(%)↑	F1-score↑
HDCA	82.60 ±4.48 ***	**80.48 ± 9.14** ***	15.29±2.93 ***	84.54 ± 2.76 ***	0.2979 ± 0.0445 ***
MDRM	**82.69 ± 6.84** ***	78.63±8.86 ***	**13.25 ± 3.72** ***	**86.43 ± 3.35** ***	**0.3217 ± 0.0600** ***
OCLNN	84.61± 4.05 ***	81.93 ± 8.06 ***	12.70 ± 0.97 ***	87.08 ± 0.97 ***	0.3369±0.0320 ***
MCNN	83.05 ± 4.95 ***	79.16±10.77 ***	13.05±2.59 ***	86.65 ± 2.33 ***	0.3241±0.0413 ***
EEGNet	79.99±3.97 ***	78.88±6.00 ***	18.89±1.27 ***	81.02 ± 1.39 ***	0.2501±0.0265 ***
Lee	85.03±4.70 ***	80.60±9.90 ***	**10.54 ± 3.23** **	**89.11 ± 3.00** **	**0.3823 ± 0.0562** ***
PLNet	79.38±4.90 ***	78.21±9.17 ***	19.45±2.48 ***	80.46 ± 2.46 ***	0.2446±0.0354 ***
PPNN	**85.48 ± 3.69** ***	**83.80 ± 7.34** *	12.83±1.27 ***	87.03 ± 1.24 ***	0.3416±0.0316 ***
Wang	86.33±4.38 ***	83.06±8.56 **	10.38±2.35 **	89.36 ± 2.62 *	0.3887±0.0521 ***
TCN-T	**86.68 ± 4.23** ***	**83.51 ± 8.47** **	**10.16 ± 2.20** *	**89.59 ± 2.09**	**0.3948 ± 0.0498** **
TFF-Former	**88.05 ± 3.73**	**85.45 ± 7.85**	**9.34 ± 2.26**	**90.46 ± 2.09**	**0.4223 ± 0.0484**

The asterisks in the table indicate significant difference between TFF-Former and the compared method by paired t-tests ($*p < 0.05, **p < 0.01, ***p < 0.001$).

4.4 Compared Methods

We compared our proposed network with the following methods:

- RSVP Task:
  - Conventional Methods: HDCA [6], MDRM [13].
  - CNN based Methods: OCLNN [18], MCNN [16], EEGNet [17], Lee [20], PLNet [34], PPNN [35].
  - Transformer based Methods: TCN-T [38], Wang [37].
- SSVEP Task:
  - Conventional Methods: CCA [21], FBCCA [22].
  - CNN based Methods: Compact-CNN [26].
  - Transformer based Methods: TCN-T [38], Wang [37].

These methods are all used for zero-training compared methods, and the experimental setup is consistent with the description in 4.3. Meanwhile, HDCA, MDRM, OCLNN, MCNN, EEGNet, PLNet and PPNN are also used for within-subject compared methods in RSVP dataset and the training setting is as follows: for each subject who has 10 blocks data, the data of b blocks are selected as the training set, and the data of the remaining $10 - b$ blocks are used as the test set ($b = 1, 2, 3, 4$).

4.5 Evaluation Metrics

We used balanced-accuracy (BA), true positive rate (TPR), false positive rate (FPR), accuracy and $F1$-score to evaluate model performance in RSVP task and accuracy in SSVEP task. The results are expressed as mean ± standard deviation for all test subjects. The calculation formulas are as follows:

$$\begin{cases} BA = \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)/2 \\ TPR = \frac{TP}{TP+FN} \\ FPR = \frac{FP}{TN+FP} \\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \\ precision = \frac{TP}{TP+FP} \\ F1 = \frac{2\times precision\times TPR}{precision+TPR} \end{cases}$$

Where TP represents the number of correctly classified positive samples, and FN represents the number of incorrectly classified positive samples. TN represents the number of correctly classified negative samples, and FP represents the number of incorrectly classified negative samples.

5 RESULTS AND DISCUSSION

The zero-training comparative experiment was conducted on RSVP dataset and SSVEP benchmark dataset. Moreover, in order to evaluate the effectiveness of TFF-Former compared to within-subject methods, we also conducted within-subject comparative experiment on RSVP dataset. Then we conducted ablation experiments to verify whether each module in TFF-Former contributed to classification. Finally, we visualized the latent features in various layers of TFF-Former and the cosine similarity matrix between temporal and frequency tokens before and after the cross-view module.

5.1 Classification Performance

The zero-training classification performance under various metric on the RSVP dataset are summarized in Table 1. The one-way repeated measures ANOVA showed a significant main effect of method in all evaluation metrics (BA : $F(10, 300) = 52.23, p < 0.001$, TPR : $F(10, 300) = 9.10, p < 0.001$, FPR : $F(10, 300) = 92.609$, $p < 0.001$, Accuracy : $F(10, 300) = 103.428, p < 0.001$ and F1-score : $F(10, 300) = 69.299$, $p < 0.001$). For BA, TPR, FPR and F1-score, the performance of our proposed TFF-Former model are significantly higher than conventional methods ($all : p < 0.001$), CNN-based methods ($all : p < 0.05$) and Transformer-based methods ($all : p < 0.05$). For accuracy, our method is significantly higher than that of compared methods ($all : p < 0.05$) except for TCN-T, but the accuracy of our method tends to be higher than TCN-T ($p < 0.1$). This demonstrates that our TFF-Former is more effective in zero-training RSVP task, because compared to conventional methods and CNN-based methods Transformer can effectively learn global interactions and adjust on the local content while modeling global relationships. And compared to Transformer-based methods, TFF-Former achieved interaction and fusion between temporal and frequency features of EEG signals.

It can be seen in Figure 3, the mean BA of within-subject methods increased with calibration data size increasing, where HDCA has the best performance for training data from one block to four blocks. However, the mean BA of TFF-Former (88.05%) is significantly higher than that of HDCA with two blocks ($all : p < 0.01$) and shows no statistical difference ($all : p > 0.4$) compared with HDCA using three and four blocks (three : 88.08%, four : 88.42%). Therefore, we reduce the calibration time of at least four blocks, which shows our zero-training method is effective in RSVP task.

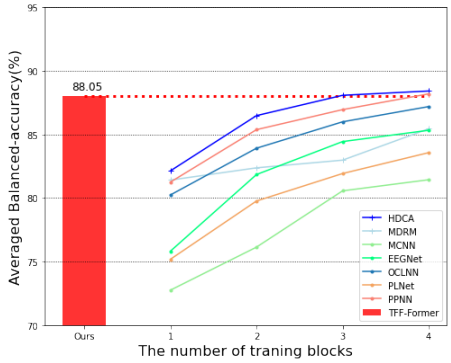


Figure 3: The comparison results between TFF-Former and within-subject methods in RSVP task.

Table 2: CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON BENCHMARK DATASET(mean ± standard deviation)

Method	Number of Classes	
	20	40
CCA	70.61±21.53 ***	63.20 ± 28.48 ***
FBCCA	82.09±17.24 ***	77.92 ± 22.90 ***
Compact-CNN	**86.24 ± 16.08** *	**82.90 ± 17.25** ***
Wang	83.02±17.05 ***	77.01±18.71 ***
TCN-T	83.39±17.55 ***	77.88 ± 18.34 ***
TFF-Former	**88.21 ± 15.03**	**84.77 ± 16.93**

Table 2. shows the classification accuracy on the SSVEP benchmark dataset with all 40 classes and 20 classes selected from 40 classes with the same frequency intervals. The one-way repeated measures ANOVA showed a significant main effect of method both in 20-class ($F(4, 132) = 14.36, p < 0.001$) and 40-class ($F(4, 132) = 19.22, p < 0.001$). In 20-class and 40-class SSVEP tasks, the performance of TFF-Former are significantly higher than those compared methods ($all : p < 0.05$). Therefore, our proposed TFF-Former model can not only achieve zero-training decoding in RSVP task, but also achieved competitive performance across tasks in SSVEP task. Because the input EEG signals contained multiple information and task-related temporal-frequency information can be captured through the two-stream architecture in both RSVP and SSVEP tasks.

5.2 Ablation Study

We conducted ablation experiments to evaluate the effectiveness of two-view structure including frequency stream and temporal stream, mean masked attention mechanism and cross-view module in TFF-Former. The results are listed in TABLE 3, where the **M1** and **M2** indicate the model only uses frequency view and temporal view of TFF-Former with mean masked attention mechanism respectively. The **M3** is TFF-Former without mean masked attention mechanism, which is used to verify that mean mask are efficient for classification. **M4** is TFF-Former without cross-view module, which directly token-wise concatenates two views tokens extracted from encoder layer as decision tokens fed to decision module. The experimental set of these four models is the same as that of compared method in TABLE 1 and TABLE 2.

For the RSVP dataset, the one-way repeated measures ANOVA showed a significant main effect of proposed modules in all evaluation metrics (BA : $F(4, 120) = 267.411, p < 0.001$, TPR : $F(4, 120) = 24.41, p < 0.001$ and FPR : $F(4, 120) = 156.497, p < 0.001$). The classification performance of our proposed method is significantly higher than that of the modals model ($all : p < 0.05$), which indicates that frequency stream and temporal stream are both useful for classification and the two-view fusion Transformer can improve the classification performance. The model with all modules significantly outperforms the model without cross-view module in BA and FPR ($p < 0.05$) and significantly outperforms the model without mean masked attention mechanism ($p < 0.05$). Therefore, each module in TFF-Former has been verified to improve the performance of the RSVP classification task. Meanwhile, the above conclusions are also tenabled on benchmark dataset, where the one-way repeated measures ANOVA showed a significant main effect of proposed modules ($F(4, 132) = 11.13, p < 0.001$) and the model with all modules also significantly outperforms the ablation models ($all : p < 0.05$). In addition, the ablation study of encoder layer sharing parameters was conducted. The T-test revealed that the mean BA of TFF-Former (88.05%) is significantly higher than that of model without sharing parameters between streams in encoder layer (87.45%) in RSVP task and the accuracy of TFF-Former (84.77%) is also significantly higher than that of ablation model (83.8%) in SSVEP task ($all : p < 0.05$), because the encoder layer sharing parameters between two streams can project two view features to the same feature subspace which is more suitable for subsequent view interaction.

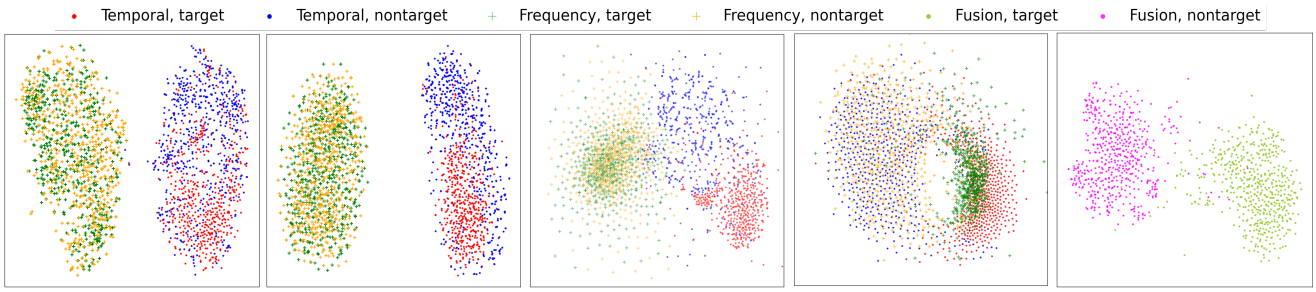
Therefore, all modules we involved in TFF-Former are verified to be effective in classification across two BCI tasks, where two-view architecture can simultaneously utilize temporal and frequency information of EEG signals, mean mask mechanism can reduce redundant aggregation of tokens by decreasing the weight of irrelevant tokens, and cross-view module can effectively extract common representations of features through view interaction.

5.3 Visualization

We applied t-distributed Stochastic Neighbor Embedding (t-SNE) to project the output of each layer into 2 dimensions and draw scatter plots. Figure 4 shows the visualization of one subject in the RSVP dataset. As revealed in Figure 4.(a), we can see maximum overlap of two classes in t-SNE visualization in each view and there are differences between the two views. After being projected by slice embedding layer respectively, the samples of each class keep the distribution form of original raw data well, which can be seen in Figure 4.(b). Figure 4.(c) shows that as the temporal tokens and frequency tokens are processed over the encoder layer sharing parameters between two streams, the distance in feature space between two views decrease. In Figure 4.(d), with the two views features interacting with each other in the cross-attention layers, the features of different views in the same class get closer. Finally, the linear separation between fusion features in two classes is clearly visible in Fig. 4.(e). Thus, the visibility of similarity in two views features and separation in two classes increase from raw data to the output of the model, which indicates that our model can effectively extract common representations of two views features and learn useful features for classification.

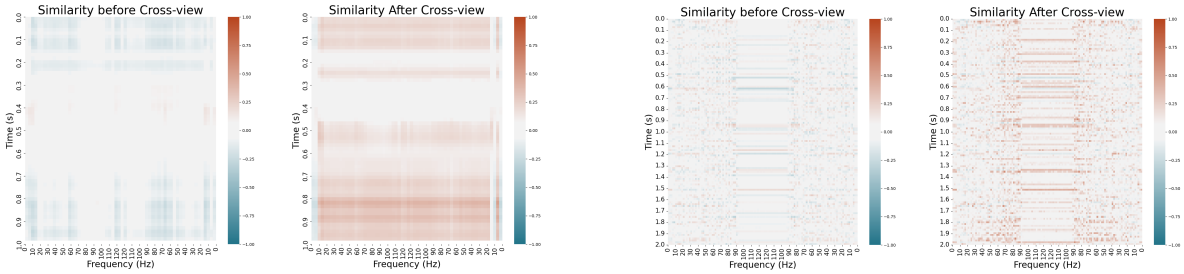
Table 3: ABLATION STUDIES ON THE TWO DATASETS.

Model	Proposed Module				RSVP			SSVEP
	Frequency-view	Temporal-view	Mean Mask	Cross-view	BA (%) ↑	TPR (%) ↑	FPR (%) ↓	Accuracy (%) ↑
M1	√	–	√	–	61.00±4.71 ***	70.23±11.92 ***	46.28±8.15 ***	68.90±20.17 ***
M2	–	√	√	–	86.73±3.80 ***	83.88±8.28 *	10.41±1.72 *	82.62 ±17.75 **
M3	√	√	√	–	87.18±3.52 *	84.98±7.31	10.60±1.79 **	83.19 ± 17.07 *
M4	√	√	–	√	87.10±3.82 **	84.53±8.21 *	10.33 ± 2.33 *	83.64 ± 17.30 *
TFF-Former	√	√	√	√	**88.05 ± 3.73**	**85.45 ± 7.85**	**9.34 ± 2.26**	**84.77 ± 16.93**



(a) Data (b) Slice Embedding Layer (c) Encoder Layer (d) Cross-attention Layer (e) Fusion Module

Figure 4: The t-SNE visualization results of different feature spaces in TFF-Former: (a) raw data, (b) the output of slice embedding layer, (c) the output of encoder layer, (d), the output of cross-attention layer, (e) the output of fusion module. The red and blue dots indicate temporal view of target samples and non-target samples respectively. The green and yellow dots indicate frequency view of target samples and non-target samples respectively.



(a) RSVP task (b) SSVEP task

Figure 5: The cosine similarity between temporal and frequency tokens before and after cross-view module: (a) RSVP task, (b) SSVEP task.

Meanwhile, to show the effect of cross-view module in view interaction, we visualized the similarity matrix between temporal tokens and frequency tokens before and after the cross-view module. We used cosine similarity to measure the correlation between each pair of temporal and frequency tokens. As revealed in Figure 5.(a) , in the RSVP classification, after view interaction, the correlation between token features from two views increased significantly. For the typical ERP frequency band ($< 15Hz$) the feature was enhanced, for the frequency band that was silenced in the preprocessing ($> 15Hz$), the features were learned with integrated information from multi-time and multi-frequency. It can be seen from Figure 5.(b) that the results also established. In particular, the temporal view were more relevant to 8-90 Hz frequency tokens which are the frequency band for SSVEP signals, which is consistent with SSVEP signals characteristics. Indeed, the ablation study also proved the effectiveness of cross-view module. Therefore, cross-view module can achieve view interaction to increase the correlation between view features and extract common representations of two views.

6 CONCLUSION

This study proposed Temporal-Frequency Fusion Transformer (TFF-Former) to realize zero-training decoding across RSVP and SSVEP tasks. We validated TFF-Former on the self-collected RSVP dataset including 31 subjects and benchmark SSVEP dataset. The experimental results showed that our model achieved significantly higher decoding performance than the compared zero-training methods and reduced calibration time of at least four blocks compared with RSVP within-subject methods. This indicates that our method can achieve superior performance in zero-training condition and further promote the application of visual evoked EEG-based BCI systems.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62020106015, Grant 61976209, and Grant U21A20388; in part by the CAS International Collaboration Key Project under Grant 173211KYSB20190024; and in part by the Strategic Priority Research Program of CAS under Grant XDB32040000; and in part by Beijing Natural Science Foundation under Grant J210010.

REFERENCES

[1] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

[2] Kai Keng Ang and Cuntai Guan. Brain-computer interface in stroke rehabilitation. *Journal of Computing Science and Engineering*, 7(2):139–146, 2013.

[3] Christopher Barngrover, Alric Althoff, Paul DeGuzman, and Ryan Kastner. A brain–computer interface (bci) for the detection of mine-like objects in sidescan sonar imagery. *IEEE journal of oceanic engineering*, 41(1):123–138, 2015.

[4] Laura Acqualagna and Benjamin Blankertz. Gaze-independent bci-spelling using rapid serial visual presentation (rsvp). *Clinical Neurophysiology*, 124(5):901–908, 2013.

[5] Zhimin Lin, Chi Zhang, Ying Zeng, Li Tong, and Bin Yan. A novel p300 bci speller based on the triple rsvp paradigm. *Scientific reports*, 8(1):1–9, 2018.

[6] A.D. Gerson, L.C. Parra, and P. Sajda. Cortically coupled computer vision for rapid image search. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):174–179, 2006.

[7] Amar R Marathe, Vernon J Lawhern, Dongrui Wu, David Slayback, and Brent J Lance. Improved neural signal classification in a rapid serial visual presentation task using active learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(3):333–343, 2015.

[8] Ming Cheng, Xiaorong Gao, Shangkai Gao, and Dingfeng Xu. Design and implementation of a brain-computer interface with high transfer rates. *IEEE transactions on biomedical engineering*, 49(10):1181–1186, 2002.

[9] Xiaogang Chen, Yijun Wang, Masaki Nakanishi, Xiaorong Gao, Tzyy-Ping Jung, and Shangkai Gao. High-speed spelling with a noninvasive brain–computer interface. *Proceedings of the national academy of sciences*, 112(44):E6058–E6067, 2015.

[10] Xiaogang Chen, Bing Zhao, Yijun Wang, and Xiaorong Gao. Combination of high-frequency ssvep-based bci and computer vision for controlling a robotic arm. *Journal of neural engineering*, 16(2):026012, 2019.

[11] Bonkon Koo, Hwan-Gon Lee, Yunjun Nam, and Seungjin Choi. Immersive bci with ssvep in vr head-mounted display. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 1103–1106. IEEE, 2015.

[12] Shih-Chung Chen, Yeou-Jiunn Chen, Ilham AE Zaeni, and Chung-Min Wu. A single-channel ssvep-based bci with a fuzzy feature threshold algorithm in a maze game. *International Journal of Fuzzy Systems*, 19(2):553–565, 2017.

[13] Alexandre Barachant and Marco Congedo. A plug&play P300 BCI using information geometry. *CoRR*, abs/1409.0107, 2014.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[15] Hubert Cecotti and Axel Graser. Convolutional neural networks for p300 detection with application to brain-computer interfaces. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):433–445, 2010.

[16] Ran Manor and Amir B Geva. Convolutional neural network for multi-category rapid serial visual presentation bci. *Frontiers in computational neuroscience*, 9:146, 2015.

[17] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[18] Hongchang Shan, Yu Liu, and Todor P Stefanov. A simple convolutional neural network for accurate p300 detection and character spelling in brain computer interface. In *IJCAI*, pages 1604–1610, 2018.

[19] Wei Wei, Shuang Qiu, Xuelin Ma, Dan Li, Bo Wang, and Huiguang He. Reducing calibration efforts in rsvp tasks with multi-source adversarial domain adaptation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2344–2355, 2020.

[20] Jongmin Lee, Kyungho Won, Moonyoung Kwon, Sung Chan Jun, and Minkyu Ahn. Cnn with large data achieves true zero-training in online p300 brain-computer interface. *IEEE Access*, 8:74385–74400, 2020.

[21] Zhonglin Lin, Changshui Zhang, Wei Wu, and Xiaorong Gao. Frequency recognition based on canonical correlation analysis for ssvep-based bcis. *IEEE transactions on biomedical engineering*, 53(12):2610–2614, 2006.

[22] Xiaogang Chen, Yijun Wang, Shangkai Gao, Tzyy-Ping Jung, and Xiaorong Gao. Filter bank canonical correlation analysis for implementing a high-speed ssvep-based brain–computer interface. *Journal of neural engineering*, 12(4):046008, 2015.

[23] Yu Zhang, Guoxu Zhou, Qibin Zhao, Akinari Onishi, Jing Jin, Xingyu Wang, and Andrzej Cichocki. Multiway canonical correlation analysis for frequency components recognition in ssvep-based bcis. In *International Conference on Neural information processing*, pages 287–295. Springer, 2011.

[24] YU Zhang, Guoxu Zhou, Jing Jin, Xingyu Wang, and Andrzej Cichocki. Frequency recognition in ssvep-based bci using multiset canonical correlation analysis. *International journal of neural systems*, 24(04):1450013, 2014.

[25] Masaki Nakanishi, Yijun Wang, Xiaogang Chen, Yu-Te Wang, Xiaorong Gao, and Tzyy-Ping Jung. Enhancing detection of ssveps for a high-speed brain speller using task-related component analysis. *IEEE Transactions on Biomedical Engineering*, 65(1):104–112, 2017.

[26] Nicholas Waytowich, Vernon J Lawhern, Javier O Garcia, Jennifer Cummings, Josef Faller, Paul Sajda, and Jean M Vettel. Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials. *Journal of neural engineering*, 15(6):066031, 2018.

[27] Yao Li, Jiayi Xiang, and Thenkurussi Kesavadas. Convolutional correlation analysis for enhancing the performance of ssvep-based brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):2681–2690, 2020.

[28] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

[30] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[31] Wei Wei, Shuang Qiu, Yukun Zhang, Jiayu Mao, and Huiguang He. Erp prototypical matching net: a meta-learning method for zero-calibration rsvp-based image retrieval. *Journal of Neural Engineering*, 19(2):026028, 2022.

[32] Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for ssvep-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2016.

[33] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[34] Boyu Zang, Yanfei Lin, Zhiwen Liu, and Xiaorong Gao. A deep learning method for single-trial eeg classification in rsvp task based on spatiotemporal features of erps. *Journal of Neural Engineering*, 18(4):0460c8, 2021.

[35] Fu Li, Chong Wang, Yang Li, Hao Wu, Boxun Fu, Youshuo Ji, Yi Niu, and Guangming Shi. Phase preservation neural network for electroencephalography classification in rapid serial visual presentation task. *IEEE Transactions on Biomedical Engineering*, 2021.

[36] Yiting Wang, Wei-Bang Jiang, Rui Li, and Bao-Liang Lu. Emotion transformer fusion: Complementary representation properties of eeg and eye movements on recognizing anger and surprise. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1575–1578. IEEE, 2021.

[37] Zhe Wang, Yongxiong Wang, Chuanfei Hu, Zhong Yin, and Yu Song. Transformers for eeg-based emotion recognition: A hierarchical spatial information learning model. *IEEE Sensors Journal*, 2022.

[38] Ramiro Casal, Leandro E Di Persia, and Gastón Schlotthauer. Temporal convolutional networks and transformers for classifying the sleep stage in awake or asleep using pulse oximetry signals. *Journal of Computational Science*, page 101544, 2022.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] A Torralba, KP Murphy, and WT Freeman. The mit-csail database of objects and scenes, 2009.