

PAPER

Decoupling representation learning for imbalanced electroencephalography classification in rapid serial visual presentation task

To cite this article: Fu Li *et al* 2022 *J. Neural Eng.* **19** 036011

View the [article online](#) for updates and enhancements.

You may also like

- [Multi-objective optimization approach for channel selection and cross-subject generalization in RSVP-based BCIs](#)
Meng Xu, Yuanfang Chen, Dan Wang et al.
- [Enhancing the EEG classification in RSVP task by combining interval model of ERPs with spatial and temporal regions of interest](#)
Bowen Li, Yanfei Lin, Xiaorong Gao et al.
- [A deep learning method for single-trial EEG classification in RSVP task based on spatiotemporal features of ERPs](#)
Boyue Zang, Yanfei Lin, Zhiwen Liu et al.



PAPER

Decoupling representation learning for imbalanced electroencephalography classification in rapid serial visual presentation task

RECEIVED
2 December 2021REVISED
24 April 2022ACCEPTED FOR PUBLICATION
25 April 2022PUBLISHED
13 May 2022Fu Li[✉], Hongxin Li, Yang Li^{*}, Hao Wu, Boxun Fu, Youshuo Ji[✉], Chong Wang and Guangming Shi

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an, People's Republic of China

^{*} Author to whom any correspondence should be addressed.E-mail: liy@xidian.edu.cn**Keywords:** electroencephalography (EEG), rapid serial visual presentation (RSVP), decoupling representation learning, class imbalance problem**Abstract**

Objective. The class imbalance problem considerably restricts the performance of electroencephalography (EEG) classification in the rapid serial visual presentation (RSVP) task. Existing solutions typically employ re-balancing strategies (e.g. re-weighting and re-sampling) to alleviate the impact of class imbalance, which enhances the classifier learning of deep networks but unexpectedly damages the representative ability of the learned deep features as original distributions become distorted. **Approach.** In this study, a novel decoupling representation learning (DRL) model, has been proposed that separates the representation learning and classification processes to capture the discriminative feature of imbalanced RSVP EEG data while classifying it accurately. The representation learning process is responsible for learning universal patterns for the classification of all samples, while the classifier determines a better bounding for the target and non-target classes. Specifically, the representation learning process adopts a dual-branch architecture, which minimizes the contrastive loss to regularize the representation space. In addition, to learn more discriminative information from RSVP EEG data, a novel multi-granular information based extractor is designed to extract spatial-temporal information. Considering the class re-balancing strategies can significantly promote classifier learning, the classifier was trained with re-balanced EEG data while freezing the parameters of the representation learning process. **Main results.** To evaluate the proposed method, experiments were conducted on two public datasets and one self-conducted dataset. The results demonstrate that the proposed DRL can achieve state-of-the-art performance for EEG classification in the RSVP task. **Significance.** This is the first study to focus on the class imbalance problem and propose a generic solution in the RSVP task. Furthermore, multi-granular data was explored to extract more complementary spatial-temporal information. The code is open-source and available at <https://github.com/Tammie-Li/DRL>.

1. Introduction

Electroencephalography (EEG), as an important technology of brain-computer interfaces (BCIs), creates a direct connection between the human brain and external devices to realize the exchange of information [1]. Some early EEG-based BCI applications, such as wheelchair control [2], text spellers [3], and prosthetic artificial limbs [4], have improved the quality of life of disabled patients. In recent years, the

application of EEG-based BCI has extended to healthy people. Rapid serial visual presentation (RSVP), has received substantial attention from researchers [5–7], as a potential method for human enhancement.

RSVP-based BCIs most used in the area of counter intelligence, police, and health care, which require professionals to review a large number of images or information. By recognizing the event-related potential (ERP) component in EEG signals, RSVP-based BCIs can detect and recognize objects, pieces of

relevant information more quickly than manual analysis, which greatly improves the work efficiency of professionals [8]. The RSVP paradigm displays images sequentially at a rate of 5–20 Hz, in which the ratio of non-target images to target images is approximately 10:1. This induces a special P300 component, which is a common ERP component in EEG signals that can be used for target image detection [9, 10].

To achieve an RSVP-based BCI, it is critical to effectively classify the EEG signals generated in the RSVP task. However, the performance of the current methods proposed for RSVP EEG classification is still far from satisfactory. Based on the literature, the following two issues need to be addressed. The first is to address the class imbalance problem. For a general RSVP paradigm, the number of non-target samples is far greater than the number of target samples in the paradigm design for the following reasons: (1) in the process of visual presentation, it is necessary to maintain an adequate time interval between two adjacent target images to induce the P300 response [8]. (2) To maintain a high refresh rate, many non-target images must be inserted into the rapid serial. Many studies have proved that this class imbalance problem seriously restricts the RSVP EEG classification performance because the majority class dominates the decision-making, which leads to high false positives [11]. Generally, some class re-balancing strategies [12], such as data re-sampling and loss re-weighting, are acceptable approaches for alleviating the extreme imbalance of the training data. However, some studies [13] have revealed that these approaches have adverse effects. For example, they can unexpectedly damage, to some extent, the representative ability of the learned deep features. Specifically, the over-sampling strategy has the risk of over-fitting the target data, while under-sampling may under-fit the overall data distributions. When re-weighting, directly changing or inverting the data presenting frequency distorts the original distributions. Therefore, it is necessary to explore a more effective solution to tackle the class imbalance problem in the RSVP task. Some recent studies revealed that class imbalance may not be an issue for high-quality representation learning, but is disastrous for classifier learning [13, 14]. Thus, decoupling the learning process for the RSVP task into representation learning and classification is expected to eliminate the impact of class imbalance.

The second major issue is how to extract a more discriminative representation from the EEG signals. Current EEG classification methods usually employ handcrafted features, such as statistical features in the time domain, band power in the frequency domain, and discrete wavelet transform in the time-frequency domain [15]. It is desirable to investigate more powerful discriminative deep features with both spatial and temporal information for EEG signals. Recent neuroscience studies [16] modeled the brain with different levels of granularity, where lower levels can

effectively extract the short-term temporal correlation of EEG signals and higher levels can retain more global information. Thus, more reasonable and complementary information can be obtained by analyzing data from multiple granularity levels [17]. In recent years, analyzing multi-granular data while mining representative information for decision-making has become the emerging computing paradigm in signal processing [18–20]. Gacek and Pedrycz [21] developed a general framework for a granular representation of electrocardiogram signals, which shares many common features with the EEG. Wang *et al* [22] further discussed the potential of this multi-granular information (MGI)-based computing paradigm for brain data. These studies are of great interest and encourage the design of new algorithms to extract more discriminative representations through multi-granular data information.

To address the aforementioned two major issues in the RSVP task, in this article, a novel decoupling representation learning (DRL) model is proposed based on the MGI, to eliminate the impact of the class imbalance problem while extracting multi-granular EEG information. To achieve the first goal, the overall learning process is decoupled into representation learning and classification to capture the discriminative feature of imbalanced RSVP EEG data while maintaining classification performance. The representation learning process is responsible for learning universal patterns for classification on all samples, while the classifier determines a better bounding for the target and non-target classes. Specifically, the representation learning process first constructs positive and negative pairs according to whether the two samples in the pair belong to the same class. Then, a dual-branch architecture is proposed to minimize contrastive loss by driving the cosine similarity in the representation space to be small for positive pairs and large for negative pairs. In the classification process, considering the re-balancing strategies can significantly promote classifier learning, the classifier was trained with the re-balanced RSVP EEG data¹ while freezing the parameters of the representation learning process. The second goal was to extract multi-granular data information. To this end, the original data were sampled in the temporal dimension with exponentially decaying sampling rates, which transform the data into multiple granularity levels. After obtaining multi-granular data information, two operations were performed to learn the representation. The first operation maps the multi-granular data information into spatial-temporal representation to extract the temporal dynamics of EEG sequences and build the channel relationship of the RSVP EEG data. In the second operation, considering the degradation

¹ Re-balanced RSVP EEG data: using over-sampling or under-sampling to make the sample numbers of different classes roughly equal.

of deep networks, the representations were fused with a residual architecture.

To the best of our knowledge, this is the first study to focus on the class imbalance problem in the RSVP task. The main contributions of this study are as follows.

- (a) A novel DRL is proposed to decouple the learning process into representation learning and classification to effectively tackle the class imbalance problem.
- (b) The MGI is first introduced into EEG signal processing in the form of a well-constructed extractor, which is then integrated into the proposed framework to boost classification performance.
- (c) The proposed DRL, which is a generic framework for dealing with the class imbalance problem, achieved state-of-the-art performance on two public datasets and one self-conducted dataset.

2. Related work

2.1. Existing traditional methods for EEG classification in the RSVP task

Over the past few years, many methods have been proposed to classify the EEG signals generated in the RSVP task. For example, Bigdely *et al* [23] adopted spatial independent component analysis and principal component analysis to extract spatial, temporal, and spectral features. These features were then combined and classified using a Fisher linear discriminant (FLD) classifier. Blankertz *et al* [24] proposed regularized linear discriminant analysis (rLDA) for the classification of ERP signals and achieved acceptable results. This model uses shrinkage estimators to form a regularized version of LDA, with performance superior to other LDA-based approaches. Sajda *et al* [25] proposed a hierarchical discriminant component analysis (HDCA), which first adopts FLD to train spatial weights and then trains a logistic regression classifier to learn temporal weights and implement classification. Xiao *et al* [26] developed an algorithm called discriminative canonical pattern matching (DCPM) to handle BCI datasets with small training sets. DCPM constructs discriminative spatial patterns as well as canonical correlation analysis patterns and then matches these two patterns to form a robust classifier.

2.2. Existing deep learning methods for EEG classification in the RSVP task

Among the various RSVP EEG classification methods, it is notable that the recently developed deep learning methods are becoming dominant in improving RSVP EEG classification. In contrast to traditional research that relies on expert-level experience and priori domain knowledge to extract representation

information, deep learning can automatically extract discriminative data representations from EEG recordings of brain activity [27]. For example, Cecotti *et al* [28] adopted a convolutional neural network (CNN) to adaptively detect P300 waves in the time domain. The network was tested on a P300 speller task [29], which displayed an improvement in performance. Later, in another study, Cecotti *et al* [30] proposed supervised spatial filtering methods to enhance discriminative information in EEG data. They proposed a CNN, with a layer dedicated to spatial filtering for the detection of ERP. The model was then trained based on the maximization of the area under the receiver operating characteristic curve (AUC). Schirrneister *et al* [31] proposed a CNN model called DeepConvNet for EEG decoding tasks in BCI. Their model consists of five convolutional layers with a softmax layer for classification. Lawhern *et al* [32] proposed a compact neural network called EEGNet for EEG-based BCI and achieved considerable performance on various EEG classification tasks. Vázquez *et al* [33] proposed EEG-Inception, which integrates inception modules to efficiently extract temporal features through different temporal scales for ERP classification. To the best of our knowledge, EEG-Inception is the state-of-the-art model for the classification of ERP signals.

2.3. Existing methods for the class imbalance problem

The class imbalance problem has always been a challenging topic in computer vision (CV) and natural language processing (NLP) [34]. Recent studies have mainly pursued the following two directions and comprise: (1) data distribution re-balance method includes over-sampling for the minority class [35, 36], and under-sampling for the majority class [37], which makes the sample numbers of different classes roughly equal. (2) Methods on class imbalance loss. Wang *et al* [38] first found that modifying the loss function can significantly improve the classification performance of an imbalanced dataset. Lin *et al* [39] proposed focal loss, which measures the contribution to the total loss from hard-to-classify samples and easy-to-classify samples. For a general RSVP task, the class imbalance problem is conspicuous and should be deeply considered. Some researchers have adopted under-sampling as a strategy to tackle this problem [31–33]. However, this strategy discards many non-target samples that may be valuable for classification. Meanwhile, under-sampling distorts the training data distribution, which results in inconsistencies within the unbalanced test set. It is important to explore a more effective solution to tackle the class imbalance problem in the RSVP task.

2.4. Contrastive learning

Contrastive learning has recently gained attention due to its excellent performance in self-supervised

representation learning [40]. Becker *et al* [41] first proposed the core idea of contrastive learning by comparing separate but related data. Subsequently, contrastive learning stagnated for a long time until LeCun *et al* [42] created the foundation for the contrastive learning framework. In the CV domain, the rapid development of contrastive learning is attributed to its excellent performance in self-supervised learning. He *et al* [43] proposed Moco, which regards contrastive learning as a dictionary to perform similarity matching between queries and keys. On this basis, Chen *et al* [44] proposed SimCLR, which simplifies Moco without requiring specialized architectures or a memory bank. Xinlei *et al* [45] proposed a Simple Siamese (SimSiam) network that achieved the best results without negative samples, large batches, and momentum encoders. In addition, contrastive learning was extended to fully supervised learning, allowing effective leveraging of label information [46]. In this work, a matching algorithm was designed to pair samples using the priori label information and adopt contrastive learning to capture the discriminative features of RSVP EEG data.

3. Method

The training process of the proposed DRL model is illustrated in figure 1, which consists of two parts, i.e. representation learning and classification. DRL model tackles the class imbalance problem by decoupling the learning process into representation learning and classification. The representation learning process is responsible for learning universal patterns for classification on all samples, while the classifier determines a better bounding for the target and non-target classes. Specifically, the representation learning process is achieved by a dual-branch architecture, which adopts three steps to capture the discriminative features of RSVP EEG data. First, positive and negative pairs were constructed according to the priori label of the training samples. Subsequently, considering the multi-granular characteristics of the EEG data, a novel MGI based extractor was developed to extract spatial-temporal data representations. Third, contrastive loss was adopted to maximize the agreement between positive pairs and minimize the agreement between negative pairs in the representation space. Finally, the DRL trained the classifier with re-balanced EEG data while freezing the parameters of the representation learning process. The overall process is described in detail as follows.

3.1. Sample pair construction

Let $\mathbf{x} \in \mathbb{R}^{N \times C \times T}$ denote the training RSVP EEG data, where N , C , and T denote the number of samples, the number of EEG channels, and the temporal length, respectively. DRL adopts a dual-branch architecture

that minimizes contrastive loss to learn representation, independently. Considering the importance of label information, the DRL makes full use of the priori label information to construct the sample pairs. More precisely, two samples $\mathbf{x}^1 \in \mathbb{R}^{C \times T}$, $\mathbf{x}^2 \in \mathbb{R}^{C \times T}$ are randomly selected from \mathbf{X} and matched as a pair. If \mathbf{x}^1 and \mathbf{x}^2 are from the same class, the pair is a positive pair, otherwise a negative pair. In other words, the target and target, non-target and non-target constitute the positive pairs, and target and non-target constitute the negative pairs. In this work, M pairs were constructed and pair labels are defined as $\mathbf{q} = [q_1, q_2, \dots, q_M] \in \mathbb{R}^M$, where $q_m \in \{+1, -1\}$ indicates a positive and negative pair.

3.2. Multi-granular spatial-temporal information extraction

To improve the EEG classification performance, an MGI based extractor was designed to represent the RSVP EEG data in a more discriminative feature space. The MGI process is depicted in figure 2. Specifically, the original data are first sampled in the temporal dimension with exponentially decaying sampling rates to transform the data into multiple granularity levels. Then, a temporal block was adopted to extract the dynamics of EEG sequences and a spatial block was adopted to capture the relationship between EEG channels. Finally, considering the degradation of deep networks, the multi-granular representations were fused with a residual architecture. The concrete process is described as follows:

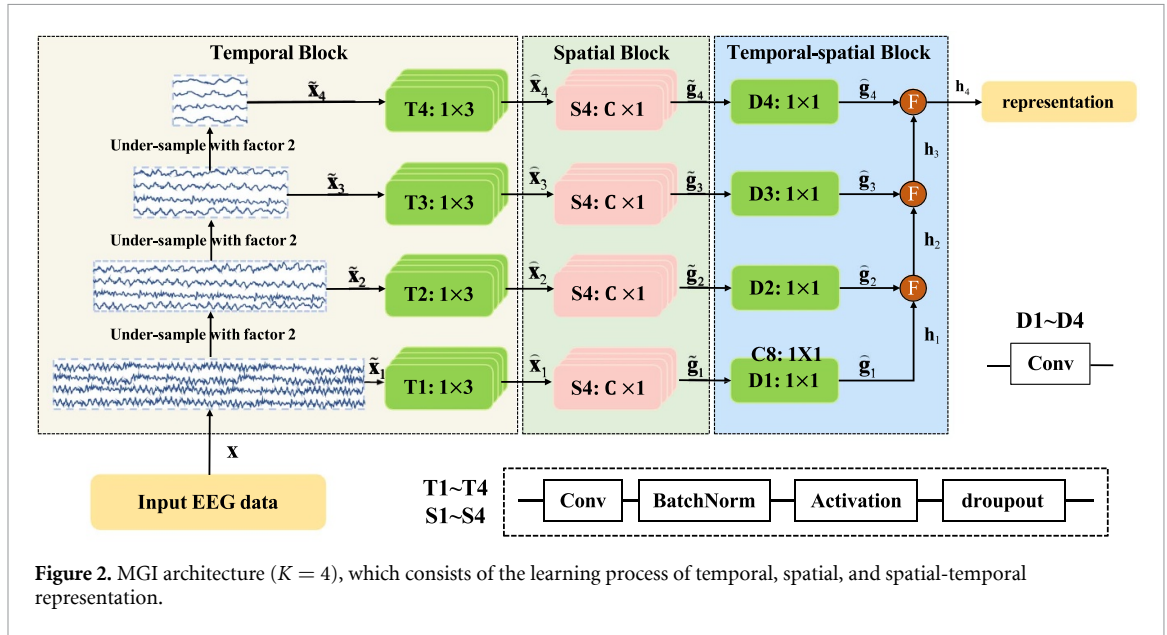
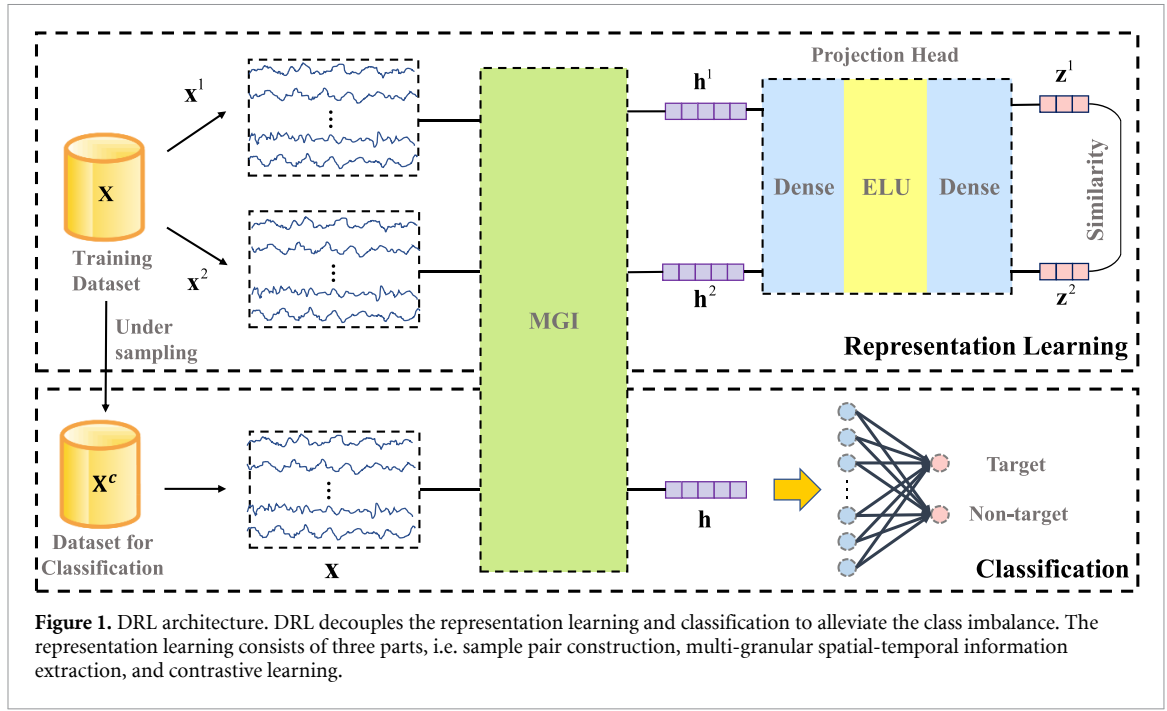
3.2.1. Temporal representation learning

For an EEG signal, analyzing the data using multiple granularities can provide complementary temporal information. To this end, multi-granular data information was obtained by sampling the raw data in the temporal dimension with exponentially decaying sampling rates. The calculation process can be followed in equation (1). MGI obtains the coarse granular data at a low temporal resolution, with the aim of extracting the temporal dynamic information under a larger receptive field, which is helpful for modeling long-term temporal dependencies. Fine granular data were obtained at a high temporal resolution to extract temporal dynamic information in detail. Formally, for a pair of EEG data $\mathbf{x}^1 = \{\mathbf{x}_1^1, \mathbf{x}_2^1, \dots, \mathbf{x}_T^1\} \in \mathbb{R}^{C \times T}$ and $\mathbf{x}^2 = \{\mathbf{x}_1^2, \mathbf{x}_2^2, \dots, \mathbf{x}_T^2\} \in \mathbb{R}^{C \times T}$, the different granular data were generated by:

$$\begin{aligned}\tilde{\mathbf{x}}_k^1 &= [\mathbf{x}_1^1, \mathbf{x}_k^1, \mathbf{x}_{2k}^1, \dots, \mathbf{x}_{T_k \times k}^1] \in \mathbb{R}^{C \times T_k}, k = 1, \dots, K, \\ \tilde{\mathbf{x}}_k^2 &= [\mathbf{x}_1^2, \mathbf{x}_k^2, \mathbf{x}_{2k}^2, \dots, \mathbf{x}_{T_k \times k}^2] \in \mathbb{R}^{C \times T_k}, k = 1, \dots, K,\end{aligned}\quad (1)$$

where $T_k = T/2^k$ denotes the temporal dimension of k th granularity level.

After obtaining multiple granularity level data, $K \times F_t$ convolution filters were employed to extract



the temporal dynamics of the EEG sequences. The calculation process can be followed as:

$$\begin{aligned} \hat{\mathbf{x}}_k^1 &= \mathcal{F}_t(\mathbf{x}_k^1) = \mathbf{x}_k^1 \otimes f \in \mathbb{R}^{F_t \times C \times T_k}, k = 1, \dots, K, \\ \hat{\mathbf{x}}_k^2 &= \mathcal{F}_t(\mathbf{x}_k^2) = \mathbf{x}_k^2 \otimes f \in \mathbb{R}^{F_t \times C \times T_k}, k = 1, \dots, K, \end{aligned} \quad (2)$$

where \mathcal{F}_t denotes the temporal convolution operation, f is the temporal convolution filter with a kernel size of 1×3 . Subsequently, a batch normalization layer was applied to accelerate the training and enhance the generalization capacity of the network. An additional exponential linear unit (ELU) [47],

which displays soft saturation characteristics when dealing with small values, served as the nonlinear activation function making the network robust to input variations and noise. In addition, a dropout layer was used to avoid over-fitting.

3.2.2. Spatial representation learning

To avoid losing the intrinsic structural relationship of electrodes, a spatial convolution operation was adopted for different EEG channels to extract the EEG spatial representation. Specifically, the relationship among all channels was constructed using a large kernel. This operation can be expressed as:

$$\begin{aligned}\hat{\mathbf{g}}_k^1 &= \mathcal{F}_s(\hat{\mathbf{x}}_k^1) = \hat{\mathbf{x}}_k^1 \otimes f \in \mathbb{R}^{F_s \times 1 \times T_k}, k = 1, \dots, K, \\ \hat{\mathbf{g}}_k^2 &= \mathcal{F}_s(\hat{\mathbf{x}}_k^2) = \hat{\mathbf{x}}_k^2 \otimes f \in \mathbb{R}^{F_s \times 1 \times T_k}, k = 1, \dots, K,\end{aligned}\quad (3)$$

where \mathcal{F}_s denotes the spatial convolution operation, f is the spatial convolution filter with a kernel size of $C \times 1$, and F_s is the number of spatial filters.

Similar to the temporal representation block, a batch normalization layer, an ELU activation function, and a dropout layer were attached sequentially to enhance the generalization capacity while maintaining the nonlinear mapping.

3.2.3. Spatial-temporal representation learning

To avoid losing granularity information, a residual connection was designed to better utilize the feature details as shown in the temporal-spatial block of figure 2. Specifically, initially $K \times F_d$ convolution filters were used to reduce the dimension. The calculation process of the dimensionality reduction operation denotes as:

$$\begin{aligned}\hat{\mathbf{g}}_k^1 &= \mathcal{F}_d(\hat{\mathbf{g}}_k^1) = \hat{\mathbf{g}}_k^1 \otimes f \in \mathbb{R}^{F_d \times 1 \times T_k}, k = 1, \dots, K, \\ \hat{\mathbf{g}}_k^2 &= \mathcal{F}_d(\hat{\mathbf{g}}_k^2) = \hat{\mathbf{g}}_k^2 \otimes f \in \mathbb{R}^{F_d \times 1 \times T_k}, k = 1, \dots, K,\end{aligned}\quad (4)$$

where \mathcal{F}_d denotes the dimensionality reduction operation, f is the convolution filter with the kernel size of 1×1 .

Subsequently, a residual connection was established between two adjacent granularity levels. The finest granularity representations were set as $\mathbf{h}_1^1 = \hat{\mathbf{g}}_1^1$ and $\mathbf{h}_1^2 = \hat{\mathbf{g}}_2^1$. As shown in figure 3, the current level fusion result was calculated through the last level fusion result and the current level representation using the following formula:

$$\begin{aligned}\mathbf{h}_k^1 &= \mathcal{D}(\mathbf{h}_{k-1}^1) + \hat{\mathbf{g}}_k^1 \in \mathbb{R}^{F_d \times 1 \times T_k}, k = 2, \dots, K, \\ \mathbf{h}_k^2 &= \mathcal{D}(\mathbf{h}_{k-1}^2) + \hat{\mathbf{g}}_k^2 \in \mathbb{R}^{F_d \times 1 \times T_k}, k = 2, \dots, K,\end{aligned}\quad (5)$$

where \mathcal{D} denotes the under-sampling operation with factor 2. This is an iterative process and terminates after approaching the final representations $\mathbf{h}_K^1 \in \mathbb{R}^{F_d \times 1 \times T_K}$ and $\mathbf{h}_K^2 \in \mathbb{R}^{F_d \times 1 \times T_K}$.

3.3. Contrastive learning

Contrastive learning, consisting of a projection head and a contrastive loss function, was adopted to measure the discriminative ability of the generated representations. The details are as follows:

3.3.1. Projection head

First a flatten layer was adopted to convert the output of the MGI into one dimension, which is denoted

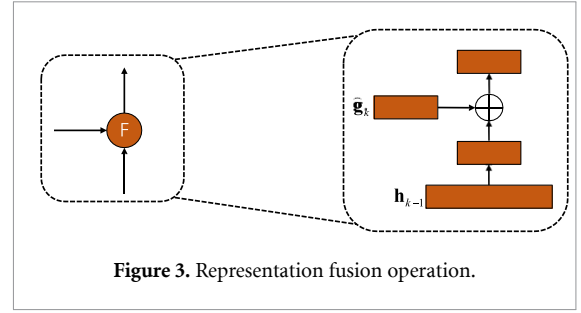


Figure 3. Representation fusion operation.

as $\mathbf{h}^1 \in \mathbb{R}^{F_d \times T_K}$ and $\mathbf{h}^2 \in \mathbb{R}^{F_d \times T_K}$. Recent studies have found that it is better to define contrastive loss in a low-dimensional space [44]. Inspired by this, we constructed the projection head module with two fully-connected layers to map the representation to the low-dimensional space. The output dimension of the two fully-connected layers is 64 and 16, respectively. This process can be formulated as:

$$\begin{aligned}\mathbf{z}^1 &= \mathbf{w}_2 \sigma(\mathbf{w}_1 \mathbf{h}^1), \\ \mathbf{z}^2 &= \mathbf{w}_2 \sigma(\mathbf{w}_1 \mathbf{h}^2),\end{aligned}\quad (6)$$

where σ is an ELU nonlinearity activation function, \mathbf{w}_1 and \mathbf{w}_2 are the learnable weights of two fully-connected layers, respectively. This activation function transforms each element of the input matrix and will not change the dimensions of the input. Then it can be further weighted by \mathbf{w}_1 and \mathbf{w}_2 .

3.3.2. Contrastive loss function

After the projection head module, a contrastive loss function was used to measure the similarity of two low-dimensional representations and a pair label. The similarity of two low-dimensional representations was calculated by cosine similarity, which is expressed by the following formula:

$$\text{sim}(\mathbf{z}^1, \mathbf{z}^2) = \mathbf{z}^1 \mathbf{z}^2 / \|\mathbf{z}^1\| \|\mathbf{z}^2\|. \quad (7)$$

Subsequently, the contrastive loss function is defined as:

$$L = -\log \frac{\sum_{m=1}^M \exp(\text{sim}(\mathbf{z}_m^1, \mathbf{z}_m^2)) \cdot \varphi(q_m = 1)}{\sum_{m=1}^M \exp(\text{sim}(\mathbf{z}_m^1, \mathbf{z}_m^2)) \cdot \varphi(q_m = -1)}, \quad (8)$$

where $\varphi \in \{0, 1\}$ is a discrimination function that evaluates to 1 if the condition in the bracket is satisfied.

Algorithm 1 summarizes the representation learning process of the proposed DRL.

Algorithm 1. Representation learning process of DRL.

Input: Training set \mathbf{X}
Output: The weight of DRL \mathbf{O}

- 1 Randomly construct M pairs $\{\mathbf{x}_m^1, \mathbf{x}_m^2\}_{m=1}^M$;
- 2 **for** $m \leftarrow 1, 2, \dots, M$ **do**
- 3 Generate multi-granular data by equation (1):
 $\tilde{\mathbf{x}}_{mk}^1, \tilde{\mathbf{x}}_{mk}^2 \leftarrow \mathbf{x}_m^1, \mathbf{x}_m^2$;
- 4 Extract temporal representation by equation (2):
 $\hat{\mathbf{x}}_{mk}^1, \hat{\mathbf{x}}_{mk}^2 \leftarrow \tilde{\mathbf{x}}_{mk}^1, \tilde{\mathbf{x}}_{mk}^2$;
- 5 Extract spatial representation by equation (3):
 $\tilde{\mathbf{g}}_{mk}^1, \tilde{\mathbf{g}}_{mk}^2 \leftarrow \hat{\mathbf{x}}_{mk}^1, \hat{\mathbf{x}}_{mk}^2$;
- 6 Reduce dimension and fusion by equations (4)
 and (5):
 $\mathbf{h}_m^1, \mathbf{h}_m^2 \leftarrow \tilde{\mathbf{g}}_{mk}^1, \tilde{\mathbf{g}}_{mk}^2$;
- 7 Through the projection head by equation (6):
 $\mathbf{z}_m^1, \mathbf{z}_m^2 \leftarrow \mathbf{h}_m^1, \mathbf{h}_m^2$;
- 8 **end**
- 9 Calculate the loss function by equation (8):
 $L \leftarrow \mathbf{z}_m^1, \mathbf{z}_m^2, \text{ and } q_m$;
- 10 Update the weight \mathbf{O} to minimize L ;
- 11 Return the weight of DRL \mathbf{O} .

3.4. Classification

In the classification process, the parameters of the feature extractor MGI was frozen and the classifier were constructed by a fully-connected layer and a softmax function. Considering unbalanced data may not be an issue for high-quality representation learning, but is disastrous for classifier learning [13], we separated the representation learning and classification processes, where the representation learning module can learn discriminative information from the full dataset. For the classifier learning process, the classifier trained on unbalanced data will boost the classification accuracy but result in a high false alarm rate, which means the majority of targets are not classified correctly. To address this problem, we selected the data distribution re-balance method including over-sampling and under-sampling. Considering that over-sampling may make the decision region of the learner smaller and more specific, which is easy to over-fit the target class, we adopted under-sampling to balance the data for classifier training. Specifically, we first under-sampled the training set \mathbf{X} to construct the training data \mathbf{X}^c for classification process. Then, the multi-granular spatial-temporal representation was extracted from a single EEG sample $\mathbf{x} \in \mathbb{R}^{C \times T}$ using the frozen feature extractor MGI. The output of MGI $\mathbf{h} \in \mathbb{R}^{F_d \times T_k}$ was then fed into a fully-connected layer by:

$$\mathbf{y} = \mathbf{h}\mathbf{Q} + \mathbf{b} = [y_1, y_2, \dots, y_E] \in \mathbb{R}^E, \quad (9)$$

where \mathbf{Q} is the transform matrix, which is used to map the data representations into label space for final classification, \mathbf{b} is a bias, which can increase the flexibility of the function and improve the fitting ability of neurons, and E is the number of classes. Subsequently, the

output vector \mathbf{y} is fed into a softmax layer for classification, which can be written as:

$$P(e|\mathbf{x}) = \exp(y_e) / \sum_{e=1}^E \exp(y_e), \quad (10)$$

where $P(e|\mathbf{x})$ denotes the probability that the input sample \mathbf{x} belongs to the e th class. Consequently, the prediction value p_v of \mathbf{x} is defined as:

$$p_v = \arg \max_e P(e|\mathbf{x}). \quad (11)$$

Finally, the classification loss is calculated by the cross-entropy loss.

4. Experiment setting**4.1. Dataset**

To evaluate the proposed DRL model, the experiments were constructed in three datasets, namely, two public datasets and one self-conducted RSVP EEG dataset. The detailed information of these datasets is described as follows.

4.1.1. Self-conducted dataset

- (a) Subjects: Eight subjects (six males and two females, aged 19–27 years) participated in the RSVP experiment. All subjects were students from Xidian University with normal or corrected-to-normal vision. None of them reported a history of neurological problems or serious diseases to affect the experimental results. The experiment was conducted according to the principles of the Declaration of Helsinki. Before the start of the experiment, the experimental process was described and the experimental task was specified in detail. All subjects signed consent forms, including permission to use their data for research.
- (b) Stimuli and procedure: RSVP-based BCIs can be used to assist users to distinguish targets and non-targets benefit from that the experimenter will have a specific brain response to the interest images in the RSVP stream. In our paradigm, we regarded images containing cars or persons as target images, and others as non-target images. The stimuli images were collected from the www.google.com/imghp website, as shown in figure 4. These images were manually resized to 800×600 pixels and images that were severely deformed were deleted. Using the above method, 500 target images and 1000 non-target images were collected for our experiment. To collect high-quality EEG signals, the subjects were seated in an electromagnetically shielded environment and maintained a suitable distance from the screen. Before the experiment, the subjects were instructed to focus on

Table 2. Number of samples in the public dataset 1.

Sub	Train		Test	
	Target	Non-target	Target	Non-target
S01	273	13 178	242	11 885
S02	192	9209	246	11 995
S03	274	13 145	278	13 582
S04	269	12 979	144	7051
S05	214	10 248	193	9441
S06	272	13 088	253	12 424

Because of the variant trial numbers, the target samples of each subject were different.

public dataset 1. The EEG data were segmented into segments of 1 s for further analysis, yielding a 256×256 resulting matrix for one sample. For each sample, we performed the filtering and normalization operation according to the preprocessing method used in the self-conducted dataset. The number of training and test samples are presented in table 2. Each subject had two session experiments: one session was used to train the model and one session to evaluate the model performance.

In our experiment, considering that the class imbalance problem exists in the RSVP task, all data were used for training or evaluation. In table 2, it is observed that the non-target samples are far more than the target samples in the public dataset 1.

4.1.3. The public dataset 2

This dataset was obtained from PhysioNet [50] and is publicly available at <https://physionet.org/content/ltrsvp/1.0.0/>. Eleven healthy subjects (seven males and four females) participated in the RSVP paradigm where one subject had data problems. Target images contained a randomly rotated and positioned airplane that had been photo-realistically superimposed, while non-target images did not contain airplanes. In each trial, 100 images were displayed at 5, 6, and 10 Hz, respectively [51], and 10 of them were target images. The EEG signals were collected from eight channels (PO7, PO8, P7, P8, PO3, PO4, O1, and O2) following the international 10–20 system with BioSemi ActiveTwo at a sampling rate of 2048 Hz. The signals were then band-pass filtered from 0.15–28 Hz and downsampled to 256 Hz, and then segmented into 1 s segments for further analysis, yielding an 8×256 matrix for one sample.

For three different circumstances with presentation rates of 5, 6, and 10 Hz, the number of target samples for each subject was 81, 68, and 45, respectively. It's not sufficient to train a traditional algorithm or a neural network because of the high dimensionality and few samples of EEG data. As such, a cross-subject method was adopted for evaluation and the entire dataset was divided into three groups of 5, 6, and 10 Hz. In each group, the first eight subjects were used as the training set and data from the rest subjects was used as the testing set. Table 3 lists the number of

Table 3. Number of samples in the public dataset 2.

Frequency	Train		Test	
	Target	Non-target	Target	Non-target
5-Hz	713	6415	178	1604
6-Hz	546	4918	137	1229
10-Hz	396	3564	99	891

Because of the variant trial numbers, the samples of each frequency were different.

Table 4. Details of multi-granular spatial-temporal information extraction, where $T_k = T/2^k$, ($k = 0, 1, \dots, K$), denotes the dimension in different granularity. F_t , F_s , and F_d are the numbers of temporal filters, spatial filters, and dimensionality reduction filters, respectively.

Block	Layer	Filter	Size	Output	Option
1	Input			(C, T)	
	Reshape			$(1, C, T)$	
	Sample			$(4, 1, C, T_k)$	
	Conv2D $\times 4$	F_t	$(1, 3)$	$(4, F_t, C, T_k)$	Same
	BN			$(4, F_t, C, T_k)$	
	ELU			$(4, F_t, C, T_k)$	
2	Dropout			$(4, F_t, C, T_k)$	$p = 0.8$
	Conv2D $\times 4$	F_s	$(C, 1)$	$(4, F_s, 1, T_k)$	same
	BN			$(4, F_s, 1, T_k)$	
	ELU			$(4, F_s, 1, T_k)$	
3	Dropout			$(4, F_s, 1, T_k)$	$p = 0.8$
	Conv2D $\times 4$	F_d	$(1, 1)$	$(4, F_p, 1, T_k)$	Same
	Flatten			$(4, F_p \times T_k)$	Same
	Fusion ^a $\times 3$			$F_p \times T_k$	

^a The operation of Fusion Layer is shown in figure 3.

samples, which shows the imbalance between targets and non-targets in the public dataset 2.

4.2. Implementation detail

The proposed DRL method was implemented using PyTorch [52]. In the learning representation stage, the model was trained using contrastive loss and the Adam [53] optimizer with a learning rate of 0.001. In the classification stage, the classifier was trained slightly differently, using the cross-entropy loss function. In addition, considering the computer memory limitations and computational issues, the number of pairs M was set to 20 000. The source code can be found at <https://github.com/Tammie-Li/DRL>.

Table 4 presents the architecture and parameters of MGI in detail. In our implementation, the number of temporal filters F_t , spatial filters F_s , dimensionality reduction filters F_d , and the granular levels were set to 8, 8, 2, and 4, respectively.

4.3. Evaluation metrics

In the experiment, the two selected metrics [54] were: unweighted average recall (UAR) and weighted average recall (WAR), which are two widely used metrics for evaluating the class imbalance problem. UAR is

the average accuracy of all classes, which can be calculated as:

$$\text{UAR} = \frac{1}{E} \sum_{e=1}^E \frac{n_e}{m_e}, \quad (12)$$

where Acc_e is the accuracy of the e th class and E is the number of classes. n_e and m_e denote the correct samples and total samples in each class, respectively. The UAR metric indicates the average accuracy of the target and non-target samples. Using the UAR results, the performance of predicting target samples can be adequately evaluated. WAR is also regarded as accurate, with respect to the recognition accuracy of the overall EEG samples. The calculating process of WAR is defined as:

$$\text{WAR} = \frac{N}{M}, \quad N = \sum_{e=1}^E n_e, \quad M = \sum_{e=1}^E m_e, \quad (13)$$

where N and M are the correct samples and total samples, respectively.

5. Experimental results and discussion

5.1. Performance in the RSVP task

To evaluate the performance of the proposed DRL, the same experiments were conducted using six comparison methods. These methods include:

- (a) Regularized LDA (rLDA) [24] is a version of LDA regularized by means of shrinkage estimators, which displays advantages over other LDA-based methods.
- (b) HDCA [25] is a commonly used two-stage machine learning method, which displays excellent performance in the RSVP task.
- (c) DeepConvNet [31] is a generic method for EEG decoding tasks in the BCI domain. Moreover, this is the first model for end-to-end EEG analysis.
- (d) EEGNet [32] has a specially designed compact and robust architecture for BCI classification tasks. Its performance has been verified in multiple paradigms, such as the ERP-based speller and sensory motor rhythms.
- (e) EEG-Inception [33] is the state-of-the-art method for ERP classification. This is the first model to integrate inception modules for ERP detection.
- (f) MGIFNet has the same structure as the model depicted in the bottom panel of figure 1. The difference is that, MGIFNet learned the weights using under-sampled data and a not frozen MGI subnetwork. This method can be used to evaluate the effect of MGI and highlight the contributions of the DRL model.

These six methods were re-implemented using MATLAB and Python, following the descriptions in

the original paper. It should be noted that the data in both the training and test sets were imbalanced in our experiment. If the original dataset is used to directly train the model without considering the impact of the class imbalance problem, the model overfits the dominant class seriously, which causes the UAR to tend to 0.5. To tackle this problem, for all comparison methods, we adopted the under-sampling, which is the most commonly used method in the RSVP task. Tables 5–7 display the UAR and WAR results of all the methods in the three RSVP datasets. From these tables, the following five observations were made:

- (a) The proposed DRL model outperforms all comparable methods in the three datasets. Specifically, from the results in table 5, the proposed method outperforms the traditional rLDA method by 10.28%/15.85% in the public dataset 1. Compared with the state-of-the-art deep learning method EEG-Inception, the UAR and WAR improved by 4.27% and 5.66%, respectively. From the results in table 6, the proposed method outperforms the traditional rLDA method by 7.70%/9.18% in the public dataset 2. Compared with the state-of-the-art deep learning method DeepConvNet, the UAR and WAR improved by 4.04% and 4.06%, respectively. In addition, the results in table 7 verify the outstanding performance of our method in the self-conducted RSVP dataset. Compared with the state-of-the-art methods in traditional and deep learning fields, the performance improved by 10.09%/9.74% (vs. rLDA) and 4.11%/6.07% (vs. DeepConvNet). These results verify the superior performance of DRL in the RSVP task.
- (b) The classification performance of deep learning methods is better than that of traditional methods in the three datasets. For the public dataset 1, compared with the two traditional methods (rLDA and HDCA), the UAR and WAR improved by 8.44% and 12.08%, respectively, using five deep learning methods (EEGNet, DeepConvNet, EEG-Inception, MGIFNet, and DRL). The gaps in the public dataset 2 and self-conducted dataset with respect to the methods were 5.38%, 6.08% and 7.76%, 5.87%, respectively. These results indicate that the deep learning method can capture more discriminative information from EEG signals.
- (c) MGIFNet achieves superior performance in the three datasets compared with other methods that adopt under-sampling to alleviate the class imbalance problem. MGIFNet improved UAR and WAR by 2.61%/1.00% compared with the state-of-the-art method EEG-Inception in the public dataset 1 and by 2.55%/2.03% compared with DeepConvNet in the public dataset 2. In addition, MGIFNet improved UAR and WAR by

Table 5. The UAR and the WAR (%) of all methods in the public dataset 1.

Subject	Method						
	rLDA [24]	HDCA [25]	DeepConvNet [31]	EEGNet [32]	EEG-Inception [33]	MGIFNet	DRL
1	78.44/80.76	70.11/76.67	87.90/91.92	85.24/89.65	85.66/92.81	86.64/92.45	90.26/94.90
2	75.98/75.29	71.48/74.67	78.86/88.86	77.16/83.99	77.19/84.14	82.32/88.14	82.85/91.07
3	74.27/77.25	72.32/76.44	77.62/88.26	79.70/86.23	81.64/86.59	84.75/88.13	86.01/91.37
4	73.18/79.69	78.63/81.01	81.92/88.01	79.05/81.85	77.96/83.84	82.41/88.15	83.26/93.46
5	72.11/75.83	68.59/70.13	73.55/76.63	71.91/74.70	77.96/80.69	78.05/79.00	80.31/87.04
6	73.25/70.22	66.98/66.85	74.50/85.28	82.98/88.65	82.87/92.14	84.80/90.31	86.22/96.31
Average	74.54/76.51	71.35/74.30	79.06/86.49	79.34/84.18	80.55/86.70	83.16/87.70	84.82/92.36
STD	2.11/3.42	3.70/4.62	4.82/4.82	4.25/4.99	3.09/4.43	2.73/4.20	3.15/3.01

Note: The left and right cell denote the UAR and WAR, respectively; STD denotes the standard deviation.

Table 6. The UAR and the WAR (%) of all methods in the public dataset 2.

Group	Method						
	rLDA [24]	HDCA [25]	DeepConvNet [31]	EEGNet [32]	EEG-Inception [33]	MGIFNet	DRL
5-Hz	68.62/65.08	65.78/66.27	68.99/70.26	70.43/70.59	68.09/70.88	74.49/73.40	75.45/74.24
6-Hz	65.52/66.54	64.99/66.18	70.80/71.96	70.10/69.55	68.97/70.42	71.49/72.62	74.01/75.50
10-Hz	59.48/59.39	58.53/58.48	64.81/64.14	64.25/64.75	62.57/62.53	66.26/66.44	67.27/68.80
Average	64.54/63.67	63.10/63.64	68.20/68.79	68.26/68.30	66.54/67.94	70.75/70.82	72.24/72.85
STD	3.80/3.08	3.25/3.65	2.51/3.36	2.84/2.54	2.83/3.83	3.40/3.11	3.57/2.91

Note: The left and right cell denote the UAR and WAR, respectively; STD denotes the standard deviation.

Table 7. The UAR and the WAR (%) of all methods in the self-conducted dataset.

Subject	Method						
	rLDA [24]	HDCA [25]	DeepConvNet [31]	EEGNet [32]	EEG-Inception [33]	MGIFNet	DRL
1	62.28/62.56	60.29/60.87	71.82/68.59	70.25/69.17	71.00/66.95	74.02/70.52	76.22/74.15
2	81.28/80.89	78.16/80.10	87.84/91.35	88.00/86.56	88.15/85.97	88.04/87.41	90.83/91.55
3	78.03/83.90	78.53/78.95	86.36/80.65	84.46/87.75	83.59/85.92	85.57/83.39	88.80/92.01
4	82.75/81.74	78.19/77.35	88.98/82.91	87.18/87.59	88.11/87.08	89.69/86.80	92.26/90.54
5	80.86/79.96	79.58/81.85	88.07/86.79	86.06/84.81	83.92/82.65	87.10/88.00	89.02/89.52
6	77.69/79.92	71.77/76.84	81.23/73.92	82.96/82.35	83.11/82.44	85.22/85.79	88.59/91.56
7	75.25/77.64	78.52/80.89	79.12/90.33	78.86/78.38	79.87/81.83	80.73/82.34	86.97/88.56
8	74.19/76.62	72.31/74.62	79.50/78.01	76.76/75.61	80.76/80.31	79.35/78.26	83.06/83.26
Average	76.88/77.90	74.66/76.43	82.86/81.57	81.81/81.53	82.31/81.64	83.71/82.81	86.97/87.64
STD	6.06/6.17	6.12/6.28	5.61/7.43	5.72/6.22	5.11/5.98	4.92/5.54	4.80/5.74

Note: The left and right cell denote the UAR and WAR, respectively; STD denotes the standard deviation.

- 0.85%/1.24% compared with the state-of-the-art method EEG-Inception in the self-conducted dataset. This verifies that analyzing the data with multiple granularities can provide more reasonable and complementary information.
- (d) WAR improved more than UAR. DRL outperformed previous state-of-the-art methods by 4.27%, 4.04%, and 4.11% in UAR and 5.66%, 4.06 %, and 6.07% in WAR in the three datasets, respectively. It is further observed that the improvement in WAR is 1.12% higher than that of UAR on average. The reason for this can be attributed to the following facts: (1) The non-target samples are far more numerous than the target samples in our test set. (2) Compared with UAR, the metric WAR tends to be dominated

- by the accuracy of the non-target samples. (3) DRL uses more non-target information than other methods included in the comparison that is based on under-sampling.
- (e) DRL has the ability to deal with the class imbalance problem. From tables 5 to 7, we can see that DRL outperforms MGIFNet by 1.66%/4.66% in the public dataset 1, 1.49%/2.03% in the public dataset 2, and 3.26%/4.83% in the self-conducted dataset, respectively. The difference between DRL and MGIFNet is that DRL adopts dual-branch architecture and the paired full data, which benefits for fitting the overall data distributions. This will be helpful for improving the performance of RSVP classification.

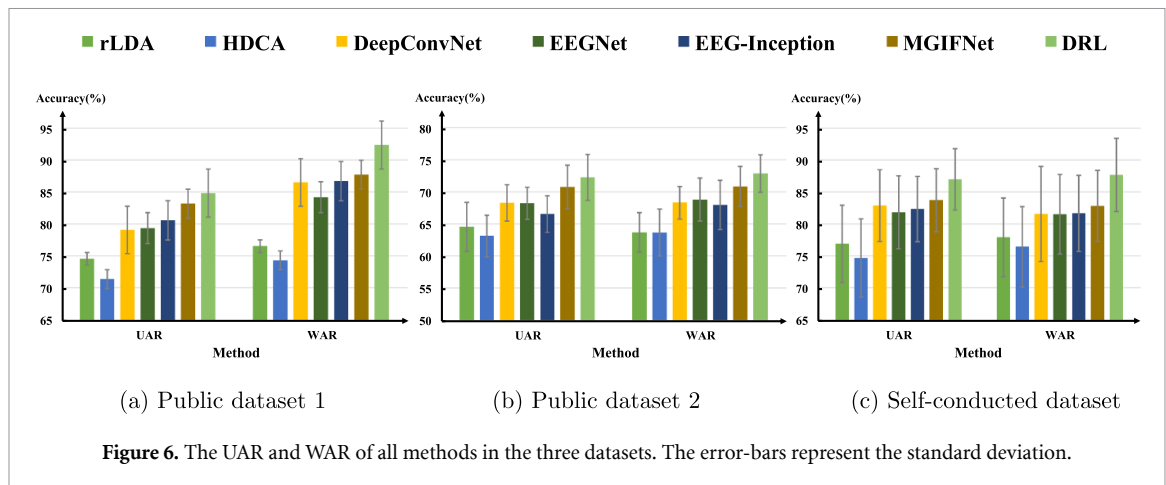


Figure 6. The UAR and WAR of all methods in the three datasets. The error-bars represent the standard deviation.

Table 8. The results of Shapro-Wilk test in three datasets. When the statistical value $p > 0.05$, which indicates the data follow the normal distribution, the results will be underlined.

Dataset	Method						
	rLDA	HDCA	DeepConvNet	EEGNet	EEGInception	MGIFNet	DRL
Public dataset 1	0.461/0.681	0.449/0.821	0.612/0.148	0.912/0.498	0.321/0.420	0.531/0.075	0.859/0.832
Public dataset 2	0.645/0.371	0.157/0.190	0.571/0.398	0.091/0.320	0.243/0.094	0.704/0.196	0.316/0.340
Self conducted	0.062/0.008	0.057/0.007	0.236/0.835	0.321/0.232	0.191/0.009	0.455/0.083	0.105/0.007

The results in figure 6 compare classification performances. It is clearly observed that the proposed DRL outperforms rLDA, HDCA, DeepConvNet, EEGNet, EEG-Inception, and MGIFNet in three datasets. It should be noted that MGIFNet achieves better results than other previously proposed methods. This is because MGIFNet has access to complementary information from the analysis of the data with multiple granularities. Compared with MGIFNet, DRL further improves the performance because DRL adopts a dual-branch architecture and paired full dataset to alleviate the class imbalance problem.

Moreover, the Shapro-Wilk test (S-W test) [55] was performed to verify the data distribution following the normal distribution hypothesis. The statistical values of the S-W test were shown in table 55. From table 55, it can be observed that the results of all algorithms follow the normal distribution (the statistical values of the S-W test $p > 0.05$), except for the WAR of rLDA, HDCA, EEGInception, and DRL in the self-conducted dataset. When the data follow the normal distribution, a repeated-measures analysis of variance [56] was conducted to analyze the main effect across two metrics, which used classification methods as the factor. For WAR in the self-conducted dataset, a Friedman test [57] was performed. The statistical analyses were conducted using the SPSS software (IBM SPSS Statistics)³ and the statistical significance was defined as $p < 0.05$. From the results in table 56, significant differences in both UAR

Table 9. The results of repeated-measures analysis of variance in three datasets. The significance level is defined as $p < 0.05$.

Dataset	Metric	
	UAR	WAR
Public dataset 1	$F(6, 30) = 16.851$, $p < 0.001$	$F(6, 30) = 22.214$, $p < 0.001$
Public dataset 2	$F(6, 12) = 28.415$, $p < 0.001$	$F(6, 12) = 54.678$, $p < 0.001$
Self conducted	$F(6, 42) = 51.803$, $p < 0.001$	$\chi^2(6)^a = 31.982$, $p < 0.001$

^a Note that the Friedman test was adopted for WAR in the self-conducted dataset.

and WAR were found among the methods in three datasets. Subsequently, pair-wise comparisons using Bonferroni correction [58] were utilized to verify whether there were statistically significant differences between DRL and other methods. From the results in table 10, except for the comparison between our DRL and our MGIFNet, 27/30 results are less than 0.05, which shows DRL statistically significantly improved the UAR and WAR compare with other methods. In summary, the proposed DRL provided significantly better classification performance than other methods.

5.2. Ability of DRL to alleviate the class imbalance

To further explore the ability of DRL to alleviate the class imbalance, additional experiments were conducted in three datasets by replacing the feature extractor MGI with other commonly used methods in the RSVP task, such as DeepConvNet [31], EEGNet [32], and EEG-Inception [33], which can be denoted

³ SPSS software will automatically adjust the p -value to control the false-positive errors.

Table 10. Pair-wise comparisons using Bonferroni correction between DRL and the other models in three datasets. SPSS Bonferroni adjusted p values are quoted. The significance level is defined as $p < 0.05$.

Dataset	DRL vs. rLDA	DRL vs. HDCA	DRL vs. DeepConvNet	DRL vs. EEGNet	DRL vs. EEGInception	DRL vs. MGIFNet
Public dataset 1	‡/*	‡/*	*/~	*/*	~/~	~/~
Public dataset 2	‡/‡	‡/‡	*/‡	*/‡	‡/‡	~/~
Self conducted	‡/‡	‡/‡	‡/‡	‡/*	‡/*	#/~

Note: ~: nonsignificant, *: $p < 0.05$, #: $p < 0.01$, ‡: $p < 0.005$, ‡: $p < 0.001$.

Table 11. The UAR and WAR (%) of three commonly methods after using DRL in the public dataset 1.

Sub	Method					
	DeepConvNet [31]	DeepConvNet + DRL	EEGNet [32]	EEGNet + DRL	EEGInception [33]	EEGInception + DRL
1	87.90/91.92	88.75/92.46	85.24/89.65	87.40/93.81	85.66/92.81	86.10/93.44
2	78.86/88.86	77.93/92.20	77.16/83.99	77.89/92.04	77.19/84.14	76.59/92.37
3	77.62/88.26	87.18/93.42	79.70/86.23	78.73/92.14	81.64/86.59	87.19/93.52
4	81.92/88.01	80.05/87.88	79.05/81.85	83.56/88.51	77.96/83.84	83.05/86.52
5	73.55/76.63	79.36/87.53	71.91/74.70	77.67/82.59	77.96/80.69	78.84/86.28
6	74.50/85.28	83.73/94.26	82.98/88.65	85.13/95.08	82.87/92.14	87.60/95.20
Avg	79.06/86.49	82.83/91.29	79.34/84.18	81.73/90.70	80.55/86.70	83.23/91.22
STD	4.82/4.82	4.05/2.62	4.25/4.99	3.81/4.15	3.09/4.43	4.21/3.51
Imp ^a		3.77/4.80		2.39/6.52		2.68/4.52

^a Note: Imp denotes classification performance improvement after using DRL.

as DeepConvNet + DRL, EEGNet + DRL, and EEG-Inception + DRL. Specifically, similar to our DRL, we first trained the feature extractor, i.e. DeepConvNet, EEGNet, and EEG-Inception, with our dual-branch architecture on the paired full dataset. Then we froze the parameters of the feature extractor and learned the classifier using the under-sampling dataset. The results are presented in tables 11–13.

In these tables, it is observed that the UAR and WAR significantly improved using our proposed DRL, which verifies that DRL can effectively alleviate the class imbalance problem in the three datasets. Compared with the aforementioned results of DeepConvNet, EEGNet, and EEG-Inception in table 11, the methods that use the proposed DRL improve UAR by 3.77%, 2.39%, 2.68%, and WAR by 4.80%, 6.52%, and 4.52% in the public dataset 1, respectively. From table 12, the methods that use the DRL improve UAR by 2.23%, 2.37%, 2.37%, and WAR by 2.51%, 2.44%, and 2.88% in the public dataset 2. Similarly, in the self-conducted dataset, table 13 shows that UAR improved by 2.75%, 3.74%, and 2.28%, and the WAR improved by 4.59%, 4.26%, and 3.88%, respectively, after using the proposed DRL. Furthermore, it was determined that the improvement of WAR was more significantly improved for the following two reasons: DRL makes full use of non-target information and non-target samples dominate the calculation of WAR.

To verify that the proposed DRL has achieved statistically significant improvement for these commonly used methods in the RSVP task, statistical analysis was performed between the methods with/without

DRL. We first adopted the S-W test to check whether the data distribution follows the normal distribution hypothesis. Except EEG-Inception ($p = 0.009$) and EEG-Inception + DRL ($p = 0.003$) in the self-conducted dataset, all statistical values p are greater than 0.05, indicating that the data conforms to the normal distribution. Subsequently, the paired t-test [59] was performed at a significance level of 0.05. For the statistical analysis between EEG-Inception and EEG-Inception + DRL, the Wilcoxon signed-rank test [60] was performed. The results are listed in table 14. In the public dataset 2 and the self-conducted dataset, the UAR and WAR of the DRL were significantly improved ($p < 0.05$). In public dataset 1, the WAR was significantly improved ($p < 0.05$), but no significant difference in the UAR (DeepConvNet: $p = 0.129$; EEGNet: $p = 0.062$; EEG-Inception: $p = 0.061$). This is because the DRL uses more non-target information, which seems difficult to directly improve the performance in UAR. In summary, the performance of the DRL achieves a statistically significant improvement on DeepConvNet, EEGNet, and EEG-Inception.

5.3. Comparisons with over-sampling methods

Considering the final loss function in DRL is obtained by comparing the similarity of sample pairs, which means the positive samples are repeatedly used. Two over-sampling methods (SMOTE [35] and ADASYN [36]) were compared with the DRL to further prove the superiority. SMOTE analyzes the minority samples and artificially synthesizes new samples

Table 12. The UAR and WAR (%) of three commonly methods after using DRL in the public dataset 2.

Group	Method					
	DeepConvNet [31]	DeepConvNet + DRL	EEGNet [32]	EEGNet + DRL	EEGInception [33]	EEGInception + DRL
5-Hz	68.99/70.26	71.26/72.49	70.43/70.59	72.68/73.08	68.09/70.88	71.44/73.40
6-Hz	70.80/71.96	73.35/73.84	70.10/69.55	73.35/72.92	68.97/70.42	71.16/74.24
10-Hz	64.81/64.14	66.67/67.56	64.25/64.75	65.86/66.22	62.57/62.53	64.14/64.81
Avg	68.20/68.79	70.43/71.30	68.26/68.30	70.63/70.74	66.54/67.94	68.91/70.82
STD	2.51/3.36	2.79/2.70	2.84/2.54	3.38/3.20	2.83/3.83	3.38/4.26
Imp ^a	2.23/2.51		2.37/2.44		2.37/2.88	

^a Note: Imp denotes classification performance improvement after using DRL.

Table 13. The UAR and WAR (%) of three commonly methods after using DRL in the self-conducted dataset.

Sub	Method					
	DeepConvNet [31]	DeepConvNet + DRL	EEGNet [32]	EEGNet + DRL	EEGInception [33]	EEGInception + DRL
1	71.82/68.59	76.16/74.49	70.25/69.17	73.90/75.13	71.00/66.95	74.00/72.00
2	87.84/91.35	88.65/88.52	88.00/86.56	87.96/85.14	88.15/85.97	88.05/87.21
3	86.36/80.65	90.06/90.84	84.46/87.75	91.48/91.79	83.59/85.92	85.61/89.06
4	88.98/82.91	90.78/91.34	87.18/87.59	88.92/88.59	88.11/87.08	89.31/90.13
5	88.07/86.79	89.12/89.91	86.06/84.81	88.76/89.37	83.92/82.65	87.11/87.92
6	81.23/73.92	87.06/88.93	82.96/82.35	85.39/84.94	83.11/82.44	86.03/87.63
7	79.12/90.33	82.24/83.27	78.86/78.38	85.65/86.14	79.87/81.83	83.52/86.49
8	79.50/78.01	80.84/82.01	76.76/75.61	82.37/85.18	80.76/80.31	83.06/83.70
Avg	82.86/81.57	85.61/86.16	81.81/81.53	85.55/85.79	82.31/81.64	84.59/85.52
STD	5.61/7.43	4.92/5.45	5.72/6.22	5.11/4.64	5.11/5.98	4.46/5.41
Imp ^a	2.75/4.59		3.74/4.26		2.28/3.88	

^a Note: Imp denotes classification performance improvement after using DRL.

Table 14. Statistics analyses between methods with/without DRL in three datasets. When the data follow to the normal distribution, paired t-test was performed, otherwise Wilcoxon signed-rank test was performed. SPSS Bonferroni adjusted *p* values are quoted. The significance level is defined as *p* < 0.05.

Dataset	DeepConvNet vs. DRL+ DeepConvNet	EEGNet vs. DRL+ EEGNet	EEGInception vs. DRL+ EEGInception
Public dataset 1	~/*	~/*	~/*
Public dataset 2	#/*	*/*	*/*
Self conducted	‡/~	‡/‡	‡/‡

Note: ~: nonsignificant, *: *p* < 0.05, #: *p* < 0.01, ‡: *p* < 0.005, ‡: *p* < 0.001.

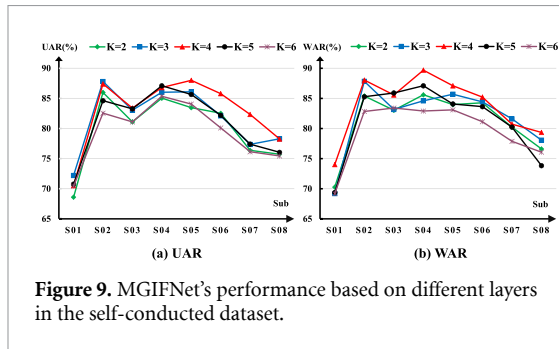
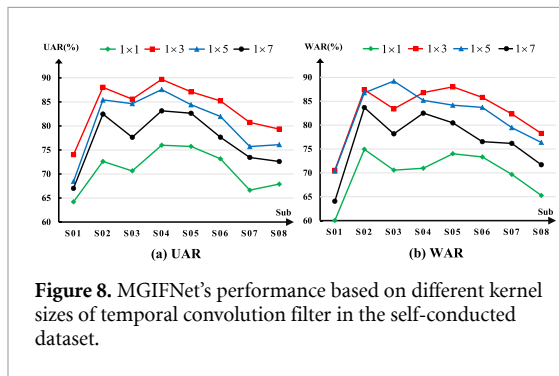
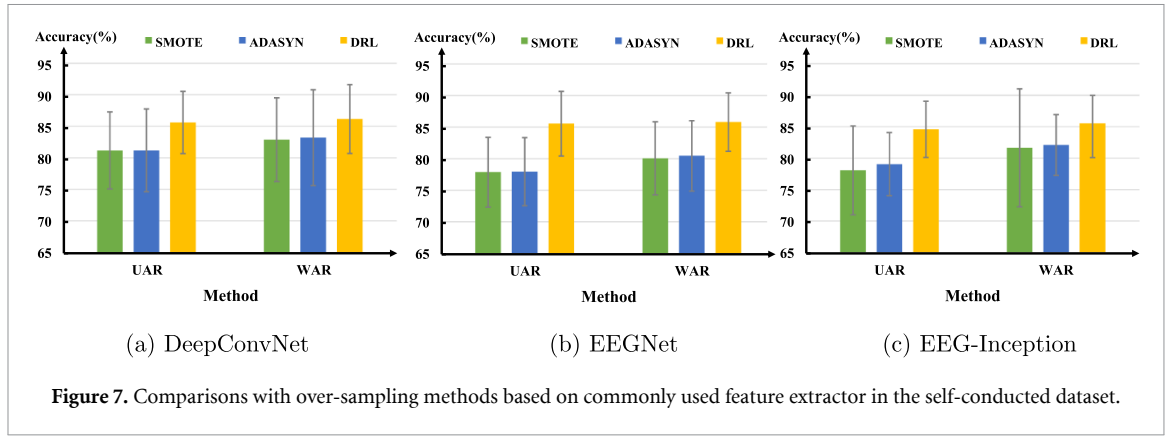
according to the minority samples, ADASYN gives different weights to different minority samples, so as to generate different numbers of samples. All experiments were based on the self-conducted dataset, and the feature extractor MGI was replaced by DeepConvNet [31], EEGNet [32], and EEG-Inception [33], respectively.

From the results in figure 7, when the feature extractor is DeepConvNet, DRL outperformed SMOTE and ADASYN by 4.44 %, 4.42 % in UAR and 3.28%, 2.96% in WAR. The gaps are 7.70%,

7.61% and 5.74%, 5.34% in EEGNet. Similarly, for EEG-Inception, the DRL improved the classification performance by 6.54%, 5.55% in UAR and 3.92%, 3.43% in WAR compared with SMOTE and ADASYN. DRL achieves obviously better performance than SMOTE and ADASYN on both UAR and WAR. This is because the wrong minority class may generate more interference samples, which is easy to cause the model to converge in the wrong direction.

5.4. Effects of temporal convolution with different kernel sizes on MGI architecture

The kernel size of the temporal convolution filter is essential, which determines the quality of the extracting dynamic information of EEG signals. To further explore the effect of temporal convolution with different kernel sizes on the performance of MGI architecture, additional experiments were conducted to evaluate MGIFNet's performance in the self-conducted dataset. Specifically, we changed the architecture and settings while varying the kernel size to 1×1 , 1×5 , and 1×7 . The results are shown in figure 8. It can be seen from figure 8 that the performance of MGIFNet achieves the best performance on kernel size of 1×3 . This is because the kernel size of 1×1 is difficult to extract the temporal dynamic information, and relatively larger kernel sizes, such as



1×5 , 1×7 , may over-smooth the EEG signal, which neglects the signal variety in the time dimension.

5.5. Effects of layer number on MGI architecture

The effect of layer number K on MGI architecture is analyzed, which determines the minimum unit of granularity in the temporal representation learning. To explore the effect of the layer number, the performance of MGIFNet was evaluated with the different number of layers from 2 to 6 with 1 interval in the self-conducted dataset. The results are shown in figure 9. It can be seen from figure 9 that the best performance of MGIFNet is achieved at $K = 4$. This is because deeper MGI layers can obtain larger receptive fields, which is beneficial for extracting the long-term temporal correlation of EEG signals. However, when the layer is too deep, it easily over-smooths the EEG signal, which may conduct a negative impact on the feature learned in the shallow layer.

5.6. Effects of under-sampling process on multi granular information extraction

In the MGI architecture, EEG signals would be progressively under-sampled with the factors 2, 2, and 2. The under-sampling process is essential to extract the MGI, which can directly influence the performance of the MGI architecture. To explore the effect of the under-sampling process of the MGI architecture, experiments were conducted based on MGIFNet in the self-conducted dataset. For EEG signals with a time length of 256, factor 4 is the maximum unit. In our experiment, the three factors were changed from the minimum value of 2 to the maximum value of 4 with 1 interval, respectively. The results are shown in figure 10 and the error-bars represent the standard deviation. It can be seen from figure 10 that the smaller factors tend to achieve better performance in both UAR and WAR. This is because the smaller factors are beneficial for extracting the short-term temporal correlation of EEG signals.

5.7. Performance in the online system

To further verify the performance of our proposed method in practical application, we invited three subjects to participate in the online experiment. Firstly, 24 blocks of data were recorded and preprocessed as the method in section 4.1.1. Then we trained the state-of-the-art method DeepConvNet and our proposed method DRL. These models were tested in the online system for six blocks. From the results in table 15, it can be seen that DRL outperforms DeepConvNet by 3.48% in UAR and 4.75% in WAR. In addition, to verify that the DRL is statistically significantly better than DeepConvNet in the online system, a paired t-test statistical analysis was performed at a significance level of 0.05. Before the t-test statistical analysis, the S-W test was adopted to ensure the data distribution follows the normal distribution hypothesis (DeepConvNet: $p = 0.881$ in UAR and $p = 0.519$ in WAR; DRL: $p = 0.577$ in UAR and $p = 0.110$ in WAR). The t-test result (UAR: $p = 0.026 < 0.05$; WAR: $p = 0.024 < 0.05$) shows the performance of DRL is significantly better than DeepConvNet in the online

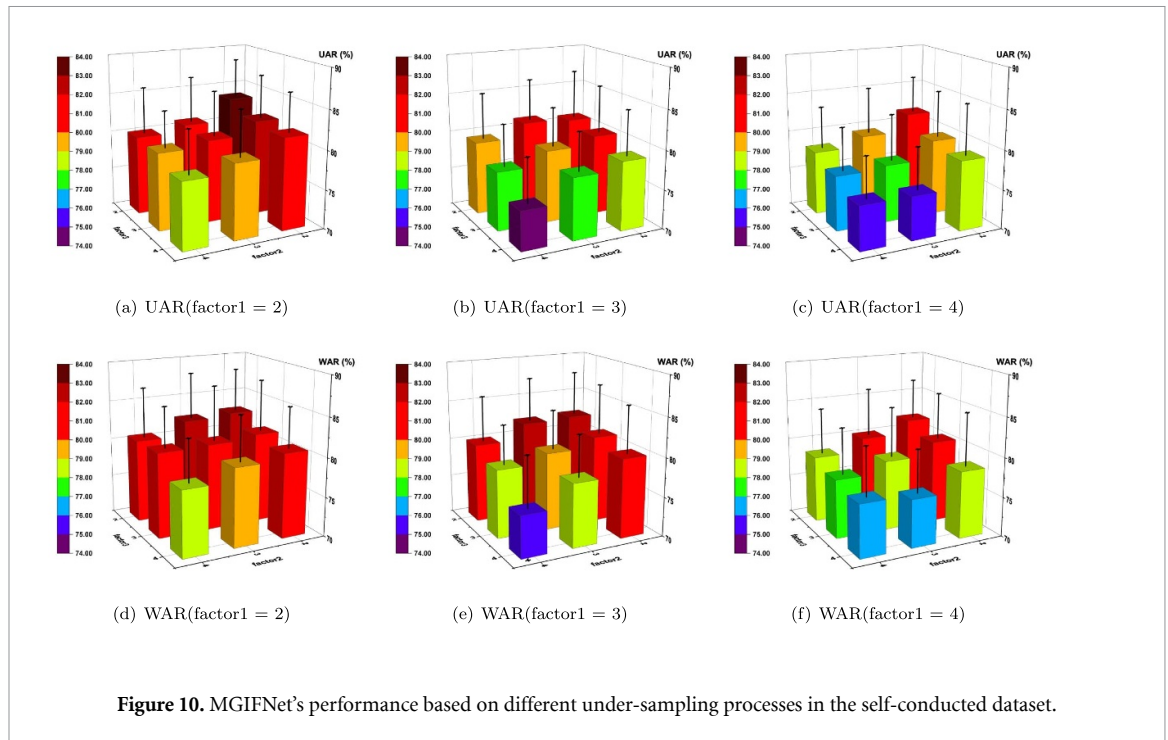


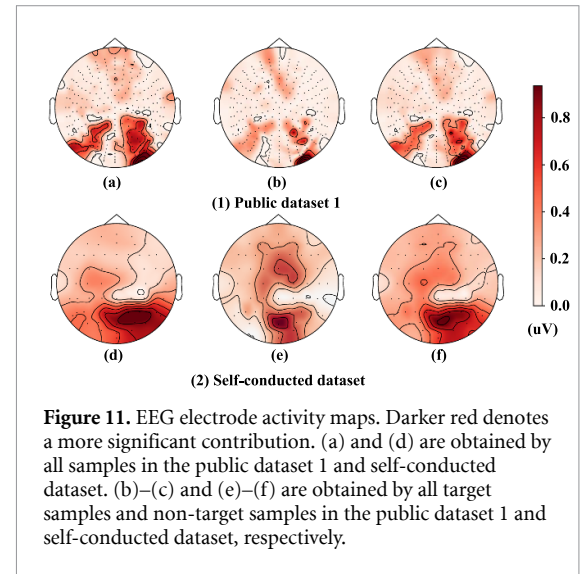
Table 15. Performance in the online system.

Subject	Method	
	DeepConvNet	DRL
S01-online	87.19/85.23	90.11/91.07
S02-online	86.06/84.87	88.96/88.17
S03-online	88.10/86.17	92.73/91.27
Average	87.12/85.42	90.60/90.17
STD	0.84/0.55	1.93/1.73

system, which indicates great potential for practical application.

5.8. Activity of EEG electrodes

To explore the contribution of different electrode sites in the decoding of RSVP EEG signals, the electrode activity maps are displayed in figure 11 on public dataset 1 and self-conducted dataset. Note that the topographical maps in figure 11 are generated by MNE-python [61]. The contribution of each electrode site was evaluated by calculating the L2 norm of the average of the multi-granular temporal representations $\tilde{\mathbf{x}}_k$ and mapping these values into the corresponding electrodes. As shown in figures 11(a) and (d), the brain activity is mainly distributed in the occipital and parietal electrode sites. The reason is that the parietal lobe is strongly associated with the integration of visual and somatosensory information [62], and the occipital lobe is mainly responsible for processing visual information [63]. Meanwhile, to investigate the difference between classes, the electrode activity maps of the target and non-target classes are further depicted in figures 11(b), (c), (e) and (f). From the results, compared with non-target



samples, it is observed that the target class displays a broader activity region and stronger intensity than the non-target class. Thus, we hypothesized that this activity represented changes in the P300 ERP component which are typically seen in response to attended targets in RSVP streams [9, 10].

5.9. Representation visualization

To verify the discriminative ability of the representations obtained by MGIFNet and DRL, the data representations were visualized using the t-distributed stochastic neighbor embedding (t-SNE) [64] method, as shown in figure 12. In our experiment, the perplexity, learning rate, and n-iter are set to 30, 1000, and 1000, respectively. The original data present a completely inseparable distribution,

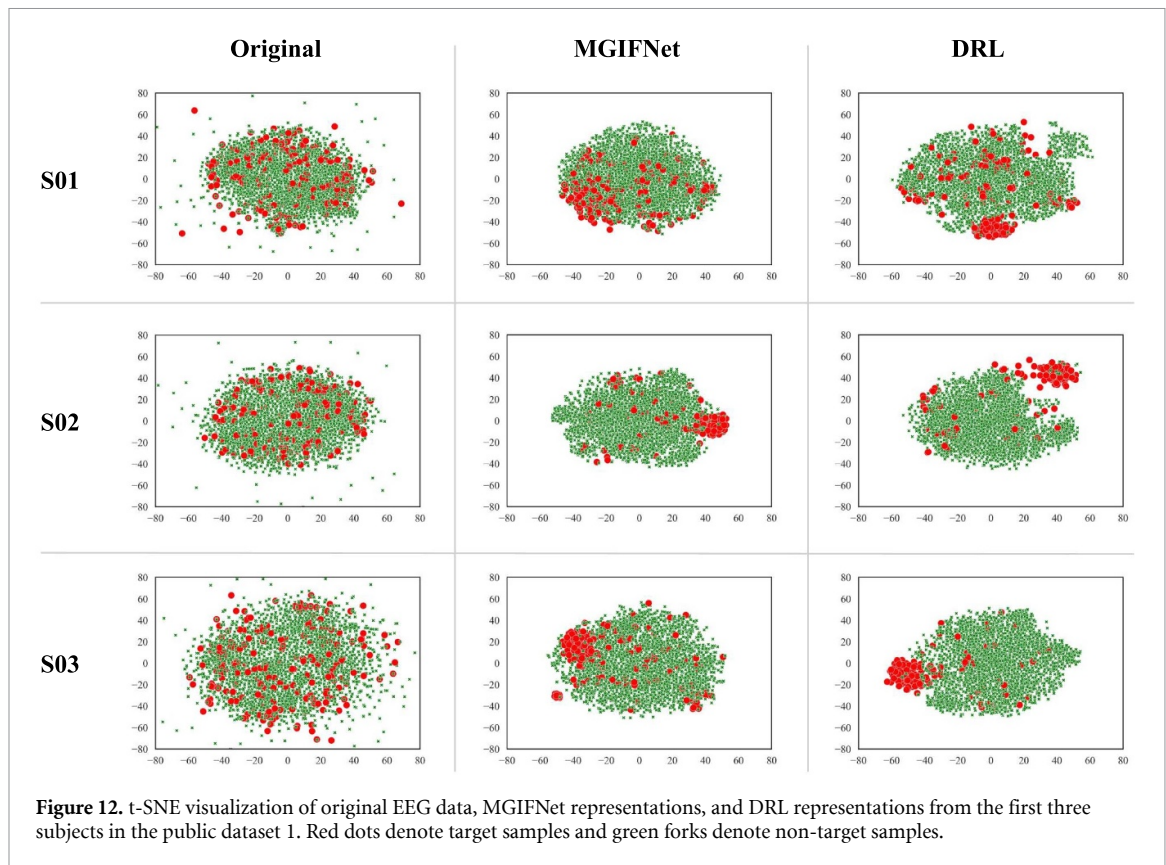


Figure 12. t-SNE visualization of original EEG data, MGIFNet representations, and DRL representations from the first three subjects in the public dataset 1. Red dots denote target samples and green forks denote non-target samples.

whereas the representations extracted by MGIFNet are separable between the target and non-target data. This is because MGIFNet has access to complementary information from the analysis of the data with multiple granularities. Furthermore, it should be noted that the representation gap is clearer after using DRL because the contrastive loss makes the positive pairs closer and pushes the negative pairs farther. In summary, the results reveal that MGIFNet and DRL have the ability to effectively extract discriminant representations for the RSVP EEG data.

5.10. Future study

On a whole, this is the first study to consider the class imbalance problem in the RSVP task. To address this problem, we provided a DRL model that separates the representation learning and classification processes. DRL model learned the universal patterns using a dual-branch architecture and paired full dataset. Subsequently, the DRL model determined a decision bounding for each class using under-sampled data. We believe this method can be applied to other neuroscience paradigms with class imbalances such as vigilance tasks and prediction tasks. The use of DRL to enhance the classification of EEG activity during these tasks would be beneficial for theoretical neuroscience studies employing these paradigms. Furthermore, there are many disorders of sustained attention and prediction, and neurofeedback paradigms harnessing these paradigms to train specific cognitive skills might be able to harness DRL training. In future

work, we will continue to focus on the class imbalance problem and explore further techniques for EEG classification.

6. Conclusion

This paper focuses on the class imbalance problem and proposes DRL, a novel DRL model, based on MGI to improve the classification performance in the RSVP task. DRL adopts a dual-branch architecture that decouples the learning process into representation learning and classification to eliminate the impact of class imbalance. Moreover, considering the multi-granular characteristics in RSVP EEG data, MGI, a novel MGI decoder based on DRL, extracts spatial-temporal representation at multiple granularity levels. The proposed framework is easy to implement, and extensive experiments on one public dataset and one self-conducted dataset demonstrate that the proposed DRL achieves state-of-the-art performance. The better classification performance of DRL is attributed to the fact that it can not only better tackle the class imbalance problem in online RSVP tasks but also extract the multi-granular spatial-temporal information from RSVP EEG signals.

Data availability statement



The data generated and/or analysed during the current study are not publicly available for legal/

ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgments

This work was supported in part by the National Key Research and Development Project of China (2018YFB2202400), NSFC (Nos. 61672404, 61875157, 61751310, 61836008 and 61632019), Science and Technology Plan of Xi'an (20191122015KYPT011JC013), Scientific Research Program Funded by Shannxi Provincial Education Department (No. 20JY022), the Fundamental Research Funds of the Central Universities of China (Nos. JC1904, JX18001 and QTZX2107).

ORCID iDs

Fu Li  <https://orcid.org/0000-0003-0319-0308>
 Yang Li  <https://orcid.org/0000-0002-5093-2151>
 Youshuo Ji  <https://orcid.org/0000-0002-6802-0759>

References

- [1] Lance B J, Kerick S E, Ries A J, Oie K S and McDowell K 2012 Brain-computer interface technologies in the coming decades *Proc. IEEE* **100** 1585–99
- [2] Galán F, Nutton M, Lew E, Ferrez P W, Vanacker G, Philips J and Millán J del R 2008 A brain-actuated wheelchair: asynchronous and non-invasive brain-computer interfaces for continuous control of robots *Clin. Neurophysiol.* **119** 2159–69
- [3] Thulasidas M, Guan C and Wu J 2006 Robust classification of EEG signal for brain-computer interface *IEEE Trans. Neural Syst. Rehabil. Eng.* **14** 24–29
- [4] Schwartz A B, Cui X T, Weber D J and Moran D W 2006 Brain-controlled interfaces: movement restoration with neural prosthetics *Neuron* **52** 205–20
- [5] Bigdely-Shamlo N, Vankov A, Ramirez R R and Makeig S 2008 Brain activity-based image classification from rapid serial visual presentation *IEEE Trans. Neural Syst. Rehabil. Eng.* **16** 432–41
- [6] Alpert G F, Manor R, Spanier A B, Deouell L Y and Geva A B 2014 Spatiotemporal representations of rapid visual target detection: a single-trial EEG classification algorithm *IEEE Trans. Biomed. Eng.* **61** 2290–303
- [7] Xu M, Han J, Wang Y, Jung T and Ming D 2020 Implementing over 100 command codes for a high-speed hybrid brain-computer interface using concurrent P300 and SSVEP features *IEEE Trans. Biomed. Eng.* **67** 3073–82
- [8] Lees S, Dayan N, Cecotti H, McCullagh P, Maguire L, Lotte F and Coyle D 2018 A review of rapid serial visual presentation-based brain-computer interfaces *J. Neural Eng.* **15** 021001
- [9] Picton T W 1992 The p300 wave of the human event-related potential *J. Clin. Neurophysiol.* **9** 456–79
- [10] Squires K C, Wickens C, Squires N K and Donchin E 1976 The effect of stimulus sequence on the waveform of the cortical event-related potential *Science* **193** 1142–6
- [11] Galar M, Fernandez A, Barrenechea E, Bustince H and Herrera F 2012 A review on ensembles for the class imbalance problem: bagging-, boosting- and hybrid-based approaches *IEEE Trans. Syst. Man Cybern. C* **42** 463–84
- [12] Japkowicz N and Stephen S 2002 The class imbalance problem: a systematic study *Intell. Data Anal.* **6** 429–49
- [13] Zhou B, Cui Q, Wei X and Chen Z M 2020 Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition *IEEE/ Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 9719–28
- [14] Kang B, Xie S, Rohrbach M, Yan Z, Gordo A, Feng J and Kalantidis Y 2019 Decoupling representation and classifier for long-tailed recognition (arXiv:1910.09217 [cs.CV])
- [15] Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A and Yger F 2018 A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update *J. Neural Eng.* **15** 031005
- [16] Iraj A et al 2019 Spatial dynamics within and between brain functional domains: a hierarchical approach to study time-varying brain function *Hum. Brain Mapp.* **40** 1969–86
- [17] Meunier D, Lambiotte R and Bullmore E T 2010 Modular and hierarchically modular organization of brain networks *Front. Neurosci.* **4** 200
- [18] Yao J, Vasilakos A V and Pedrycz W 2013 Granular computing: perspectives and challenges *IEEE Trans. Cybern.* **43** 1977–89
- [19] Yao Y 2000 Granular computing: basic issues and possible solutions *2000 5th Proc. Conf. Information Sciences (JCIS)* pp 186–9
- [20] Chen Y and Yao Y 2008 A multiview approach for intelligent data analysis based on data operators *Inf. Sci.* **178** 1–20
- [21] Gacek A and Pedrycz W 2006 A granular description of ECG signals *IEEE Trans. Biomed. Eng.* **53** 1972–82
- [22] Wang G and Xu J 2014 Granular computing with multiple granular layers for brain big data processing *Brain Informatics* **1** 1–10
- [23] Bigdely-Shamlo N, Vankov A, Ramirez R R and Makeig S 2008 Brain activity-based image classification from rapid serial visual presentation *IEEE Trans. Neural Syst. Rehabil. Eng.* **16** 432–41
- [24] Blankertz B, Lemm S, Treder M, Haufe S and Müller K-R 2011 Single-trial analysis and classification of ERP components—a tutorial *NeuroImage* **56** 814–25
- [25] Sajda P, Pohlmeier E, Wang J, Parra L C, Christoforou C, Dmochowski J, Hanna B, Bahlmann C, Singh M K and Chang S-F 2010 In a blink of an eye and a switch of a transistor: cortically coupled computer vision *Proc. IEEE* **98** 462–78
- [26] Xiao X, Xu M, Jin J, Wang Y, Jung T and Ming D 2020 Discriminative canonical pattern matching for single-trial classification of ERP components *IEEE Trans. Biomed. Eng.* **67** 2266–75
- [27] Craik A, He Y and Contreras-Vidal J L 2019 Deep learning for electroencephalogram (EEG) classification tasks: a review *J. Neural Eng.* **16** 031001
- [28] Cecotti H and Graser A 2011 Convolutional neural networks for p300 detection with application to brain-computer interfaces *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 433–45
- [29] Blankertz B et al 2004 The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials *IEEE Trans. Biomed. Eng.* **51** 1044–51
- [30] Cecotti H, Eckstein M P and Giesbrecht B 2014 Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering *IEEE Trans. Neural Netw. Learn. Syst.* **25** 2030–42
- [31] Schirrmester R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggersperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420
- [32] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces *J. Neural Eng.* **15** 056013
- [33] Santamaria-Vázquez E, Martínez-Cagigal V, Vaquerizo-Villar F and Hornero R 2020 EEG-inception: a novel deep convolutional neural network for assistive ERP-based brain-computer interfaces *IEEE Trans. Neural Syst. Rehabil. Eng.* **28** 2773–82

- [34] Oksuz K, Cam B C, Kalkan S and Akbas E 2021 Imbalance problems in object detection: a review *IEEE Trans. Pattern Anal. Mach. Intell.* **43** 3388–415
- [35] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique *J. Artif. Intell. Res.* **16** 321–57
- [36] He H, Bai Y, Garcia E A and Li S 2008 ADASYN: adaptive synthetic sampling approach for imbalanced learning *Int. Conf. on Neural Networks (IJCNN)* pp 1322–28
- [37] Liu X, Wu J and Zhou Z 2008 Exploratory undersampling for class-imbalance learning *IEEE Trans. Syst. Man Cybern. B* **39** 539–50
- [38] Wang S, Liu W, Wu J, Cao L, Meng Q and Kennedy P J 2016 Training deep neural networks on imbalanced data sets *Int. Conf. on Neural Networks (IJCNN)* pp 4368–74
- [39] Lin T, Goyal P, Girshick R, He K and Dollar P 2020 Focal loss for dense object detection *IEEE Trans. Pattern Anal. Mach. Intell.* **42** 318–27
- [40] Le-Khac P H, Healy G and Smeaton A F 2020 Contrastive representation learning: a framework and review *IEEE Access* **8** 193907–34
- [41] Becker S and Hinton G E 1992 Self-organizing neural network that discovers surfaces in random-dot stereograms *Nature* **355** 161–3
- [42] Chopra S, Hadsell R and LeCun Y 2005 Learning a similarity metric discriminatively, with application to face verification *IEEE/ Conf. on Computer Vision and Pattern Recognition (CVPR)* vol 1 pp 539–46
- [43] He K, Fan H, Wu Y, Xie S and Girshick R 2020 Momentum contrast for unsupervised visual representation learning *IEEE/ Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 9729–38
- [44] Chen T, Kornblith S, Norouzi M and Hinton G 2020 A simple framework for contrastive learning of visual representations *Int. Conf. on Machine Learning (ICML)* pp 1597–607
- [45] Chen X and He K 2021 Exploring simple siamese representation learning *IEEE/ Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 15750–8
- [46] Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C and Krishnan D Supervised contrastive learning (arXiv:2004.11362 [cs.LG])
- [47] Clevert D A, Unterthiner T and Hochreiter S 2015 Fast and accurate deep network learning by exponential linear units (elus) (arXiv:1511.07289 [cs.LG])
- [48] Peirce J W 2007 PsychoPy-Psychophysics software in Python *J. Neurosci. Methods* **162** 8–13
- [49] Rivet B, Cecotti H, Souloumiac A, Maby E and Mattout J 2011 Theoretical analysis of xDAWN algorithm: application to an efficient sensor selection in a p300 BCI *19th Eur. Signal Process. Conf. (EUSIPCO)* pp 1382–86
- [50] Goldberger A L, Amaral L A N, Glass L, Hausdorff J M, Ivanov P C, Mark R G, Mietus J E, Moody G B, Peng C-K and Stanley H E 2000 Physiobank, physiotoolkit and physionet: components of a new research resource for complex physiologic signals *Circulation* **101** e215–20
- [51] Matran-Fernandez A, Poli R and Hu D 2017 Towards the automated localisation of targets in rapid image-sifting by collaborative brain-computer interfaces *PLoS One* **12** 1–28
- [52] Paszke A et al 2019 Pytorch: an imperative style, high-performance deep learning library *Proc. Adv. Neural Inf. Process. Syst* pp 8026–37
- [53] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980 [cs.LG])
- [54] Schuller B, Vlasenko B, Eyben F, Wollmer M, Stuhlsatz A, Wendemuth A and Rigoll G 2010 Cross-corpus acoustic emotion recognition: variances and strategies *IEEE Trans. Affective Comput.* **1** 119–31
- [55] Hanusz Z, Tarasinska J and Zielinski W 2016 Shapiro-Wilk test with known mean *REVSTAT-Stat. J.* **14** 89–100
- [56] Park E, Cho M and Ki C S 2009 Correct use of repeated measures analysis of variance *Korean J. Lab. Med* **29** 1–9
- [57] Sheldon M R, Fillyaw M J and Thompson W D 1996 The use and interpretation of the friedman test in the analysis of ordinal-scale data in repeated measures designs *Physiotherapy Res. Int.* **1** 221–28
- [58] Bland J M and Altman D G 1995 Statistics notes: multiple significance tests: the Bonferroni method *Bmj* **310** 170
- [59] Semenick D 1990 Tests and measurements: the t-test *NSCA J.* **12** 36–37
- [60] Woolson R F 2007 Wilcoxon signed-rank test *Wiley Encyclopedia Clin. Trials* **1** 1–3
- [61] Gramfort A, Luessi M, Larson E, Engemann D A, Strohmeier D, Brodbeck C, Parkkonen L and Hämäläinen M S 2014 MNE software for processing MEG and EEG data *NeuroImage* **86** 446–60
- [62] Lynch J, Mountcastle V, Talbot W and Yin T 1977 Parietal lobe mechanisms for directed visual attention *J. Neurophysiol.* **40** 362–89
- [63] Grossberg S 1999 How does the cerebral cortex work? Learning, attention and grouping by the laminar circuits of visual cortex *Spatial Vis.* **12** 163–85
- [64] Van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605