A Grassmannian Manifold Self-Attention Network for Signal Classification

Rui Wang[1,2] , **Chen Hu**[1,2] , **Ziheng Chen**[3*] , **Xiao-Jun Wu**[1,2*] and **Xiaoning Song**[1,2]
[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
[2]Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence,
Jiangnan University, Wuxi, China
[3]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
{cs_wr, wu_xiaojun, x.song}@jiangnan.edu.cn, 6233112017@stu.jiangnan.edu.cn, ziheng_ch@163.com

Abstract

In the community of artificial intelligence, significant progress has been made in encoding sequential data using deep learning techniques. Nevertheless, how to effectively mine useful information from channel dimensions remains a major challenge, as these features have a submanifold structure. Linear subspace, the basic element of the Grassmannian manifold, has proven to be an effective manifold-valued feature descriptor in statistical representation. Besides, the Euclidean self-attention mechanism has shown great success in capturing long-range relationships of data. Inspired by these facts, we extend the self-attention mechanism to the Grassmannian manifold. Our framework can effectively characterize the spatiotemporal fluctuations of sequential data encoded in the Grassmannian manifold. Extensive experimental results on three benchmarking datasets (a drone recognition dataset and two EEG signal classification datasets) demonstrate the superiority of our method over the state-of-the-art. The code and supplementary material for this work can be found at https://github.com/ChenHu-ML/GDLNet.

1 Introduction

In recent years, the emerging field of computer vision and pattern recognition (CV&PR) has witnessed remarkable advancements, especially in the realm of sequential data (video clip, Electroencephalograph (EEG) signal, image set, *etc*) analysis [Nguyen, 2021; Chen *et al.*, 2021; Ingolfsson *et al.*, 2020]. Compared with the existing Euclidean feature learning methods, geometry-aware approaches that leverage manifold structures have gained prominence, mainly because they can capture appropriate statistical representations [Huang and Van Gool, 2017; Huang *et al.*, 2018; Nguyen *et al.*, 2019; Chakraborty *et al.*, 2020; Nguyen, 2021; Chen *et al.*, 2023]. One such fundamental latent space is the Grassmannian manifold [Edelman *et al.*, 1998a], the space of linear subspaces. The Riemannian geometry of linear subspaces provides a solid foundation for the characterization and analysis of sequential data, offering a powerful paradigm for capturing spatiotemporal relationships. In the field of medical imaging, linear subspace finds application in the classification of time-series data for Brain-Computer Interfaces (BCI) [Gao *et al.*, 2022] and in the analysis of magnetic resonance imaging [Chakraborty *et al.*, 2020]. For visual classification, its effectiveness has been well-substantiated across a spectrum of practical scenarios, such as dynamic scene classification [Wang *et al.*, 2021; Wei *et al.*, 2022b], facial emotion recognition [Huang *et al.*, 2018; Wang *et al.*, 2022b], face recognition [Wang and Wu, 2020; Wei *et al.*, 2022a], and action recognition [Nguyen and Yang, 2023; Chen *et al.*, 2024b].

While linear subspaces can accommodate the influence of data variations and demonstrate comparatively higher computational efficiency, their intrinsic Riemannian geometry hinders the direct generalization of Euclidean methods to the Grassmannian manifolds. Fortunately, the exploitation of projection operator [Huang *et al.*, 2015; Harandi *et al.*, 2013] to represent each Grassmannian element can bridge the gap mentioned above. In this scenario, the associated measurement on the Grassmannian manifold is known as the Projection Metric (PM) [Edelman *et al.*, 1998b]. Based on the PM, some methods utilize the kernel functions [Hamm and Lee, 2008; Harandi *et al.*, 2013] to achieve discriminative transformation of manifold data points to Euclidean representations, while others directly learn an embedding mapping between two Grassmasnnian manifolds [Huang *et al.*, 2015; Wang *et al.*, 2022b]. Although the latter could yield more discriminative features, there exists an inherent shortcoming that weakens its representational capacity, *i.e.*, feature learning on the nonlinear manifolds utilizing a linear embedding function.

Convolutional neural networks (ConvNets) [He *et al.*, 2016; Simonyan and Zisserman, 2014] are widely acknowledged for their superior performance compared to traditional shallow learning architectures in acquiring potent features. This stems not only from their ability to conduct multi-stage nonlinear computations but also from the effectiveness and scalability of the gradient-descent training procedure. Building upon this insight, certain researchers have undertaken efforts to generalize the ConvNets paradigm to the scenario of Riemannian manifolds [Huang and Van Gool, 2017; Huang

[*]Corresponding author

et al., 2018; Wang *et al.*, 2022a; Chakraborty *et al.*, 2020; Chen *et al.*, 2024b; Chen *et al.*, 2024a], injecting new dynamism into the realms of data modeling, learning, and classification. GrNet [Huang *et al.*, 2018] is a Riemannian neural network designed with an end-to-end architecture, introducing a deep and nonlinear learning mechanism tailored for linear subspaces. It consists of two fundamental trainable blocks, a Projection block and a Pooling block, for the implementation of data compression, nonlinear activation, and Grassmannian feature pooling.

This innovative architecture guarantees the preservation of Grassmannian properties for the input data at each layer. Subsequently, an Output block is designed to project the learned manifold representations into an Euclidean space for classification. This design has laid the groundwork for further developments in refining and tailoring the existing building blocks [Wang and Wu, 2020] to suit a variety of CV&PR tasks better.

The utilization of deep learning techniques in the field of Grassmannian manifolds is promising, but it remains in its infancy. The existing Grassmannian neural networks (GrasNets) encounter two major limitations: 1) the network inputs, *i.e.*, orthonormal basis matrices, are the global spatiotemporal representation of the raw sequential data, which may be underinformed; 2) there is no explicit statistical correlation established between channel dimensions. Both of these problems will have a potential negative impact on the learning ability of GrasNets. Given the great success of the self-attention mechanism defined in the Euclidean space in characterizing the long-range relationships between features [Dosovitskiy *et al.*, 2020; Huang *et al.*, 2022; Li *et al.*, 2022], we propose a self-attention mechanism on the Grassmannian manifold (GMSA) by referring to some geometric operators, including Riemannian metric, Riemannian mean, and Riemannian optimization. Based on GMSA, a geometric deep learning network, referred to as GDLNet, is proposed to address the aforementioned issues. Specifically, GDLNet first stacks several convolutional layers as an extractor of spatiotemporal representations w.r.t the raw data. Then, a manifold modeling module is attached, which maps the resulting features onto the Grassmannian manifold in two steps. Firstly, each input tensor data is grouped into m sections in the channel dimension, where m stands for the number of divisions. Secondly, we model each section as a Grassmannian element using eigenvalue decomposition. This is followed by a Riemannian self-attention module with the purpose of mining statistical complementarity between different channels.

In contrast to regular self-attention that operates on vectors, our approach proposes query, key, and value for the Grassmannian manifold. In such a case, to effectively mine and aggregate the geometrical dependencies between different channels, the Riemannian metric (PM in this paper) instead of the commonly used dot-product is exploited to measure the similarity between query and key. Based on the attention matrix computed, the PM-based weighted Fréchet mean (wFM) is naturally utilized to obtain the final outputs. The main reasons for using the PM-based wFM are threefold: 1) it is faithful to the Riemannian geometry of Grassmannian manifolds; 2) the Fréchet mean has shown theoretical and practical advantages in Riemannian data analysis [Chakraborty *et al.*, 2020]; 3) the PM has exhibited success in many applications [Huang *et al.*, 2018; Wang *et al.*, 2021]. Our main contributions can be summarized into the following three aspects:

- A self-attention mechanism is proposed on the Grassmannian manifold.
- A lightweight geometric deep learning network is designed for learning vibrant spatiotemporal representations of sequential data across Euclidean and Riemannian spaces.
- Extensive empirical validations of our model on three benchmarking datasets certify its effectiveness.

2 Preliminary

This section provides a brief review of the Grassmannian geometry. The Grassmannian manifold $\mathcal{G}(q, d)$ is comprised of a set of q-dimensional linear subspaces of the $\mathbb{R}^d$. Each linear subspace can be naturally represented by its orthonormal basis $\mathbf{Y}$ of size $d \times q$ ($\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \mathbf{I}_q$ and $\mathbf{I}_q$ is an identity matrix of size $q \times q$). Therefore, the matrix representation of Grassmannian is constituted by the equivalence class of orthonormal basis

$$[\mathbf{Y}] = \{\widetilde{\mathbf{Y}} \mid \widetilde{\mathbf{Y}} = \mathbf{Y}\mathbf{O}, \mathbf{O} \in \mathrm{O}(q)\}, \tag{1}$$

This definition is known as the Orthonormal Basis (ONB) perspective [Bendokat *et al.*, 2024]. By abuse of notation, we use $[\mathbf{Y}]$ or $\mathbf{Y}$ interchangeably.

As shown in [Bendokat *et al.*, 2024], each Grassmannian point can also be represented as an idempotent symmetric matrix of rank q by $\Phi(\mathbf{Y}) = \mathbf{Y}\mathbf{Y}^{\mathrm{T}}$, which is known as the projector perspective. This representation indicates that the Grassmannian is a submanifold of the Euclidean space of symmetric matrices. Therefore, an extrinsic distance can be induced by the ambient Euclidean space, which is also known as the Projection Metric (PM) [Hamm and Lee, 2008]:

$$\mathrm{d}_{\mathrm{PM}}(\mathbf{Y}_1, \mathbf{Y}_2) = 2^{-1/2}\|\mathbf{Y}_1\mathbf{Y}_1^{\mathrm{T}} - \mathbf{Y}_2\mathbf{Y}_2^{\mathrm{T}}\|_{\mathrm{F}}, \tag{2}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. As demonstrated in [Harandi *et al.*, 2013], the distance computed by the PM deviates from the true geodesic distance on the Grassmannian manifold up to a scale of $\sqrt{2}$, thus making it a widely used Grassmannian metric.

3 Proposed Method

As shown in Fig. 1, our GDLNet is composed of three components: a feature extraction module, a manifold modeling module, and a Grassmannian self-attention module. This section will give a detailed introduction to each of them.

3.1 Feature Extraction Module (FEM)

In this module, a stack of convolutional layers is tailored to different classification tasks to extract task-specific information from raw sequential data. For the RADAR dataset, the FEM is composed of a convolutional layer meant for extracting spatiotemporal information. For the EEG datasets, we follow [Wei *et al.*, 2019] to make the FEM contain two convolutional layers, one of which is used to impose spatial filtering to the multi-channel EEG signals, while the other extracts spatiotemporal features.

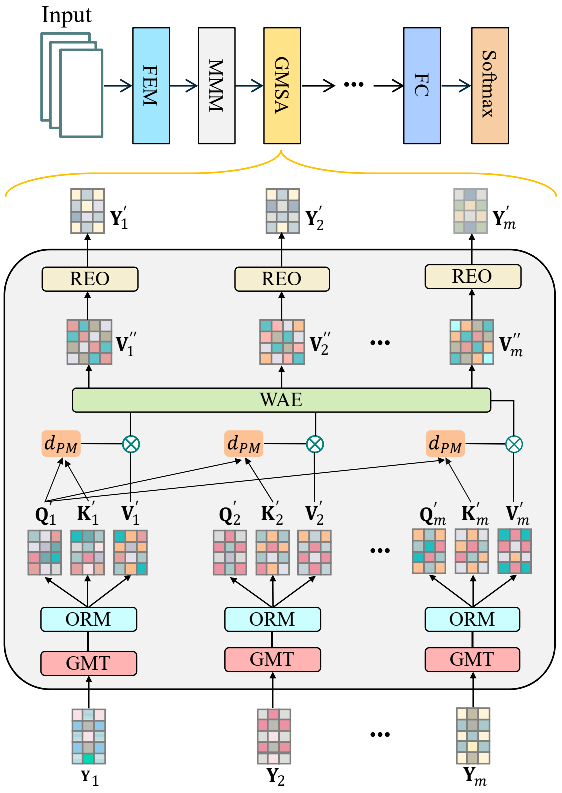

Figure 1: The architecture of the proposed GDLNet and GMSA, where $\mathbf{Q}_r', \mathbf{K}_r', \mathbf{V}_r'$ correspond to the query, key, and value for the r-th ($r \in \{1, 2, \cdots, m\}$) input matrix $\mathbf{Y}_r$, and $\mathbf{Y}_r'$ is the r-th output of GMSA. For classification, we follow GrNet to first exploit the projection map (Φ) to transform each $\mathbf{Y}_r'$ into a symmetric matrix (denoted by $\mathbf{Y}_r''$). Then, we vectorize each $\mathbf{Y}_r''$ and concatenate the resulting vectors. Finally, a FC layer followed by a Softmax function is applied for decision making.

3.2 Manifold Modeling Module (MMM)

Let $\mathbf{E}_i \in \mathbb{R}^{c \times l}$ be the i-th feature matrix generated by FEM w.r.t the i-th input data sequence. Here, c represents the number of channels, while l indicates the dimensionality of a channel. Since each point of $\mathcal{G}(q, d)$ represents a q-dimentional linear subspace of the d-dimentional vector space $\mathbb{R}^d$ (see Section 2), the Grassmannian manifold $\mathcal{G}(q, d)$ thus becomes a reasonable and efficient tool for parametrizing the q-dimentional real vector subspace embedded in $\mathbf{E}_i$ [Turaga *et al.*, 2011; Harandi *et al.*, 2011]. To capture complementary statistical information embodied in different channel features, as illustrated in Fig. 2, we first partition each $\mathbf{E}_i$ into m sections along the channel dimension, denoted as $\tilde{\mathbf{E}}_{i1}, \tilde{\mathbf{E}}_{i2}, \cdots, \tilde{\mathbf{E}}_{im}$. Then, a similarity matrix is computed for each $\tilde{\mathbf{E}}_{ir}$, followed by the SVD operation to obtain a q-dimentional linear subspace spanned by an orthonormal matrix $\mathbf{Y}_{ir} \in \mathbb{R}^{d \times q}$, s.t. $\tilde{\mathbf{E}}_{ir}\tilde{\mathbf{E}}_{ir}^{\mathrm{T}} \simeq \mathbf{Y}_{ir}\mathbf{\Sigma}_{ir}\mathbf{Y}_{ir}^{\mathrm{T}}$. Wherein, $\mathbf{Y}_{ir}$ and $\mathbf{\Sigma}_{ir}$ are two matrices consisting of q leading eigenvalues and the corresponding eigenvectors, respectively. For simplicity, we abbreviate $\mathbf{Y}_{ir}$ as $\mathbf{Y}_r$ in the following. Now, the resulting Grassmannian representation of $\mathbf{E}_i$ is denoted as $\Upsilon = [\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_m]$.

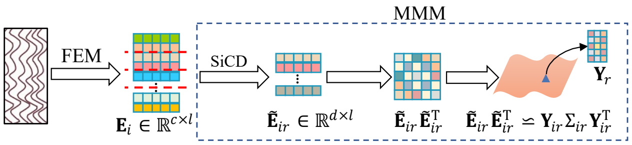
MMM
FEM
SiCD
$\mathbf{E}_i \in \mathbb{R}^{c \times l}$
$\tilde{\mathbf{E}}_{ir} \in \mathbb{R}^{d \times l}$
$\tilde{\mathbf{E}}_{ir}\tilde{\mathbf{E}}_{ir}^{\mathrm{T}}$
$\tilde{\mathbf{E}}_{ir}\tilde{\mathbf{E}}_{ir}^{\mathrm{T}} \simeq \mathbf{Y}_{ir}\Sigma_{ir}\mathbf{Y}_{ir}^{\mathrm{T}}$
$\mathbf{Y}_r$

Figure 2: An overview of Grassmannian modeling, where 'SiCD' means split in the channel dimension, $d = c/m$, c denotes the number of channels, and m signifies the number of segments. Here, the MMM takes the r-th ($r \in \{1, 2, \cdots, m\}$) segment as an example.

3.3 Grassmannian Manifold Self-Attention Module (GMSA)

An overview of the designed GMSA is illustrated in Fig. 1. It can be seen that the input to this module is a series of orthonormal matrices, *i.e.*, $\Upsilon = [\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_m]$. To capture the geometric relationship within Υ, we extend the Euclidean multihead attention mechanism [Vaswani *et al.*, 2017; Dosovitskiy *et al.*, 2020] to the Grassmannian manifolds, establishing query, key, and value based on the Grassmannian geometry. To intuitively articulate the forward propagation of each orthonormal matrix involved in GMSA, we design seven auxiliary layers as follows.

GMT layer. In contrast to the traditional linear transformation of vectors in Euclidean space, the generation of query, key, and value in our approach relies on the orthonormal matrices. Inspired by [Huang *et al.*, 2018], we design a Grassmannian transformation (GMT) layer to produce $\mathbf{Q}_r$, $\mathbf{K}_r$, and $\mathbf{V}_r$ from $\mathbf{Y}_r$ via the mapping function f_{gmt}:

$$\mathbf{Q}_r = f_{gmt}(\mathbf{W}_q, \mathbf{Y}_r) = \mathbf{W}_q\mathbf{Y}_r, \quad (3)$$

$$\mathbf{K}_r = f_{gmt}(\mathbf{W}_k, \mathbf{Y}_r) = \mathbf{W}_k\mathbf{Y}_r, \quad (4)$$

$$\mathbf{V}_r = f_{gmt}(\mathbf{W}_v, \mathbf{Y}_r) = \mathbf{W}_v\mathbf{Y}_r, \quad (5)$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d_t \times d}$ ($d_t < d$) are three projection matrices of row full rank. Since the underlying space of these weight matrices is a non-compact Stiefel manifold and the geodesic distance has no upper bound [Huang *et al.*, 2018; Edelman *et al.*, 1998b], it is infeasible to directly optimize them on such manifold. To tackle this issue, we follow [Huang *et al.*, 2018] to impose orthogonality constraint on each weight matrix $\mathbf{W}_b$ ($b \in \{q, k, v\}$), such that the weight space $\mathbb{R}_*^{d_t \times d}$ becomes a compact Stiefel manifold $St(d_t, d)$ [Absil *et al.*, 2009].

ORM layer. Inspired by [Huang *et al.*, 2018], we design the orthonormal maintaining (ORM) layer to impose orthonormality, preventing the matrices from degeneracy. Taking $\mathbf{K}_r$ as an example, since it is not an orthonormal basis matrix, a typical solution, *i.e.*, QR decomposition, is adopted to make the space spanned by the output feature matrices of the previous GMT layer be a valid Grassmannian manifold. Specifically, the QR decomposition w.r.t $\mathbf{K}_r$ is given below:

$$\mathbf{K}_r = \mathbf{\Omega}_r\mathbf{R}_r \quad (6)$$

where $\mathbf{\Omega}_r \in \mathbb{R}^{d_t \times q}$ and $\mathbf{R}_r \in \mathbb{R}^{q \times q}$ are the orthonormal and invertible upper-triangular matrices, respectively. Therefore, $\mathbf{K}_r$ can be normalized into an orthonormal basis matrix via the mapping f_{orm}:

$$\mathbf{K}_r' = f_{orm}(\mathbf{K}_r) = \mathbf{K}_r\mathbf{R}_r^{-1} = \mathbf{\Omega}_r. \quad (7)$$

Similarly, we can obtain $\mathbf{Q}'_r$ and $\mathbf{V}'_r$.

PM layer. In Euclidean space, the inner product is commonly used to measure the similarity between the query and key vectors. In contrast, the query, key, and value in our method are orthonormal matrices residing on the Grassmannian manifold. Therefore, in the designed Projection Metric (PM) layer (denoted as f_{pm}), we utilize Eq. (2) to compute the distance between $\mathbf{Q}_r$ and $\mathbf{K}_j$, given below:

$$\mathcal{D}_{rj} = f_{pm}(\mathbf{Q}'_r, \mathbf{K}'_j) = ||\mathbf{Q}'_r {\mathbf{Q}'_r}^T - \mathbf{K}'_j {\mathbf{K}'_j}^T||_{\mathrm{F}}^2, \qquad (8)$$

where $r, j \in \{1, 2, \cdots, m\}$.

SIM layer. However, the distances computed in the previous PM layer cannot be directly used as attention weights, because an increased similarity between any two samples invariably results in a decrease in their corresponding distance. To this end, we design the similarity measurement (SIM) layer to convert $\mathcal{D}_{rj}$ to be a valid form using the transformation function f_{sim}:

$$\mathcal{D}'_{rj} = f_{sim}(\mathcal{D}_{rj}) = \frac{1}{1 + \log(1 + \mathcal{D}_{rj})}. \qquad (9)$$

It can be seen that f_{sim} is a decreasing function w.r.t $\mathrm{d}_{\mathrm{PM}}(,)$. Now, we denote the attention matrix as $\boldsymbol{\mathcal{A}} = [\mathcal{D}'_{rj}]_{m \times m}$.

SMX layer. Considering that the values in each row of $\boldsymbol{\mathcal{A}}$ do not necessarily satisfy convexity constraint, we use the Softmax (SMX) function, denoted as f_{smx}, to compress its value range along the row direction in the designed SMX layer:

$$\mathcal{D}''_{rj} = f_{smx}(\boldsymbol{\mathcal{A}}) = \frac{\exp(\mathcal{D}'_{rj})}{\sum_{e=1}^{m} \exp(\mathcal{D}'_{re})}. \qquad (10)$$

Now, we use $\hat{\boldsymbol{\mathcal{A}}} = [\mathcal{D}''_{rj}]_{m \times m}$ to represent the final attention probability matrix.

WAE layer. As stated in Section 1, this article employs the weighted Fréchet mean (wFM) to realize the weighted average (WAE) operation involved in GMSA. Before introducing the designed WAE layer, we first give the definition of wFM on the Grassmannian manifolds.

Given a batch of Grassmannian activations $\{\mathbf{Y}_r\}_{r=1}^m$, their weighted Fréchet mean ($\boldsymbol{\mathcal{P}}^*$) can be defined as:

$$\boldsymbol{\mathcal{P}}^* = \underset{\boldsymbol{\mathcal{P}} \in \mathcal{G}(q,d)}{\arg\min} \sum_{r=1}^{m} w_r \mathrm{d}_{\mathcal{G}}^2(\mathbf{Y}_r, \boldsymbol{\mathcal{P}}), \qquad (11)$$

where $\mathrm{d}_{\mathcal{G}}$ is a distance on the Grassmannian manifold, and w_r is the weight assigned to each $\mathbf{Y}_r$, satisfying $w_r > 0$ for all $r \in \{1, 2, \cdots, m\}$ and $\sum_r w_r = 1$.

As noted in [Helmke *et al.*, 2007], $\mathcal{G}(q, d)$ is isometric to the $d \times d$ idempotent symmetric matrices of rank q, denoted as $\mathcal{IS}(q, d)$, which is a submanifold of the Euclidean space $\mathcal{S}^d$ of $d \times d$ symmetric matrices [Bendokat *et al.*, 2024]. The famous PM is the distance from the ambient space $\mathcal{S}^d$. Inspired by this, the PM is utilized to compute the wFM on $\mathcal{IS}(q, d)$:

$$\min_{\breve{\boldsymbol{\mathcal{P}}} \in \mathcal{IS}(q,d)} \sum_{r=1}^{m} w_r ||\breve{\mathbf{Y}}_r - \breve{\boldsymbol{\mathcal{P}}}||_{\mathrm{F}}^2, \qquad (12)$$

where $\breve{\mathbf{Y}}_r = \mathbf{Y}_r \mathbf{Y}_r^{\mathrm{T}}$ and $\breve{\boldsymbol{\mathcal{P}}} = \boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{P}}^{\mathrm{T}}$. However, Eq. (12) is a non-convex problem, as $\breve{\boldsymbol{\mathcal{P}}} \in \mathcal{IS}(q, d)$ implies a non-convex constraint, *i.e.*, $\breve{\boldsymbol{\mathcal{P}}}^2 = \breve{\boldsymbol{\mathcal{P}}}$. Considering that the essential purpose of designing the WAE layer is to achieve feature fusion on the Grassmannian manifold, we remove the idempotent constraint from Eq. (12). In this way, the above formula is reduced to the following form:

$$\min_{\breve{\boldsymbol{\mathcal{P}}} \in \mathcal{S}^d} \sum_{r=1}^{m} w_r ||\breve{\mathbf{Y}}_r - \breve{\boldsymbol{\mathcal{P}}}||_{\mathrm{F}}^2, \qquad (13)$$

where $\mathcal{S}^d$ denotes a set of real symmetric matrices. Then, the solution to Eq. (13) can be derived directly through the Euclidean weighted average:

$$\boldsymbol{\mathcal{P}}^* = \sum_{r=1}^{m} w_r \mathbf{Y}_r \mathbf{Y}_r^{\mathrm{T}}. \qquad (14)$$

Replacing w_r and $\mathbf{Y}_r$ with $\mathcal{D}''_{rj}$ and $\mathbf{V}'_j$ respectively, the WAE layer can be expressed as:

$$\mathbf{V}''_r = f_{wae}(\hat{\boldsymbol{\mathcal{A}}}, \boldsymbol{\mathcal{V}}) = \sum_{j=1}^{m} \mathcal{D}''_{rj} \cdot (\mathbf{V}'_j {\mathbf{V}'_j}^{\mathrm{T}}), \qquad (15)$$

where $\boldsymbol{\mathcal{V}} = \{\mathbf{V}'_1, \cdots, \mathbf{V}'_m\}$ is produced by the ORM layer.

REO layer. Finally, we design the reorthonormalization (REO) layer, denoted as f_{reo}, to map each $\mathbf{V}''_r \in \mathbb{R}^{d_t \times d_t}$ (symmetric matrix) output by the WAE layer back onto the Grassmannian manifold:

$$\mathbf{V}''_r = \mathbf{Z}\mathbf{S}\mathbf{Z}^{\mathrm{T}}, \qquad (16)$$

$$\mathbf{Y}'_r = f_{reo}(\mathbf{V}''_r) = \mathbf{Z}_{1:q}, \qquad (17)$$

where Eq. (16) represents the SVD operation, and $\boldsymbol{Z}_{1:q}$ is a matrix composed by the eigenvectors corresponding to the first q largest eigenvalues.

Now, the embedding mapping of GMSA can be expressed as: $\boldsymbol{\Upsilon}' = \phi_{GMSA}(\boldsymbol{\Upsilon})$, where $\boldsymbol{\Upsilon}' = \{\mathbf{Y}'_1, \mathbf{Y}'_2, \cdots, \mathbf{Y}'_m\}$. The forward pass of the proposed GMSA is summarized in Algorithm 1.

Remark 1. *Generally speaking, the computed Riemannian mean should be a valid point on the Grassmannian manifold, that is, it meets the following condition:*

$$\mathrm{wFM}(\{\mathbf{Y}_r\}) = f^{-1}(\mathrm{wFM}(\{f(\mathbf{Y}_r)\})), \qquad (18)$$

where $f(\mathbf{Y}_r) = \mathbf{Y}_r \mathbf{Y}_r^{\mathrm{T}}$, and f is a bijection from $\mathcal{G}(q, d)$ to $\mathcal{IS}(q, d)$. Easy computation shows that f^{-1} is defined by Eqs. (16-17). However, the learned $\boldsymbol{\mathcal{P}}^* \in \mathcal{S}^d$ in the WAE layer (Eq. (14)) is not idempotent. Nevertheless, we treat the designed REO function (f_{reo}) as an approximation to f^{-1}. Although this approach may not be able to render the optimal Riemannian mean, the composition of the functions of WAE and REO provides a feasible pathway for the realization of the weighted average on the Grassmannian manifolds. Besides, when $w_r = \frac{1}{m}$ for all r in Eq. (13), $f^{-1}(\boldsymbol{\mathcal{P}}^*)$ is the extrinsic mean [Srivastava and Klassen, 2004].

Algorithm 1 Grassmannian Manifold Self-Attention Module

```
Input: A sequence of Grassmannian data $\{\mathbf{Y}_r\}_{r=1}^m$
Parameter: The projection matrices: $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$
Output: A sequence of Grassmannian data $\{\mathbf{Y}'_r\}_{r=1}^m$
for $r \leftarrow 1$ to $m$ do
    $\mathbf{Q}'_r = f_{orm}(\mathbf{Q}_r) = f_{orm}(\mathbf{W}_q\mathbf{Y}_r)$
    $\mathbf{K}'_r = f_{orm}(\mathbf{K}_r) = f_{orm}(\mathbf{W}_k\mathbf{Y}_r)$
    $\mathbf{V}'_r = f_{orm}(\mathbf{V}_r) = f_{orm}(\mathbf{W}_v\mathbf{Y}_r)$
end for
$\forall r, j \in \{1, 2, \cdots, m\}$:
        $\boldsymbol{\mathcal{A}} := [\mathcal{D}'_{rj}]_{m\times m} = \frac{1}{1+\log(1+\mathrm{d}^2_{\mathrm{PM}}(\mathbf{Q}'_r,\mathbf{K}'_j))}$
$\hat{\boldsymbol{\mathcal{A}}} := \mathrm{Softmax}(\boldsymbol{\mathcal{A}}) = [\mathcal{D}''_{rj}]_{m\times m}$
for $r \leftarrow 1$ to $m$ do
    $\mathbf{V}''_r = \sum_{j=1}^m \mathcal{D}''_{rj} \cdot (\mathbf{V}'_j{\mathbf{V}'_j}^{\mathrm{T}})$
    $\mathbf{Y}'_r = f_{reo}(\mathbf{V}''_r) = \mathbf{Z}_{1:q}$
end for
```

3.4 Backward Propagation

Due to space limitation, this section just provides the computed gradients in the layers of PM, WAE, and REO. For details on other layers, please refer to our *supp.*. To facilitate expression, we use the symbol k to represent any layer in the designed attention module.

REO layer. As Eq. (16) involves SVD operation, we refer to [Ionescu *et al.*, 2015] to first compute the partial derivative of $L^{(k)}$ w.r.t $\mathbf{V}''_r$ in this layer, given below:

$$\frac{\partial L^{(k)}}{\partial \mathbf{V}''_r} = \mathbf{Z}(2(\mathbf{G}^{\mathrm{T}} \circ (\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Lambda}_z)_{\mathrm{sym}}))\mathbf{Z}^{\mathrm{T}} + \mathbf{Z}(\boldsymbol{\Lambda}_s)_{\mathrm{diag}}\mathbf{Z}^{\mathrm{T}}, \quad (19)$$

where $\boldsymbol{\Lambda}_z = \frac{\partial L^{(k')}}{\partial \mathbf{Z}}$, $\boldsymbol{\Lambda}_s = \frac{\partial L^{(k')}}{\partial \mathbf{S}}$, $L^{(k)} = \ell \circ f^{(K)} \circ ... \circ f^{(k)}$ (ℓ is the cross-entropy loss) represents the loss function of the k-th layer, k' denotes a virtual transition layer, regarding $\mathbf{V}''_r$ as its input and outputs $\mathbf{Z}$ and $\mathbf{S}$, and $\mathbf{G}$ is defined by:

$$\mathbf{G}_{rj} = \begin{cases} \frac{1}{\sigma_r - \sigma_j}, & \text{if } \sigma_r \neq \sigma_j, \\ 0, & \text{otherwise} , \end{cases} \quad (20)$$

where σ_r signifies the r-th eigenvalue in $\mathbf{S}$. According to the *invariance of the first-order differential*, which is also the basic criterion for deducing Eq. (19), the following two partial derivatives can be obtained:

$$\frac{\partial L^{(k')}}{\partial \mathbf{Z}} = \left[\frac{\partial L^{(k+1)}}{\partial \mathbf{Y}'_r} \quad \mathbf{0}\right], \qquad \frac{\partial L^{(k')}}{\partial \mathbf{S}} = 0. \quad (21)$$

WAE layer. According to Eq. (15), the partial derivatives of $L^{(k)}$ w.r.t $\mathcal{D}''_{rj}$ and $\mathbf{V}_j$ can be computed by:

$$\frac{\partial L^{(k)}}{\partial \mathbf{V}'_j} = 2\mathcal{D}''_{rj} \cdot \frac{\partial L^{(k+1)}}{\partial \mathbf{V}''_r}\mathbf{V}'_j, \quad (22)$$

$$\frac{\partial L^{(k)}}{\partial \mathcal{D}''_{rj}} = \mathrm{trace}\left[\left(\mathbf{V}'_j{\mathbf{V}'_j}^{\mathrm{T}}\right)\frac{\partial L^{(k+1)}}{\partial \mathbf{V}''_r}\right]. \quad (23)$$

PM layer. On the basis of Eq. (8), the partial derivatives of $L^{(k)}$ w.r.t $\mathbf{Q}'_r$ and $\mathbf{K}'_j$ are shown below:

$$\frac{\partial L^{(k)}}{\partial \mathbf{Q}'_r} = 4[\mathbf{Q}'_r{\mathbf{Q}'_r}^{\mathrm{T}} - \mathbf{K}'_j{\mathbf{K}'_j}^{\mathrm{T}}]\mathbf{Q}'_r \cdot \frac{\partial L^{(k+1)}}{\partial \mathcal{D}_{rj}}, \quad (24)$$

$$\frac{\partial L^{(k)}}{\partial \mathbf{K}'_j} = -4[\mathbf{Q}'_r{\mathbf{Q}'_r}^{\mathrm{T}} - \mathbf{K}'_j{\mathbf{K}'_j}^{\mathrm{T}}]\mathbf{K}'_j \cdot \frac{\partial L^{(k+1)}}{\partial \mathcal{D}_{rj}}. \quad (25)$$

Based on the aforementioned components, the training and inference of the designed network can be unclogged.

4 Experiments

In this section, we assess the performance of the proposed GDLNet in two distinct classification tasks: drone recognition using the RADAR dataset [Brooks *et al.*, 2019] and EEG signal classification employing the MAMEM-SSVEP-II dataset [Pan *et al.*, 2022] and the BCI-ERN dataset [Margaux *et al.*, 2012], respectively. In this article, we execute the publicly accessible source codes of all the involved comparative methods and report their best results across all the used datasets. The number of GMSAs is configured as one in the designed GDLNet, which determines that the following six layers are required for the construction of GMSA: $f_{gmt} \rightarrow f_{orm} \rightarrow f_{pm} \rightarrow f_{sim} \rightarrow f_{smx} \rightarrow f_{wae}$. Besides, our model is trained on a PC equipped with an i7-13700H CPU and 32GB of RAM.

4.1 Drone Recognition

The RADAR dataset encompasses 3,000 synthetic radar signals, distributed among three distinct categories. Each radar signal is segmented into windows of length 20, resulting in an orthonormal matrix of size 23 × 10 for the representation of one sequential data. For a fair comparison, we follow [Brooks *et al.*, 2019] to designate 50%, 25%, and 25% of the obtained 3,000 orthonormal matrices to the training set, validation set, and test set, respectively. On this dataset, the dimensionality q of the generated linear subspaces and the size of the projection matrices of the GMT layer are set to 10 and 18 × 23, respectively. Besides, the learning rate of GDLNet is configured as $5\mathrm{e}^{-3}$, and the batch size is fixed to 50.

The 10-fold experimental results obtained by different Riemannian neural networks (RiemNets) on this dataset are listed in Table 1. It can be seen that the classification score of GDLNet surpasses those of GrNet, SPDNet, and SPDNetBN by 4.57%, 4.79%, and 0.30% respectively, while the lowest standard deviation (SD) additionally provides the initial evidence for the robustness of the designed model. Following this, we reduced the volume of training data by 95% for a more comprehensive evaluation of the robustness of these four networks. As presented in the last column of Table 1, all the competitors lag behind GDLNet in terms of SD. Moreover, our method still stands out in terms of classification ability compared with GrNet and SPDNet. At the same time, Fig. 3 illustrates that compared with the baseline model (GrNet), the convergence performance of the proposed GDLNet is also quite good. All in all, these experimental results not only attest to the efficacy of GDLNet in capturing useful spatiotemporal statistics of sequential data, but also demonstrate its robustness in the scenario of data scarcity.

Models	Acc. (all data)	Acc. (5% data)
GrNet [Huang *et al.*, 2018]	90.11 ± 1.45	77.29 ± 2.23
SPDNet [Huang and Van Gool, 2017]	89.89 ± 1.21	78.84 ± 4.51
SPDNetBN [Brooks *et al.*, 2019]	94.38 ± 3.10	**80.49** ± 4.70
GDLNet	**94.68** ± **0.90**	79.52 ± **1.99**

Table 1: Accuracy (%) comparison on the RADAR dataset.

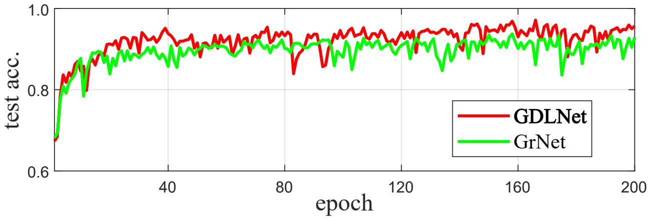

Figure 3: Performance comparison on the RADAR dataset.

4.2 EEG Signal Classification

EEG is a sophisticated neuromonitoring technique that can precisely measure the electric fields generated by cortical neurons, enabling the non-invasive capture of various rhythmic activities occurring within the brain. Several modalities of BCI are based on EEG signals, such as the steady-state visual evoked potential (SSVEP) paradigm and the error-related negativity (ERN) paradigm. The SSVEP paradigm entails external stimuli leading to changes in brain potentials, and the ERN paradigm captures event-related potential (ERP) in the brain's electrical activity when individuals make errors.

However, several special characteristics of EEG data, such as non-linearity, non-stationarity, and a high susceptibility to external interference, make meaningful feature extraction a substantial challenge within this domain. In this part, we apply the designed GDLNet to the task of EEG decoding to further validate its effectiveness, selecting the MAMEM-SSVEP-II and BCI-ERN datasets as two typical examples.

SSVEP. This Dataset (MAMEM-SSVEP-II) was collected using the EGI 300 Geodesic EEG System, which features 256 channels and a data sampling rate of 250 Hz. It comprises the data garnered from 11 participants, who partook in five identical yet independent sessions. In each session, the participants were instructed to concentrate on a visual stimulus enduring for 5 seconds, each oscillating at different frequencies: 6.66, 7.50, 8.57, 10.00, and 12.00 Hz. Each subject performed five trials, corresponding to each of the five stimulus frequencies, as prompted. Moreover, each session involved 100 trials, with each trial segment crafted within 1-5 seconds post-prompt, and further split evenly into four one-second segments.

ERN. This dataset (BCI-ERN) utilizes a P300-based BCI, with a total of 26 participants partaking in a spell-check task. The incorrect inputs from the participants are monitored and exploited for measuring the ERP. The primary objective of this challenge is to determine and detect the types of signal disturbances elicited by BCI spelling error feedback to form a judgment regarding its robustness. Since the quantity of correct inputs detected by the BCI speller greatly surpasses that of incorrect inputs, this task is established as an unbalanced binary classification task.

Models	SSVEP	ERN
EEGNet [Lawhern *et al.*, 2018]	53.72 ± 7.23	74.28 ± 2.47
ShallowCNet [Schirrmeister *et al.*, 2017]	56.93 ± 6.97	71.86 ± 2.64
SCCNet [Wei *et al.*, 2019]	62.11 ± 7.70	70.93 ± 2.31
EEG-TCNet [Ingolfsson *et al.*, 2020]	55.45 ± 7.66	77.05 ± 2.46
FBCNet [Mane *et al.*, 2021]	53.09 ± 5.67	60.47 ± 3.06
TCNet-Fusion [Musallam *et al.*, 2021]	45.00 ± 6.45	70.46 ± 2.94
MBEEGSE [Altuwaijri *et al.*, 2022]	56.45 ± 7.27	75.46 ± 2.34
GrNet [Huang and Van Gool, 2017]	61.23 ± 3.56	72.23 ± 4.56
SPDNet [Huang and Van Gool, 2017]	62.30 ± 3.12	72.05 ± 4.43
SPDNetBN [Brooks *et al.*, 2019]	62.76 ± 3.01	72.34 ± 3.46
MAtt [Pan *et al.*, 2022]	65.19 ± 3.14	75.68 ± **2.23**
GDLNet	**65.52** ± **2.86**	**78.23** ± 2.52

Table 2: Accuracy comparison (%) on the SSVEP and ERN datasets.

We comply with the established standards in [Mane *et al.*, 2021] for data preprocessing and performance evaluation. To be specific, the initial four sessions of each subject serve as the training set, in which one out of four (*i.e.*, session 4) is used for validation, and the remaining session 5 is allotted for testing. The maximum number of training epochs of the proposed GDLNet is set to 150 and 130 on the SSVEP and ERN datasets, respectively. Besides, the model obtaining the lowest validation loss during training is selected for testing on the 5-th session of the same participant. Here, we take the mean accuracy of ten repetitions per subject as the performance indicator of GDLNet on the SSVEP dataset. Due to data imbalance, we follow the criterion of [Lawhern *et al.*, 2018] to utilize the area under the curve (AUC) to estimate the performance of our model on the ERN dataset. Besides, on the SSVEP dataset, the size of each transformation matrix in the GMT layer, learning rate of GDLNet, and batch size are respectively configured as 19×21, $5e^{-3}$, and 64, while those on the ERN dataset are set to 12×15, $4e^{-2}$, and 30, respectively. By grouping the feature maps output by FEM in the channel dimension, a number of 3 orthonormal matrices of size $21 \times q$ can be generated in MMM for the characterization of each EEG sample on the SSVEP dataset. Here, the value of q varies for each participant, due to the variability of participants. Similarly, the number and size of the produced orthonormal matrices in MMM are 3 and $15 \times q$ on the ERN dataset, respectively.

It can be seen from Table. 2 that the classification ability (CA) of RiemNets is superior to most of the Euclidean deep learning (EuDL)-based EEG models. This provides a good demonstration of the effectiveness of Riemannian geometry in encoding the nonlinear structure of sequential signals. More importantly, the proposed GDLNet is the best performer on the SSVEP and ERN datasets, further certifying its efficacy in learning useful spatiotemporal representations. It is noteworthy that although MAtt yields similar results to GDLNet on the SSVEP dataset, the training time of GDLNet averages at 1.23 seconds per epoch, which is 0.64 seconds faster than MAtt. This discrepancy arises from the fact that MAtt, a SPD manifold self-attention method, involves a greater number of eigenvalue operations than our GDLNet.

4.3 Ablation Study

In this part, we take experiments to further study the significance of each primary component involved in GDLNet.

Methods	RADAR	SSVEP	ERN
FEM	83.21 ± 0.89	24.08 ± 2.65	70.33 ± 3.97
GMSA	80.34 ± 0.82	31.71 ± 3.38	56.76 ± 7.88
FEM+EuSA	88.56 ± 4.72	32.39 ± 3.34	71.25 ± 4.25
FEM+GMSA (GDLNet)	94.68 ± 0.90	65.52 ± 2.86	78.23 ± 2.52

Table 3: Accuracy (%) comparison on the RADAR, SSVEP, anc ERN datasets.

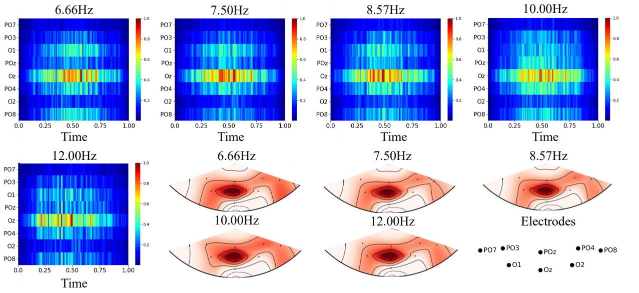

Figure 4: The presentation of the heatmaps and spatial topomaps shows the absolute gradients of the S11 model across five different frequencies on the SSVEP dataset. In the heatmap, the x-axis and y-axis denote the time and various EEG channels, respectively.

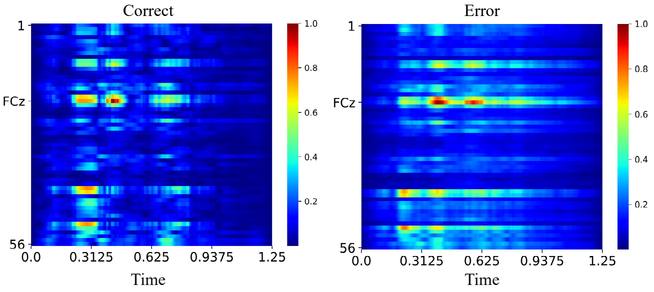

Figure 5: Utilizing the S7 model from the ERN dataset as an illustration. The heatmaps provide a visual representation of the BCI spellers in response to the 'correct' and 'error' feedback categories.

(1) From Table. 3, it is evident that excluding any module from the proposed GDLNet results in a significant decrease in classification accuracy, implying that there are no surplus components included in our framework. Since MMM acted as the input part of GMSA, it is omitted from Table 3.

(2) The fourth line in Table 3 gives the accuracy of FEM+EuSA on the three used datasets, where EuSA represents the Euclidean self-attention module and is computed by: $\mathrm{Softmax}\left(\mathbf{QK}^{\mathrm{T}}/\sqrt{d_k}\right)\mathbf{V}$ [Dosovitskiy *et al.*, 2020]. The comparison between GDLNet and FEM+EuSA demonstrates the necessity and effectiveness of Riemannian computations in the design of a manifold attention module.

(3) Please refer to our *supp.* for other ablation studies.

4.4 EEG Model Interpretation

Through comprehensive analysis of GDLNet, fundamental features captured from EEG data can be elucidated. On the SSVEP dataset and as illustrated in Fig. 4, across five stimulus frequencies, the predominant gradient responses are located on the Oz, exhibiting a focused presence between 0.4 and 0.7 seconds. An abundance of gradient responses underscores the pivotal role of Oz in the visual cortex. These findings are resolutely in line with existing research on the correlation between SSVEP and Oz in EEG recordings [Herrmann, 2001; Han *et al.*, 2018]. This is attributable to Oz location at the heart of the primary visual cortex, possessing more robust induced potential amplitudes and a superior signal-to-noise ratio.

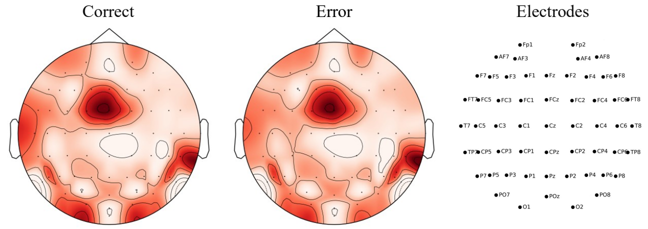

Figure 6: Spatial topography of the time-averaged absolute gradients of the S7 model for two classes (error and correct) on the ERN dataset. Brain regions showing strong gradient activation in FCz in the visual cortex are marked in dark red.

On the ERN dataset, the strong gradient responses in the classification of 'correct' and 'error' predominantly center around the FCz, as demonstrated in Fig. 5 and Fig. 6. This is consistent with a large body of empirical evidence that the ERN is generated in the anterior cingulate cortex. This area, being a part of the medial prefrontal cortex, boasts of rich connections to the limbic and frontal brain regions. Additionally, the FCz at the frontal-central midline effectively captures the ERP of ERN. Special attention must be given to the consistent gradient responses exhibited by both feedback types at the FCz location, particularly around the 0.1 and 0.4 seconds timeframe. Notably, these findings strongly conforming to the differences in ERP waveforms observed between correct and incorrect stimuli as indicated by [Hajcak, 2012].

Please refer to our *supp.* for other experimental results.

5 Conclusion

In this article, we propose a novel geometric deep learning network for more effective signal representation. By Riemannian computations, the proposed Grassmannian self-attention module is qualified to encode and learn useful manifold-valued spatiotemporal patterns of input sequential features, within a lightweight, end-to-end architecture. The experimental results achieved on three benchmarking datasets demonstrate the superiority of GDLNet over some leading comparative methods in both RDL- and EuDL-based signal classification. In addition, the ablation studies confirm the effectiveness of each ingredient in GDLNet. In summary, the proposed network, especially our GMSA block, is a competitive candidate in the modeling, learning, and classification of sequential data across Euclidean and Riemannian spaces. Besides, GMSA provides a feasible and effective path for explicitly mining statistical correlations between Grassmannian features, which brings a new ideology into Grassmannian approaches for signal classification.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62306127, 62020106012, 62332008, U1836218, 62336004), the Natural Science Foundation of Jiangsu Province (BK20231040, BK20221535), the Fundamental Research Funds for the Central Universities (JUSRP124015), the Key Project of Wuxi Municipal Health Commission (Z202318), the National Social Science Foundation of China (21&ZD166), and the National Key R&D Program of China (2023YFF1105102, 2023YFF1105105).

References

[Absil *et al.*, 2009] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. *Princeton University Press*, 2009.

[Altuwaijri *et al.*, 2022] Ghadir Ali Altuwaijri, Ghulam Muhammad, Hamdi Altaheri, and Mansour Alsulaiman. A multi-branch convolutional neural network with squeeze-and-excitation attention blocks for EEG-based motor imagery signals classification. *Diagnostics*, page 995, 2022.

[Bendokat *et al.*, 2024] Thomas Bendokat, Ralf Zimmermann, and P-A Absil. A grassmann manifold handbook: Basic geometry and computational aspects. *Adv. Comput. Math.*, pages 1–51, 2024.

[Brooks *et al.*, 2019] Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for SPD neural networks. *In: NeurIPS*, pages 15463–15474, 2019.

[Chakraborty *et al.*, 2020] Rudrasis Chakraborty, Jose Bouza, Jonathan Manton, and Baba C Vemuri. Manifoldnet: A deep neural network for manifold-valued data with applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 799–810, 2020.

[Chen *et al.*, 2021] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. *In: ICCV*, pages 13359–13368, 2021.

[Chen *et al.*, 2023] Ziheng Chen, Tianyang Xu, Xiao-Jun Wu, Rui Wang, Zhiwu Huang, and Josef Kittler. Riemannian local mechanism for SPD neural networks. *In: AAAI*, pages 7104–7112, 2023.

[Chen *et al.*, 2024a] Ziheng Chen, Yue Song, Gaowen Liu, Ramana Rao Kompella, Xiaojun Wu, and Nicu Sebe. Riemannian multiclass logistics regression for SPD neural networks. *In: CVPR*, 2024.

[Chen *et al.*, 2024b] Ziheng Chen, Yue Song, Yunmei Liu, and Nicu Sebe. A Lie group approach to riemannian batch normalization. *In: ICLR*, 2024.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Edelman *et al.*, 1998a] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, pages 303–353, 1998.

[Edelman *et al.*, 1998b] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, pages 303–353, 1998.

[Gao *et al.*, 2022] Yunyuan Gao, Yici Liu, Qingshan She, and Jianhai Zhang. Domain adaptive algorithm based on multi-manifold embedded distributed alignment for brain-computer interfaces. *IEEE J. Biomed. Health Inform.*, pages 296–307, 2022.

[Hajcak, 2012] Greg Hajcak. What we've learned from mistakes: Insights from error-related brain activity. *Curr. Dir. Psychol.*, pages 101–106, 2012.

[Hamm and Lee, 2008] Jihun Hamm and Daniel D Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. *In: ICML*, pages 376–383, 2008.

[Han *et al.*, 2018] Chengcheng Han, Guanghua Xu, Jun Xie, Chaoyang Chen, and Sicong Zhang. Highly interactive brain-computer interface based on flicker-free steady-state motion visual evoked potential. *Sci. Rep.*, page 5835, 2018.

[Harandi *et al.*, 2011] Mehrtash T Harandi, Conrad Sanderson, Sareh Shirazi, and Brian C Lovell. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. *In: CVPR*, pages 2705–2712, 2011.

[Harandi *et al.*, 2013] Mehrtash Harandi, Conrad Sanderson, Chunhua Shen, and Brian C Lovell. Dictionary learning and sparse coding on Grassmann manifolds: An extrinsic solution. *In: ICCV*, pages 3120–3127, 2013.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *In: CVPR*, pages 770–778, 2016.

[Helmke *et al.*, 2007] Uwe Helmke, Knut Hüper, and Jochen Trumpf. Newton's method on Grassmann manifolds. *arXiv preprint arXiv:0709.2205*, 2007.

[Herrmann, 2001] Christoph S Herrmann. Human EEG responses to 1–100 hz flicker: resonance phenomena in visual cortex and their potential correlation to cognitive phenomena. *Exp. Brain Res.*, pages 346–353, 2001.

[Huang and Van Gool, 2017] Zhiwu Huang and Luc Van Gool. A Riemannian network for SPD matrix learning. *In: AAAI*, pages 2036–2042, 2017.

[Huang *et al.*, 2015] Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Projection metric learning on Grassmann manifold with application to video based face recognition. *In: CVPR*, pages 140–149, 2015.

[Huang *et al.*, 2018] Zhiwu Huang, Jiqing Wu, and Luc Van Gool. Building deep networks on Grassmann manifolds. *In: AAAI*, pages 1137–1145, 2018.

[Huang *et al.*, 2022] Huaibo Huang, Xiaoqiang Zhou, and Ran He. Orthogonal transformer: An efficient vision transformer backbone with token orthogonalization. *In: NeurPIS*, pages 14596–14607, 2022.

[Ingolfsson *et al.*, 2020] Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. *IEEE International Conference on Systems, Man, and Cybernetics*, pages 2958–2965, 2020.

[Ionescu *et al.*, 2015] Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. *In: ICCV*, pages 2965–2973, 2015.

[Lawhern *et al.*, 2018] 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. *J. Neural Eng.*, page 056013, 2018.

[Li *et al.*, 2022] Zhengyu Li, XUAN TANG, Zihao Xu, Xihao Wang, Hui Yu, Mingsong Chen, and xian wei. Geodesic self-attention for 3d point clouds. *In: NeurIPS*, pages 6190–6203, 2022.

[Mane *et al.*, 2021] Ravikiran Mane, Effie Chew, Karen Chua, Kai Keng Ang, Neethu Robinson, A Prasad Vinod, Seong-Whan Lee, and Cuntai Guan. FBCNet: A multi-view convolutional neural network for brain-computer interface. *arXiv preprint arXiv:2104.01233*, 2021.

[Margaux *et al.*, 2012] Perrin Margaux, Maby Emmanuel, Daligault Sébastien, Bertrand Olivier, and Mattout Jérémie. Objective and subjective evaluation of online error correction during P300-based spelling. *Adv. Hum-Comput. Interact.*, pages 1–13, 2012.

[Musallam *et al.*, 2021] Yazeed K Musallam, Nasser I AlFassam, Ghulam Muhammad, Syed Umar Amin, Mansour Alsulaiman, Wadood Abdul, Hamdi Altaheri, Mohamed A Bencherif, and Mohammed Algabri. Electroencephalography-based motor imagery classification using temporal convolutional network fusion. *Biomed. Signal Process. Control*, page 102826, 2021.

[Nguyen and Yang, 2023] Xuan Son Nguyen and Shuo Yang. Building neural networks on matrix manifolds: A gyrovector space approach. *In: ICML*, pages 26031–26062, 2023.

[Nguyen *et al.*, 2019] Xuan Son Nguyen, Luc Brun, Olivier Lézoray, and Sébastien Bougleux. A neural network based on SPD manifold learning for skeleton-based hand gesture recognition. *In: CVPR*, pages 12036–12045, 2019.

[Nguyen, 2021] Xuan Son Nguyen. Geomnet: A neural network based on Riemannian geometries of SPD matrix space and cholesky space for 3d skeleton-based interaction recognition. *In: ICCV*, pages 13379–13389, 2021.

[Pan *et al.*, 2022] Yue-Ting Pan, Jing-Lun Chou, and Chun-Shu Wei. MAtt: A manifold attention network for EEG decoding. *In: NeurIPS*, pages 31116–31129, 2022.

[Schirrmeister *et al.*, 2017] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.*, pages 5391–5420, 2017.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Srivastava and Klassen, 2004] Anuj Srivastava and Eric Klassen. Bayesian and geometric subspace tracking. *Adv. Appl. Prob.*, pages 43–56, 2004.

[Turaga *et al.*, 2011] Pavan Turaga, Ashok Veeraraghavan, Anuj Srivastava, and Rama Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 2273–2286, 2011.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *In: NeurIPS*, pages 6000–6010, 2017.

[Wang and Wu, 2020] Rui Wang and Xiao-Jun Wu. Grasnet: A simple Grassmannian network for image set classification. *Neural Process. Lett.*, pages 693–711, 2020.

[Wang *et al.*, 2021] Rui Wang, Xiao-Jun Wu, and Josef Kittler. Graph embedding multi-kernel metric learning for image set classification with Grassmannian manifold-valued features. *IEEE Trans. Multimedia*, pages 228–242, 2021.

[Wang *et al.*, 2022a] Rui Wang, Xiao-Jun Wu, and Josef Kittler. Symnet: A simple symmetric positive definite manifold deep learning method for image set classification. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 2208–2222, 2022.

[Wang *et al.*, 2022b] Rui Wang, Xiao-Jun Wu, Zhen Liu, and Josef Kittler. Geometry-aware graph embedding projection metric learning for image set classification. *IEEE Trans. Cogn. Dev. Syst.*, pages 957–970, 2022.

[Wei *et al.*, 2019] Chun-Shu Wei, Toshiaki Koike-Akino, and Ye Wang. Spatial component-wise convolutional network (sccnet) for motor-imagery EEG classification. *IEEE/EMBS Conference on Neural Engineering*, pages 328–331, 2019.

[Wei *et al.*, 2022a] Dong Wei, Xiaobo Shen, Quansen Sun, Xizhan Gao, and Zhenwen Ren. Neighborhood preserving embedding on Grassmann manifold for image-set analysis. *Pattern Recognit.*, page 108335, 2022.

[Wei *et al.*, 2022b] Dong Wei, Xiaobo Shen, Quansen Sun, Xizhan Gao, and Zhenwen Ren. Sparse representation classifier guided Grassmann reconstruction metric learning with applications to image set analysis. *IEEE Trans. Multimedia*, pages 4307–4322, 2022.