Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

A Correlation Manifold Self-Attention Network for EEG Decoding

Chen Hu[1] , **Rui Wang**[1*] , **Xiaoning Song**[1] , **Tao Zhou**[1] ,
Xiao-Jun Wu[1] , **Nicu Sebe**[2] and **Ziheng Chen**[2]
[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
[2]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
6233112017@stu.jiangnan.edu.cn, {cs_wr, x.song, taozhou.ai, wu_xiaojun}@jiangnan.edu.cn,
niculae.sebe@unitn.it, ziheng_ch@163.com

Abstract

Riemannian neural networks, which generalize the deep learning paradigm to non-Euclidean geometries, have garnered widespread attention across diverse applications in artificial intelligence. Among these, the representative attention models have been studied on various non-Euclidean spaces to geometrically capture the spatiotemporal dependencies inherent in time series data, *e.g.*, electroencephalography (EEG). Recent studies have highlighted the full-rank correlation matrix as an advantageous alternative to the covariance matrix for data representation, owing to its invariance to the scale of variables. Motivated by these advancements, we propose the Correlation Attention Network (CorAtt) tailored for full-rank correlation matrices and implement it under the permutation-invariant and computationally efficient Off-Log and Log-Scaled geometries, respectively. Extensive evaluations on three benchmarking EEG datasets provide substantial evidence for the effectiveness of our introduced CorAtt. The code and supplementary material can be found at https://github.com/ChenHu-ML/CorAtt.

1 Introduction

Deep neural networks (DNNs) have significantly progressed across a broad range of applications [Simonyan and Zisserman, 2015; He *et al.*, 2016; Vaswani *et al.*, 2017; Zeng *et al.*, 2024; Tang *et al.*, 2024]. However, most existing methods assume that the data adheres to a vector space structure, whereas many of them emerge from latent spaces governed by non-Euclidean geometries, such as Riemannian geometries. Building on this insight, researchers have made notable strides in generalizing different types of DNNs to manifolds, known as Riemannian neural networks [Huang and Van Gool, 2017; Gulcehre *et al.*, 2018; Chen *et al.*, 2023; Wang *et al.*, 2024b; Wang *et al.*, 2024a; Chen *et al.*, 2024b; Chen *et al.*, 2024d; Chen *et al.*, 2024a; Chen *et al.*, 2024c; Wang *et al.*, 2025; Chen *et al.*, 2025a; Chen *et al.*, 2025b].

Drawing inspiration from the effectiveness of the attention mechanism in capturing correlations between different feature regions [Vaswani *et al.*, 2017; Hu *et al.*, 2018; Dosovitskiy, 2020], the investigation of the Riemannian attention mechanism has gained increasing interest. Notably, the hyperbolic attention network [Gulcehre *et al.*, 2018] represents a pioneering effort in this area, designed based on the Hyperboloid and Klein models. Building on this, [Pan *et al.*, 2022] extended the attention mechanism to Symmetric Positive Definite (SPD) manifolds, implemented under the Log-Euclidean geometry. Subsequently, [Wang *et al.*, 2024a] adapted this approach to Grassmannian manifolds, utilizing an extrinsic mean within the Projection Metric.

The correlation matrix, which is scale-invariant [David and Gu, 2019], serves as a compact and normalized alternative to the covariance matrix for data representation. A slice of research fields in artificial intelligence, such as Diffusion Tensor Imaging (DTI) [Pennec *et al.*, 2006], Brain-Computer Interfaces (BCI) [Jalili and Knyazeva, 2011], and Gaussian graphical models [Epskamp and Fried, 2018] have particularly benefited from the utilization of correlation matrices in place of covariance matrices. The basic reason is that eliminating the influence of variable scales is particularly effective for the handled problems where the scales are irrelevant [Thanwerdas, 2024]. In particular, non-invasive BCI systems rely heavily on effectively decoding EEG signals to enable direct communication between the brain and external devices. EEG records neural activity with high temporal resolution by measuring electrical potentials on the scalp [Subha *et al.*, 2010], but the resulting signals are often noisy and lack specificity [Hine *et al.*, 2017]. To address these challenges, correlation matrices have emerged as a suitable representation for EEG analysis, as they emphasize statistical dependencies over absolute magnitudes. This is particularly advantageous since strong inter-channel correlations could remain stable despite substantial variations in electrode signal strengths.

Recently, several Riemannian metrics have been proposed for the manifolds of full-rank correlation matrices, including the Off-Log Metric (OLM) and Log-Scaled Metric (LSM) [Thanwerdas, 2024]. The Riemannian operators associated with them, such as geodesics, Fréchet Means, and exponential & logarithmic maps, are not only permutation-invariant but also computationally efficient. This provides a theoretical possibility for further exploration of attention mechanisms on

*Corresponding author: Rui Wang

Preprint

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

full-rank correlation matrices.

Designing the attention mechanism for full-rank correlation matrices presents a unique challenge, primarily due to the absence of corresponding transformation layers. The main difficulties stem from the following two aspects. On the one hand, the designed transformation function should preserve the characteristics of full-rank correlation matrices, making the traditional linear layers or their manifold counterparts, *e.g.*, bilinear mapping (BiMap) function [Huang and Van Gool, 2017], for covariance matrices unsuitable. On the other hand, to the best of our knowledge, there is no prior knowledge for constructing neural networks on the manifolds of full-rank correlation matrices, preventing the generation of manifold-valued queries, keys, and values. Another important problem to be solved is the lack of classification layers defined on the Correlation manifolds. To address these challenges, we introduce two novel transformation layers based on the Lie group homomorphisms, explicitly tailored for the OLM and LSM within the Riemannian geometry of Correlation manifolds. Moreover, building upon [Thanwerdas, 2024], we derive the Weighted Fréchet Mean (WFM) [Karcher, 1977], a more general Fréchet Mean, for feature aggregation under OLM and LSM. Additionally, we harness the Riemannian logarithm function to develop two tangent mapping layers under the framework of OLM and LSM to enable the classification of the Correlation manifolds. With these preparations, we propose a Correlation Attention Network (CorAtt) for learning effective spatiotemporal statistical information of EEG signals. In summary, our key contributions are as follows:

- **Two novel transformation layers based on Lie group homomorphisms.** We design two transformation layers explicitly tailored to preserve the geometric structure of full-rank correlation matrices under the OLM and LSM.
- **Two attention models are established on the Correlation manifolds.** This article proposes two attention models for the full-rank correlation matrices based on the permutation-invariant and computationally efficient OLM and LSM, respectively.
- **Two tangent mapping layers are proposed under the Correlation geometry.** Two tangent mapping layers are induced by LSM and OLM to project the full-rank correlation matrices into a flat space for classification.
- **Empirical validations in three EEG decoding tasks**. Experimental results achieved on three benchmarking EEG datasets validate the effectiveness of our proposed CorAtt and each of the designed components.

2 Preliminary

This section briefly reviews the Lie group and the geometry of full-rank correlation matrices. For more in-depth discussions, please refer to [Do Carmo and Flaherty Francis, 1992; Tu, 2011; David and Gu, 2019; Thanwerdas, 2024].

Definition 2.1 (**Lie Groups).** A smooth manifold is a Lie group if it is endowed with a group operation $\odot$ such that both mappings, $m(x, y) \mapsto x \odot y$ and $i(x) \mapsto x_{\odot}^{-1}$, are smooth. Here, $x_{\odot}^{-1}$ denotes the group inverse.

A Lie group is both a group and a manifold, which motivates the study of smooth maps that preserve these structures.

Definition 2.2 (**Lie Homomorphisms).** Let $\{\mathcal{M}, \odot_{\mathcal{M}}\}$ and $\{\mathcal{N}, \odot_{\mathcal{N}}\}$ be two Lie groups. A smooth map $f(\cdot) : \{\mathcal{M}, \odot_{\mathcal{M}}\} \to \{\mathcal{N}, \odot_{\mathcal{N}}\}$ forms a Lie group homomorphism if it preserves the group structure:

$$f(x \odot_{\mathcal{M}} y) = f(x) \odot_{\mathcal{N}} f(y), \quad \forall x, y \in \mathcal{M}. \tag{1}$$

Next, we briefly review the manifolds of full-rank correlation matrices. Any correlation matrix is derived by normalizing the covariance matrix with its variances. Let X be a random variable with an invertible covariance matrix $P = (\mathrm{Cov}(X_i, X_j))_{1 \leq i,j \leq n}$, the corresponding correlation matrix C is defined as:

$$C = \mathrm{Cor}(P) = \mathrm{Diag}(P)^{-\frac{1}{2}} P \ \mathrm{Diag}(P)^{-\frac{1}{2}}, \tag{2}$$

where $\mathrm{Diag}(P)$ is the diagonal matrix of P. The set of all full-rank correlation matrices is denoted as $\mathcal{C}_{++}^n$.

Recent studies have discovered that $\mathcal{C}_{++}^n$ has a smooth structure and developed several different Riemannian metrics [David and Gu, 2019; Thanwerdas, 2024] over it. This paper focuses on two permutation-invariant and simple metrics, which are the OLM and LSM [Thanwerdas, 2024]. Here, the permutation-invariant property ensures that the analysis is unaffected by arbitrary choices in ordering. In the following, we first introduce four key maps (diffeomorphisms) used to construct these metrics. Wherein, the OLM is associated with the two maps given below:

$$\begin{cases} \mathrm{Log}^{\mathrm{o}} : C \in \mathcal{C}_{++}^n \mapsto \mathrm{Off}\left(\mathrm{mlog}(C)\right) \in \mathrm{Hol}(n), \\ \mathrm{Exp}^{\mathrm{o}} : S \in \mathrm{Hol}(n) \mapsto \mathrm{mexp}\left(S + \mathcal{D}^{o}(S)\right) \in \mathcal{C}_{++}^n, \end{cases} \tag{3}$$

where $\mathrm{mlog}(\cdot)$ and $\mathrm{mexp}(\cdot)$ denote the matrix logarithm and exponential, $\mathrm{Hol}(n) = \{X \in \mathbb{R}^{n \times n} \mid X = X^{\top}, \mathrm{Diag}(X) = \mathbf{0}\}$, while $\mathrm{Off}(X)$ denotes the off-diagonal part of X. As demonstrated by [Archakov and Hansen, 2021][Sec. 3.3], $\mathcal{D}^{o}(S)$ is a diagonal matrix satisfying

$$\mathrm{mlog}\left(\mathrm{Diag}\left(\mathrm{mexp}(S + \mathcal{D}^{o}(S))\right)\right) = \mathbf{0}. \tag{4}$$

This can be solved via the fixed-point iteration.

The following two maps are associated with LSM:

$$\begin{cases} \mathrm{Log}^{\star} : C \in \mathcal{C}_{++}^n \mapsto \mathrm{mlog}\left(\mathcal{D}^{\star}(C) C \mathcal{D}^{\star}(C)\right) \in \mathrm{Row}_0(n), \\ \mathrm{Exp}^{\star} : S \in \mathrm{Row}_0(n) \mapsto \mathrm{Cor}\left(\mathrm{mexp}(S)\right) \in \mathcal{C}_{++}^n. \end{cases} \tag{5}$$

Here, $\mathrm{Row}_0(n) = \{X \in \mathbb{R}^{n \times n} \mid X = X^{\top}, X\mathbb{1} = \mathbf{0}\}$ and $\mathbb{1} \in \mathbb{R}^n$ is a vector of all ones. The diagonal matrix $\mathcal{D}^{\star}(C)$ is the unique zero of

$$f : x \in \mathbb{R}_{++}^n \mapsto Cx - \frac{1}{x}, \tag{6}$$

where x represents a positive vector and $\frac{1}{x} = \left(\frac{1}{x_1}, \ldots, \frac{1}{x_n}\right)$. Eq. (6) can be solved via the damped Newton's method [Thanwerdas, 2024][Sec. 3.5].

Actually, the **OLM** is induced from $\mathrm{Hol(n)}$ via the map $\mathrm{Log}^{\mathrm{o}}(\cdot)$, while **LSM** is induced from $\mathrm{Row_0(n)}$ via the map $\mathrm{Log}^{\star}(\cdot)$. Additionally, as demonstrated by [Thanwerdas, 2024], both correlation manifolds form Lie groups under OLM and LSM. The geodesic distances and group operations for these two metrics are summarized in Tab. 1.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Metric	$\mathrm{d}(C_1, C_2)$	Group operation $\odot$
OLM	$\lVert \mathrm{Log}^{o}(C_1) - \mathrm{Log}^{o}(C_2) \rVert_{\mathrm{F}}$	$\mathrm{Exp}^{o}\left(\mathrm{Log}^{o}(C_1) + \mathrm{Log}^{o}(C_2)\right)$
LSM	$\lVert \mathrm{Log}^{\star}(C_1) - \mathrm{Log}^{\star}(C_2) \rVert_{\mathrm{F}}$	$\mathrm{Exp}^{\star}\left(\mathrm{Log}^{\star}(C_1) + \mathrm{Log}^{\star}(C_2)\right)$

Table 1: Summary of the geodesic distances and group operations under OLM and LSM.

3 Proposed Method

In this section, we provide the technical details of the proposed attention mechanism for full-rank correlation matrices. To be specific, we first introduce the main framework of the suggested correlation attention mechanism in Sec. 3.1. This is followed by the specific implementations under OLM and LSM in Secs. 3.2 and 3.3, respectively. Finally, the classification method defined on the Correlation manifolds is detailed in Sec. 3.4.

3.1 Correlation Attention Mechanism

This section showcases how to leverage the correlation matrix geometry to generalize the core operations of transformation mapping, attention computation, and feature aggregation.

Correlation Transformations. In the Euclidean attention mechanism, the linear map $\mathrm{Linear}(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ is commonly employed to generate q_i, k_i, and v_i. This transformation preserves the vector space structure, as shown by the following property:

$$\mathrm{Linear}(x_1 + x_2) = \mathrm{Linear}(x_1) + \mathrm{Linear}(x_2). \tag{7}$$

This indicates that the transformation is a homomorphism over vector spaces. Since correlation matrices lie in a manifold with non-Euclidean geometry, directly applying linear transformations will compromise their inherent geometric properties. However, the Lie group homomorphism (Def. 2.2) generalizes this concept from vector spaces to Lie groups. Therefore, it appears to be a possible and natural choice to define transformation layer on the Correlation manifold using Lie homomorphism $\mathrm{hom}(\cdot)$.

Correlation Attention. Let X_i, Q_i, K_i, V_i, R_i be the input correlation matrices, queries, keys, values, and output data, respectively. One of the key ideas in attention is to compute the similarity-based score between Q_i and K_j for each pair of $\{V_i, V_j\}$. In contrast to the commonly used dot product for vectors, the most natural way to compute the similarity between Correlation manifold-valued points is the utilization of geodesic distance. However, a function is needed to map the computed geodesic distance into a valid form, as higher similarity corresponds to smaller distance. Specifically, given that both Q_i and K_j reside on the Correlation manifolds, the attention weight is computed by:

$$\mathcal{A}_{ij} = \mathrm{Softmax}\left(\left(1 + \log(1 + \mathrm{d}(Q_i, K_j))\right)^{-1}\right), \tag{8}$$

where $\mathrm{d}(Q_i, K_j)$ represents the geodesic distance between the correlation matrices as shown in Tab. 1.

Correlation Aggregation. Corresponding to the weighted average in Euclidean space, the WFM is a principled mathematical tool for feature aggregation in manifolds [Karcher, 1977; Ginestet *et al.*, 2012]. It minimizes the weighted sum of squared geodesic distances. Given the geodesic distance $\mathrm{d}(\cdot, \cdot)$, a set of points $P_{i\ldots N} \in \mathcal{M}$, and the corresponding weights $\{w_{1\ldots N}\}$ that satisfy the convexity constraint, *i.e.,*$\forall i, w_i > 0$ and $\sum_i w_i = 1$, the WFM is defined as:

$$\mathrm{WFM}(\{w_i\}, \{P_i\}) = \underset{G \in \mathcal{M}}{\mathrm{argmin}} \sum_{i=1}^{N} w_i \, \mathrm{d}^2\left(P_i, G\right). \tag{9}$$

With the computed attention matrix $\mathcal{A}$ and a set of values $\{V_{i\ldots N} \in \mathcal{C}^n_{++}\}$, the i-th aggregated output $R_i \in \mathcal{C}^n_{++}$ in the built correlation attention model is formulated as:

$$R_i = \mathrm{WFM}(\{\mathcal{A}_{ij}\}_{j=1\ldots N}, \{V_j\}_{j=1\ldots N}). \tag{10}$$

With these basic components in place, we summarize the forward pass of the proposed attention mechanism on the Correlation manifolds in Alg. 1.

Algorithm 1: Cor Attention (CorAtt) over full-rank correlation matrix manifolds

Input : A set of correlation matrices $\{X_{1\ldots N}\}$
Output : A set of correlation matrices $\{R_{1\ldots N}\}$

for $i \leftarrow 1$ **to** N **do**
  Queries: $Q_i = \mathrm{hom}(X_i)$
  Keys: $K_i = \mathrm{hom}(X_i)$
  Values: $V_i = \mathrm{hom}(X_i)$
end
for $i \leftarrow 1$ **to** N **do**
  for $j \leftarrow 1$ **to** N **do**
    $\mathcal{S}_{ij} = (1 + \log(1 + \mathrm{d}(Q_i, K_j)))^{-1}$
  end
  Attention weight: $\mathcal{A}_{ij} = \mathrm{Softmax}(\mathcal{S}_{ij})$
  Aggregation: $R_i = \mathrm{WFM}(\{\mathcal{A}_{ij}\}_{j=1}^{N}, \{V_j\}_{j=1}^{N})$
end

3.2 Correlation Attention Based on OLM

This section details the implementation of Alg. 1 under OLM. While the geodesic distance for similarity computation is summarized in Tab. 1, we focus here on deriving the expressions for the Lie group homomorphism and the WFM. By defining the group operation $\odot_{ol}$ for OLM, we present the expression for the Lie group homomorphism over the Correlation manifolds as follows:

Theorem 3.1 (OLM Lie Homomorphism). *For any* $C \in \{\mathcal{C}^n_{++}, \odot_{ol}\}$*, and* $M \in \mathbb{R}^{n \times m}$*. The transformation mapping* $\mathrm{hom}^{ol}(\cdot) : \{\mathcal{C}^n_{++}, \odot_{ol}\} \to \{\mathcal{C}^m_{++}, \odot_{ol}\}$ *is defined as:*

$$\mathrm{hom}^{ol}(C) = \mathrm{Exp}^{o}\left(\mathrm{Off}\left(M^{\top} \mathrm{Log}^{o}(C) M\right)\right). \tag{11}$$

It can be proved that $\mathrm{hom}^{ol}(\cdot)$ *is a Lie group homomorphism.*

Proof. The proof is presented in App. C.1 □

As discussed in [Thanwerdas, 2024], the Correlation manifold enjoys closed-form expressions of Fréchet mean under OLM. For attention computation, we present a more general version, the WFM under OLM.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Theorem 3.2 (**The WFM under OLM**). *For $C_{1\ldots N} \in \mathcal{C}^n_{++}$, $w_{1\ldots N} > 0$ satisfying $\sum_i w_i = 1$, the expression of WFM has a closed form shown below:*

$$G = \operatorname{Exp}^{\circ}\left(\sum_{i=1}^{N} w_i \operatorname{Log}^{\circ}(C_i)\right). \tag{12}$$

Proof. The proof is given in App. C.2. □

It is evident that G corresponds to the OLM-based Fréchet mean, when $w_i = \frac{1}{N}$ for all i in Eq. (12).

3.3 Correlation Attention Based on LSM

Similarly, we derive the expressions for the Lie group homomorphism and the WFM over the Correlation manifolds under LSM.

Theorem 3.3 (**LSM Lie Homomorphism**). *For any $C \in \{\mathcal{C}^n_{++}, \odot_{ls}\}$, and $M \in \mathbb{R}^{n\times m}$, the transformation mapping $\hom^{ls}(\cdot) : \{\mathcal{C}^n_{++}, \odot_{ls}\} \to \{\mathcal{C}^m_{++}, \odot_{ls}\}$ is formulated as:*

$$\hom^{ls}(C) = \operatorname{Exp}^{\star}\left(\phi\left(M^{\top}\operatorname{Log}^{\star}(C)M\right)\right), \tag{13}$$

where $\odot_{ls}$ denotes the LSM-based group operation and $\phi(X)$ is expressed as:

$$\phi(X) = X - \operatorname{diag}(X\mathbb{1}). \tag{14}$$

Wherein, $\operatorname{diag}(\cdot)$ *creates a diagonal matrix from a vector. We can prove that* $\hom^{ls}(\cdot)$ *is a Lie group homomorphism.*

Proof. The proof is presented in App. C.3. □

Under LSM, the WFM has a closed-form expression.

Theorem 3.4 (**The WFM under LSM**). *For $C_{1\ldots N} \in \mathcal{C}^n_{++}$, $w_{1\ldots N} > 0$ satisfying $\sum_i w_i = 1$, the WFM under LSM can be described as:*

$$G = \operatorname{Exp}^{\star}\left(\sum_{i=1}^{N} w_i \operatorname{Log}^{\star}(C_i)\right), \tag{15}$$

Proof. The proof is presented in App. C.4 □

When $\forall i, w_i = \frac{1}{N}$ in Eq. (15), G corresponds to the LSM-based Fréchet mean, as indicated by [Thanwerdas, 2024].

3.4 Classification

This section presents the design of the classification layer on the Correlation manifolds. Since the underlying space of the correlation matrices is a non-Euclidean manifold, a manifold-to-Euclidean embedding mapping is required to convert the learned manifold data into the corresponding Euclidean representation. To this end, the tangent mapping layer is designed to project the refined correlation matrices onto the tangent space of the Correlation manifold at the identity matrix using the Riemannian logarithm function. With the following two propositions, the tangent mapping operations under OLM and LSM can be defined.

Proposition 3.5. *For any $C \in \mathcal{C}^n_{++}$, the OLM-based Riemannian logarithm at the identity matrix $\operatorname{Log}^{ol}_{I_n}(C)$ can be formulated as:*

$$\operatorname{Log}^{ol}_{I_n}(C) = \operatorname{Off}\left(\operatorname{mlog}(C)\right). \tag{16}$$

Proof. The proof is detailed in App. C.5. □

Proposition 3.6. *For any $C \in \mathcal{C}^n_{++}$, the LSM-based Riemannian logarithm at the identity matrix $\operatorname{Log}^{ls}_{I_n}(C)$ can be expressed as:*

$$\operatorname{Log}^{ls}_{I_n}(C) = \operatorname{Off}\left(\operatorname{Log}^{\star}(C)\right). \tag{17}$$

Proof. The proof is presented in App. C.6. □

Now, the tangent mapping operations under OLM and LSM are defined by Eqs. (16) and (17), respectively.

Since each output of the tangent mapping layer is a symmetric matrix with all zeros on the main diagonal, we extract its strictly lower triangular part, vectorize it, and concatenate all vectors w.r.t the i-th input data. The obtained data points are then passed through a fully connected (FC) layer, followed by a Softmax function for the final classification.

4 Experiments

This section tests the proposed CorAtt in two specific forms, called CorAtt-OLM and CorAtt-LSM. To ensure a comprehensive assessment, we apply the two models to three typical BCI tasks, which are the Mental Imagery (MI) decoding on the BCIC-IV-2a dataset [Brunner *et al.*, 2008], Steady-State Visual Evoked Potential (SSVEP) decoding on the MAMEM-SSVEP-II dataset [Nikolopoulos, 2016], and Error-Related Negativity (ERN) decoding on the BCI-ERN dataset [Margaux *et al.*, 2012]. For comparison, the following state-of-the-art (SOTA) deep learning methods are included: ShallowConvNet [Schirrmeister *et al.*, 2017], EEGNet [Lawhern *et al.*, 2018], SCCNet [Wei *et al.*, 2019], MBEEGSE [Altuwaijri *et al.*, 2022], TCNet-Fusion [Musallam *et al.*, 2021], and FBCNet [Mane *et al.*, 2021]. In addition, we incorporate several representative geometric deep learning models, such as SPDNet [Huang and Van Gool, 2017], SPDNetBN [Brooks *et al.*, 2019], MAtt [Pan *et al.*, 2022], and GDLNet [Wang *et al.*, 2024a], to provide a more convincing comparison. All experiments were conducted on an i9-14900 CPU with 64GB RAM and two NVIDIA RTX4080 Super GPUs.

Motor Imagery. The BCIC-IV-2a dataset [Brunner *et al.*, 2008] is a widely recognized public EEG resource, containing signals from 9 subjects performing a four-class motor imagery task. Each subject completed two sessions, with each trial involving four seconds of imagined movement (right hand, left hand, feet, or tongue). Following the protocol in [Pan *et al.*, 2022], the first session of BCIC-IV-2a is used for training, reserving one-eighth of it for validation. Besides, the performance indicator is based on classification accuracy.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

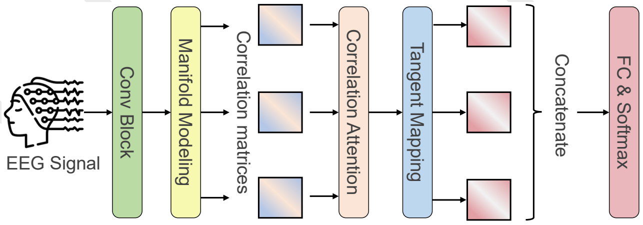

Figure 1: An overview of the proposed CorAtt architecture.

SSVEP. The MAMEM-SSVEP-II dataset [Nikolopoulos, 2016] includes EEG data from 11 subjects, each contributing five sessions. In each session, subjects focused on a 5-second visual stimulus oscillating at one of five frequencies: 6.66, 7.50, 8.57, 10.00, or 12.00 Hz. Each subject completed five trials, one for each frequency, yielding 100 trials per session. Each trial lasted between 1 to 5 seconds after the prompt, divided into four one-second segments. Following the protocol in [Pan *et al.*, 2022], we train on the first four sessions, with session 4 used for validation, and tested on the fifth session.

ERN. The BCI-ERN dataset [Margaux *et al.*, 2012] originates from a Kaggle BCI Challenge and contains recordings from 26 subjects who participated in a P300-based spelling task. ERN was measured in response to mistakes made by the BCI speller, leading to a binary, imbalanced classification problem, as correct inputs significantly outnumber erroneous ones. Following the criterion in [Pan *et al.*, 2022; Wang *et al.*, 2024a], we adopt the same dataset partitioning as in the MAMEM-SSVEP-II dataset and employ the Area Under the Curve (AUC) to measure the model performance.

4.1 Proposed Network

Block	MI	MAMEM	ERN
Input data	$1 \times 22 \times 438$	$1 \times 8 \times 125$	$1 \times 56 \times 160$
SpatConv	$22 \times 1 \times 438$	$125 \times 1 \times 125$	$14 \times 1 \times 160$
SpatTempConv	$20 \times 1 \times 439$	$15 \times 1 \times 126$	$42 \times 1 \times 161$
Split & Correlation	$3 \times 20 \times 20$	$7 \times 15 \times 15$	$3 \times 14 \times 14$
Correlation Attention	$3 \times 20 \times 20$	$7 \times 15 \times 15$	$3 \times 14 \times 14$
Tangent mapping	$3 \times 20 \times 20$	$7 \times 15 \times 15$	$3 \times 14 \times 14$
Vectorization	570	735	273
FC + Softmax	4	5	2

Table 2: CorAtt architectures across three datasets. Where SpatConv and SpatTempConv denote spatial and spatiotemporal convolution layers. The attention block represents the Correlation Attention block under the corresponding metric.

Network Architecture. As shown in Fig. 1, the architecture of the proposed CorAtt consists of four main components: a Feature Extraction Module (FEM), a Manifold Modeling Module (MMM), a Correlation Attention Module, and a classification module. We follow [Wei *et al.*, 2019] to make the FEM contain two convolutional layers: one for applying spatial filtering to the multi-channel EEG signals and the other for extracting spatiotemporal features. The MMM is applied to split and transform data points onto the Correlation manifold. We impose segmentation on the output data of FEM, generating s non-overlapping subparts. Then, a correlation matrix is computed for each subpart using Eq. (2).

Models	MI	SSVEP	ERN
EEGNet	61.84 ± 6.39	53.72 ± 7.23	74.28 ± 2.47
ShallowCNet	57.43 ± 6.25	56.93 ± 6.97	71.86 ± 2.64
SCCNet	71.95 ± 5.05	62.11 ± 7.70	70.93 ± 2.31
FBCNet	56.52 ± 3.07	53.09 ± 5.67	60.47 ± 3.06
TCNet-Fusion	71.45 ± 4.45	45.00 ± 6.45	70.46 ± 2.94
MBEEGSE	64.58 ± 6.07	56.45 ± 7.27	75.46 ± 2.34
SPDNet	72.93 ± 4.33	62.30 ± 3.12	72.05 ± 4.43
SPDNetBN	73.02 ± 3.67	62.76 ± 3.01	72.34 ± 3.46
MAtt	**74.71 ± 5.01**	65.19 ± 3.14	75.68 ± 2.23
GDLNet	69.32 ± 2.89	65.52 ± 2.86	78.23 ± 2.52
CorAtt-OLM	**75.01 ± 2.78**	**67.39 ± 3.22**	**78.78 ± 3.40**
CorAtt-LSM	74.47 ± 2.43	**67.74 ± 2.44**	**78.63 ± 3.31**
CorAtt-MIX	**75.56 ± 1.58**	**68.27 ± 2.50**	**79.04 ± 2.91**

Table 3: Average performance (± standard deviation) over 10 runs, comparing CorAtt with SOTA methods on three EEG datasets. CorAtt-MIX indicates that the Attention Block and Tangent Mapping use different metrics. The best three results are highlighted with **red**, **blue**, **cyan**.

This is followed by utilizing the correlation attention block, as shown in Alg. 1, to capture the long-range dependencies between different features on the Correlation manifold. whereafter, the classification layer (introduced in Sec. 3.4), incorporated with FC & Softmax, to realize EEG classification.

Implementation Details. Considering that orthogonal constraint can serve as an implicit regularization to improve the network's generalization [Lezcano-Casado and Martınez-Rubio, 2019], we impose orthogonality on M in both $\hom^{ol}(\cdot)$ and $\hom^{ls}(\cdot)$. As orthogonal matrices lie in special orthogonal groups, their optimization requires a Riemannian optimizer, which we implement by generating a parameter $A \in \mathbb{R}^{n\times n}$ and computing its skew-symmetric matrix as $S = A - A^\top$. Under this parameterization, the orthogonal matrix O can be obtained by:

$$O = (I_n - S)\,(I_n + S)^{-1}. \tag{18}$$

This approach optimizes all parameters within Euclidean spaces. For the BCIC-IV-2a dataset, the number of subparts, the size of the transformation matrix in CorAtt, the learning rate, and the batch size are respectively set to 3, 25×25, $5e^{-4}$, and 128, while those for the MAMEM-SSVEP-II dataset are configured as 7, 15×15, $5e^{-3}$, and 64, respectively. In comparison, these values are respectively set to 3, 14×14, $1e^{-3}$, and 32 on the BCI-ERN dataset. For convenience, we summarize the specific network configurations of CorAtt across all the used datasets in Tab. 2.

4.2 Performance Comparison

Tab. 3 lists the experimental results of CorAtts and the selected competitors on the three used EEG datasets. Here, CorAtt-OLM and -LSM represent a unified metric for both the Attention Block and Tangent Mapping layers (OLM and LSM, respectively), whereas CorAtt-MIX adopts different metrics, specifically OLM for attention and LSM for classification. Note that with identical parameter settings for CorAtt-LSM, -OLM, and -MIX. Overall, standard deep-learning

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

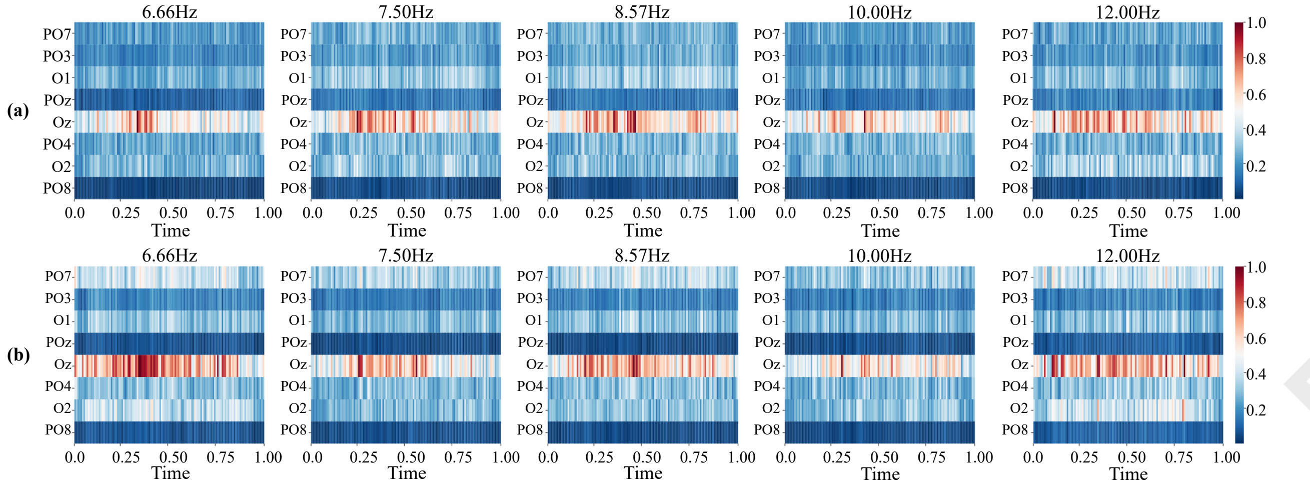

Figure 2: Heatmaps of CorAtt-OLM (a) and CorAtt-LSM (b) for the S11 subject across five different frequencies on the MAMEM-SSVEP-II dataset. The x-axis and y-axis represent time and EEG channels, respectively.

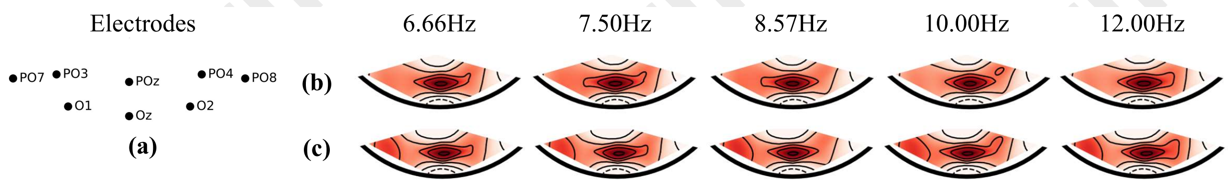

Figure 3: The diagram of electrode distribution (a) and the spatial topo-maps of CorAtt-OLM (b) and CorAtt-LSM (c) for the S11 subject across five different frequencies on the SSVEP dataset. Strong gradient activations are marked in dark red.

Method	MI	SSVEP	ERN
FEM	26.32 ± 0.92	20.33 ± 1.28	73.27 ± 2.87
Attention-OLM	56.67 ± 0.83	29.04 ± 2.51	58.77 ± 3.24
Attention-LSM	57.32 ± 0.97	29.63 ± 2.79	59.03 ± 3.76
FEM+ESA	49.32 ± 5.43	22.92 ± 2.13	64.32 ± 1.81
CorAtt-OLM	75.01 ± 2.78	67.39 ± 3.22	78.78 ± 3.40
CorAtt-LSM	74.47 ± 2.43	67.74 ± 2.44	78.63 ± 3.31

Table 4: Ablations of CorAtt components (ten-fold mean ± std) across all three datasets.

TanMap	Att Metric	MI	SSVEP	ERN
w/o	OLM	68.55 ± 2.85	63.75 ± 2.88	71.17 ± 4.41
OLM	OLM	75.01 ± 2.78	67.39 ± 3.22	78.78 ± 3.40
LSM	OLM	$\mathbf{75.56 \pm 1.58}$	$\mathbf{68.27 \pm 2.50}$	$\mathbf{79.04 \pm 2.91}$
w/o	LSM	64.51 ± 2.89	62.49 ± 3.01	70.63 ± 4.76
OLM	LSM	71.13 ± 1.78	67.37 ± 3.00	76.77 ± 2.63
LSM	LSM	74.47 ± 2.43	67.74 ± 2.44	78.63 ± 3.31

Table 5: Ablations of Tangent Mapping (ten-fold mean ± std) across all three datasets, where TanMap denotes the tangent mapping, Att Metric is the metric of Attention block.

models (EEGNet, ShallowCNet) exhibit relatively lower accuracy than geometry-based approaches (MAtt, GDLNet). CorAtt-MIX achieves the highest performance across all three tasks. This suggests that employing different metrics between attention and classification layers can more effectively adapt the geometric properties of correlation matrices, further demonstrating the flexibility and effectiveness of the proposed approach. Meanwhile, CorAtt-OLM and CorAtt-LSM consistently outperform MAtt, showing gains of approximately 0.30% and 0.24% on the MI dataset, 2.20% and 2.55% on SSVEP, and 3.10% and 2.95% on ERN. We attribute this performance gap between CorAtt and MAtt to two key factors: (1) CorAtt focuses on correlation matrix modeling, inherently addressing scale-invariant features of signals; (2) CorAtt transformation layer preserves the Lie group structure of Correlation manifolds, providing a more natural extension of attention mechanisms to non-Euclidean geometries.

4.3 Ablations

Ablations of the main components. As shown in Tab. 4, removing any module from the proposed CorAtts significantly drops the classification accuracy, confirming that all the components are essential. The fourth row in Tab. 4 reports the performance of FEM combined with a Euclidean self-attention (ESA) module. The comparison between CorAtts and FEM+ESA highlights the necessity of incorporating Riemannian computations into manifold attention design.

Ablations of the tangent mapping layers. In this subsection, we investigate the impact of the tangent mapping layer on the classification performance of the proposed CorAtt. From Tab. 5, it is evident that when the tangent mapping layer is omitted, the learning ability of CorAtt significantly drops. For example, the accuracy of CorAtt-LSM decreased by 9.94%, 5.25%, and 8.00% on the MI, SSVEP, and ERN

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

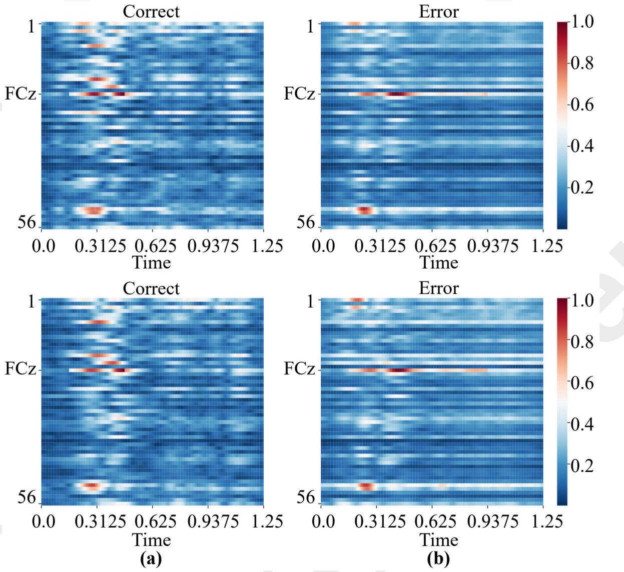

Figure 4: Heatmaps of CorAtt-OLM (a) and CorAtt-LSM (b) for two classes on the BCI-ERN datasets. The x-axis represents time, and the y-axis represents EEG channels.

Number	Metric	MI	SSVEP	ERN
1	OLM	75.01 ± 2.78	67.39 ± 3.22	78.78 ± 3.40
1	LSM	74.47 ± 2.43	67.74 ± 2.44	78.63 ± 3.31
2	OLM	**75.12 ± 2.69**	67.48 ± 2.95	**78.94 ± 2.93**
2	LSM	74.48 ± 1.76	67.76 ± 2.75	77.97 ± 2.71
3	OLM	74.40 ± 2.55	**68.10 ± 2.55**	78.76 ± 2.67
3	LSM	73.43 ± 2.64	67.00 ± 2.86	77.06 ± 3.09

Table 6: Ablations of Number of Attention Blocks. (ten-fold mean ± std) across all three datasets.

datasets, respectively. Furthermore, when the attention block is equipped with OLM and the classification layer is realized by LSM, CorAtt consistently achieves the highest accuracy on all the used datasets. This suggests that selecting appropriate Riemannian metrics for different layers is beneficial to enhancing performance, further revealing the flexibility and adaptability of our method.

Ablations for the number of attention blocks. We investigate the impact of using one, two, or three correlation attention blocks under both OLM and LSM. As shown in Tab. 6, two blocks occasionally offer slight gains (*e.g.*, OLM for MI and ERN), but three blocks generally degrade performance (*e.g.*, OLM for MI and LSM for all tasks). Notably, three blocks yield a marginal improvement in the SSVEP task under OLM yet produce an overall decline in accuracy for MI and ERN. This suggests that a single correlation attention block is sufficient for low-dimensional EEG data.

4.4 EEG Model Interpretation

For the MAMEM-SSVEP-II dataset, as shown in Figs. 2 and 3, across five stimulus frequencies, both CorAtt-OLM and CorAtt-LSM primarily exhibit heightened gradient responses around the Oz electrode. These responses appear most prominently between 0.25 and 0.75 seconds, indicating the crucial role of Oz in the visual cortex. Such findings are highly consistent with existing literature on the correlation between SSVEP and Oz in EEG recordings [Herrmann, 2001; Han *et al.*, 2018], likely due to the electrode's central location in the primary visual cortex, resulting in more pronounced induced potentials and an improved signal-to-noise ratio.

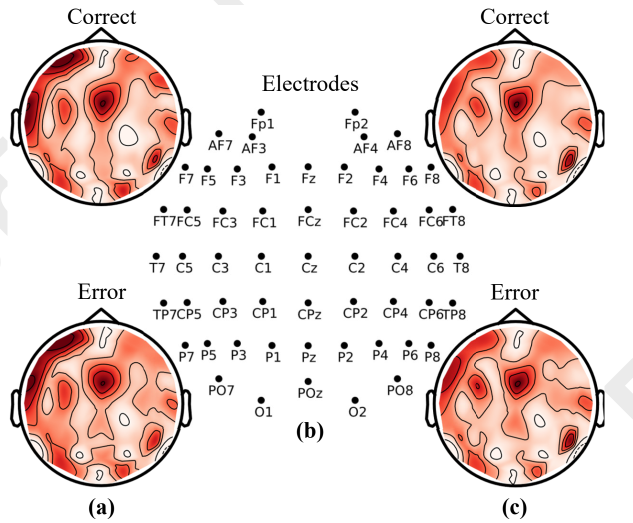

Figure 5: (a) and (c) display the visualization results of CorAtt-OLM and CorAtt-LSM on the BCI-ERN datasets S7 model, respectively, while (b) presents a diagram of the electrode distribution.

As shown in Figs. 4 and 5, for the BCI-ERN dataset, gradient responses for distinguishing 'correct' versus 'error' trials predominantly centre around the FCz. This observation aligns with substantial empirical evidence that the anterior cingulate cortex, a central medial prefrontal cortex region connected to limbic and frontal areas, underlies ERN generation. The consistent gradient responses for both feedback types around the FCz electrode should be noted, particularly in the 0.1 to 0.4-second interval. These findings strongly corroborate the differences in ERP waveforms between correct and incorrect stimuli reported by [Hajcak, 2012]. For the BCIC-IV-2a dataset, please refer to our App. A.

5 Conclusion

This paper proposes the correlation attention mechanism, which generalizes the Euclidean paradigm to the context of Correlation manifolds. Besides, we define the tangent mapping operations for classification over the Correlation manifolds under two Riemannian metrics. Extensive experimental results achieved on three EEG datasets certify the effectiveness and versatility of the proposed CorAtt. In summary, this is the first work to design a deep learning model (attention model in this article) on the Correlation manifolds to the best of our knowledge. The exploration of CorAtt is expected to help the emergence of more geometric deep learning methods for the correlation matrices in the future.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62306127, 62020106012, 62332008, 62172228), the Natural Science Foundation of Jiangsu Province (BK20231040, BK20221535), the Fundamental Research Funds for the Central Universities (JUSRP124015), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJCX25_1319), the Key Project of Wuxi Municipal Health Commission (Z202318), and the National Key R&D Program of China (2023YFF1105102, 2023YFF1105105).

Contribution Statement

Rui Wang and Ziheng Chen contributed equally to the supervision of this work.

References

[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*, 2022.

[Archakov and Hansen, 2021] Ilya Archakov and Peter Reinhard Hansen. A new parametrization of correlation matrices. *Econometrica*, 2021.

[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*, 2019.

[Brunner *et al.*, 2008] Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Graz University of Technology, Austria*, 2008.

[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*, 2023.

[Chen *et al.*, 2024a] Ziheng Chen, Yue Song, Gaowen Liu, Ramana Rao Kompella, Xiao-Jun Wu, and Nicu Sebe. Riemannian multinomial 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.

[Chen *et al.*, 2024c] Ziheng Chen, Yue Song, Xiaojun, and Nicu Sebe. RMLR: Extending multinomial logistic regression into general geometries. In *NeurIPS*, 2024.

[Chen *et al.*, 2024d] Ziheng Chen, Yue Song, Tianyang Xu, Zhiwu Huang, Xiao-Jun Wu, and Nicu Sebe. Adaptive Log-Euclidean metrics for SPD matrix learning. *IEEE TIP*, 2024.

[Chen *et al.*, 2025a] Ziheng Chen, Yue Song, Xiao-Jun Wu, Gaowen Liu, and Nicu Sebe. Understanding matrix function normalizations in covariance pooling through the lens of Riemannian geometry. In *ICLR*, 2025.

[Chen *et al.*, 2025b] Ziheng Chen, Yue Song, Xiaojun Wu, and Nicu Sebe. Gyrogroup batch normalization. In *ICLR*, 2025.

[David and Gu, 2019] Paul David and Weiqing Gu. A riemannian structure for correlation matrices. *Oper. Matrices*, 2019.

[Do Carmo and Flaherty Francis, 1992] Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian Geometry*. Springer, 1992.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[Epskamp and Fried, 2018] Sacha Epskamp and Eiko I Fried. A tutorial on regularized partial correlation networks. *Psychol. Methods*, 2018.

[Garcia-Hernando *et al.*, 2018] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *CVPR*, 2018.

[Ginestet *et al.*, 2012] Cedric E Ginestet, Andrew Simmons, and Eric D Kolaczyk. Weighted frechet means as convex combinations in metric spaces: properties and generalized median inequalities. *Stat. Probab. Lett.*, 2012.

[Gulcehre *et al.*, 2018] Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. Hyperbolic attention networks. In *ICLR*, 2018.

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

[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.*, 2018.

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

[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.*, 2001.

[Hine *et al.*, 2017] Gabriel Emile Hine, Emanuele Maiorana, and Patrizio Campisi. Resting-state EEG: A study on its non-stationarity for biometric applications. In *BIOSIG*, 2017.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

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

[Jalili and Knyazeva, 2011] Mahdi Jalili and Maria G Knyazeva. Constructing brain functional networks from eeg: partial and unpartial correlations. *J. Integr. Neurosci.*, 2011.

[Karcher, 1977] Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Commun. Pure Appl. Math.*, 1977.

[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.*, 2018.

[Lezcano-Casado and Martınez-Rubio, 2019] Mario Lezcano-Casado and David Martınez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *ICML*, 2019.

[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.*, 2012.

[Müller *et al.*, 2007] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Documentation mocap database HDM05. Technical report, Universität Bonn, 2007.

[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*, 2021.

[Nguyen *et al.*, 2024] Xuan Son Nguyen, Shuo Yang, and Aymeric Histace. Matrix manifold neural networks++. In *ICLR*, 2024.

[Nikolopoulos, 2016] Spiros Nikolopoulos. MAMEM EEG SSVEP Dataset II (256 channels, 11 subjects, 5 frequencies presented simultaneously). Dataset, 2016.

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

[Pennec *et al.*, 2006] Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *IJCV*, 2006.

[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.*, 2017.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Subha *et al.*, 2010] D Puthankattil Subha, Paul K Joseph, Rajendra Acharya U, and Choo Min Lim. EEG signal analysis: a survey. *J. Med. Syst.*, 2010.

[Tang *et al.*, 2024] Lichun Tang, Zhaoxia Yin, Hang Su, Wanli Lyu, and Bin Luo. Wfss: weighted fusion of spectral transformer and spatial self-attention for robust hyperspectral image classification against adversarial attacks. *Vis. Intell.*, 2024.

[Thanwerdas, 2024] Yann Thanwerdas. Permutation-invariant log-euclidean geometries on full-rank correlation matrices. *SIAM J. Matrix Anal. Appl.*, 2024.

[Tu, 2011] Loring W.. Tu. *An introduction to manifolds*. Springer, 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*, 2017.

[Wang *et al.*, 2024a] Rui Wang, Chen Hu, Ziheng Chen, Xiao-Jun Wu, and Xiaoning Song. A Grassmannian manifold self-attention network for signal classification. In *IJCAI*, 2024.

[Wang *et al.*, 2024b] Rui Wang, Xiao-Jun Wu, Ziheng Chen, Cong Hu, and Josef Kittler. SPD manifold deep metric learning for image set classification. *IEEE TNNLS*, 2024.

[Wang *et al.*, 2025] Rui Wang, Shaocheng Jin, Ziheng Chen, Xiaoqing Luo, and Xiao-Jun Wu. Learning to normalize on the SPD manifold under Bures-Wasserstein geometry. In *CVPR*, 2025.

[Wei *et al.*, 2019] Chun-Shu Wei, Toshiaki Koike-Akino, and Ye Wang. Spatial component-wise convolutional network (SCCNet) for motor-imagery EEG classification. In *NER*, 2019.

[Zeng *et al.*, 2024] Zelong Zeng, Fan Yang, Hong Liu, and Shin'ichi Satoh. Improving deep metric learning via self-distillation and online batch diffusion process. *Vis. Intell.*, 2024.