A Grassmannian Manifold Self-Attention Network for Signal Classification (Supplementary Material)

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

Table 1: The definition of each operation layer in the GMSA.

GMSA	Operation functions
GMT	$\mathbf{Q}_r = f_{gmt}(\mathbf{W}_q, \mathbf{Y}_r) = \mathbf{W}_q\mathbf{Y}_r$
	$\mathbf{K}_r = f_{gmt}(\mathbf{W}_k, \mathbf{Y}_r) = \mathbf{W}_k\mathbf{Y}_r$
	$\mathbf{V}_r = f_{gmt}(\mathbf{W}_v, \mathbf{Y}_r) = \mathbf{W}_v\mathbf{Y}_r$
ORM	$\mathbf{K}_r = \mathbf{\Omega}_r\mathbf{R}_r$
	$\mathbf{K}'_r = f_{orm}(\mathbf{K}_r) = \mathbf{K}_r\mathbf{R}_r^{-1} = \mathbf{\Omega}_r$
PM	$\mathcal{D}_{rj} = f_{pm}(\mathbf{Q}'_r, \mathbf{K}'_j) = \lvert\lvert\mathbf{Q}'_r{\mathbf{Q}'_r}^{\mathrm{T}} - \mathbf{K}'_j{\mathbf{K}'_j}^{\mathrm{T}}\rvert\rvert_{\mathrm{F}}^2$
SIM	$\mathcal{D}'_{rj} = f_{sim}(\mathcal{D}_{rj}) = \frac{1}{1+\log(1+\mathcal{D}_{rj})}.$
SMX	$\mathcal{D}''_{rj} = f_{smx}(\boldsymbol{\mathcal{A}}) = \frac{\exp(\mathcal{D}'_{rj})}{\sum_{t=1}^{m}\exp(\mathcal{D}'_{rt})}$
WAE	$\mathbf{V}''_r = f_{wae}(\hat{\boldsymbol{\mathcal{A}}}, \boldsymbol{\mathcal{V}}) = \sum_{j=1}^{m}\mathcal{D}''_{rj}\cdot(\mathbf{V}'_j{\mathbf{V}'_j}^{\mathrm{T}})$
REO	$\mathbf{V}''_r = \mathbf{Z}\mathbf{S}\mathbf{Z}^{\mathrm{T}}$
	$\mathbf{Y}'_r = f_{reo}(\mathbf{V}''_r) = \mathbf{Z}_{1:q_i}$

In this supplementary material, we first show the details of gradient computation in the proposed GMSA in Section 1. In Section 2, some additional experiments are conducted to confirm the necessity of some primary components contained in GMSA. Finally, we present an additional interpretation of the EEG model in Section 3.

1 Backward Propagation

For the proposed GMSA, a series of successive function compositions $f = f^{(\rho)} \circ f^{(\rho-1)} \circ f^{(\rho-2)} \circ ... \circ f^{(2)} \circ f^{(1)}$ with parameters $\boldsymbol{\mathcal{W}} = \{\mathbf{W}_\rho, \mathbf{W}_{\rho-1}, ..., \mathbf{W}_1\}$ can be considered as the data embedding model, which satisfies the properties of metric spaces. Here, $f^{(k)}$ and $\mathbf{W}_k$ are the operation function and weight parameter of the k-th layer respectively, and ρ denotes the number of layers of GMSA. The loss of the k-th layer can be signified as: $L^{(k)} = \ell \circ f^{(\rho)} \circ ... \circ f^{(k)}$, where ℓ is the loss function, *i.e.*, cross-entropy loss, of the output layer.

GMT layer: Due to the weight space of GMT layers being a compact Stiefel manifold $St(d_z, d_c)$, we refer to the method studied in [Huang *et al.*, 2018] to update the weight parameter by generalizing the traditional stochastic gradient descent (SGD) settings to the context of Stiefel manifolds. The updating rule for $\mathbf{W}_k$ on the $St(d_z, d_c)$ is given below:

According to the GMT function shown in Table 1, we rewrite it as: $\mathbf{Y}_k = f^{(k)}(\mathbf{W}_k, \mathbf{Y}_{k-1}) = \mathbf{W}_k\mathbf{Y}_{k-1}$. Then, the following variation of $\mathbf{Y}_k$ can be obtained:

$$d\mathbf{Y}_k = d\mathbf{W}_k\mathbf{Y}_{k-1} + \mathbf{W}_k d\mathbf{Y}_{k-1}. \tag{1}$$

Based on the invariance of the first-order differential, we can have the following chain rule:

$$\frac{\partial L^{(k+1)}}{\partial \mathbf{Y}_k} : d\mathbf{Y}_k = \frac{\partial L^{(k)}}{\partial \mathbf{W}_k} : d\mathbf{W}_k + \frac{\partial L^{(k)}}{\partial \mathbf{Y}_{k-1}} : d\mathbf{Y}_{k-1}. \tag{2}$$

By replacing the left-hand side of Eq. (2) with Eq. (1) and exploiting the matrix inter product ":" property, the following two formulas can be derived:

$$\frac{\partial L^{(k+1)}}{\partial \mathbf{Y}_k} : d\mathbf{W}_k\mathbf{Y}_{k-1} = \frac{\partial L^{(k+1)}}{\partial \mathbf{Y}_k}\mathbf{Y}_{k-1}^{\mathrm{T}} : d\mathbf{W}_k, \tag{3}$$

$$\frac{\partial L^{(k+1)}}{\partial \mathbf{Y}_k} : \mathbf{W}_k d\mathbf{Y}_{k-1} = \mathbf{W}_k^{\mathrm{T}}\frac{\partial L^{(k+1)}}{\partial \mathbf{Y}_k} : d\mathbf{Y}_{k-1}. \tag{4}$$

Combining Eqs. (2-4), the partial derivatives of $L^{(k)}$ w.r.t $\mathbf{W}_k$ and $\mathbf{Y}_{k-1}$ can be computed by:

$$\frac{\partial L^{(k)}}{\partial \mathbf{W}_k} = \frac{\partial L^{(k+1)}}{\partial \mathbf{Y}_k}\mathbf{Y}_{k-1}^{\mathrm{T}}, \quad \frac{\partial L^{(k)}}{\partial \mathbf{Y}_{k-1}} = \mathbf{W}_k^{\mathrm{T}}\frac{\partial L^{(k+1)}}{\partial \mathbf{Y}_k}. \tag{5}$$

At this time, the updating criteria of $\boldsymbol{W}_k$ on the Stiefel manifold are given below:

$$\mathbf{W}_k^{t+1} = \mathcal{R}_{\mathbf{W}_k^t}(-\eta\Pi_{\mathbf{W}_k^t}(\nabla L^{(k)}_{\boldsymbol{W}_k^t})), \tag{6}$$

where $\mathcal{R}$ signifies the retraction operation used to map the optimized parameter back onto the Stiefel manifold, η is the learning rate, Π represents the projection operator used to convert the Euclidean gradient into the corresponding Riemannian counterpart:

$$\tilde{\nabla} L^{(k)}_{\boldsymbol{W}_k^t} = \Pi_{\mathbf{W}_k^t}(\nabla L^{(k)}_{\boldsymbol{W}_k^t}) = \nabla L^{(k)}_{\boldsymbol{W}_k^t} - \boldsymbol{W}_k^t(\nabla L^{(k)}_{\boldsymbol{W}_k^t})^{\mathrm{T}}\boldsymbol{W}_k^t, \tag{7}$$

where $\nabla L^{(k)}_{\boldsymbol{W}_k^t}$ is the Euclidean gradient, computed by the first term of Eq. (5), $\tilde{\nabla} L^{(k)}_{\boldsymbol{W}_k^t}$ denotes the obtained Riemannian

[*] Corresponding author

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

Methods	RADAR	SSVEP	ERN
'w/o' ORM	85.48 ± 1.87	57.15 ± 3.14	74.28 ± 4.41
EuWAE	87.60 ± 1.82	61.62 ± 2.61	74.25 ± 3.21
EM	87.50 ± 2.24	62.81 ± 3.13	71.32 ± 4.23
FEM+GMSA (GDLNet)	94.68 ± 0.90	65.52 ± 2.86	78.23 ± 2.52

Table 3: Acc. (%) under different NoGs on the SSVEP dataset.

NoGs=1	NoGs=2	NoGs=3
65.52 ± 2.86	65.88 ± 3.11	63.78 ± 3.34

gradient. After that, the weight parameter can be updated by: $\boldsymbol{W}_k^{t+1} = \mathcal{R}(\boldsymbol{W}_k^t - \eta \tilde{\nabla} L^{(k)}_{\boldsymbol{W}_k^t})$. For detailed information about the Riemannian geometry of the Stiefel manifold and its corresponding retraction operation, please kindly refer to [Absil *et al.*, 2009].

SMX layer: According to Table 1, the partial derivative of $L^{(k)}$ w.r.t $\mathcal{D}'_{rj}$ is computed by:

$$\frac{\partial L^{(k)}}{\partial \mathcal{D}'_{rj}} = \frac{\exp(\mathcal{D}_{rj}) \cdot \lambda}{[\sum_{t=1}^{m} \exp(\mathcal{D}'_{rt})]^2} \cdot \frac{\partial L^{(k+1)}}{\partial \mathcal{D}''_{rj}}, \quad (8)$$

where $\lambda = [\sum_{t=1}^{m} \exp(\mathcal{D}'_{rt}) - \exp(\mathcal{D}'_{rj})]$.

SIM layer: based on Table 1, the partial derivative of $L^{(k)}$ w.r.t $\mathcal{D}_{rj}$ is given below:

$$\frac{\partial L^{(k)}}{\partial \mathcal{D}_{rj}} = \frac{-1}{[1 + \log(1 + \mathcal{D}_{rj})]^2} \cdot \frac{1}{(1 + \mathcal{D}_{rj})} \cdot \frac{\partial L^{(k+1)}}{\partial \mathcal{D}'_{rj}}. \quad (9)$$

2 Ablation Study

In this section, we evaluate the significance of each primary component contained in GMSA through a sequence of ablation studies.

(1) From Table. 2, we can see that the classification performance of our model decreases significantly when removing the ORM layer. The fundamental reason is that the Grassmannian properties of the input feature matrices generated by the GMT layer can not be maintained, demonstrating the necessity of Riemannian computations.

(2) To verify the effectiveness of the WAE layer, we use the Euclidean-like weighted average, denoted as EuWAE ($\mathbf{V}''_r = \sum_{j=1}^{m} \mathcal{D}''_{rj} \cdot \mathbf{V}'_j$), to replace the manifold-valued WAE operation. Considering that EuWAE is not actually an geometric operator defined on the Grassmannian manifold, it will induce a decline in classification accuracy of our model. The experimental results listed in Table 2 also verify this.

(3) Considering that the PM layer could also be realized using the Euclidean distance, we use $\mathcal{D}_{rj} = ||\mathbf{Q}_r - \mathbf{K}_j||_{\mathrm{F}}^2$ to rewrite Eq. (8) of the main paper. For simplicity, we denote it as EM in Table 2. The fundamental reason of the reduced accuracy caused by EM is that it fails to accurately estimate the geodesic similarity between any two Grassmannian elements, making the obtained Riemannian barycenter will deviate from the actual one.

(4) Number of GMSAs (NoGs): Table. 3 shows that NoGs=2 only brings marginal improvement, while NoGs=3 exhibits reduced performance. These indicate that 1-block GMSA is almost saturated for the low-dimensional signal data.

Table 4: Acc. comparison on the NW-UCLA dataset.

Models	Acc.
CTR-GCN [a]	92.67%
CTR-GCN+EuSA	93.58%
CTR-GCN+GMSA	94.18%
CTR-GCN+GMSA+EuSA	**94.40%**

All in all, the aforementioned experimental results demonstrate the importance of Riemannian geometric operators in the design of manifold attention network.

3 EEG Model Interpretation

This section highlights the characteristics that GDLNet extracts from EEG signals. Fig. 1 displays the temporal variation in spatial distribution for each form of visual stimulus contained in the MAMEM-SSVEP-II dataset. In the experiment, we observe that brain topomaps at varying times show diverse yet analogous brain activity. A persisting notable activation at Oz is witnessed across all temporal sequences. Similarly, the obvious gradient responses on the scalp during each epoch exhibit consistency for every visual stimulus frequencies. In conclusion, the proposed GDLNet is able to capture small differences in similar spatial distributions present in the topomaps of each frequency, which helps in decoding the SSVEP-EEG signals at all frequencies.

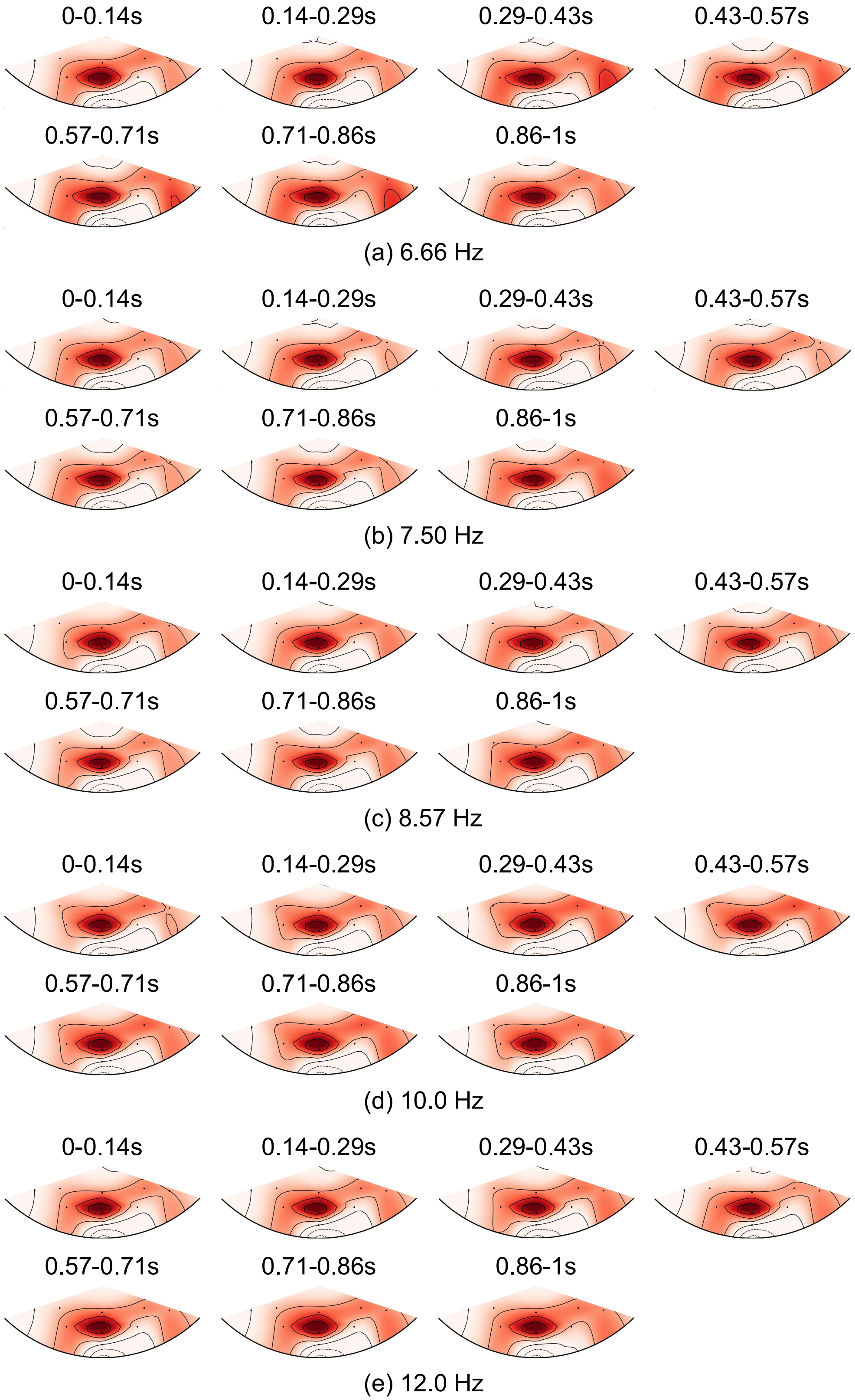
Fig. 2 depicts the detailed all-channel gradient responses of model S7 in the BCI-ERN dataset. As stated in the main paper, the FCz located in the frontal-central midline can effectively capture the ERP of ERN. Furthermore, under both types of stimulation, the temporo-parietal junction also exhibit moderate activation.

Collectively, our results clearly underline the effectiveness and potential of the proposed GDLNet in capturing the intricate spatiotemporal variations of EEG signals.

4 More Experience

The Northwestern-UCLA dataset [Wang *et al.*, 2013], attained by employing three Kinect cameras from a multitude of perspectives, encompasses 1494 video segments that represent 10 distinct categories of actions. Each action is performed by 10 different subjects. We follow the same evaluation protocol in [Chen *et al.*, 2021]: training data from the first two cameras, and testing data from the other camera.

The proposed GDLNet is further extended to the skeleton-based action recognition task using the NW-UCLA dataset. The results are listed in. We adopt CTR-GCN [Chen *et al.*, 2021] as FEM, and feed the resulting features into our GDLNet. Specifically, we transform the input features of the pre-trained CTR-GCN FC layer into $\mathbb{R}^{256 \times 260}$, where 256 is the CTR-GCN input channels and 260 is the feature map of each channel reshaped into vectors. We set the number of epochs to 3 and feed the resulting features into MMM. For simplicity, we let CTR-GCN just contain the joint stream. The results are listed in 4. It can be seen that CTR-GCN+GMSA reaches 1.51% and 0.60% higher accuracy than that of CTR-GCN and CTR-GCN+EuSA, demonstrating the effectiveness



(a) 6.66 Hz

(b) 7.50 Hz

(c) 8.57 Hz

(d) 10.0 Hz

(e) 12.0 Hz

Figure 1: For the MAMEM-SSVEP-II dataset, Spatial topomaps of the S11 model over epoch and at different frequencies of visual stimulation. The SSVEP intense average absolute gradient activation within the visual cortex's distinct regions is marked in dark red.

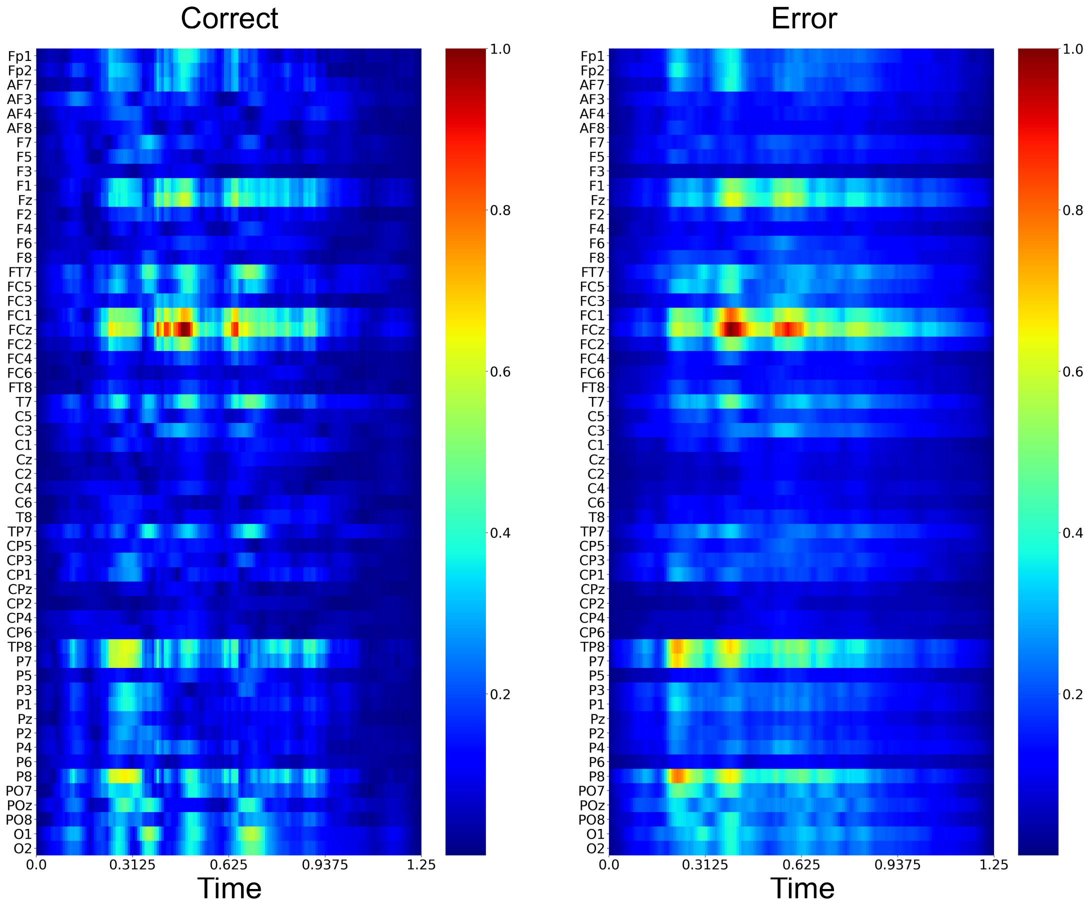


Figure 2: Gradient responses to 'correct' and 'error' feedback categories for all channels in the BCI-ERN dataset.

of our method. This part will be added to the final version, and more details will be presented.

References

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

[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*, 2021.

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

[Wang *et al.*, 2013] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Learning actionlet ensemble for 3d human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.