A Spatial-Temporal Transformer based on Domain Generalization for Motor Imagery Classification

Shaozhe Liu, Leike An, Chi Zhang, and Ziyu Jia*

***Abstract*— Motor imagery (MI) has emerged as a classical paradigm in brain-computer interface (BCI) research. In recent years, advancements in deep learning techniques, such as the application of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have enabled the use of MI classification. Despite their success, CNNs and RNNs are not capable of effectively extracting brain spatial and temporal information necessary for MI classification. Additionally, differences in individual subjects further complicate the classification process. To address these limitations, a novel Spatial-Temporal Transformer based on Domain Generalization (ST-DG) has been proposed for MI classification using EEG signals. This framework utilizes a spatial-temporal transformer architecture to capture essential spatiotemporal characteristics of the brain, while also employing Domain Generalization techniques to account for cross-subject variability and improve the model's generalization performance. Experimental results on two public datasets demonstrate the state-of-the-art classification performance.**

I. INTRODUCTION

The classification of motor imagery (MI) relies heavily on the accurate extraction of electroencephalography (EEG) signals from experimental subjects engaged in motor-like imagery behaviors [1]. These signals are then utilized to determine the user's intention, allowing for the control of a variety of devices, including wheelchairs, robots, and autonomous vehicles [2]–[4]. Consequently, it is imperative that techniques for robust information extraction be employed in order to ensure accurate signal processing and reliable user control.

The classification of MI has historically relied on the utilization of manually engineered features extracted from EEG signals. These features, such as the common spatial pattern (CSP) and filter bank common spatial pattern (FBCSP), have been extensively employed in traditional methods [5], [6]. However, these methods are limited by their reliance on prior knowledge of brain-computer interfaces (BCIs) and fail to fully capture the complex spatiotemporal patterns of EEG signals, leading to decreased accuracy in MI classification. In the current era, deep-learning models, specifically Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have garnered substantial attention as highly effective methodologies for the classification of MI in BCI applications. By capitalizing on their adeptness in representation learning, CNNs and RNNs possess the capability to autonomously acquire intricate and discerning characteristics from unprocessed EEG signals. Consequently, these models exhibit enhanced performance and durability in the classification of MI within real-life scenarios. Nonetheless, despite the notable accomplishments, challenges persist in the realm of MI classification.

The first challenge encountered in MI classification pertains to the insufficient consideration of spatial-temporal characteristics. Addressing this difficulty necessitates the formulation of pertinent methodology to encode the topology connecting brain regions alongside the temporal development of EEG signals. CNN-based approaches are widely adopted for effective learning of data representations from EEG channels, enabling extraction of spatially invariant and localized features [7]–[10]. For instance, EEGNet [11] proposes the incorporation of depthwise and separable convolutions within the EEG classification framework. Conversely, Sakhavi et al. [12], [13] employ spatial filtering and temporal sampling techniques to augment the input processing of EEG signals. Additionally, Zhao [14] utilizes a 3D representation to express the spatial distribution of channels, which permits the exploitation of original data features by a 3D CNN. In contrast to convolutional approaches, RNN-based methods are adept at capturing the temporal information inherent in EEG trials [15], [16], which are critical for modeling the temporal dependencies of the EEG signals. Some mothods [17], [18] employ LSTM networks to extract spatiotemporal features separately. Zhang [19] integrates the advantages of CNN and LSTM via concatenation, producing an advanced hybrid model for MI classification. However, it is important to acknowledge that CNNs may encounter difficulties in preserving certain temporal details inherent in EEG time-series data. Consequently, this limitation can impede their capacity to comprehend a broad spectrum of intrinsic relationships within the signal. Simultaneously, RNNs only take into account the preceding states and the current situation, thereby potentially missing out on the entire temporal sequence of actions embedded in the EEG signals. In recent years, the transformer has gained significant traction in processing sequential data by quantifying the associations between varying positions while correlating the sequence's dependencies [20]. Intriguingly, the transformer appears capable of selecting

Shaozhe Liu is with Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China; (email:liushaozhe@pku.org.cn)

Leike An is with College of Engineering, Peking University, Beijing, China; (email:anleike721@pku.edu.cn)

Chi Zhang is with Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China; (email:chiizhang@pku.edu.cn)

Ziyu Jia is with Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China; (email:jia.ziyu@outlook.com)

*Corresponding author

crucial information in an unbiased and automated manner while encoding practical data from EEG signals [21]. As a result, researchers have developed techniques that integrate the transformer with CNNs and RNNs to explore and decipher more discriminative information from EEG signals, which can help improve MI classification [22]–[24].

An additional challenge in EEG signals analysis lies in the substantial variation across subjects, posing significant hindrances to the model's generalizability. This cross-subject variability problem is commonly recognized as a domain shift issue [25]. Despite attempts to mitigate the cross-subject variability problem through fine-tuning the model on new subject data [26], [27], this methodology is often encumbered by its time-consuming nature. Consequently, transfer learning-based approaches have gained prominence as a viable solution to this issue. By utilizing pre-trained models or pre-processing techniques, transfer learning enables researchers to transfer knowledge across disparate subjects and tasks, thereby enhancing accuracy and robustness in MI classification. This approach is expected to remain a prominent research area in the field of BCI. Domain Adaptation (DA) is a transfer learning framework that utilizes multi-domain source data along with a limited amount of target data to improve classification accuracy. [28]. However, the requirement of target domain data in DA approaches poses a challenge in the context of MI classification, as obtaining data from new subjects is a time-consuming and costly process. Conversely, Domain Generalization (DG) strives to create a model capable of generalizing to unseen data from several subjects, without relying on specific information from a target domain [29]–[31]. This renders DG an apt choice for addressing cross-subject variation challenges in the MI classification context, where a model ought to perform well across numerous subjects. Through exclusive reliance on annotated data from heterogeneous subjects, DG approaches facilitate improved classification accuracy and a more robust framework for EEG signal modeling. Therefore, DG-based methods hold promising applications for addressing challenges related to cross-subject variability in MI classification and may offer valuable contributions to the field of BCI research.

This paper presents a novel methodology for MI classification that effectively tackles the previously mentioned challenges. The proposed approach is based on two key modules: the ST-Transformer module and the Domain Generalization module. The former employs spatial and temporal transformer blocks to extract spatial and temporal information from EEG signals, ultimately enabling the model to acquire essential high-level features capable of facilitating motor imagery classification. The latter module is specifically designed to overcome the issue of cross-subject variability that frequently arises in EEG analysis. By leveraging solely on domain-labeled data, our model exhibits higher generalization abilities towards novel subjects and improved robustness across diverse experimental conditions. The primary contributions of this research are the following:

- To our knowledge, this study represents the inaugural endeavor to integrate the Spatial-Temporal Transformer and Domain Generalization techniques within a cohesive framework to enhance MI classification.
- We devise the ST-Transformer to proficiently capture the spatial and temporal characteristics of the EEG signals.
- In our study, we employ the technique of Domain Generalization to extract subject-invariant features, thereby enhancing the generalization capabilities of the proposed model.
- The proposed ST-DG achieves the state-of-the-art performance on BCI-2A and 2B datasets in leave-one-subject-out (LOSO) validation.

II. METHODOLOGY

This research introduces a pioneering method for encoding EEG signals through the utilization of a spatial-temporal transformer, which capitalizes on the principles of domain generalization. The schematic diagram of the overall framework is depicted in Fig. 1. The EEG signals undergo an initial processing step performed by a multi-convolutional-pooling embedding layer. The spatial transformer employs feature-channel attention to enhance the model's ability to discriminate relevant channels from irrelevant ones. In the temporal transformer, we employ multi-head attention to encode temporal information. Finally, the encoded feature is fed into the domain generalization module, and the outcome is achieved via multi-fully connected layers.

A. Spatial-Temporal Transformer

The architectural overview of the ST-Transformer is visually presented in Fig. 2. The ST-Transformer consists of L stacked Transformer blocks, where each block incorporates attention units, a feed-forward layer, and layer normalization along with the residual connection. The spatial and temporal information is acquired through a process of calculating the importance score, which is utilized to assign appropriate weights to the values of the EEG signals.

1) Spatial Transformer: Within this module, the linear projections of the embedded sequence of variable X are denoted as Q, K, and V. These projections are mathematically represented in the implementation as follows:

$$Q, K, V = \text{Linear}(X), X \in R^{C \times T} \tag{1}$$

$$\text{Attention}_{\text{s}}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_s}})V \tag{2}$$

where Q signifies the EEG channels that serves as a reference for comparison, while K denotes all other EEG channels involved in the process, computed by means of dot product. The obtained result is subsequently divided by a scaling factor equivalent to the square root of the spatial dimension, denoted as $\sqrt{d_s}$. Finally, the weight score for the ultimate representation is determined by assigning the dot product with V. $\text{Attention}_{\text{s}}(Q, K, V) \in R^{C \times T}$ is the embedded feature weighted by the attention score in Equation 2. The attention map visualizes the correlation between individual

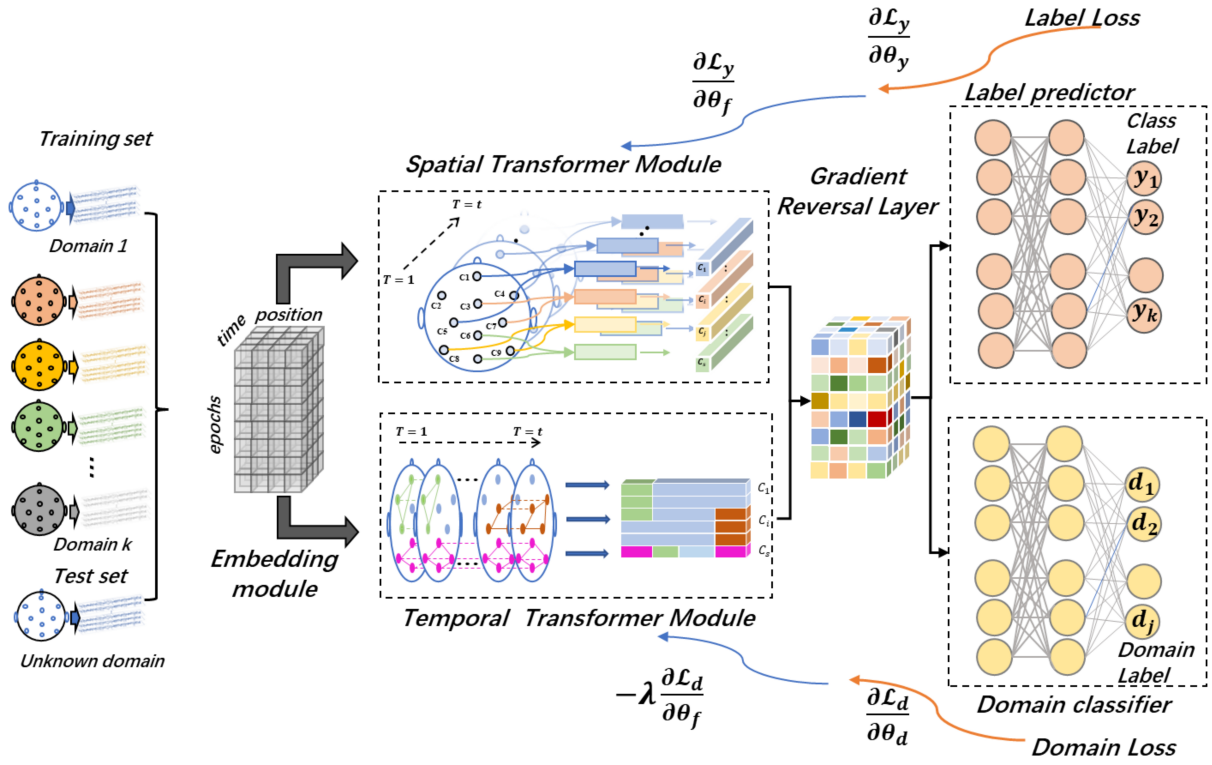

Fig. 1. The overall architecture of ST-DG. Initially, the architecture incorporates an embedding layer composed of convolutional layers that perform feature compression. Subsequently, an attention-based spatial-temporal transformer is developed to extract the most relevant spatial-temporal features for classification purposes. To enhance the generalizability performance of the model, a gradient reversal layer is introduced, facilitating domain generalization, and enabling the extraction of subject-invariant features. Notably, in the domain generalization process, each subject in the training dataset is treated as a specific source domain. This approach distinguishes itself from other transfer learning methods by not requiring any prior knowledge or input from the unknown target domain. Additionally, domain generalization serves as an optimal strategy for effectively addressing cross-subject variability issues.

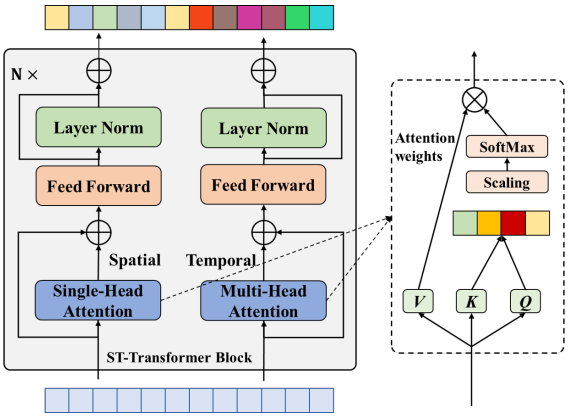

Fig. 2. The ST-Transformer architecture.

EEG channels, enabling our model to attend more to salient spatial features while disregarding less relevant ones.

2) Temporal Transformer: Different from the spatial transformer approach, in this study, we utilize multi-head attention (MHA) [20] to capture the interdependencies present in the temporal evolution. The input to the temporal transformer is a time series $X \in R^{C\times T}$, which is split into h parts and calculates the attention score in parallel. After the weighted dot product operation, the multi-head attention outputs are concatenated. The process can be mathematically defined as:

$$\text{Attention}_\text{t}(Q,K,V) = \text{softmax}(\frac{Q^T K}{\sqrt{d_t}})V^T \quad (3)$$

$$\text{head}_\text{t} = \text{Attention}_\text{t}(QW_i{}^Q, KW_i{}^K, VW_i{}^V) \quad (4)$$

$$\text{MultiHead} = \text{Concat}(\text{head}_1, \ldots \text{head}_\text{h})\, W^O \quad (5)$$

where $Q^T K \in R^{T\times T}$ represent the attention map in Equation 3. The linear mapping functions denoted as $W_i{}^Q$, $W_i{}^K$, and $W_i{}^V$, are utilized to obtain the query sequence, key sequence, and value sequence of each head. Additionally, the linear transformation function W^o is employed to derive the output. In order to augment the model's non-linear learning capability, two fully-connected layers with residual connections and the ELU activation function are connected to the multi-head attention mechanism.

B. Domain Generalization

In order to alleviate the influence of inter-individual variations, an adversarial domain generalization approach is adopted to improve the resilience of our model. Specifically, DG aims to train the model in a manner that makes it impossible to differentiate the source domain of the sample data. Simultaneously, it strives to improve the classification

performance of MI as much as possible. For instance, the model cannot discern whether the samples from domain i correspond to its own domain, yet it can still accurately recognize the motor imagery. Thus, the model has not learned personalized features specific to each domain, but rather subject-invariant features associated with MI classification.

Domain generalization comprises three main components: a feature extractor $\mathcal{G}_f$, a label predictor $\mathcal{G}_l$, and a domain classifier $\mathcal{G}_d$. As described in Section II-A, the ST-Transformer serves as the feature extractor $\mathcal{G}_f$ that transforms the input data into a feature space that is invariant to domain shifts.

$$\mathbb{X} = \mathcal{G}_f(X; \theta_f) \tag{6}$$

where X refers to the raw EEG data input, θ_f represents the model's trainable parameters, and $\mathbb{X}$ corresponds to the feature vector that has been embedded with both temporal and spatial information.

The extracted sequence is subsequently inputted into both the label predictor, denoted as $\mathcal{G}_l$, and the domain classifier, denoted as $\mathcal{G}_d$. A softmax function is employed for both of these stages:

$$\hat{y_i} = \mathrm{softmax}(\mathcal{G}_l(\mathbb{X}_i; \theta_y)) \tag{7}$$

$$\hat{d}_i = \mathrm{softmax}(\mathcal{G}_d(\mathbb{X}_i; \theta_d)) \tag{8}$$

where $\mathbb{X}_i$ refers to the embedded sequence originating from the i^{th} sample. The predicted results of $\mathcal{G}_l$ and $\mathcal{G}_d$ denoted as $\hat{y_i}$ and $\hat{d}_i$ respectively are both multi-class classifiers. The loss functions can be defined as follows:

$$\mathcal{L}_y = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C_y} y_{i,j} log \hat{y_{i,j}} \tag{9}$$

$$\mathcal{L}_d = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C_d} d_{i,j} log \hat{d_{i,j}} \tag{10}$$

where $\mathcal{L}_y$ represents the cross entropy loss function applied to the multi-classification problem, N refers to the total number of samples within the dataset, whereas C_y and C_d signify the number of classes and domains respectively. The variable y denotes the true label associated with a sample, while d represents the true domain of that sample.

In order to establish an adversarial relationship between the label predictor and the domain classifier, we introduce a Gradient Reversal Layer (GRL) [25] between the feature extractor $\mathcal{G}_f$ and the domain classifier $\mathcal{G}_d$. By incorporating GRL, we are able to effectively merge feature learning and domain generalization into a cohesive framework that enables seamless execution of backpropagation algorithms. The overall loss function and optimization procedure for our proposed model are formulated as follows:

$$\mathcal{L}_{DG} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C_y} y_{i,j} log \hat{y_{i,j}} + \lambda \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C_d} d_{i,j} log \hat{d_{i,j}} \tag{11}$$

$$\left(\hat{\theta_f}, \hat{\theta_y}\right) = \arg\min_{\theta_f, \theta_y} \mathcal{L}\left(\theta_f, \theta_y, \hat{\theta_d}\right) \tag{12}$$

$$\left(\hat{\theta_d}\right) = \arg\max_{\theta_d} \mathcal{L}\left(\hat{\theta_f}, \hat{\theta_y}, \theta_d\right) \tag{13}$$

Optimizing the loss function can enable the feature extractor $\mathcal{G}_f$ to realize the objective of identifying the subject-invariant feature space. To achieve this, we utilize θ_y and θ_d as parameters for minimizing the loss of $\mathcal{G}_y$ and $\mathcal{G}_d$, respectively. Specifically, we perform optimization on the parameter θ_y in a manner that entails minimizing the loss associated with $\mathcal{G}_y$ while concurrently maximizing the loss associated with $\mathcal{G}_d$.

III. EXPERIMENT

A. Baseline and experiment setting

We conduct a comparative analysis of our model against the following approaches:

- EEGNET [12]: A classical compact fully convolutional network that incorporates depth-wise and separable convolutions for EEG classification via CNNs.
- ConvNet [32]: The ConvNet consists of four convolution-pooling blocks, which greatly enhance its capability to extract high-level features from raw EEG signals.
- MMCNN [33]: An end-to-end deep learning model encompasses a composition of five parallel EEG Inception Networks, each comprised of an EEG Inception block, a Residual block, and a Squeeze and Excitation block.

In order to evaluate the accuracy across subjects, we perform leave-one-subject-out (LOSO) validation for all the algorithms employed in our experimental setup. Our approach is coded using Python 3.7 and the PyTorch library, running on a Geforce 2080Ti GPU. The batch size is set to 250, while the learning rate is configured to 0.001.

B. Comparison and Analysis of Experiment Results

In order to assess the performance of our proposed model for MI classification, we conduct a comparative analysis with several baseline models using the BCI-2A and BCI-2B datasets, as shown in Table I. The experimental outcomes demonstrate that our proposed model, referred to as ST-DG, surpasses the performance of the baseline methods by up to 5% in terms of accuracy, spanning across all subjects. To be more precise, ConvNet and EEGNet manifest the effectiveness of CNN-based deep architectures for accomplishing accurate classification of EEG signals. However, these models encounter limitations in capturing global dependencies due to their reliance on convolutions at different scales, thereby resulting in suboptimal performance. In contrast, the proposed model addresses this drawback by incorporating domain generalization, allowing the acquisition of subject-invariant spatial and temporal features. This distinctive approach contributes to achieving state-of-the-art performance in MI classification.

Table II indeed shows the results of different models for each subject in the 2B dataset, and our proposed model exhibits good generalization ability in comparison to other models. ST-DG achieves the highest accuracy for several subjects, despite being trained on a different subset of subjects. Thus, our approach is effective in addressing the issue of cross-subject variability in EEG signals analysis.

TABLE I

THE PERFORMANCE COMPARISON ON DATASETS 2A AND 2B

Dataset	Methods	Accuracy(%)	AUC
BCI-2A	EEGNet	51.639±0.017	0.777±0.013
	ConvNet	53.003±0.017	0.799±0.014
	MMCNN	52.135±0.015	0.789±0.013
	ST-DG	57.705±0.014	0.823±0.014
BCI-2B	EEGNet	70.441±0.017	0.785±0.013
	ConvNet	69.134±0.018	0.765±0.014
	MMCNN	70.613±0.015	0.782±0.013
	ST-DG	75.089±0.014	0.834±0.013

TABLE II

CLASSIFICATION ACCURACY(%) ON DIFFERENT SUBJECT

Model	Subject									Avg.
	1	2	3	4	5	6	7	8	9	
ConvNet	65.2	66.0	63.0	81.1	65.9	65.7	75.5	71.8	67.7	69.1
EEGNet	66.7	71.2	64.5	83.8	68.3	70.0	75.7	71.3	68.5	71.1
MMCNN	72.0	71.0	64.5	81.9	70.4	76.7	76.0	75.2	68.0	72.8
ST-DG	75.0	71.7	67.0	85.4	70.7	79.7	77.7	73.8	74.5	75.0

Fig. 3 indeed shows the results of our sensitivity analysis, where we investigate how different hyperparameter settings and network configurations impact the performance of our model. Interestingly, we found that our model is relatively insensitive to changes in hyperparameters. The ST-DG is quite robust and can perform well across a wide range of settings.

C. Ablation experiments

In order to comprehensively assess the impact of various components within the ST-DG framework, we conduct a series of ablation experiments. The experimental results are subsequently analyzed and evaluated using the BCI-2B dataset.

- **S-DG**: In order to investigate the advantages of incorporating temporal information from EEG signals, the Temporal Transformer module is removed.
- **T-DG**: In order to investigate the advantages of incorporating spatial information from EEG signals, the Spatial Transformer module is removed.
- **ST**: In order to investigate the advantages of incorporating subject-invariant information from EEG signals, the DG architecture is removed.

All variant models utilized identical configurations as the ST-DG model, with the exceptions previously delineated.

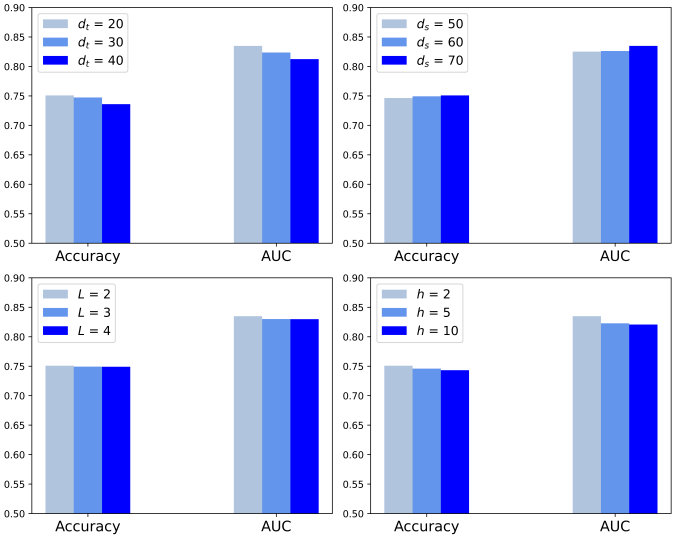

Fig. 3. The sensitivity test results. Four figures are provided, each illustrating the impact of different configurations on the motor imagery classification performance. These configurations include the temporal model dimension (d_t), the spatial model dimension (d_s), the number of encoder layers (L), and the number of attention heads (h).

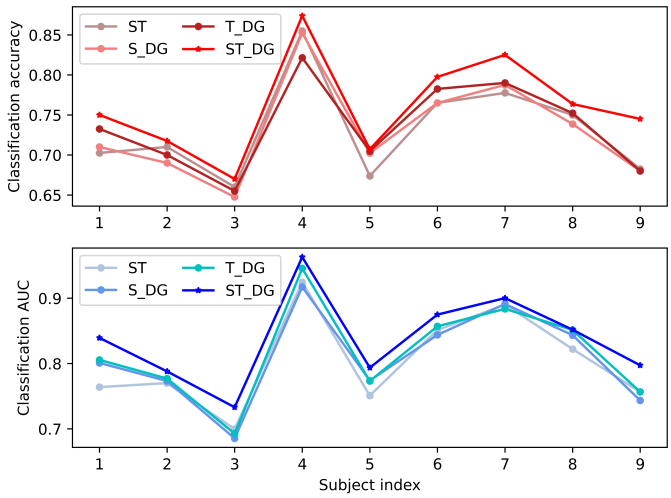

Fig. 4. Ablation experiments results in each subject on BCI-2B dataset.

Results are illustrated in Fig. 4. A comparison between S-DG, T-DG, and ST-DG highlights the pivotal role of the spatial and temporal transformer module. Worse results were obtained by ST, which employed transformer without domain generalization, on nearly all subjects, implying the effectiveness of domain generalization in modeling subject-invariant features. This outcome provides empirical evidence supporting the importance of modeling spatial-temporal and subject-invariant features by implementing the ST-DG model for MI classification.

IV. CONCLUSION

This paper presents an innovative technique, namely ST-DG, for MI classification. ST-DG not only considers the spatial and temporal dynamics of EEG signals but also accounts for subject-specific variations. Specifically, a spatial-temporal

transformer is devised to effectively capture the most pertinent spatial-temporal features for MI classification. Furthermore, by integrating domain generalization and the spatial-temporal transformer within a unified framework, ST-DG successfully extracts subject-invariant features. Through thorough experimental evaluations conducted on two publicly available datasets, the state-of-the-art performance of ST-DG is demonstrated. Implementing a generalizable framework for analyzing multivariate physiological time series, our proposed approach holds promise for various applications.

ACKNOWLEDGMENTS

This project is funded by China Postdoctoral Science Foundation through Grant No. 2023M733738.

REFERENCES

[1] Z. Li, J. Wang, Z. Jia, and Y. Lin, "Learning space-time-frequency representation with two-stream attention based 3d network for motor imagery classification," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 1124–1129.

[2] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Bencherif, and M. S. Hossain, "Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification," *IEEE access*, 2019.

[3] C. Maswanganyi, C. Tu, P. Owolawi, and S. Du, "Discrimination of motor imagery task using wavelet based EEG signal features," in *International Conference on Intelligent and Innovative Computing Applications*, 2019.

[4] Z. Jia, J. Ji, X. Zhou, and Y. Zhou, "Hybrid spiking neural network for sleep electroencephalogram signals," *Science China Information Sciences*, vol. 65, no. 4, p. 140403, 2022.

[5] C. Liu, H. Wang, and Z. Lu, "EEG classification for multiclass motor imagery BCI," in *2013 25th Chinese Control and Decision Conference (CCDC)*, 2013.

[6] H. J. Zhang and X. J. Wang, "Research on the classification and recognition of multi-channel EEG signal based on the RBF kernel support vector machine classification," *Mechanical Electrical Engineering Technology*, 2008.

[7] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain-computer interface using convolutional neural networks," *IEEE transaction on neural networks and learning systems*, vol. 29, no. 11, pp. 5619–5629, 2018.

[8] D. Huang, S. Chen, C. Liu, L. Zheng, Z. Tian, and D. Jiang, "Differences first in asymmetric brain: A bi-hemisphere discrepancy convolutional neural network for eeg emotion recognition," *Neurocomputing (Amsterdam)*, vol. 448, pp. 140–151, 2021.

[9] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "Salientsleepnet: Multimodal salient wave detection network for sleep staging," *arXiv preprint arXiv:2105.13864*, 2021.

[10] Y. Liu and Z. Jia, "Bstt: A bayesian spatial-temporal transformer for sleep staging," in *The Eleventh International Conference on Learning Representations*, 2023.

[11] J. Chen, L. Wang, X. Jia, and P. Zhang, "EEG-based emotion recognition using deep convolutional neural network," *Computer Engineering and Applications*, 2019.

[12] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional network for eeg-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, pp. 056 013.1–056 013.17, 2018.

[13] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain-computer interface using convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2018.

[14] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun, "A multi-branch 3d convolutional neural network for EEG-based motor imagery classification," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 27, no. 10, pp. 2164–2177, 2019.

[15] C. Wei, L.-l. Chen, Z.-z. Song, X.-g. Lou, and D.-d. Li, "EEG-based emotion recognition using simple recurrent units network and ensemble learning," *Biomedical signal processing and control*, vol. 58, p. 101756, 2020.

[16] L. Xiang, D. Song, Z. Peng, G. Yu, and B. Hu, "Emotion recognition from multi-channel EEG data through convolutional recurrent neural network," in *IEEE International Conference on Bioinformatics Biomedicine*, 2017.

[17] W. Ping, A. Jiang, X. Liu, S. Jing, and Z. Li, "LSTM-based EEG classification in motor imagery tasks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. PP, no. 11, pp. 1–1, 2018.

[18] X. Ma, Q. Shuang, C. Du, J. Xing, and H. He, "Improving EEG-based motor imagery classification via spatial and temporal recurrent neural networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018.

[19] R. Zhang, Q. Zong, L. Dou, X. Zhao, and Z. Li, "Hybrid deep neural network using transfer learning for EEG motor imagery decoding," *Biomedical Signal Processing and Control*, vol. 63, p. 102144, 2021.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," vol. 30. Neural Information Processing Systems (Nips), 2017.

[21] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, and R. Faulkner, "Relational inductive biases, deep learning, and graph networks," 2018.

[22] X. Zheng and W. Chen, "An attention-based Bi-LSTM method for visual object classification via EEG," *Biomedical Signal Processing and Control*, vol. 63, p. 102174, 2021.

[23] W. Tao, C. Li, R. Song, J. Cheng, and X. Chen, "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, 2020.

[24] D. Zhang, L. Yao, K. Chen, and J. Monaghan, "A convolutional recurrent attention model for subject-independent EEG signal analysis," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 715–719, 2019.

[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[26] Y. Yang, S. Chevallier, J. Wiart, and I. Bloch, "Subject-specific time-frequency selection for multi-class motor imagery-based BCIs using few laplacian EEG channels," *Biomedical signal processing and control*, vol. 38, pp. 302–311, 2017.

[27] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE transactions on biomedical engineering*, vol. 58, no. 2, pp. 355–362, 2011.

[28] E. Jeon, W. Ko, and H. I. Suk, "Domain adaptation with source selection for motor-imagery based BCI," in *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*, 2019.

[29] Y. Cui, Y. Xu, and D. Wu, "EEG-based driver drowsiness estimation using feature weighted episodic training," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 27, no. 11, pp. 2263–2273, 2019.

[30] Z. Jia, Y. Lin, J. Wang, X. Ning, Y. He, R. Zhou, Y. Zhou, and H. L. Liwei, "Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1977–1986, 2021.

[31] Z. Jia, X. Cai, and Z. Jiao, "Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging," *IEEE Sensors Journal*, vol. 22, no. 4, pp. 3464–3471, 2022.

[32] R. T. Schirrmeister, J. T. Springenberg, L. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," 2017.

[33] Z. Jia, Y. Lin, J. Wang, K. Yang, T. Liu, and X. Zhang, "MMCNN: A multi-branch multi-scale convolutional neural network for motor imagery classification," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*. Springer, 2021, pp. 736–751.