Enhancing Eye-Tracking Performance through Multi-Task Learning Transformer

Weigeng Li, Neng Zhou, and Xiaodong Qu

The George Washington University, Washington, DC 20052, US
{weigengli, nengzhou, x.qu}@gwu.edu

Abstract. In this study, we introduce an innovative EEG signal reconstruction sub-module designed to enhance the performance of deep learning models on EEG eye-tracking tasks. This sub-module can integrate with all Encoder-Classifier-based deep learning models and achieve end-to-end training within a multi-task learning framework. Additionally, as the module operates under unsupervised learning, it is versatile and applicable to various tasks. We demonstrate its effectiveness by incorporating it into advanced deep-learning models, including Transformers and pre-trained Transformers. Our results indicate a significant enhancement in feature representation capabilities, evidenced by a Root Mean Squared Error (RMSE) of 54.1mm. This represents a notable improvement over existing methods, showcasing the sub-module's potential in refining EEG-based model performance.

The success of this approach suggests that this reconstruction sub-module is capable of enhancing the feature extraction ability of the encoder. Due to the sub-module being mounted as a sub-task under the main task and maintained through a multi-task learning framework, our model preserves the end-to-end training process of the original model. In contrast to pre-training methods like autoencoder, our model saves computational costs associated with pre-training and exhibits greater flexibility in adapting to various model structures. Benefiting from the unsupervised nature of the sub-module, it can be applied across diverse tasks. We believe it represents a novel paradigm for improving the performance of deep learning models in EEG-related challenges.

Keywords: EEG Eye-Tracking · Hybrid Vision Transformers · Multi-Task Learning · Signal Reconstruction · Unsupervised Learning · Spatio-Temporal Data Processing · Feature Extraction · Neuroscience

1 Introduction

Electroencephalography (EEG) stands as a crucial neuroimaging tool for comprehending the complex workings of brain activity and neural interactions [52]. EEG captures the electrical signals produced by neurons, providing a distinctive view of the brain's dynamic processes with exceptional temporal precision. A wide range of machine learning and deep learning techniques have been applied

to EEG data, facilitating advanced understanding and applications across multiple domains [14, 47, 1, 18, 20, 25, 29, 26, 37, 41, 56, 17, 60, 54, 39, 43, 44, 42, 45, 40].

EEG has been employed in various tasks, reflecting its versatility. Researchers have harnessed EEG data for purposes such as brain-computer interfaces (BCIs) [4], sleep analysis [36], and more recently, eye movement prediction [24]. These tasks have revealed different aspects of brain function and have contributed to our understanding of the neural mechanisms underlying cognition, behavior, and sensory processing.

The challenges of EEG-based tasks are multiple, including issues related to data quality, computational complexity, and model generalization [46]. EEG signals are vulnerable to noise and artifacts, which can affect the reliability of results. In addition, the high dimension of EEG data poses computational challenges, requiring sophisticated preprocessing and feature extraction techniques.

In recent years, convolutional neural networks (CNNs) have become a powerful tool in the field of EEG research [14]. Originally developed for image analysis, convolutional neural networks have now been applied to process and interpret EEG data. These neural networks can automatically learn complex spatial and temporal patterns in EEG signals, providing a new dimension in the analysis of brain activity.

Multi-task Learning (MTL) [8], in contrast to single-task learning (STL), involves simultaneous consideration of multiple related tasks, leveraging shared information to address complex challenges. This approach capitalizes on task connections to extract complementary information, enhancing decoding model accuracy and reliability. Previous research has highlighted the advantages of multi-task EEG analysis, revealing its applications in emotion recognition [28] [13], classification [6] [48], and disease prediction [34].

1.1 Research Questions

Decoding EEG signals typically involves a series of steps, including preprocessing, feature extraction, and classification. Achieving successful EEG decoding in open-world scenarios necessitates careful consideration at each stage. Even when recorded under the most stringent conditions [10], EEG signals are susceptible to various artifacts such as eye blinks, muscle interference, cardiac disturbances, and electromagnetic interference.

In this context, MTL emerges as a valuable strategy for improving the feature-extracting ability of EEG decoding. By harnessing the power of multiple related tasks, MTL enhances the generalization capabilities of EEG models and mitigates the risk of overfitting, thereby contributing to more effective EEG signal analysis in open-world environments. This approach leverages the inherent connections between tasks and allows for the extraction of complementary information, ultimately enhancing the accuracy and reliability of EEG decoding models.

Our research aims to address the following questions at the intersection of Machine Learning and EEG eye-tracking:

1. Can we use EEG Signal Reconstruction as a sub-task to enhance the Transformer encoder's feature-extracting ability?
2. Which aspects of the prediction results, like specific regional accuracy or the overall prediction pattern, improved after integrating our framework?

2 Related work

2.1 Deep Learning for EEG Tasks

Early studies highlighted the potential of Convolutional Neural Networks (CNNs) in EEG analysis. For instance, a CNN-based approach [35] is introduced for epileptic seizure classification on EEG data, utilizing the continuous wavelet transform (CWT) to convert EEG data into time-frequency domain images. Similarly, Transformer-based models [49] have shown their superiority over CNNs, RNNs, and DBNs in EEG classification, indicating the promise of the hybrid Transformer-CNN approach.

2.2 MTL for EEG Tasks

Multi-task Learning (MTL) [8] has been leveraged in various EEG signal analysis applications, including emotion recognition [28] [13], classification [6] [48], and disease prediction [34]. DMTL-BCI [48] employed an MTL framework to jointly optimize three modules (representation, classification, and reconstruction), outperforming state-of-the-art methods by 3.0% on BCI Competition IV dataset 2a. MIN2Net [6] utilized deep metric learning and autoencoder for subject-independent motor imagery EEG signal classification, outperforming state-of-the-art techniques by 6.72% and 2.23% on the SMR-BCI and OpenBMI datasets. Choo et al. [13] investigated the effectiveness of MTL in raw EEG-based emotion recognition, demonstrating significant classification accuracy improvements with their MTL-ShallowConvNet architecture. Furthermore, EEG-DEMTL [11] is a computation-based MTL network for assessing railway passenger comfort through EEG signals, improving the evaluation performance by 6.3% in field experiments.

2.3 Vision Transformers (ViTs)

The transformer model [53] is a deep learning model based on the self-attention mechanism, primarily used for processing sequential data. The core idea of the Transformer model is that through the attention mechanism, the model can focus on any part of the input sequence, thereby more effectively capturing long-distance dependencies within the sequence. This mechanism has led to tremendous success for the Transformer in the field of natural language processing (NLP), especially in tasks such as machine translation, text generation, and comprehension.

Building on the success of the Transformer model, Dosovitskiy and others proposed the Vision Transformer (ViT) in 2020 [16] Vision Transformer. ViT

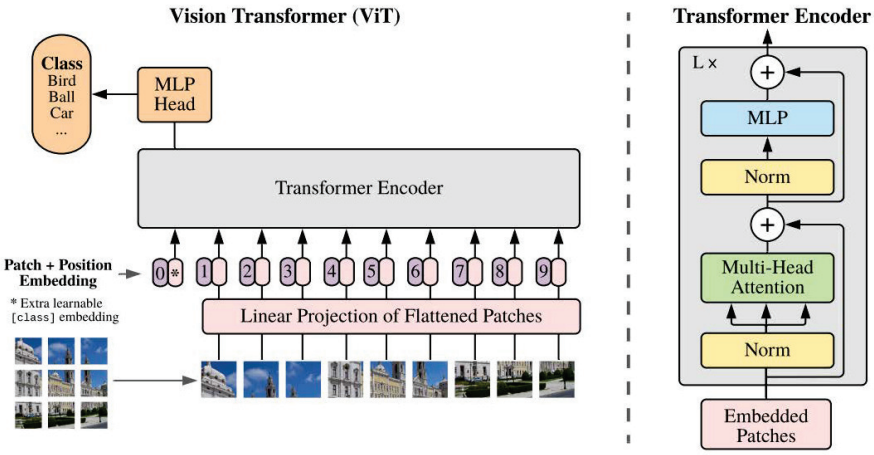

Fig. 1. Vision Transformer Encoder proposed by [16]

applies the concept of the Transformer to the field of computer vision, dividing images into a series of small patches and feeding these patches as a sequence into a self-attention-based Transformer network. This approach allows ViT to process image data effectively, capturing complex patterns and relationships within images, thus achieving excellent performance in image classification and other visual tasks. Subsequently, ViT has also shown great potential in other areas, such as EEG data analysis, demonstrating its effectiveness in processing non-traditional visual data.

Several studies have demonstrated their effectiveness regarding the application of Vision Transformers (ViT) in EEG tasks. Yang and Modesitt demonstrated the application of a hybrid ViT model, pre-trained on ImageNet, in an EEG regression task. Additionally, a bi-branch Vision Transformer-based EEG emotion recognition model, Bi-ViTNet, integrating spatial-temporal and spatial-frequency feature representations, has shown ViT's potential in handling complex EEG data[30]. EEG-ConvTransformer demonstrated improved classification accuracy over state-of-the-art techniques in five different visual stimuli classification tasks. This further proves the effectiveness of ViT models in EEG signal processing.[7] Finally, the importance of the attention mechanism in EEG signals was introduced through two ViT-based methods for the classification of EEG signals based on emotions.[5]

These studies indicate that ViT models can effectively process EEG data, especially in complex tasks such as emotion recognition and visual stimuli classification. These findings support the use of ViT as a base model for multi-task learning (MTL) in EEG tasks.

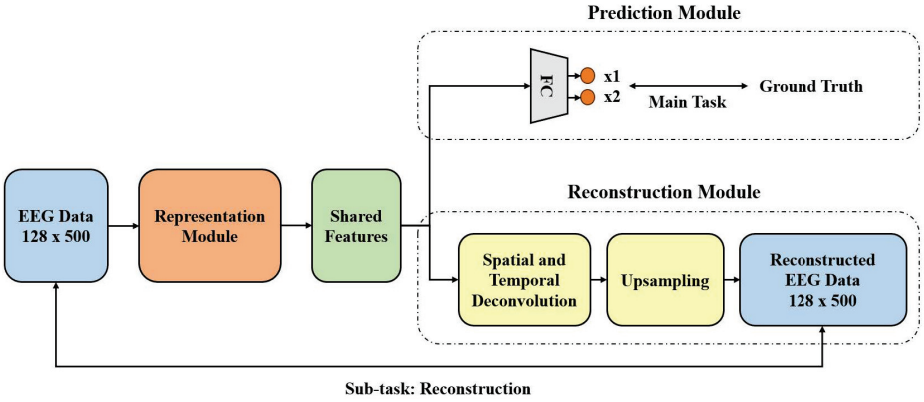

Fig. 2. Proposed MTL-Transformer Architecture: Eye Tracking and Data Reconstruction

3 Methods

In this paper, we plan to combine multi-task learning and Vision Transformer [15] to enhance the performance of the EEGEyeNet dataset's eye-tracking task [24]. By simultaneously addressing multiple related tasks within the dataset, we aim to improve the model's performance on the eye-tracking task. Our approach holds the potential to uncover novel connections and enhance the overall understanding of eye-tracking patterns in the context of EEG signals.

3.1 Model Architecture

Our model architecture is specifically designed to enhance performance in EEG eye-tracking tasks. The cornerstone of our approach is the introduction of a multi-task framework, which handles various sub-tasks simultaneously. This design choice is motivated by the need to capture the diverse aspects of EEG data more effectively.

Drawing inspiration from the work of Song et al. [48], our Multi-task Learning Transformer uniquely combines classification and reconstruction tasks within its architecture. By doing so, it efficiently leverages the representation module to maintain dual capabilities in feature extraction. This multi-task learning approach significantly boosts the model's ability to generalize across different EEG data scenarios.

The processing flow of our model, particularly highlighting the interaction between its different components within the multi-task framework, is depicted in Figure 2. This illustration provides a clear visual representation of how the model integrates and processes various sub-tasks, contributing to its enhanced performance.

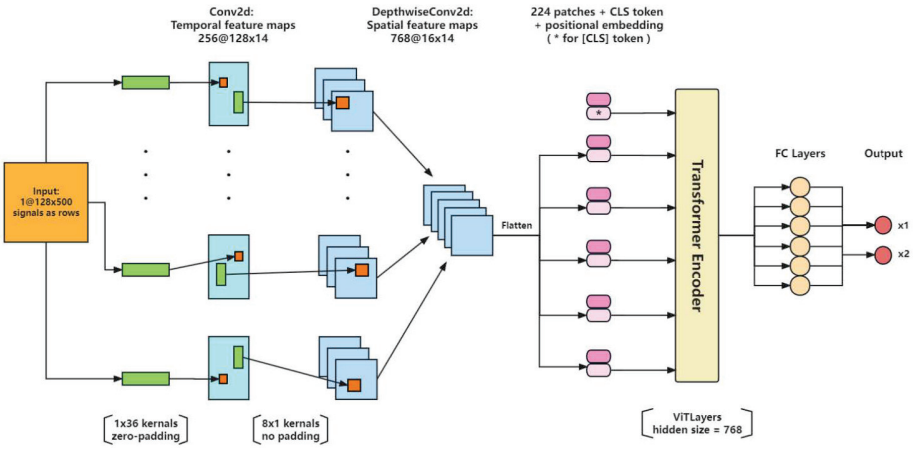

Fig. 3. Model architecture of EEG2VIT [55]. In this paper, we use the Convolution Layer and Transformer Encoder as the Representation Module

3.2 Representation Module

The ViT2EEG model, proposed by Yang and Modesitt [55], utilizes a hybrid Vision Transformer architecture pre-trained on ImageNet for EEG data regression tasks. It outperforms other models, including a non-trained ViT, demonstrating that models pre-trained on image data can be effectively fine-tuned for EEG tasks. In our model architecture, we employ the Convolution Layer and pre-trained Vision Transformer (ViT) encoder, the same as the ViT2EEG model in figure 3. This setup has been shown to effectively capture complex patterns in data, which is essential for the reconstruction sub-module.

The input EEG data, with dimensions $1 \times 128 \times 500$, undergo a convolutional process yielding temporal feature maps of $256 \times 128 \times 14$. This is followed by depthwise convolutional layers, which further refine the spatial characteristics into feature maps sized $768 \times 16 \times 14$. The Conv2d layers utilize 1×36 kernels with zero padding, while the DepthwiseConv2d layers apply 8×1 kernels without padding, ensuring an effective spatial-temporal feature representation.

These features are then transformed into a sequence of 224 flattened patches, each integrated with a unique positional embedding. An additional [CLS] token embedding is also included, a common practice in ViT architectures to facilitate classification. The resulting embeddings are processed through a Transformer encoder, equipped with a hidden size of 768 to capture complex dependencies within the data.

The architecture concludes with fully connected layers, outputting two distinct values, which are the final inference results of the model. This innovative design leverages the strengths of both convolutional operations and transformer-based modeling to handle the intricacies of EEG signal analysis effectively.

3.3 Prediction Module

The prediction module in our architecture is designed as a sequence of interconnected layers. It comprises a fully connected layer, followed by a dropout layer for regularization, and concludes with another fully connected layer. The output of this module is articulated as follows:

$$\hat{y} = FC(dropout(FC(H^{(l)})) \tag{1}$$

In this formulation, $H^{(l)}$ represents the output of the last layer in the encoder. The notation $H^{(l)}$ signifies the hidden representation obtained after the input data has undergone a series of transformations through the layers of the encoder neural network. Each layer in the encoder, denoted by l, contributes to shaping this representation, and $H^{(l)}$ captures the information learned up to that point. FC denotes the fully connected layers. The *dropout* function represents the dropout layer, a crucial component for preventing overfitting by randomly dropping units from the neural network during training.

For the main task of our model, the Mean Squared Error (MSE) loss is employed. This loss function is defined as:

$$Loss_{MSE}(\hat{y}, y_1) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_{1i})^2 \tag{2}$$

Here, $\hat{y}$ is the predicted output of the network, and y_1 is the actual label for the main task. The MSE loss function computes the average of the squares of the differences between the predicted and actual values, providing a measure of the model's accuracy.

This structure ensures a streamlined flow of data through the layers, facilitating effective feature extraction and subsequent prediction.

3.4 Reconstruction Module

The reconstruction module plays a pivotal role in our system, consisting of a series of spatial and temporal deconvolution blocks designed to incrementally expand the dimensionality of shared features to reconstruct the input data effectively.

The spatial deconvolution block is crucial for spatial feature reconstruction and is defined by the following equation:

$$H_{decoder\_spatial} = Deconv\_spatial(H^{(l)}) \tag{3}$$

In this block, *Deconv_spatial* is composed of a three-layer structure: starting with a 1D Deconvolution layer with a kernel size of 1×36. This specific configuration mirrors the first convolution layer in the encoder, ensuring symmetry in feature extraction and reconstruction. It is followed by an InstanceNorm layer, enhancing the normalization of features, and a ReLU activation layer, introducing non-linearity for better feature representation.

Similarly, the temporal deconvolution block, essential for time-series data reconstruction, is formulated as:

$$H_{decoder\_temporal} = Deconv\_temporal(H_{decoder\_spatial}) \tag{4}$$

The *Deconv_temporal* block also includes three layers. It begins with a 1D Deconvolution layer, this time with a kernel size of 8×1. This dimensionality aligns with the patch size used in the Vision Transformer (ViT) encoder block, allowing for a consistent approach to handling spatial-temporal data. This layer is followed by an InstanceNorm layer and a ReLU activation layer, similar to the spatial deconvolution block.

The final step in the reconstruction process is the upsampling block, defined as:

$$\hat{x} = Upsampling(H_{decoder\_temporal}) \tag{5}$$

This block efficiently transforms the decoder output to match the original input size, ensuring the reconstructed data $\hat{x}$ is comparable to the original input x.

Lastly, we define our loss function using Mean Squared Error (MSE) to quantify the reconstruction accuracy:

$$loss\_MSE(\hat{x}, x) = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}^{(i)} - x^{(i)})^2 \tag{6}$$

Where $\hat{x}$ is the reconstructed input, x is the original input, and N represents the total number of elements in x. MSE is chosen for its effectiveness in emphasizing larger errors and its suitability in scenarios where maintaining the fidelity of the reconstructed data is crucial.

3.5 Multi-Task Learning Framework

In our multi-task learning framework, we aim to enhance the training of the primary eye-tracking task by integrating the losses from sub-tasks. The overall loss L of the framework is computed using the following equation:

$$L(\theta) = Loss_{MSE}(x, y_1) + \sum \alpha_i Loss(x, y_i) + \lambda||\theta||^2 \tag{7}$$

Where x is the input EEG signal, represented as a 2D matrix. y_1 is the label of the eye-tracking task, which is also the main task. $Loss_{MSE}(x, y_1)$ denotes the MSE loss for the eye-tracking task, and $Loss(x, y_i)$ is the loss of other sub-task. Hyper-parameter α_i is utilized to balance the relative importance of the supervised and unsupervised loss. We apply l_2 regularization term with coefficient to alleviate overfitting. Our task is to minimize (θ). All trainable parameters of the network are trained in an end-to-end manner.

To evaluate different strategies within our proposed multi-task learning framework, we developed two distinct model architectures, each focusing on a separate

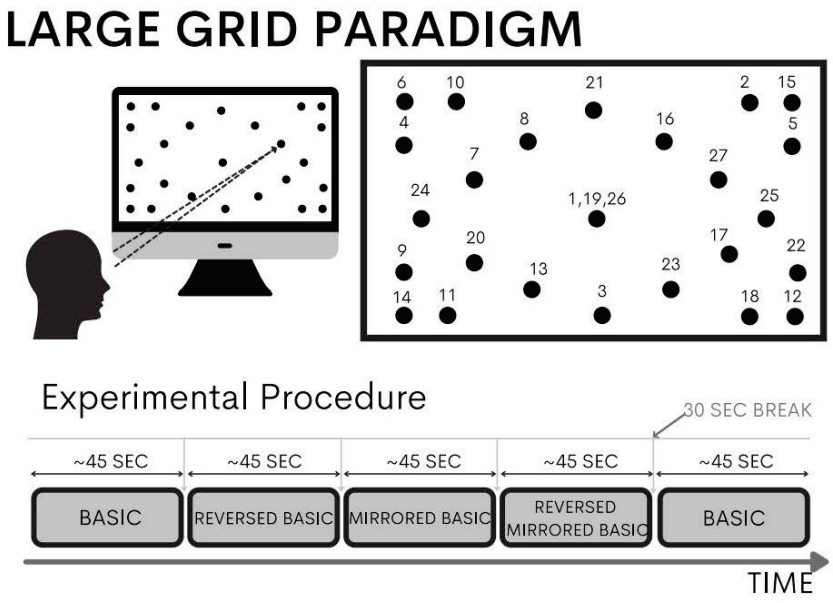

Fig. 4. EEGEyeNet Large Grid Paradigm [24]. Participants are asked to fixate on particular dots in a given period

sub-task. The first model, named MTL-Transformer, employs a reconstruction sub-task, as introduced earlier in the paper. This model aims to reconstruct the original EEG data. The second model, MTL-Transformer2, diverges by replacing the reconstruction module with a pupil size prediction module. This auxiliary subtask was introduced to explore the relevance of pupil size to the eye-tracking task. To accommodate this, we reorganized our dataset to include pupil size for each sample. Both models were measured using the Root Mean Square Error (RMSE) metric to ensure a consistent and objective evaluation of their performance.

4 Experiments

4.1 Dataset

The EEGEyeNet dataset [24] offers a comprehensive resource for EEG research, featuring 47 hours of high-density 128-channel EEG data, which provides detailed neural activity recordings synchronized with eye-tracking data from 356 adults. This dataset is particularly suited for our study, which aims to leverage the rich EEG data to predict behavioral responses in eye-tracking tasks. Our focus on the eye-tracking task stems from its potential to reveal how neural

Paradigm	# Fixations	# Saccades	# Blinks
Pro-Antisac.	357115	358384	56179
Large Grid	68075	68245	11108
VSS	43384	43443	971
Total	468574	470072	68258

Table 1. Eyes event label distribution in EEGEyeNet dataset (minimal preprocessing) [24]

patterns correlate with visual attention and eye movement behaviors. Detailed information about the dataset, including the specificities of the eye-tracking tasks and participant demographics, is elaborately presented in Table 1 and Figure 4. For our experiment, we propose a dataset split of 70% for the training set, 15% for the validation set, and 15% for the test set. This distribution is designed to maximize learning from the EEG data while ensuring robust validation and testing of our predictive models.

4.2 Baseline Models

Machine Learning We employed a range of traditional machine-learning algorithms as baseline models. These include K-Nearest Neighbors (KNN), Support Vector Machines with Radial Basis Function (RBF SVC/SVR), Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Random Forest, Gradient Boosting, AdaBoost, and XGBoost. While these methods provide solid benchmarks, they have limitations in handling the high dimensionality and complex temporal dynamics inherent in EEG data.

Deep Learning Convolutional Neural Network (CNN) CNNs were effective in capturing spatial patterns in EEG data but less adept at modeling temporal dynamics.

PyramidalCNN The PyramidalCNN, with its unique structure, offered improved performance in capturing hierarchical features, leading to better generalization [23].

EEGNet EEGNet, designed for EEG data analysis, showed proficiency in handling both spatial and temporal features but may struggle with very large datasets [27].

InceptionTime InceptionTime's modular architecture allowed for a robust capture of temporal dynamics, surpassing traditional CNNs [21].

Xception Xception's depthwise separable convolutions were efficient, though they may not fully exploit the multi-channel nature of EEG data [12].

EEGViT The EEGViT, adapting the Vision Transformer for EEG data, presented an innovative approach in modeling long-range dependencies, a common challenge in EEG analysis [55].

Model	AbsolutePosition RMSE (mm)
Naive Baseline	123.3 ± 0
KNN	119.7 ± 0
RBF SVC/SVR	123 ± 0
Linear Regression	118.3 ± 0
Ridge Regression	118.2 ± 0
Lasso Regression	118 ± 0
Elastic Net	118.1 ± 0
Random Forest	116.7 ± 0.1
Gradient Boost	117 ± 0.1
AdaBoost	119.4 ± 0.1
XGBoost	118 ± 0
CNN	70.2 ± 1.1
PyramidalCNN	73.6 ± 1.9
EEGNet	81.7 ± 1.0
InceptionTime	70.8 ± 0.8
Xception	78.7 ± 1.6
ViT-Base	61.5 ± 0.6
ViT-Base Pre-trained	58.1 ± 0.6
EEGViT	61.7 ± 0.6
EEGViT Pre-trained	55.4 ± 0.2
MTL-Transformer(Ours)	**54.1 ± 0.2**
MTL-Transformer2(Ours)	**57.4 ± 0.3**

Table 2. Root Mean Squared Error (RMSE) Comparison of Baseline Models on the EEGEyeNet eye-tracking task [24]. The primary model, MTL-Transformer, demonstrates significant performance improvement, utilizing EEGViT Pre-trained as its base model. Additionally, MTL-Transformer2, which includes pupil size prediction as an auxiliary sub-task, is presented to demonstrate the scope of our experimental exploration despite its lesser impact on RMSE reduction.

4.3 Implementation Details

All models in our study were trained for 15 epochs on an RTX 4090 GPU. For deep learning models, we set an initial learning rate of 10^{-4} and implemented a decay strategy, reducing the learning rate by 10% every 6 epochs. This approach is designed to balance the rate of convergence and ensure effective learning over the training period.

In our proposed model, we integrated two dropout layers with a dropout rate of 0.3, specifically in the prediction module. This rate is higher than typical settings, chosen to mitigate overfitting while dealing with the complex nature of EEG data. This dropout strategy is particularly crucial given the model's architecture and the high-dimensional feature space of the EEG signals.

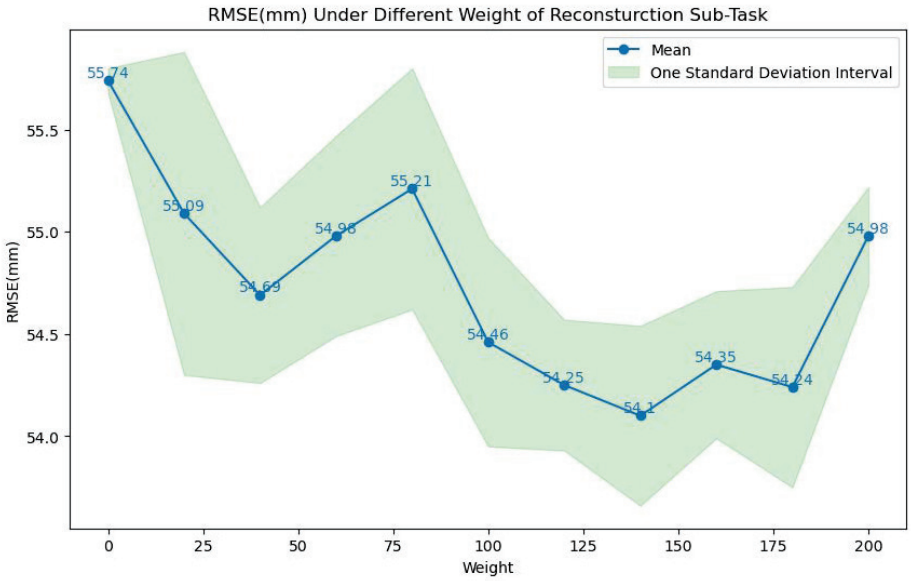

Fig. 5. RMSE(mm) Under Different Weight of Reconstruction Sub-Task

5 Results

The performance of various models on the EEG eye-tracking task is summarized in Table 2. Notably, our proposed model achieved a Root Mean Square Error (RMSE) of 54.1mm, which slightly surpasses the current State-Of-The-Art (SOTA) model's RMSE of 55.4mm. This improvement, although marginal, indicates the effectiveness of our model's architecture and the methodologies employed, especially in handling the complexities of EEG data in eye-tracking tasks.

Figure 5 illustrates the RMSE of our model under varying weights assigned to the reconstruction sub-task. At a weight of 0, the reconstruction sub-module does not participate in the gradient computation, and our model's results align with the EEGViT Pre-trained model, which is the base model. This parallel performance indicates that the enhancements in accuracy are not attributable to alterations in the model's structure. As the weight increases to 140, there is a discernible improvement in model accuracy, suggesting that the reconstruction sub-module contributes positively to the base model's performance. However, beyond a weight of 140, the excessive emphasis on the reconstruction sub-task seems to detract from the sub-task at hand, as evidenced by a decline in accuracy. This trend demonstrates a critical balance between the reconstruction weight and the model's focus on the primary task, underscoring the need for optimal weight tuning to harness the reconstruction sub-module's benefits without compromising the main objective.

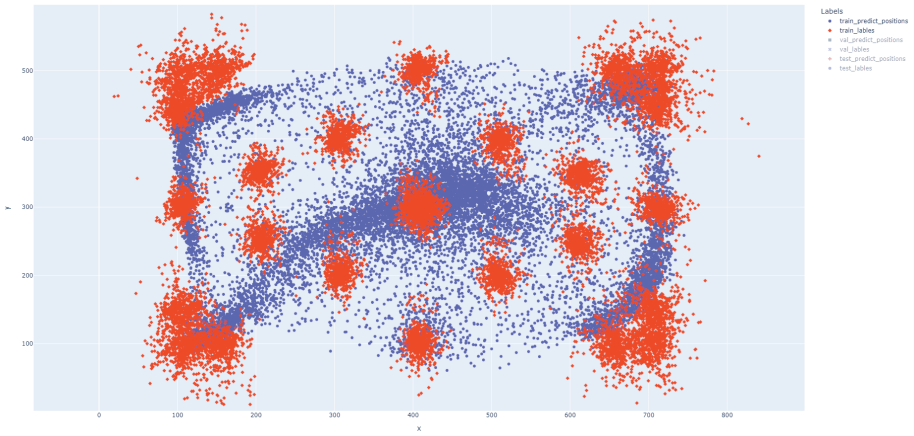

Fig. 6. Predict Gazing Position and Real Gazing Position on Training Dataset

6 Discussion

Our research presents promising implications for the field of EEG-based eye tracking. The slight yet significant improvement in accuracy provided by our model paves the way for more precise and reliable EEG eye-tracking systems. This advancement is particularly relevant in applications where minute differences in eye movement can have substantial implications, such as in neuromarketing or neurological disorder diagnosis.

Figure 6 shows the eye-tracking prediction on the training dataset. Our model demonstrates a commendable capacity to discriminate between central and peripheral points. However, it exhibits limitations in accurately distinguishing points within the intermediate regions, with a predominant aggregation of data at the center. This phenomenon may be attributed to a frequency bias towards central points, resulting in an imbalance within the dataset. To address this, future iterations of our model could incorporate weighted learning for different regions, facilitating a more balanced and nuanced understanding and thereby enhancing the model's predictive accuracy.

Looking ahead, we aim to broaden the applicability of our model by testing it on various other EEG datasets. Such an expansion will not only validate the model's effectiveness across different data types but also enhance its robustness and generalizability. This step is crucial for asserting the model's utility in diverse real-world scenarios.

Additionally, the versatility of our Multi-task Learning module is a notable aspect of our architecture. Its design allows it to be integrated as a separate module into any EEG-based task. This modular approach offers a flexible solution for improving existing EEG analysis systems, potentially transforming how EEG data is processed and interpreted in various applications.

Moreover, we attempted to leverage pre-trained language Transformer models, such as GPT and BERT, which are typically used for time-series or language tasks. However, these models generally demand substantial GPU memory capacity, which exceeds the capabilities of our personal workstations. This limitation constrained the scope of our experiments. Future work will, therefore, focus on optimizing computational efficiency, perhaps through model distillation or pruning techniques that can reduce the memory footprint of these large models. Future studies can also investigate other potential deep learning approaches on various datasets for comparative analysis [2, 3, 22, 19, 31, 9, 32, 33, 51, 50, 38, 59, 58, 57].

In conclusion, while our current focus has been on EEG eye-tracking tasks, the broader impact of our work lies in its potential to revolutionize various aspects of EEG data analysis and application. Future research will delve deeper into these possibilities, continually pushing the boundaries of what is achievable in this domain.

7 Conclusion

In this study, we have demonstrated the effectiveness of integrating multi-task learning with Vision Transformers in the domain of EEG eye-tracking. Our approach has successfully employed an innovative EEG signal reconstruction submodule, enhancing the feature extraction capabilities of deep learning models applied to this task. This sub-module, adaptable to various Encoder-Classifier-based models, facilitates end-to-end training within a multi-task learning framework and operates effectively under unsupervised learning conditions.

Our experimental results, particularly the achieved RMSE of 54.1mm, which surpasses the previous state-of-the-art model, underscore the potential of our method in improving EEG-based eye-tracking systems. This advancement is not only significant in terms of model performance but also in its application potential across various EEG datasets and tasks.

Looking forward, the adaptability and versatility of our Multi-task Learning module open new avenues for enhancing EEG data processing and interpretation. This work lays the groundwork for future research in this area, aiming to further explore and expand the capabilities of deep learning models in the realm of neuroscience and cognitive research. We believe that our approach represents a novel paradigm in EEG data analysis, with the potential to contribute significantly to various EEG-related challenges and applications.

References

1. Altaheri, H., Muhammad, G., Alsulaiman, M., Amin, S.U., Altuwaijri, G.A., Abdul, W., Bencherif, M.A., Faisal, M.: Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. Neural Computing and Applications **35**(20), 14681–14722 (2023)

2. An, S., Bhat, G., Gumussoy, S., Ogras, U.: Transfer learning for human activity recognition using representational analysis of neural networks. ACM Transactions on Computing for Healthcare **4**(1), 1–21 (2023)
3. An, S., Tuncel, Y., Basaklar, T., Ogras, U.Y.: A survey of embedded machine learning for smart and sustainable healthcare applications. In: Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Use Cases and Emerging Challenges, pp. 127–150. Springer (2023)
4. Ang, K.K., Guan, C.: Brain-computer interface in stroke rehabilitation. Journal of Computing Science and Engineering **7**(2), 139–146 (2013)
5. Arjun, A., Rajpoot, A.S., Raveendranatha Panicker, M.: Introducing attention mechanism for eeg signals: Emotion recognition with vision transformers. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 5723–5726 (2021). https://doi.org/10.1109/EMBC46164.2021.9629837
6. Autthasan, P., Chaisaen, R., Sudhawiyangkul, T., Rangpong, P., Kiatthaveephong, S., Dilokthanakul, N., Bhakdisongkhram, G., Phan, H., Guan, C., Wilaiprasitporn, T.: Min2net: End-to-end multi-task learning for subject-independent motor imagery eeg classification. IEEE Transactions on Biomedical Engineering **69**(6), 2105–2118 (2021)
7. Bagchi, S., Bathula, D.R.: Eeg-convtransformer for single-trial eeg based visual stimuli classification (2021)
8. Caruana, R.: Multitask learning. Machine learning **28**, 41–75 (1997)
9. Chen, J., Hu, Y., Wang, Y., Lu, Y., Cao, X., Lin, M., Xu, H., Wu, J., Xiao, C., Sun, J., et al.: Trialbench: Multi-modal artificial intelligence-ready clinical trial datasets. arXiv preprint arXiv:2407.00631 (2024)
10. Chen, X., Li, C., Liu, A., McKeown, M.J., Qian, R., Wang, Z.J.: Toward open-world electroencephalogram decoding via deep learning: A comprehensive survey. IEEE Signal Processing Magazine **39**(2), 117–134 (2022)
11. Cheng, B., Fu, H., Li, T., Zhang, H., Huang, J., Peng, Y., Chen, H., Fan, C.: Evolutionary computation-based multitask learning network for railway passenger comfort evaluation from eeg signals. Applied Soft Computing **136**, 110079 (2023)
12. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
13. Choo, S., Park, H., Kim, S., Park, D., Jung, J.Y., Lee, S., Nam, C.S.: Effectiveness of multi-task deep learning framework for eeg-based emotion and context recognition. Expert Systems with Applications **227**, 120348 (2023)
14. Craik, A., He, Y., Contreras-Vidal, J.L.: Deep learning for electroencephalogram (eeg) classification tasks: a review. Journal of neural engineering **16**(3), 031001 (2019)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T.: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Dou, G., Zhou, Z., Qu, X.: Time majority voting, a pc-based eeg classifier for non-expert users. In: International Conference on Human-Computer Interaction. pp. 415–428. Springer (2022)

18. Gao, Z., Dang, W., Wang, X., Hong, X., Hou, L., Ma, K., Perc, M.: Complex networks and deep learning for eeg signal analysis. Cognitive Neurodynamics **15**(3), 369–388 (2021)
19. Gui, S., Song, S., Qin, R., Tang, Y.: Remote sensing object detection in the deep learning era—a review. Remote Sensing **16**(2), 327 (2024)
20. Hossain, K.M., Islam, M.A., Hossain, S., Nijholt, A., Ahad, M.A.R.: Status of deep learning for eeg-based brain–computer interface applications. Frontiers in computational neuroscience **16**, 1006763 (2023)
21. Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.A., Petitjean, F.: Inceptiontime: Finding alexnet for time series classification. Data Mining and Knowledge Discovery **34**(6), 1936–1962 (2020)
22. Jiang, C., Hui, B., Liu, B., Yan, D.: Successfully applying lottery ticket hypothesis to diffusion model. arXiv preprint arXiv:2310.18823 (2023)
23. Johnson, R., Zhang, T.: Deep pyramid convolutional neural networks for text categorization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 562–570 (2017)
24. Kastrati, A., Płomecka, M.B., Pascual, D., Wolf, L., Gillioz, V., Wattenhofer, R., Langer, N.: Eegeyenet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. arXiv preprint arXiv:2111.05100 (2021)
25. Key, M.L., Mehtiyev, T., Qu, X.: Advancing eeg-based gaze prediction using depthwise separable convolution and enhanced pre-processing. In: International Conference on Human-Computer Interaction. pp. 3–17. Springer (2024)
26. Koome Murungi, N., Pham, M.V., Dai, X., Qu, X.: Trends in machine learning and electroencephalogram (eeg): A review for undergraduate researchers. arXiv e-prints pp. arXiv–2307 (2023)
27. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering **15**(5), 056013 (2018)
28. Li, C., Wang, B., Zhang, S., Liu, Y., Song, R., Cheng, J., Chen, X.: Emotion recognition from eeg based on multi-task learning with capsule network and attention mechanism. Computers in Biology and Medicine **143**, 105303 (2022)
29. Li, W., Zhou, N., Qu, X.: Enhancing eye-tracking performance through multi-task learning transformer. In: International Conference on Human-Computer Interaction. pp. 31–46. Springer (2024)
30. Lu, W., Tan, T.P., Ma, H.: Bi-branch vision transformer network for eeg emotion recognition. IEEE Access **11**, 36233–36243 (2023). https://doi.org/10.1109/ACCESS.2023.3266117
31. Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Wei, W.: Machine learning for synthetic data generation: a review. arXiv preprint arXiv:2302.04062 (2023)
32. Ma, X.: Traffic performance evaluation using statistical and machine learning methods. Ph.D. thesis, The University of Arizona (2022)
33. Ma, X., Karimpour, A., Wu, Y.J.: Data-driven transfer learning framework for estimating on-ramp and off-ramp traffic flows. Journal of Intelligent Transportation Systems pp. 1–14 (2024)
34. Ma, X., Qiu, S., Zhang, Y., Lian, X., He, H.: Predicting epileptic seizures from intracranial eeg using lstm-based multi-task learning. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 157–167. Springer (2018)

35. Mao, W., Fathurrahman, H., Lee, Y., Chang, T.: Eeg dataset classification using cnn method. In: Journal of physics: conference series. vol. 1456, p. 012017. IOP Publishing (2020)
36. Motamedi-Fakhr, S., Moshrefi-Torbati, M., Hill, M., Hill, C.M., White, P.R.: Signal processing techniques applied to human sleep eeg signals—a review. Biomedical Signal Processing and Control **10**, 21–33 (2014)
37. Murungi, N.K., Pham, M.V., Dai, X.C., Qu, X.: Empowering computer science students in electroencephalography (eeg) analysis: A review of machine learning algorithms for eeg datasets. SIGKDD (2023)
38. Qiu, Y., Zhao, Z., Yao, H., Chen, D., Wang, Z.: Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3228–3239 (2023)
39. Qu, X.: Time Continuity Voting for Electroencephalography (EEG) Classification. Ph.D. thesis, Brandeis University (2022)
40. Qu, X., Hall, M., Sun, Y., Sekuler, R., Hickey, T.J.: A personalized reading coach using wearable eeg sensors. CSEDU (2019)
41. Qu, X., Hickey, T.J.: Eeg4home: A human-in-the-loop machine learning model for eeg-based bci. In: Augmented Cognition: 16th International Conference, AC 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings. pp. 162–172. Springer (2022)
42. Qu, X., Liu, P., Li, Z., Hickey, T.: Multi-class time continuity voting for eeg classification. In: Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2. pp. 24–33. Springer (2020)
43. Qu, X., Liukasemsarn, S., Tu, J., Higgins, A., Hickey, T.J., Hall, M.H.: Identifying clinically and functionally distinct groups among healthy controls and first episode psychosis patients by clustering on eeg patterns. Frontiers in psychiatry **11**, 541659 (2020)
44. Qu, X., Mei, Q., Liu, P., Hickey, T.: Using eeg to distinguish between writing and typing for the same cognitive task. In: Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2. pp. 66–74. Springer (2020)
45. Qu, X., Sun, Y., Sekuler, R., Hickey, T.: Eeg markers of stem learning. In: 2018 IEEE Frontiers in Education Conference (FIE). pp. 1–9. IEEE (2018)
46. Rashid, M., Sulaiman, N., PP Abdul Majeed, A., Musa, R.M., Ab Nasir, A.F., Bari, B.S., Khatun, S.: Current status, challenges, and possible solutions of eeg-based brain-computer interface: a comprehensive review. Frontiers in neurorobotics p. 25 (2020)
47. Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J.: Deep learning-based electroencephalography analysis: a systematic review. Journal of neural engineering **16**(5), 051001 (2019)
48. Song, Y., Wang, D., Yue, K., Zheng, N., Shen, Z.J.M.: Eeg-based motor imagery classification with deep multi-task learning. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2019)
49. Sun, J., Xie, J., Zhou, H.: Eeg classification with transformer-based models. In: 2021 ieee 3rd global conference on life sciences and technologies (lifetech). pp. 92–93. IEEE (2021)
50. Tan, J., Shen, X., Zhang, X., Wang, Y.: Multivariate encoding analysis of medial prefrontal cortex cortical activity during task learning. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 6699–6702. IEEE (2021)

51. Tan, J., Zhang, X., Wu, S., Song, Z., Chen, S., Huang, Y., Wang, Y.: Audio-induced medial prefrontal cortical dynamics enhances coadaptive learning in brain–machine interfaces. Journal of Neural Engineering **20**(5), 056035 (2023)
52. Teplan, M., et al.: Fundamentals of eeg measurement. Measurement science review **2**(2), 1–11 (2002)
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
54. Wang, R., Qu, X.: Eeg daydreaming, a machine learning approach to detect daydreaming activities. In: International Conference on Human-Computer Interaction. pp. 202–212. Springer (2022)
55. Yang, R., Modesitt, E.: Vit2eeg: Leveraging hybrid pretrained vision transformers for eeg data. arXiv preprint arXiv:2308.00454 (2023)
56. Yi, L., Qu, X.: Attention-based cnn capturing eeg recording's average voltage and local change. In: Artificial Intelligence in HCI: 3rd International Conference, AI-HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings. pp. 448–459. Springer (2022)
57. Zhang, Z., Tian, R., Ding, Z.: Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3534–3542 (2023)
58. Zhang, Z., Tian, R., Sherony, R., Domeyer, J., Ding, Z.: Attention-based interrelation modeling for explainable automated driving. IEEE Transactions on Intelligent Vehicles **8**(2), 1564–1573 (2022)
59. Zhao, S., Yang, X., Zeng, Z., Qian, P., Zhao, Z., Dai, L., Prabhu, N., Nordlund, P., Tam, W.L.: Deep learning based cetsa feature prediction cross multiple cell lines with latent space representation. Scientific Reports **14**(1), 1878 (2024)
60. Zhou, Z., Dou, G., Qu, X.: Brainactivity1: A framework of eeg data collection and machine learning analysis for college students. In: International Conference on Human-Computer Interaction. pp. 119–127. Springer (2022)