Effect of Kernel Size on CNN-Vision-Transformer-Based Gaze Prediction Using Electroencephalography Data

Chuhui Qiu(✉), Bugao Liang, and Matthew L. Key

The George Washington University, Washington DC 20052, USA
{chqiu,bliang271,matthewlkey}@gwmail.gwu.edu

Abstract. In this paper, we present an algorithm of gaze prediction from Electroencephalography (EEG) data. EEG-based gaze prediction is a new research topic that can serve as an alternative to traditional video-based eye-tracking. Compared to the existing state-of-the-art (SOTA) method, we improved the root mean-squared-error of EEG-based gaze prediction to 53.06 mm, while reducing the training time to less than 33% of its original duration. Our source code can be found at https://github.com/AmCh-Q/CSCI6907Project.

Keywords: Machine Learning · Deep Learning · Brain-Computer Interfaces · BCI · Electroencephalography · EEG · Gaze Prediction · Eye Tracking · Transformer

1 Introduction

Electroencephalography (EEG) is a non-invasive technique used to record the electrical activity generated by the brain. Owing to its relative accessibility, non-invasiveness, superior temporal resolution compared to other neuroimaging techniques such as positron emission tomography (PET) or functional magnetic resonance imaging (fMRI), EEG's potential extends to many different fields. One such application is the complimentary application in eye-tracking. As existing video-based eye-tracking methods rely on setting up fixed cameras and pointing them directly toward the subject's eyes, EEG-based eye-tracking may lead to a promising alternative solution that does not necessarily require fixed cameras within the subject's field-of-view.

EEGViT [16] is the current state-of-the-art (SOTA) model on EEG-based gaze prediction accuracy on the EEGEyeNet dataset [5]. It employs a hybrid transformer model fine-tuned with EEG data [7,12].

1.1 Research Question

In this paper, we propose a method that answer the following questions:

D. D. Schmorrow and C. M. Fidopiastis (Eds.): HCII 2024, LNAI 14695, pp. 60–71, 2024.
https://doi.org/10.1007/978-3-031-61572-6_5

- In CNN-transformer hybrid models, how do different convolution kernel sizes over the EEG spatial features (channels) affect the accuracy of the CNN-transformer hybrid models?
- How does this compare against a convolution over all EEG channels?

By answering this question, we investigate the effects of convolution kernels on the CNN-transformer hybrid networks.

2 Related Work

While EEG and Eye-tracking have each been studied individually for over a century, their combined use has only seen an increased interest in recent years with the aid of convolutional neural networks (CNNs) and transformers.

2.1 Dataset

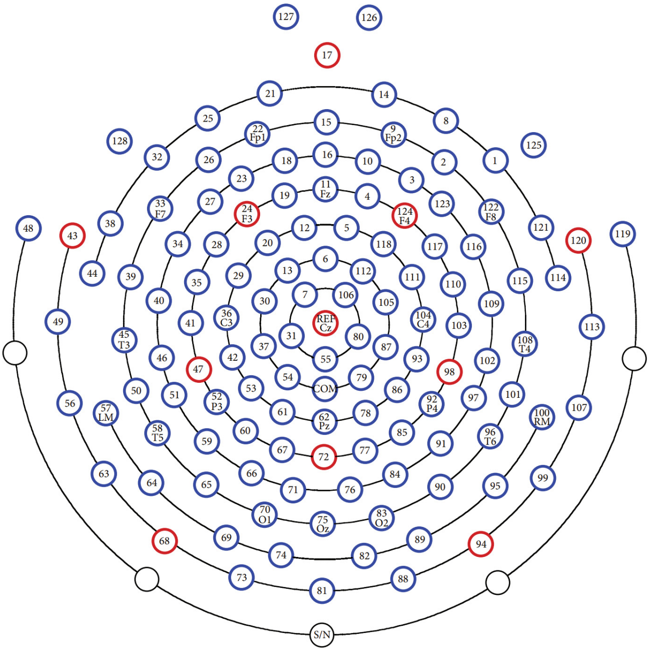

Fig. 1. Electrode Layout of the 128-channel EEG Geodesic Hydrocel system [1]

The EEGEyeNet dataset [5] offers EEG and eye tracking data that were collected simultaneously as well as benchmarks for eye movement and gaze position prediction. The EEG data of EEGEyeNet are collected from 356 participants using a 128-channel EEG Geodesic Hydrocel system, where the EEG channels are individually numbered from 1 to 128 as shown in Fig. 1. An additional reference electrode in the center make up a total of 129 EEG channels in the raw dataset.

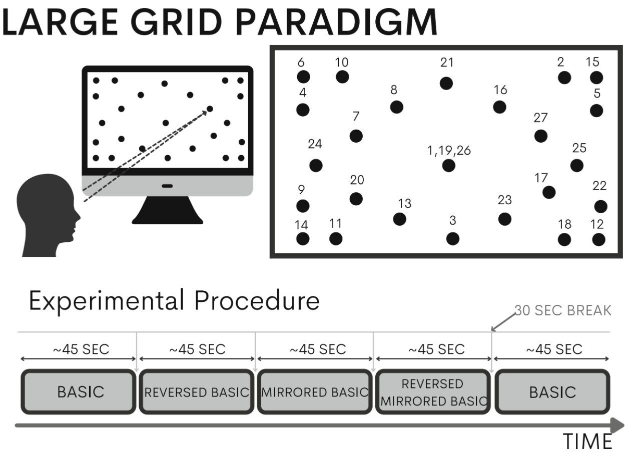

Fig. 2. The Large Grid Paradigm of EEGEyeNet [5]

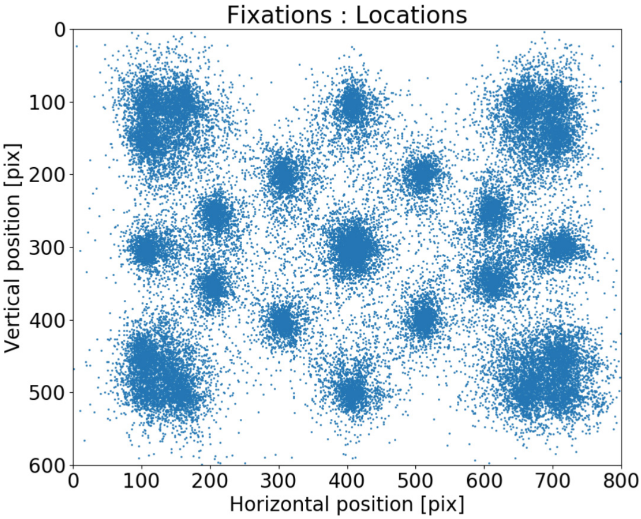

Fig. 3. Distribution of the Fixation Positions in the Large Grid Paradigm [5]

Experimental Paradigm. In one of EEGEyeNet's experimental paradigms, the participants are asked to fixate on specific dots on an "large grid" on the screen for a period as seen in Fig. 2. At the same time of recording EEG data, the participants' gaze positions are recorded. The gaze position distributions of 21464 samples can be seen in Fig. 3 [5].

2.2 State-of-the-Art

Since the publication of EEGEyeNet, several follow-up works have been made, often focusing on classification tasks (left-right or events such as blinking) [13–15]. The current state-of-the-art model in predicting gaze position is EEGViT [16], a hybrid vision transformer model fine-tuned with EEG data as shown in Fig. 4. EEGViT combines a two-level convolution feature extraction method, previously proposed in EEGNet [9] and Filter Bank Common Spatial Patterns [11] which enables efficient extraction of spatial (EEG electrodes) features for each temporal (frequency) channel, and a vision transformer using the ViT-Base model [3] pre-trained with ImageNet [2,10], to achieve a reported RMSE of 55.4 ± 0.2 mm on the EEGEyeNet dataset [16].

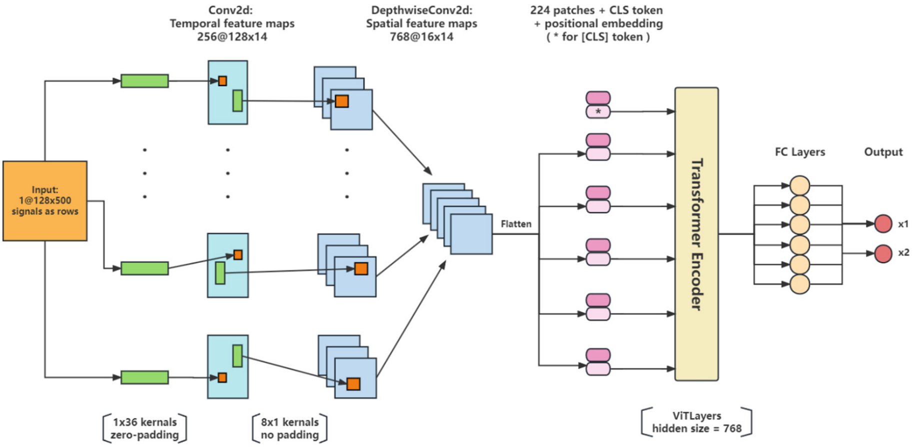

Fig. 4. EEGViT Model Architecture [16]

Table 1. Detailed Description of Our Model Architecture

Layer	Description
0	Input Size 129 × 500 × 1, Zero-padded to 129 × 512 × 1 on both sides
1	256 Temporal Convolution size 1 × 16 for Kernel and Stride, Batchnorm
2	768 Spatial Convolution size 129 × 1
3	ViT Model transformer, image size 129 × 32, patch size 129 × 1
4	Linear layer with 768 neurons on top of the final hidden CLS token
5	Linear layer with 1000 neurons, Dropout $p = 0.1$
6	Linear layer with 2 neurons (output)

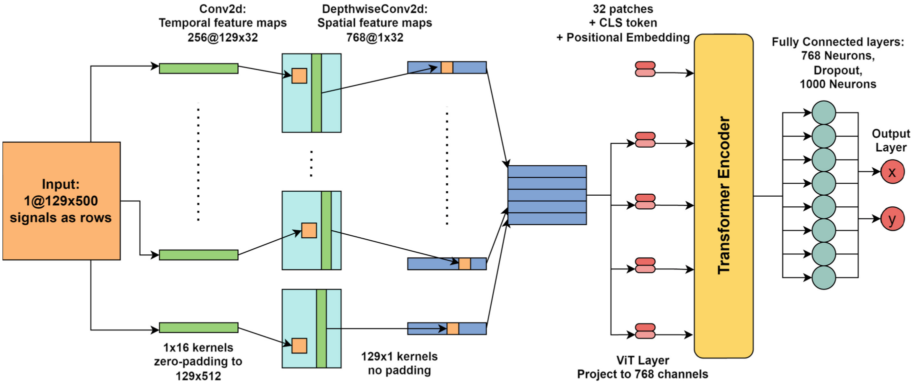

Fig. 5. Our Model Architecture, modified from [16]

3 Experiment

3.1 Model

The architecture of our Method can be seen in Fig. 5 and Table 1. Similar to prior works [9,11,16], we employ two convolution layers which filter the temporal and spatial (EEG channels) dimensions respectively.

In the first layer, a 1 × 16 kernel scans across the 1-s 129 × 500 input which is zero-padded to 129 × 512. The kernels effectively function as band-pass filters on the raw input signals. Our choice of 1 × 16 kernel is smaller than that of EEGViT at 1 × 36 [16] and that of EEGNet at 1 × 64 [9]. This provides a greater resolution of temporal features to be learned. Batch normalization is then applied on the 128 × 32 output [4].

In the second layer, a depth-wise 129 × 1 kernel scans over all EEG channels of each temporal filter. This is in contrast to EEGViT's approach, where a kernel of shape $(8, 1)$ is used [16].

Then, similar to EEGViT [16], the result is passed through a ViT transformer model, with the only difference being the shape of the input data. The base-ViT model [3] was pre-trained on ImageNet-21k and ImageNet 2012 [2,10] for image classification tasks. EEGViT [16] has previously shown that a ViT model pre-trained for image classification offers surprisingly good results when fine tuned with EEG data.

Lastly, two linear layers on top of the hidden CLS token of the ViT model output the x, y coordinates of predicted gaze position. We have additionally introduced a dropout layer to improve the robustness of the model.

3.2 Training Parameters and Software Implementation

We split the EEGEyeNet dataset into 0.7:0.15:0.15 for training, validation, and testing, and the model epoch with the lowest validation RMSE is used for testing. The split is by participant id in the original EEGEyeNet dataset to avoid leakage due to one participant's data samples appearing in more than one of training, validation, testing sets.

We included baseline ML implementations made public by the EEGEyeNet authors to be tested [5]. For EEGViT [16], we ported the authors' implementation match the setup of EEGEyeNet for training and testing in order to have the closest comparisons.

Our model and EEGViT are trained for 15 epochs in batches of 64 samples, with the Adam Optimizer [8] and an initial learning rate of 1e−4, which is dropped by a factor of 10 every 6 epochs. The model with the lowest validation error is used for testing. An example of the MSE loss during training in one of the runs can be seen in Fig. 6 and the resulting model's predictions on the testing set can be seen in Fig. 7.

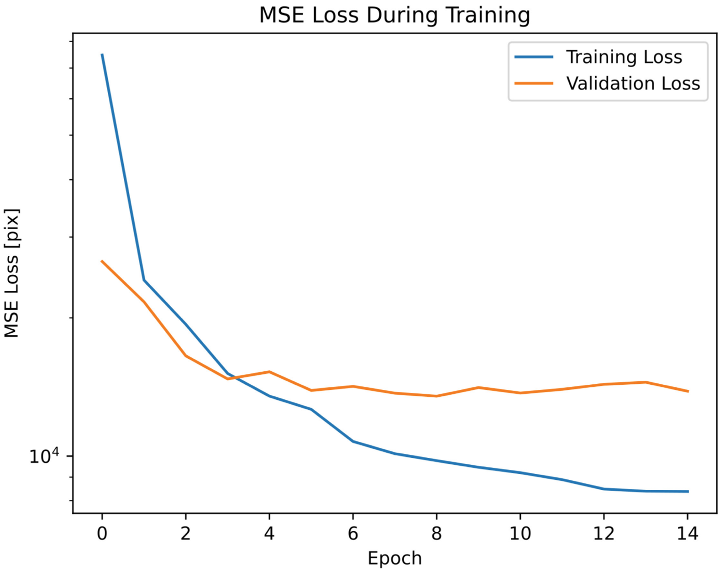

Fig. 6. Our Method's MSE Loss During Training

The full set of source code can be found at https://github.com/AmCh-Q/CSCI6907Project.

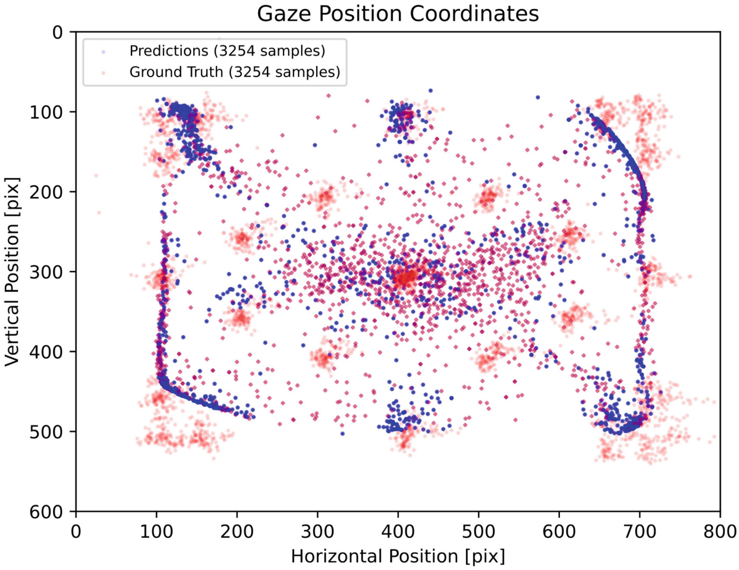

Fig. 7. Our method Gaze Position Coordinates, where predictions are colored blue, and the ground truths are colored red [6] (Color figure online).

3.3 Environment Setup

We performed all training, validation, and testing using Google Colab. We used an Intel Xeon Processor at 2.20 GHz, 51 GB of RAM, and a NVIDIA V100 GPU. For CUDA we used version 12.2, for PyTorch we used version 2.1.0, and for Scikit-learn we used version 1.2.2.

3.4 Evaluation

EEGEyeNet includes a benchmark where, given samples of shape $(129, 500)$ collected from 129 EEG channels at 500 Hz for 1 s when the participant fixates on one location, a machine learning model is to be trained to predict the 2-dimensional gaze position (in pixels) of the participant, and the accuracy may be evaluated as either the root mean-squared error (RMSE: Eq. 1) or mean Euclidean distance (MED: Eq. 2) in pixels or millimeters.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}((x_{i,\mathrm{truth}} - x_{i,\mathrm{pred}})^2 + (y_{i,\mathrm{truth}} - y_{i,\mathrm{pred}})^2)}{2n}} \tag{1}$$

$$\mathrm{MED} = \frac{\sum_{i=1}^{n} \sqrt{(x_{i,\mathrm{truth}} - x_{i,\mathrm{pred}})^2 + (y_{i,\mathrm{truth}} - y_{i,\mathrm{pred}})^2}}{n} \tag{2}$$

Here $(x_{i,\mathrm{truth}}, y_{i,\mathrm{truth}}) \in \mathbb{R}^2$ is the coordinates of the gaze position collected with a video-based eye-tracker in the i-th sample, and $(x_{i,\mathrm{pred}}, y_{i,\mathrm{pred}}) \in \mathbb{R}^2$ is the coordinates of the gaze position predicted by machine learning models from EEG data in the i-th sample, and n is the number of 1-s data samples collected.

Five runs were run for each of the two metrics above, the mean and standard deviation of the runs were recorded, and the results can be seen in Table 2 and Table 3.

4 Discussion

In this work, we presented an algorithm for predicting gaze position from EEG signals, and Table 2 shows the comparison of accuracy against various models including the SOTA (EEGViT). As can be seen in both the root mean-squared-error (Eq. 1) and mean euclidean distance (Eq. 2) metric, our method outperforms the SOTA. This is due to the use of a spatial filtering convolution kernel of shape $(129, 1)$, spanning all EEG channels, because the electrode layout, as seen Fig. 1, appear to be unordered and thus unlikely to be able to be learned through convolution with a smaller kernel as employed by EEGViT, and a kernel spanning all EEG channels would be able to better learn any spatial relationships between any two EEG channels at the same point in time.

We have also inspected the effect of permutation of the EEG channels and found that permuting the order of the EEG channels, either by shuffling or reordering the channels in spiral or z-order, yielded no noticeable difference in accuracy with either our method or EEGViT. We believe this means that the interactions between EEG channel signals is likely too complex and cannot be captured by convolution with a small receptive field.

Table 2. EEGEyeNet Gaze Position Scores and Standard Deviation across 5 Runs of EEGEyeNet baseline methods and EEGViT, compared against our method. Lower is better. All values are in millimeters and rounded to two decimal places. The column "Reported" contains the RMSE values that were originally reported from the respective studies.

Model	Reported RMSE	Bench RMSE	Bench MED	Study
Naive Center	–	95.85 ± 0	123.43 ± 0	[5]
Naive Mean	123.3 ± 0	95.81 ± 0	123.31 ± 0	[5]
Naive Median	–	95.79 ± 0	123.23 ± 0	[5]
KNN ($K = 100$)	119.7 ± 0	92.21 ± 0	119.67 ± 0	[5]
RBF SVR	123 ± 0	95.56 ± 0	123.00 ± 0	[5]
Linear Regression	118.3 ± 0	91.08 ± 0	118.37 ± 0	[5]
Ridge Regression	118.2 ± 0	90.91 ± 0	118.25 ± 0	[5]
Lasso Regression	118 ± 0	90.80 ± 0	118.04 ± 0	[5]
Elastic Net	118.1 ± 0	90.83 ± 0	118.13 ± 0	[5]
Random Forest	116.7 ± 0.1	90.09 ± 0.08	116.71 ± 0.08	[5]
Gradient Boost	117 ± 0.1	91.01 ± 0.06	117.50 ± 0.05	[5]
AdaBoost	119.4 ± 0.1	91.98 ± 0.07	119.39 ± 0.06	[5]
XGBoost	118 ± 0	91.73 ± 0	118.00 ± 0	[5]
CNN	70.2 ± 1.1	59.39 ± 0.63	70.11 ± 1.56	[5]
PyramidalCNN	73.6 ± 1.9	60.32 ± 1.67	70.86 ± 0.87	[5]
EEGNet	81.7 ± 1.0	61.92 ± 0.37	76.93 ± 0.73	[5]
InceptionTime	70.8 ± 0.8	60.32 ± 0.74	69.37 ± 0.90	[5]
Xception	78.7 ± 1.6	66.44 ± 0.80	76.77 ± 1.20	[5]
EEGViT	55.4 ± 0.2	54.41 ± 0.76	63.44 ± 0.83	[16]
Ours	–	**53.06** ± 0.73	**60.50** ± 0.93	–

In addition to measuring the accuracy of the models. We have also measured the run time of each of the models, the result of which are shown in Table 3. While slower than simpler methods such as CNN and even more considerably slower than methods such as KNN or linear regression, our method still offers an approximately 3.2 times speedup compared to the SOTA. This is due to our algorithm utilizing a much large spatial (channel) kernel, reducing the amount of trainable parameters in the model.

We were also able to confirm the findings of EEGEyeNet [5] that simple Machine learning models such as KNN, linear regression, and random forest were unable to gather meaningful information from EEG data and yielded no significant difference to naive center (where the model naively predicts the center of the screen), naive mean or naive median (where the model naively predicts the mean or median location of the training set's gaze position), while deep learning models such as CNN and EEGNet were able to yield significantly better

Table 3. EEGEyeNet Gaze Position run time (model training and validation of 21464 data samples) across 5 runs of EEGEyeNet baseline methods and EEGViT, compared against our method.

Model	Runtime [seconds]	Study
Naive Center	<0.01	[5]
Naive Mean	< 0.01	[5]
Naive Median	<0.01	[5]
KNN	0.71 ± 0.02	[5]
RBF SVR	13.23 ± 0.28	[5]
LinearReg	0.40 ± 0.07	[5]
Ridge	0.16 ± 0.01	[5]
Lasso	1.12 ± 0.02	[5]
ElasticNet	1.27 ± 0.01	[5]
RandomForest	355.90 ± 4.36	[5]
GradientBoost	816.69 ± 6.95	[5]
AdaBoost	113.31 ± 0.08	[5]
XGBoost	44.69 ± 0.43	[5]
CNN	362.71 ± 21.52	[5]
PyramidalCNN	281.84 ± 15.27	[5]
EEGNet	1696.90 ± 0.97	[5]
Xception	563.30 ± 10.59	[5]
EEGViT	2629.97 ± 5.79	[16]
Ours	812.33 ± 0.88	–

results than the naive baselines. We've also discovered that EEGEyeNet may have wrongly reported their results as "root mean-squared-error" when they may have in fact measured the mean euclidean distance error of the models, because in EEGEyeNet's source code we found that they have commented out the codes using RMSE and replaced it with MED, and that the resulting "RMSE" differs significantly with the RMSE result from our experiments, while appearing nearly identical to our "MED" measurements. Since our measured RMSE results on EEGEyeNet's models are significantly lower than reported by the authors of EEGEyeNet, the improvement made from models such as the SOTA, while still noticeable, may be smaller than what may have been believed previously.

4.1 Limitations

While the proposed method improves the accuracy and speed compared to the SOTA, the RMSE remains at approximately 5.3 cm and the mean euclidean distance remains at 6.1 cm, and training and validating the model takes an order of hundreds of seconds. This is considerable worse than commercially available video-based eye-tracking solutions in terms of both accuracy and run time. Moreover, EEGEyeNet's data was recorded in a laboratory setting and the participants were asked to stay still and have their gaze fixated on one spot on a screen, which is not reflective of most real-world application environments of eye-tracking [5]. The EEG setup is also more complex than most commercially available video-based solutions.

5 Conclusion

In this paper, we proposed an algorithm of EEG-based gaze prediction that outperforms the SOTA in both accuracy and speed. Our method improves the root mean-squared-error of the tracking to approximately 5.3 cm, and we found that having a large depth-wise convolution kernel for all EEG channels had the greatest impact. Nonetheless, EEG-based eye-tracing still has way to go and further research is needed for it to be comparable to the accuracy of traditional video-based eye tracking solutions.

Acknowledgments. This study was part of the authors' work in the course "CSCI 6907 Applied Machine Learning" in The George Washington University.

Disclosure of Interests. The authors declare no competing interests.

References

1. Bamatraf, S., et al.: A system for true and false memory prediction based on 2d and 3d educational contents and EEG brain signals. Comput. Intell. Neurosci. **2016**, 45–45 (2016)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
3. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. PMLR (2015)
5. Kastrati, A., et al.: EEGEyeNet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. arXiv preprint arXiv:2111.05100 (2021)
6. Key, M.L., Mehtiyev, T., Qu, X.: Advancing EEG-based gaze prediction using depthwise separable convolution and enhanced pre-processing, preprint (2024)
7. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: a survey. ACM Comput. Surv. (CSUR) **54**(10s), 1–41 (2022)
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. 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. J. Neural Eng. **15**(5), 056013 (2018)

10. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: ImageNet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972 (2021)
11. Schirrmeister, R.T., et al.: Deep learning with convolutional neural networks for EEG decoding and visualization. Hum. Brain Mapp. **38**(11), 5391–5420 (2017)
12. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
13. Wolf, L., et al.: A deep learning approach for the segmentation of electroencephalography data in eye tracking applications. arXiv preprint arXiv:2206.08672 (2022)
14. Xiang, B., Abdelmonsef, A.: Too fine or too coarse? The goldilocks composition of data complexity for robust left-right eye-tracking classifiers. arXiv preprint arXiv:2209.03761 (2022)
15. Xiang, B., Abdelmonsef, A.: Vector-based data improves left-right eye-tracking classifier performance after a covariate distributional shift. In: Kurosu, M., et al. (eds.) HCII 2022. LNCS, vol. 13516, pp. 617–632. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17615-9_44
16. Yang, R., Modesitt, E.: ViT2EEG: leveraging hybrid pretrained vision transformers for EEG data. arXiv preprint arXiv:2308.00454 (2023)