JCSE Journal of Computing Science and Engineering

Regular Paper

*Journal of Computing Science and Engineering,*
*Vol. 17, No. 2, June 2023, pp. 41-50*

# Looking to Personalize Gaze Estimation Using Transformers

**Seung Hoon Choi, Donghyun Son, Yunjong Ha, Yonggyu Kim, and Seonghun Hong**
VisualCamp, Seoul, South Korea
**{simon, ryan, michael, aiden, jeff}@visual.camp**

**Taejung Park***
Department of Cybersecurity, Duksung Women's University, Seoul, South Korea
**tjpark@duksung.ac.kr**

## Abstract

Anatomical differences between people restrain the accuracy of appearance-based gaze estimation. These differences can be taken into account with few-shot approaches for further optimization. However, these approaches come with additional computational complexity cost and are vulnerable to corrupt data inputs. Consequently, the use of accurate gaze estimation in real-world scenarios is restricted. To solve this problem, we introduce a novel and robust gaze estimation calibration framework called personal transformer-based gaze estimation (PTGE), utilizing a deep learning network that is separate from the gaze estimation model to adapt to new users. This network learns to model and estimate person-specific differences in gaze estimation as a low-dimensional latent vector from image features, head pose information, and gaze point labels. The expensive computational optimization process in few-shot approaches is removed in PTGE through our separate network. This separate network is composed of transformers, allowing self-attention to weigh the quality of calibration samples and mitigate the negative effects of corrupt inputs. PTGE achieves near state-of-the-art performance of 1.49 cm on GazeCapture with a small number of calibration samples (≤16) and no optimization when adapting to a new user, only a 2% decrease from the state-of-the-art achieved without the hour-long optimization process.

**Category:** Human-Computer Interaction

**Keywords:** Gaze estimation; Transformer; Artificial intelligence; Human computer interaction; Personal calibration

## I. INTRODUCTION

Gaze tracking is the process of measuring where a person is looking at. Gaze estimation is a valuable tool in several fields, including human-computer interaction, augmented/virtual reality, and behavioral analysis. The recent performance and popularity of machine learning models has caused the field of gaze estimation to adopt appearance-based approaches in favor of the conventional model-based approaches. Appearance-based approaches provide better results for in-the-wild settings, yet still have room for improvement in providing highly accurate predictions.

One of the main difficulties in accurately estimating gaze is representing person-specific differences. Not only do people look different from one another, but the structure of their eyes are all unique. This uniqueness impacts the line of sight and eye appearance. Anatomically, factors

Open Access http://dx.doi.org/10.5626/JCSE.2023.17.2.41 http://jcse.kiise.org



**Received** 26 March 2023; **Accepted** 22 May 2023
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

such as the size and shape of the eye, kappa (the angle between the visual and pupillary axis), and weights are different from person to person. These factors must be considered to deliver a more consistent and accurate gaze estimation model. In other words, a gaze estimation model must be calibrated to the user.

Recent approaches to personal calibration employ few-shot learning to adapt to a new person with minimal information. For example, some researches show that the difference in characteristics between people can be modeled as a low-dimensional latent vector space and adjusted for through few-shot learning in [1, 2]. However, the computational cost of meta-learning/fine-tuning alone places limitations on the use of accurate gaze estimation tools on mobile devices. Moreover, these approaches are vulnerable to contaminated samples because the loss function they target is very sensitive to such samples.

To overcome these challenges, we propose a light and robust calibration framework, called personal transformer-based gaze estimation (PTGE). In PTGE, we employ a concept of *preference vector*, a low-dimensional vector representing user-specific differences. In particular, to model the preference vector in the training phase, we train subject-wise embedding which is concatenated with image features before fully connected layers. At the same time, we train a *calibration model* which predicts preference vector from a small number of user image features and gaze label pairs. In the inference phase, only one forward pass in the calibration model is needed to estimate the preference vector for a user, which requires no special optimization process unlike previous approaches, e.g., [2].

We adopted transformer [3] as an architecture for the calibration model because of its outstanding properties: permutation invariance, ability to solve relationships between inputs, ability to handle arbitrary length of inputs and robustness. In particular, we presume the robustness that the attention mechanism exhibits can greatly reduce dependency on corrupted samples (e.g., mislabeled images, closed eyes) by suppressing them with low weights.

We conducted two main experiments to prove the excellence of our proposed method. First, we evaluated PTGE in two public gaze estimation datasets—GazeCapture [4], MPIIGaze [5]—where PTGE achieves near state-of-the-art performance despite its computational efficiency. Also, we evaluated PTGE in partially corrupted scenarios, where some of the calibration samples are corrupted, to check if our method handles such samples in a robust manner. The results showed that the performance drop is small when even a large portion of the calibration samples are corrupted, proving our conjecture that transformers can efficiently handle corrupted samples.

Our contributions can be summarized as follows:

(1) We suggested a gaze calibration framework called PTGE that does not require any optimization process during inference, which makes it feasible to be deployed in a low-resource environment (e.g., mobile devices). To achieve this, we train a calibration model which predicts user-specific differences in one forward pass. To the best of our knowledge, this is the first work to adopt an additional neural network for the calibration phase.

(2) By exploiting the robust property of a transformer, our proposed method achieves remarkable performance in partially corrupted scenarios. This is meaningful because such scenarios are similar to real-world calibration where users are prone to make mistakes.

## II. RELATED WORKS

### A. Appearance-Based Method

Prior approaches to gaze estimation were predominantly based on model-based methods [6-8] which utilize a geometric model of the eye. Initial approaches generated corneal reflections (glints) from an external infrared light source to measure geometric information including eye orientation and eyeball radius [9]. This approach relies on dedicated infrared devices, decreasing its availability to the general public. To address this issue, model-based approaches adopted RGB cameras to obtain features of the eye, such as the region of the iris and its center [10, 11]. However, these methods require high-resolution images of the eye to accurately extract eye features and are vulnerable to a multitude of factors, including head pose variation and illumination.

In contrast, appearance-based methods directly predict gaze estimation through images of the eye and/or face. Recent improvements have allowed for accurate results from a wide range of cameras and in-the-wild conditions [12]. The adoption of deep neural networks in appearance-based methods has contributed to this improvement and has led to a 5°–6° increase in accuracy [13, 14].

As appearance-based methods use neural networks to infer gaze points, it needs a large amount of dataset to leverage its general prediction performance compared to the model-based methods. With the introduction of large real-world gaze datasets like [12] which contains images from 15 participants (MPIIGaze) and [15] from 110 participants with various head poses and gaze angles (ETH-XGaze), the robustness of gaze estimation models was enhanced.

In previous appearance-based approaches, eye and face images are directly applied to train neural networks assuming that the raw facial image contains information of head pose variations. Krafka et al. [4] proposed a convolutional neural network (CNN) architecture that inputs cropped eyes images, face images and face location masks. Zhang et al. [14] have proposed a network that takes the face image only to estimate gaze.

On the other hand, learning head pose variations from

hidden layers can lead to overfitting of its dataset. Other studies consider combining or removing geometric information like head pose vectors and eyeball rotation angles. Zhu and Deng [16] have designed a network that gets head pose and eyeball rotation angles as input. Zhang et al. [17] have proposed a face image normalization method to remove head pose variations and scale factors which improve model robustness in wild condition.

## B. Personalization Approach

Previous model-based gaze estimation methods usually include the personal parameters as a part of the three-dimensional geometric model. They estimate the subject-specific factors like eyeball radius [9] or fovea offsets [18] through a personalized calibration to enhance the performance of these unified models.

When the appearance-based methods are executing the calibration, they need to incorporate the process without the explicit eye model [2]. Usually, we consider the calibration problem as domain adaptation problem, where the models trained from the source domain should be adapted to the target domain [19]. The calibration samples from the target domain are comprised of images from unseen environmental conditions or of unknown subjects. Therefore, we categorize the personalized calibration process in terms of how to adapt the general gaze estimation model to the test set properly and how to get sufficient calibration samples of unseen subjects efficiently.

***1) Domain Adaptation***

Domain adaptation methods can be classified into two types.

***a) Domain adaptation via fine-tuning to a target domain:*** The simplest approach to calibration is to train the general gaze estimation model into a person-specific model. Zhang et al. [20] fine-tune their CNN model in each target domain to enable the adapted model to track the gaze direction in multiple devices. On the other hand, Krafka et al. [4] used a general model as a feature extractor and train an additional support vector regression to predict each subject's gaze points. These methods improve performance and are easy to implement, but they cannot fully use the potential of the personalized calibration.

***b) Domain adaptation via the use of user embedding vector:*** To process the calibration by adapting the whole model or adding a post-processing procedure means a high-dimensional calibration parameter space, followed by a requirement for a several calibration samples. So, recent studies introduced the low-dimensional vectors which represent person-specific features. They show that the personal features can be learned during fine-tuning these user embedding vectors and stabilize the calibration process with little resources. Linden et al. [2] proposed a spatial adaptive model and show that personal variations are well-modeled by a three-dimensional latent vector for each eye.

***2) Collection of Calibration Samples from Unknown Subjects***

We also consider how to obtain an adequate number of calibration samples efficiently. If a gaze estimation application requires many test sets for a subject, it can degrade the user experience and lower the availability. Park et al. [1] make their model learn a rotation-aware latent representation of gaze and adapt to a new subject with very few calibration samples. From an empirical point of view, various studies [21-23] propose calibration data collection methods. Although using a human-labeled calibration sample can be accurate, it is challenging to acquire enough samples in a practical situation. Salvalaio and de Oliveira Ramos [21] collect samples when the subject uses an input device like a mouse. They enable their model to self-adapt to a new person rapidly with these calibration samples by online transfer learning method.

## C. Transformer and Attention

Since transformer networks were introduced as an efficient approach to processing sequential problems in natural language processing [3, 24-27], they have expanded their scope to alternatives to traditional structures, including CNNs and recurrent neural networks (RNNs). One of the notable advantages of a transformer is that it is parallel and has a longer learning range than CNNs and RNNs.

As a result, not only has transformer become mainstream in the field of natural language processing, but the field of computer vision has also seen benefits with transformer-based models for better expressiveness. Dosovitskiy et al. [28] introduced vision transformer (ViT) to apply the transformer to image classification tasks without relying on CNNs. Touvron et al. [29] have achieved similar performances to those of the ViT with their approach, data-efficient image transformer (DeiT), using only a much smaller data set (around 1/300). Liu et al. [30] proposed shifted window (Swin) transformer to improve the ViT by applying hierarchy with shifted windows.

Park and Kim [31] analyze the opposite characteristics of ViT and CNN. They show that conventional CNNs work as high-pass filters and ViTs as low-pass filters. From this observation, the authors propose a novel combined network structure (AlterNet).

One of the transformer's essential engines for implementing the mentioned benefits is the multi-head attention (MHA) mechanism [3]. The attention value is defined as a linear combination of each *value* vector and its probability. The probabilities are calculated as the *softmax* values of the *similarities* between a given *query* and multiple *keys*. The similarity is defined as a dot product operation between a *query* and a *key* (often after a linear transformation

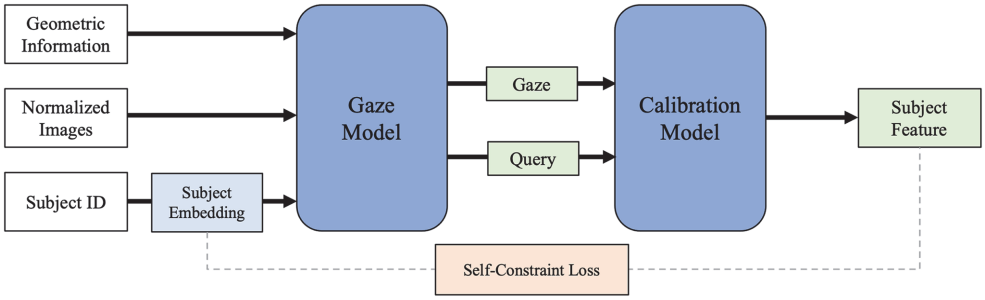


**Fig. 1.** Overview of the entire gaze-calibration pipeline.

operation).

As many researchers and practitioners understand that attention mechanisms make important differences in transformer networks, several researchers have tried to unravel the underlying principles of attention mechanisms. To better understand the principle of attention, Tsai et al. [32] interpreted the Attention operation as a process of applying a kernel that calculates the similarity between two input data to design an improved Attention mechanism by better understanding the principle of Attention. Mickus et al. [33] mathematically analyze the transformer network, focusing on additional properties of the MHA.

## III. METHOD

We present a new method, illustrated in Fig. 1, to accurately estimate gaze by calibrating new users to a general gaze estimation model utilizing transformers. The method consists of two main parts. The gaze model that takes in normalized images, head pose and rotation information, and subject-specific embeddings to predict the gaze point and create the input for the calibration model. The calibration model takes in both the gaze point and image features extracted within the gaze model to predict calibration parameters used to adapt the gaze model. When in use, the calibration parameters are estimated once by the calibration model and used by the gaze model afterward.

### *A. Gaze Model*

With the premise that an adaptable gaze estimation model requires a base model to adapt, we began our process with a person-independent gaze estimation model [1]. We created the person-independent gaze model to recognize basic patterns from people's facial features that translate to where the users are looking at.

Taking inspiration from previous approaches, the gaze model consists of two main CNN models, defined as $g_{eye}$ for images of the eyes and $g_{face}$ for images of the entire face. Given an input image $x$, we extracted the normalized face image $x_{face}$ and normalized eye images, $x_{righteye}$ and $x_{lefteye}$. We also estimate the head rotation matrix **R**, and three-dimensional eye coordinates, $p_{eye}$. This process is described in the experiments section. As shown in Fig. 2, the feature extractor creates both the face features $g_{face}(x_{face})$ and eye features $g_{eye}(x_{eye})$. We concatenate the extracted features with the head rotation matrix and three-dimensional eye coordinates along with subject-wise embeddings, $E$ to input an multi-layer perceptron (MLP) we call Gaze Extractor, $g_{ex}$, to get the final gaze estimation result, G.

The subject-wise embeddings are a separate layer created to represent person-specific differences and separated them from the gaze model. Each subject has a separate embedding vector of length $n$. This vector is trained to minimize the Euclidean distance between the estimated gaze point and ground truth per person. Note that the embeddings found during the training process are not used in the final evaluation.

Most works calculate the translation of the yaw and pitch estimation to a two-dimensional gaze coordinate

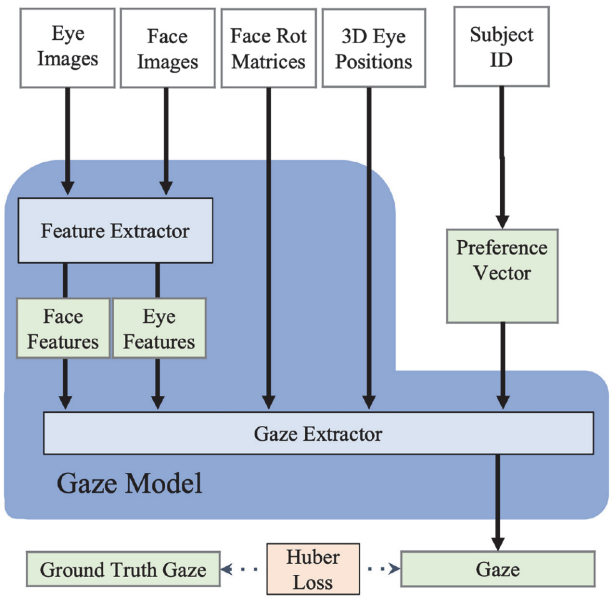


**Fig. 2.** Gaze model architecture.

separately without the help of a deep learning model. However, we input the head rotation matrix along with the three-dimensional coordinates for each eye and lets the model learn the process. We found the head pose estimation algorithm in the image preprocessing process to be fragile to outside noise. We hypothesized that the head pose estimation and three-dimensional eye coordinates calculated are also affected by person-specific differences. Experimentally, we found that inputting these values into the gaze extractor allowed the subject-wise embeddings and model to tweak the results as needed. Hence, instead of predicting a gaze vector and geometrically calculating the gaze point on the screen with the head rotation and three-dimensional eye coordinate information, we input all of these values into our gaze model.

We optimized this model with the following loss function,

$$\mathcal{L}_{gaze} = \frac{1}{B}\sum_{n=1}^{B}\mathcal{L}_H(G,\hat{G}) \tag{1}$$

where $\hat{G}$ is the ground truth gaze point and $B$ is the number of batches. The Hubber loss $\mathcal{L}_{\mathrm{H}}$ is defined as follows:

$$\mathcal{L}_H(G,\hat{G}) = \begin{cases} \frac{1}{2}\|G-\hat{G}\|^2, & \text{if } \|G-\hat{G}\| \le \delta \\ \frac{1}{2}\delta^2 + \delta(\|G-\hat{G}\|-\delta), & \text{otherwise} \end{cases} \tag{2}$$

The parameter $\delta$ in (2) determines where two functions (linear and quadratic) are exchanged. We set $\delta$ = 1.5 for our implementation.

## B. Calibration Model

The objective of the calibration model is learning to predict person-specific calibration parameters from the image features extracted by the gaze model, head pose information and ground truth gaze labels. These person-specific calibration parameters are set as inputs to the gaze estimation model to personalize it to a user and increase its overall accuracy.

Previous approaches use few-shot learning, a meta-learning method to train models to make predictions given a limited number of examples, to adapt the gaze estimation model to different people. However, the few-shot adaptation process has a large computational cost which is not feasible for use on mobile devices. We show that gaze estimation does not require the few-shot process to adapt to new people due to the degree of similarity between different subjects. Instead, the person-specific calibration parameters can be learned for estimation given the features extracted from our gaze model, giving inspiration to the idea of our calibration model. To the best of our knowledge, this is the first approach to estimate person-specific calibration parameters without any optimization using a deep learning model for gaze estimation.

### *1) Architecture Overview*

As illustrated in Fig. 3, our calibration model consists of a stack of transformer encoders between MLPs. This model learns to predict an embedding vector that represents person-specific differences given a small number of calibration samples.

We concatenate the features extracted by the gaze model, $g_{face}$ and $g_{eye}$, geometric information, **R** and $p_{eye}$, and gaze ground truth to create the calibration query, defined at $q$. This query is set as the input to an MLP to create our calibration model transformer input. Our query is then put into a stack of transformer encoders. The outputs of the transformer are then passed into another MLP to get the final subject-wise calibration parameter estimation.

We choose the transformer as our main architecture in the calibration model to increase speed and improve calibration robustness. This approach speeds up the calibration process by removing any optimization when seeing a new user and creates a more robust calibration result by being permutation-invariant and utilizing self-attention. We believe that the performance of the calibration should not be dependent on the order of the inputs and

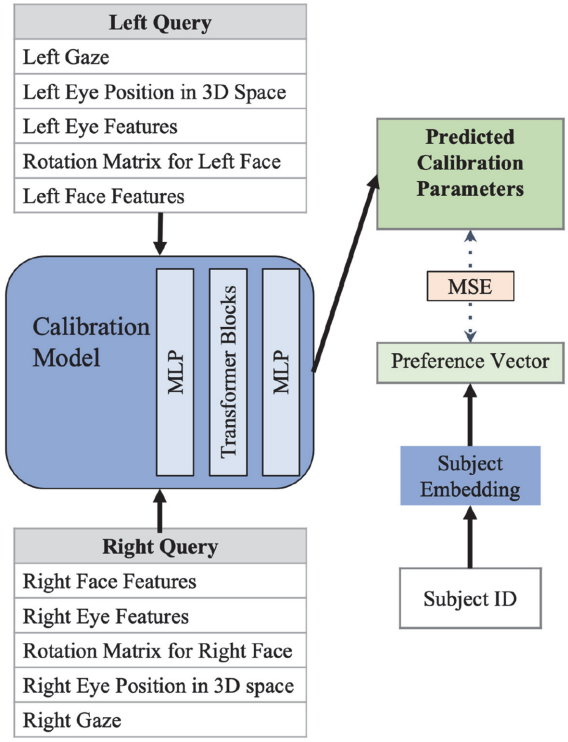


**Fig. 3.** Overview of the calibration model.

should recognize corrupted input data (e.g., closed eye images, mislabeled data, etc.). Few-shot learning approaches are vulnerable to corrupted inputs as they base their predictions on the new, limited inputs and do not have the capability to filter out corrupted ones. In contrast, self-attention has the capability to weigh the importance of its inputs, mitigating the negative effects of those that are corrupted.

***2) Training the Model***

We trained the calibration model to estimate the learned preference vector in the gaze model. We input $B$ different batches of $s$ calibration samples of one user's features to get $B$ results from the calibration model. To train a calibration that is cohesive with the gaze model, we set the ground truth of the calibration parameters as the preference vector trained in the gaze model. We set the self-constraint loss function as the mean squared error between the two.

$$\mathcal{L}_{accuracy} = \frac{1}{B}\sum_{n=1}^{B} \|C(q_n) - \hat{C}\|^2 \tag{3}$$

Here, $C$ is our calibration model, $\hat{C}$ is the preference vector layer in the gaze model, and $q_n$ is the calibration query for batch $n$.

We also encourage our calibration model to find the same embedding for one person. To achieve this, we also utilized an embedding consistency loss similar to the one proposed by [1]. This loss takes the sum of the angle difference of each preference vector estimation.

$$\mathcal{L}_{consistency} = \frac{1}{B}\sum_{i=2}^{B} d(C(q_1), C(q_i)) \tag{4}$$

Here, $d$ is a function measuring the angular difference between the two inputs.

$$d(a,b) = \arccos\left(\frac{a \cdot b}{\|a\|\|b\|}\right) \tag{5}$$

***3) Final Adaptation***

When calibrating, both the gaze model and calibration model are used. The gaze model must provide the query for the calibration model. With just a few calibration samples, the gaze model creates $s$ calibration queries for the calibration model. The calibration model takes in the $s$ calibration inputs and one forward step, estimates the subject's calibration parameters. Then, the estimated calibration parameters are set as the preference vector input for the gaze model. The gaze model has adapted to the new user and the calibration model is no longer needed after this. Our method provides a fast and accurate calibration with no optimization.

# IV. EXPERIMENTS

## A. Datasets

***1) MPIIGaze***

MPIIGaze [5] is the most widely used benchmark for appearance-based gaze estimation. It consists of around 2,500 images from 15 subjects each taken randomly for several days. We used the images and data from the MPIIFaceGaze subset for training and evaluation. We train on MPIIFaceGaze using the leave-one-out strategy. Following previous approaches, we used the last 500 images of each for evaluation and selected k calibration samples for the calibration model from the remaining images.

***2) GazeCapture***

GazeCapture [4] is the largest public gaze estimation dataset that contains around 2.5 million photos from over 1,450 people. To use the preprocessing pipeline proposed in [17], we gathered the camera intrinsic parameters of the iPhones and iPads used from the web. We change the face detector to BlazeFace for a model that can be used outside of iOS devices. Following [4], we report our gaze estimation error in centimeters.

## B. Image Preprocessing

From the normalization approach proposed in [17], we retrieved normalized images and geometric information from an RGB image to promote easier learning and generalization. We flipped the left eye image to match the overall shape of the right eye and train the feature extractor with shared weights. To create a symmetric network for the left and right eye, we paired the flipped left eye image with a flipped face image. In other words, the gaze model has four image inputs: right eye, flipped left eye, cropped face, and flipped cropped face.

## C. Implementation Details

***1) Gaze Model***

All models, including the gaze model, are trained with an Adam optimizer with default $\beta$ and $\varepsilon$ values ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-7}$).

We used a base learning rate of $3 \times 10^{-4}$ for the first 40 epochs. During this phase, gradients are not passed to the preference vector layer embedding. After the first 40 person-independent gaze training epoch is completed, the learning rate is decreased to $10^{-4}$ with a cosine decay schedule. Moreover, the preference vector layer embedding uses $l_2$ regularization of 0.01.

*2) Calibration Model*

The calibration model is also set with a learning rate of $10^{-4}$ with a cosine decay schedule. There are a total of six transformer blocks in the calibration model, each with four heads. The batch size $N$ is set to 8 and $s$ is set to 16. We set the preference vector to a length of 6 as we experimentally found that it provides the most optimal performance.

## D. Comparison with state of the art

*1) GazeCapture*

As shown in Fig. 4, PTGE accuracy is comparable to SPAZE [2], the current state-of-the-art appearance-based gaze estimation model. SPAZE is a CNN-based model that adapts to new users using BFGS optimization. SPAZE shows better uncalibrated accuracy compared to PTGE. However, on a small number of calibration samples (≤4), PTGE shows almost identical performance. With four calibration samples, both SPAZE and PTGE reach a mean test error of 1.61 centimeters. This, combined with the fact that PTGE requires no optimization to calibrate lowers the computational cost substantially compared to previous methods. In other words, PTGE allows for state-of-the-art performance gaze estimation on mobile devices.

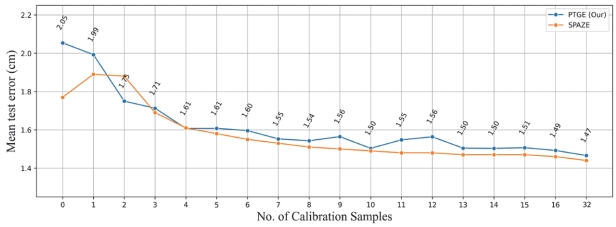


**Fig. 4.** Accuracy of GazeCapture.

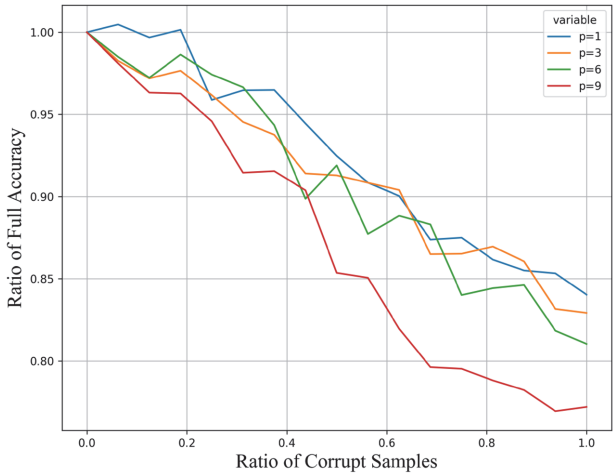


**Fig. 5.** Robustness of calibration model on MPIIGaze with varying preference vector length.

Even when 16 calibration points are utilized, the difference of 0.02 cm, or a 2% drop in accuracy is almost negligible.

*2) MPIIGaze*

On MPIIGaze [5], we compare our model with FAZE and SPAZE, the two most recent state-of-the-art adaptable gaze estimation models. Both FAZE and SPAZE require further optimization during calibration, MAML for FAZE and BFGS for SPAZE. Table 1 displays the comparison of our results along with FAZE and SPAZE. For within-MPIIGaze training, PTGE shows an angular error of 3.76°, a 3% improvement compared to FAZE. However, due to the small number of subjects and amount of data, PTGE is unable to separate person-specific embeddings to the fullest extent. Training with GazeCapture offers a solution to this problem. GazeCapture is a much larger dataset compared to MPIIGaze and contains almost 1,450 more subjects.

Training on MPIIGaze along with GazeCapture offers the best performance of PTGE, decreasing its angular error to 2.81°. This metric has a difference of only 4% from SPAZE.

## E. Robustness

Experimenting on MPIIGaze, we found that PTGE is a robust model that can filter out corrupt data. Illustrated in Fig. 5, we see that our model can retain up to 95% of its original performance when even 20% of the calibration samples are corrupt. Experiments with preference vectors of length 1, 3, 6 all achieve similar results while those with a length of 9 become less robust. We hypothesize that this is because the larger the preference vector, the more information can capture. The preference vector in the gaze model begins to capture person-independent factors.

## F. Number of Calibration Parameters

To find the optimal preference vector length, we ran experiments on MPIIGaze. As shown in Table 2, a preference vector of length 6 appears to be the most optimal. As stated above, we hypothesize that a length of 9 performs worse as it begins to take over the role of the person-independent gaze estimation model. Hence, in all the experiments where the preference vector length is not explicitly stated the preference vector length is 6.

**Table 1.** Comparison of PTGE with other calibration models on MPIIGaze [5]

| Method | Error (°) |
|---|---|
| FAZE | 3.88 |
| PTGE | 3.76 |
| PTGE (GazeCapture pretrained) | 2.81 |
| SPAZE | 2.70 |

**Table 2.** MPIIGaze [5] subject-wise angular error based on the length of the calibration vector

| Subject ID | Preference vector length | | | |
|---|---|---|---|---|
| | 1 | 3 | 6 | 9 |
| p00 | 2.98 | 2.76 | 2.92 | 2.99 |
| p01 | 2.80 | 2.74 | 2.79 | 2.89 |
| p02 | 3.05 | 2.90 | 2.60 | 3.21 |
| p03 | 5.52 | 5.82 | 5.68 | 4.99 |
| p04 | 2.72 | 2.78 | 2.70 | 2.88 |
| p05 | 2.21 | 2.30 | 2.17 | 2.20 |
| p06 | 3.84 | 3.91 | 3.67 | 3.74 |
| p07 | 4.99 | 4.57 | 4.83 | 5.03 |
| p08 | 4.88 | 4.52 | 4.87 | 4.54 |
| p09 | 4.86 | 5.72 | 5.45 | 5.19 |
| p10 | 5.32 | 5.61 | 4.82 | 5.18 |
| p11 | 3.67 | 3.30 | 3.23 | 3.53 |
| p12 | 2.56 | 2.51 | 2.58 | 2.72 |
| p13 | 5.02 | 4.04 | 4.24 | 4.43 |
| p14 | 3.10 | 3.38 | 3.93 | 4.19 |
| Avg. | 3.83 | 3.79 | 3.76 | 3.85 |

## V. CONCLUSION

In this paper, we presented PTGE, the first calibration approach in gaze estimation using deep learning models, to predict calibration parameters. PTGE achieves near state-of-the-art performance without any optimization during calibration, providing almost the same performance while decreasing the computational cost significantly. We display that the personal variations can be represented as a latent vector and predicted from image features and geometric information.

## ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A4A5028907).

## Conflict of Interest(COI)

The authors have declared that no competing interests exist.

## REFERENCES

1. S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, "Few-shot adaptive gaze estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 9367-9376.
2. E. Linden, J. Sjostrand, and A. Proutiere, "Learning to personalize in appearance-based gaze tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, South Korea, 2019, pp. 1140-1148.
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5598-6008, 2017.
4. K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 2176-2184.
5. X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Mpiigaze: real-world dataset and deep appearance-based gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 162-175, 2019.
6. E. Wood and A. Bulling, "EyeTab: model-based gaze estimation on unmodified tablet computers," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, Safety Harbor, FL, 2014, pp. 207-210.
7. Z. Zhu and Q. Ji, "Eye gaze tracking under natural head movements," in *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Sam Diego, CA, 2005, pp. 918-923.
8. D. H. Yoo and M. J. Chung, "A novel non-intrusive eye gaze estimation using cross-ratio under large head motion," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 25-51, 2005.
9. E. D. Guestrin and M. Eizenman, M. (2006). General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1124-1133, 2006.
10. R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 802-815, 2012.
11. L. Sun, M. Song, Z. Liu, and M. T. Sun, "Real-time gaze estimation with online calibration," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Chengdu, China, 2014, pp. 1-6.
12. F. Lu, T. Okabe, Y. Sugano, and Y. Sato, "A head pose-free approach for appearance-based gaze estimation," in *Proceedings of the British Machine Vision Conference (BMVC)*, Dundee, UK, 2011, pp. 1-11.
13. A. Ali and Y. G. Kim, "Deep fusion for 3D gaze estimation from natural face images using multi-stream CNNs," *IEEE Access*, vol. 8, pp. 69212-69221, 2020.
14. X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: full-face appearance-based gaze estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, HI, 2017, pp. 2299-2308.

15. X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "ETH-XGaze: a large scale dataset for gaze estimation under extreme head pose and gaze variation," in *Computer Vision–ECCV 2020*. Cham, Switzerland: Springer, 2020, pp. 365-381.
16. W. Zhu and H. Deng, "Monocular free-head 3D gaze tracking with deep learning and geometry constraints," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 3162-3171.
17. X. Zhang, Y. Sugano, and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, Warsaw, Poland, 2018, pp. 1-9.
18. D. W. Hansen and Q. Ji, "In the eye of the beholder: a survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478-500, 2010.
19. Y. Cheng, H. Wang, Y. Bao, and F. Lu, "Appearance-based gaze estimation with deep learning: a review and benchmark," 2021 [Online]. Available: https://arxiv.org/abs/2104.12668.
20. X. Zhang, M. X. Huang, Y. Sugano, and A. Bulling, "Training person-specific gaze estimators from user interactions with multiple devices," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal, Canada, 2018, pp. 1-12.
21. B. K. Salvalaio and G. de Oliveira Ramos, "Self-adaptive appearance-based eye-tracking with online transfer learning," in *Proceedings of 2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, Salvador, Brazil, 2019, pp. 383-388.
22. Z. Chang, J. Matias Di Martino, Q. Qiu, S. Espinosa, and G. Sapiro, "Salgaze: personalizing gaze estimation using visual saliency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, South Korea, 2019, pp. 1169-1178.
23. K. Wang, S. Wang, and Q. Ji, "Deep eye fixation map learning for calibration-free eye gaze tracking," in *Proceedings of the 9th Biennial ACM Symposium on Eye Tracking Research & Applications*, Charleston, SC, 2016, pp. 47-55.
24. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018 [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.
25. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019 [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
26. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, 2019, pp. 4171-4186.
27. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.
28. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An image is worth 16x16 words: transformers for image recognition at scale," 2020 [Online]. Available: https://arxiv.org/abs/2010.11929.
29. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 10347-10357, 2021.
30. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, 2021, pp. 9992-10002.
31. N. Park and S. Kim, "How do vision transformers work?," 2022 [Online]. Available: https://arxiv.org/abs/2202.06709.
32. Y. H. H. Tsai, S. Bai, M. Yamada, L. P. Morency, and R. Salakhutdinov, "Transformer dissection: a unified understanding of transformer's attention via the lens of kernel," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, 2019,pp. 4344-4353.
33. T. Mickus, D. Paperno, and M. Constant, "How to dissect a muppet: the structure of transformer embedding spaces," 2022 [Online]. Available: https://arxiv.org/abs/2206.03529.

## Seung Hoon Choi

Seung Hoon Choi is an undergraduate student studying computer science at the University of Illinois Urbana-Champaign. He has previously worked at VisualCamp as a Machine Learning Research Engineer for two years. At VisualCamp, he conducted research on improving the accuracy and robustness of appearance-based gaze estimation models. His current research interests include multimodal learning and exploring adversarial examples.

## Donghyun Son

Donghyun Son is an undergraduate student at Department of Computer Science and Engineering in Seoul National University. He worked at VisualCamp as an AI Researcher for two years and is working at Hyperconnect as a Machine Learning Engineer. His research interests include gaze estimation, domain generalization, and self-supervised learning for large-scale models.

## Yunjong Ha

Yunjong Ha is a computer vision researcher at RnD team in VisualCamp since 2021. He received B.A degree from the University of Seoul in 2018 and M.A degree in computer engineering from the Seoul National University in 2020. He studied simple MAC protocol under sensor networks as an undergraduate researcher. He researched about GPGPU compute resource utilization enhancement and developed a medical image processing s/w for his master's degree. After working at Hallym Medical AI center, he currently researches to improve gaze estimation accuracy through better appearance-based methods in VisualCamp.

## Yonggyu Kim

Yonggyu Kim is a researcher at RnD team in VisualCamp since 2021. He received B.A and M.S. degrees in computer science engineer from KoreaTech in 2013 and 2019, respectively. He designed the progressive occupancy network architecture for 3D reconstruction for his master's degrees. His current research interests include gaze estimation in computer vision area.

## Seonghun Hong

Seonghun Hong is an undergraduate student at the Department of Electrical and Computer Engineering in Seoul National University since 2017. He worked as a machine learning engineer in 2022 at VisualCamp. He conducted research at VisualCamp on domain generalization and image processing. His current interests include multi-modal learning and building large-scale machine learning systems.

## Taejung Park

Taejung Park is an associate professor at Department of Cybersecurity/IT Media in Duksung Women's University since 2013. He received B.A. and M.S. degrees in electrical and computer engineering from Seoul National University in 1997 and 1999, respectively. He designed a data structure for a direct matrix solver for FEM simulation for semiconductor devices for his master's degree. After working with two small start-up technology businesses in Korea in 1999 and 2002, he received a Ph.D. from the Department of Electrical and Computer Engineering in Seoul National University for 3D mesh compression in 2006. His current research interests include parallel numerical simulation techniques and nonlinear interpolation methods from the viewpoint of information technology.