Self-Attention with Convolution and Deconvolution for Efficient Eye Gaze Estimation from a Full Face Image

Jun O Oh
Computer Engineering
Dankook University
72200119@dankook.ac.kr

Hyung Jin Chang
School of Computer Science
University of Birmingham
h.j.chang@bham.ac.uk

Sang-Il Choi*
Computer Engineering
Dankook University
choisi@dankook.ac.kr

Abstract

This paper proposes a whole new face image-based eye gaze estimation network to solve low generalization performance. Due to the high variance of facial appearance and environmental conditions, conventional methods in gaze estimation have low generalization performance and are easily overfitted to training subjects. To solve this problem, we adopt a self-attention mechanism that has better generalization performance. Nevertheless, applying self-attention directly to an image incurs a high computational cost. Thus, we introduce a new projection that uses convolution in the entire face image to accurate model the local context and reduce the computational cost of self-attention. The proposed model also includes deconvolution that transforms the down-sampled global context to the same size as the input so that spatial information is not lost. We confirmed through observations that the new method achieved state of the art on the EYEDIAP, MPIIFaceGaze, Gaze360 and RT-GENE datasets and achieved a performance increase of 0.02° to 0.30° compared to the other state of the art model. In addition, we show the generalization performance of the proposed model through a cross-dataset evaluation.

1. Introduction

Gaze is a typical nonverbal expression that is used as an effective clue to read a person's intention and social relationships. Estimating the gaze from an image has been trying to be used as an interface in HCI [25], AR/VR [16], or autonomous driving [18].

Among the various approaches [14, 31, 32] to estimate the gaze of an image, several effective methods [5, 22, 33] have recently been proposed based on facial appearance using deep learning techniques. In particular, methods based on the holistic appearance of the face learn the function of mapping the gaze from the appearance of the whole face [22, 33]. However, the faces are diverse according to the individual or environmental factors such as gender, race, head, and lighting, so the model is affected by the unique feature of the individual, making it difficult to generalize the estimation of gaze [30].

To improve the generalization performance of gaze estimation, many methods have attempted to estimate gaze by separating the eyes from the face [2, 3, 20]. These methods consist of a module for segmenting an eye area from a face and a module for inferring a gaze from the segmented eye region. Although this minimizes the influence from personal appearance and environmental factors when inferring gaze, the following problems remain: 1) To segment the eye in the image, it requires additional labeling such as eye position or head posture, as well as ground truth about the gaze [8, 14, 32]. 2) If the module to isolate the eye area from the face cannot work properly (i.e., extreme head posture, dark shading, etc.), the subsequent gaze inference module will also not work [2, 8, 20]. 3) Since the eye region segmentation module and the gaze inference module are independently trained, and then each module is sequentially combined to construct a system, it does not guarantee that the training result is the globally optimal solution for the entire gaze estimation system [30].

On the other hand, methods have been proposed to infer gaze by using the information of the entire face, such as head posture and global illumination, without having to separate the eyes from the face [4, 33]. In [4], gaze inference was attempted using a transformer capable of dynamic attention to effectively learn data with a high variance while considering the entire face's global context, and it showed higher performance than existing CNN-based methods. However, although the transformer can effectively reflect various information from the face related to gaze inference, there is still a problem in directly applying it to this field. When applying the transformer to the vision field, it is difficult to solve the high computational cost of calculating self-attention in the image. A method for embedding images [6] was proposed to address this problem, but they

*Corresponding author.

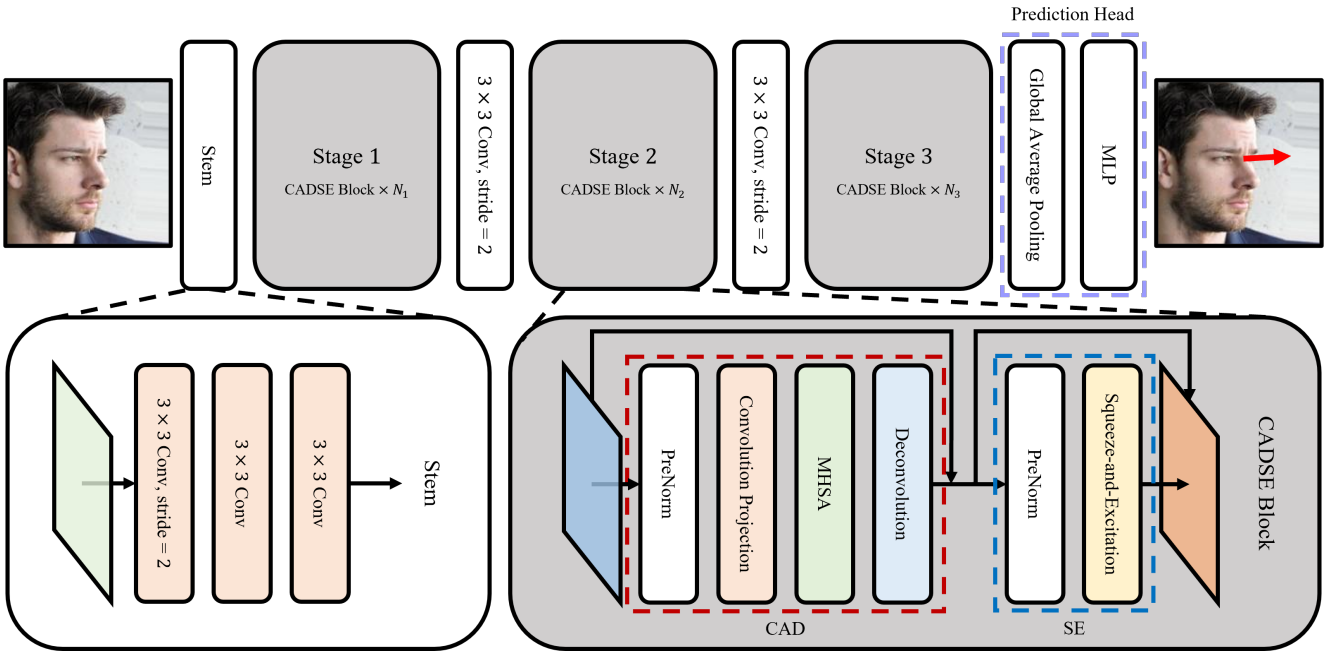

Figure 1. **The overall framework of the proposed method.** The Stem is a layer introduced for the convenience of training similarly to [26]. In the early stages, the eyes and face information are extracted, and in the following later stages, the extracted information is getting abstracted.

may inevitably lose the information necessary to infer gaze from the image.

Since gaze is largely affected by the spatial interpretation where eyes are located on the face, the self-attention module [27] that utilizes the spatial context of an image effectively improves the generalization performance of gaze inference. Some methods include a down-sampling layer by modeling local information to reduce the amount of computation in self-attention [6]. However, the existing embedding that applied a linear function after dividing the image into non-overlapping patches does not consider the positional information of the internal pixels when projecting the patch. Although these methods can be suitable for general classification problems where the similarity between patches is important, they are not effective for gaze inference problems where important information such as the location of the iris may exist inside the patch.

In this paper, we propose a new face image-based eye gaze estimation method using spatial context information, which effectively solves low generalization performance issues. Before applying self-attention directly to the input image, we obtain a global context in which the local context is effectively modeled while significantly reducing the amount of computation through *convolution projection*. The convolution projection removes individual features irrelevant to gaze inference in the local context to focus on information for gaze inference. Then, we apply *deconvolution* to the result of self-attention to have the same dimensions as the input and keep original with the skip connection. As a result, we could largely reduce the computational cost compared to the direct application of self-attention. and preserve fine features that may have been lost in embedding. Moreover, since all these processes are performed through end-to-end learning, a globally optimal solution for gaze inference can be learned.

Our main contributions are as follows:

- We propose a new self-attention module with convolution projection and deconvolution layer to improve the generalization performance of gaze estimation without being affected by the individual characteristics of the face. Furthermore, it can solve the problem of high computational cost caused by applying self-attention naively to image data for gaze estimation.
- We design the convolution to filter irrelevant information for gaze estimation effectively and the deconvolution for high-resolution activation, which improves the accuracy of gaze estimation by maintaining detailed image features.

2. Related Work

Several deep learning-based methods have been proposed to automatically estimate gaze from facial images.

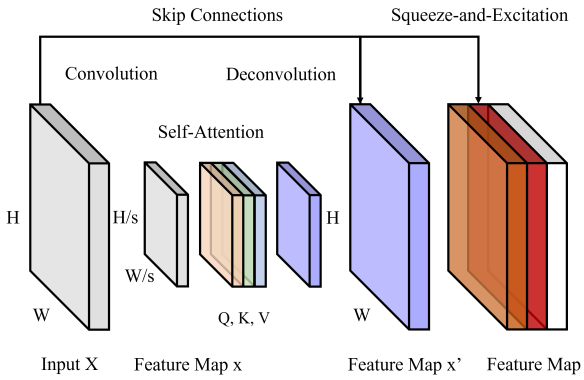

Figure 2. **CADSE block.** In the CADSE block, CAD learns the local context by convolution layer achieving a significant computation reduction and learns the global context by self-attention. The SE explicitly applies channel attention to spatial attention learned in CAD.

Appearance-based methods can be largely divided into two paradigms, depending on what information the network is based on to infer eye gaze. One is a way to pre-process the eye areas or the information that is considered essential for gaze inference in the face, then transfer the information to the gaze estimation network. The other is how the gaze estimation network attempts to infer eye gaze directly from the face.

Methods of learning the eye estimation model by separating the eye area in the face image to prevent the gaze estimation model from over-fitting to the training data set and to achieve a high estimation accuracy. In [14], they explored the face, both eyes, and the face grid from the input image and provided useful information to the gaze estimation network. The method in [8] added independently trained networks to extract eye patches and head positions in front of eye estimation networks, reflecting information about head positions lost due to eye separation. However, these models require a face alignment module [7,10], which increases the computational costs of the overall framework and results in delays in data delivery between the modules. Furthermore, the range (yaw from -90 ° to 90 ° as reported by Krafka et al. [15]) of the head posture in the dataset [15,23] used separately to learn this module is limited, and the performance of the model becomes vulnerable in different environments from the training data set.

End-to-end learning-based methods have also been proposed to perform eye estimation from facial images in a single framework. In [13], the estimation of the gaze from the image using LSTM was proposed assuming that the information over time was essential for gaze inference. However, it is computationally high and difficult to process in real-time due to the input type of video unit and the characteristics of RNNs that are processed sequentially. In [30], they incorporated the existing face-eye separation paradigm into end-to-end learning by allowing the model to learn how to extract areas critical to gaze inference from images. However, the global context of the face image cannot be utilized for gaze inference because gaze estimation still uses information from patches separated from the image. Some methods have also improved performance by applying the concepts of attention [33] and self-attention [4], which are actively utilized in computer vision tasks. However, these methods also have limitations in effectively reflecting the global context of images in gaze estimation, and the computation for gaze estimation increases significantly.

3. Methodology

A complete pipeline of the proposed structure is shown in Figure 1. The proposed method comprises a convolution stem [26] for convenience of training, three stages, each consisting of N_i $(i = 1, 2, 3)$ CADSE blocks and a prediction head. The CADSE block is a component that includes a Convolution-Attention-Deconvolution (CAD) block and a Squeeze-and-Excitation (SE) block. When a face image is given as an input to the model, low-level features are extracted in the first CADSE, and high-level features abstracted for gaze estimation are extracted through subsequent CADSE blocks. The feature maps extracted from the last CADSE are subjected to global average pooling (GAP) to construct a flattened feature vector. Then, the feature vector is used as an input to the fully connected layer for gaze prediction.

3.1. Preliminaries

The transformer responds well to long sequences by using the global computation and memory of the self-attention layer, showing outstanding performance in natural language processing [27]. Self-attention is a module proposed to use information from the temporal context of natural language effectively. ViT(Vision Transformer) [6] applied this idea of self-attention to image data and showed effective results in classification and object detection problems. Among the encoders and decoders that make up the transformer, the encoder is mainly used in computer vision task. The transformer encoder has three main components: multi-head self-attention (MHSA), multi-layer perceptron (MLP), and layer normalization (LN).

The self-attention is the essential module of the transformer, and it models the relationship between $m \cdot n$ patches that embedded from a given image. Let us denote an input as $x \in R^{m \times n \times d}$ consists of $m \cdot n$ patches, where d is the number of channels of each patch. Self-attention can be represented as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (1)$$

where Q, K and, V are vectors that transform input x with trainable weight matrices W_Q, W_K and, W_V, respectively. d_k means the dimension of K.

MHSA is an extension of the self-attention module for multiple subspaces. MLP, located immediately after the MHSA module, adjusted the dimensions and incorporated nonlinearities into the network. In addition, LN is deployed for stable and fast network learning. The transformer uses the skip connection [11] for residual learning and its layer can be formulated as:

$$x = LN(MHSA(X) + X) \tag{2}$$

$$x' = LN(MLP(x) + x) \tag{3}$$

This module is easy to extend to multi-layer because the output x' usually takes the same dimension as input X, i.e., $x \in R^{m \times n \times d}$. We applied the ability to extract the global spatial context of self-attention to deal with spatial context information from face images and extract features for estimating gaze.

3.2. CAD Block: Self-Attention with Convolution and Deconvolution

Figure 2 shows the structure of the proposed CADSE block: i) self-attention with convolution projection and deconvolution architecture (CAD) and ii) squeeze-and-excitation (SE). As shown in the figure, we place the self-attention layer in between convolution and decovolution layer. When a transformer is applied to an image, there is a limit to reflecting the spatial context of the entire image because the amount of computation is hugely increased, and the local information extracted from the individual pixel is limited.

So, we reduced the amount of computation by performing convolution projection before applying self-attention. In addition, we extracted the global spatial context that effectively reflects local connectivity by moving the convolution filters to overlap each other. For example, as shown in Figure 2, if a convolution with kernel size k and stride s is applied to a feature map X with a size of $H \times W$, the down-sampled query, key and value has a size of $H/s \times W/s$. Therefore, the operation cost in the self-attention can be saved by approximately s^2. Finally, the input X is converted to x as follows:

$$x = Conv2D(X, k, s) \tag{4}$$

Convolution, considering the inherent inductive bias in the image, brings faster convergence and improved performance than the linear function and helps eliminate positional encoding (PE) [28]. Furthermore, through the convolution layer, it can be learnt which features are related to estimating gaze during training to improve the performance. This prevents individual characteristics from being affected in the self-attention process. The global context modeled through convolution projection contains refined information related to gaze, and self-attention is applied to this to obtain similarity between entities and give weight to important information.

After computed self-attention, we add a deconvolution layer to make the spatial dimension equal to the input and extend the weighted features from the global to the local level. In general, the transformer configures the shape of the output of the layer to be the same as the input to stack into multiple layers. However, in our model, since the input of the self-attention module is down-sampled, the result becomes as downsized. When down-sampling the query, key, and value to reduce self-attention operations [28], they preserve the query as the original size so that the result is the same size as the input. However, eyes, which contain important information for gaze estimation, occupy a small part of the entire face. If they are pooled before they have sufficiently abstracted, the feature of eyes may disappear from low-resolution images. Therefore, to efficiently manage the computational cost while maintaining the highest possible resolution while abstracting the eyes, we down-sampled the query and the attention map computed at the global level through the deconvolution layer. Also, as shown in Figure 1, the input value is maintained by adding a skip connection. The following layer is designed to check both the abstracted information and the information before it is lost. We adopted the transposed convolution as follows to implement deconvolution layer:

$$x' = TransposedConv2D(MHSA(x), k, s) + X \tag{5}$$

We can learn how to extend the attention calculated at the global level to the same dimension as the input through this process. Moreover, since the inductive bias of the image is maintained in all processes, we could remove PE without degrading performance [28].

3.3. SE Block: Squeeze-and-Excitation

The feature undergoes convolution and deconvolution layer within the CAD block, and modeling related to spatial information is performed, and the shape is adjusted. By placing the SE [12] block after the CAD block, we weighted for multiple channels created through the multi-head and refined the feature map by reflecting the relative importance of each channel.

The SE applies global average pooling to a feature map to extract the values of every channel. The extracted values are passed through the multi-layer perceptron (MLP) and then converted into weights that represent the importance of the individual channel. The weights are multiplied by the feature map to decide which values are passed to the next block.

Table 1. **Overview of the datasets used in experiments.** We show the number of subjects, the maximum head poses and gaze in horizontal and vertical directions in the camera coordinate systems, the amount of data (number of images or duration of the video), and image resolution.

	Subjects	Head Pose	Gaze	Data	Resolution
EYEDIAP [9]	16	$\pm 15°$, $30°$	$\pm 25°$, $20°$	237 min	HD & VGA
Gaze360 [13]	238	$\pm 90°$, unknown	$\pm 140°$, $-50°$	172,000	4096×3382
MPIIFaceGaze [33]	15	$\pm 15°$, $30°$	$\pm 20°$, $\pm 20°$	45,000	1280×270
RT-GENE [8]	15	$\pm 40°$, $\pm 40°$	$\pm 40°$, $-40°$	122,531	1920×1080

Table 2. **Gaze estimation errors in degrees on cross dataset evaluations.** For direct comparisons, we put the results of ResNet50 and the proposed model side by side: (ResNet50/Proposed)

Train \ Test	EYEDIAP	Gaze360	MPIIFaceGaze	RT-GENE
EYEDIAP	-	33.7°/24.6°	15.7°/12.9°	15.6°/12.3°
Gaze360	11.3°/7.10°	-	10.0°/8.31°	26.6°/17.7°

3.4. CADSE Block

The overall transformer structure composed of CAD and SE modules can be expressed as follows:

$$x' = CAD(PreNorm(X)) + X \tag{6}$$

$$x'' = SE(PreNorm(x')) + x' \tag{7}$$

In addition, as in Figure 2, it is designed to learn residuals and preserve the originals by adding skip connections after each component.

We use layernorm [1] to normalize the input features. The first transformer method [27] takes the post normalization strategy which can be formulated as follows:

$$f' = LayerNorm(F(f) + f) \tag{8}$$

where f, f' and $LayerNorm(\cdot)$ denotes feature map, normalized feature map and layernorm, respectively. However, pre-normalization [19] shows better performance and could resolve the learning instability that may occur in a post normalization due to the residual term. The pre-normalization can be formed as:

$$f' = F(LayerNorm(f)) + f. \tag{9}$$

Each stage of the network comprises N_i CADSE layers, where N_i is the number of layers in the *i-th* stage. In the early stages of the model, we maintain a high resolution to sufficiently abstract information about the eyes from the face image. In the subsequent stage, we reduce spatial information by pooling and increase channels to more abstract features and increase the expressive power of the model. Finally the features extracted in the last stage are flattened through GAP and then transferred to a prediction head composed of multi-layer perceptron.

4. Experimental Results

We performed three types of experiments to evaluate the proposed algorithm. First, we performed a cross-data set evaluation to check the generalization performance of the CAD block. As a result, the generalization performance of the proposed model is improved compared to CNN. Second, we conducted an ablation study to confirm the effects of the convolution and deconvolution layer to extract spatial context using local connectivity in CAD. Finally, we evaluated the gaze estimation performance of the proposed model through a direct comparison with the state of the art methods for gaze estimation.

Dataset for pre-training We used the ETH-XGaze [29] dataset for pre-training. ETH-XGaze consists of 1.1M images obtained from 110 subjects. It is divided into training and evaluation sets, and we used the training set containing 765K images of 80 subjects to pre-train the model. The evaluation set is divided into within-dataset and person-specific evaluations, each including 15 people. We used the within-dataset as the test set for pre-training validation. The dataset provides normalized data, and we fed it directly into the model.

Dataset for evaluation To evaluate the gaze estimation performance, we selected four datasets among the public released datasets: EYEDIAP [9], Gaze360 [13], MPIIFaceGaze [33], and RT-GENE [8]. All datasets are labeled for 3D gaze estimation. For more information about each data set (see Table 1). The EYEDIAP dataset consists of 94 videos with a 237-minute duration obtained from 16 subjects, and we used four-fold cross-validation to evaluate the performance with this dataset. The Gaze360 dataset collected 172K images from 238 subjects and has the most

Table 3. **Ablation Study.** We can clearly see that deconvolution layer contributes a significant performance improvement, and convolution projection layer is also crucial.

Method \ Dataset	EYEDIAP	Gaze360	MPIIFaceGaze	RT-GENE
Proposed method (*w/* Conv/Deconv)	5.25°	10.70°	4.04°	7.00°
w/o convolution projection	5.62°	12.57°	4.76°	8.12°
w/o deconvolution	6.73°	14.80°	4.99°	9.21°

Table 4. **Comparison with state-of-the-art methods.** The proposed method achieves state-of-the-art results

Category	Method \ Dataset	EYEDIAP	Gaze360	MPIIFaceGaze	RT-GENE
A	FullFace [33]	6.53°	14.99°	4.93°	10.00°
	RT-GENE [8]	6.02°	12.26°	4.66°	8.00°
	Dilated-Net [2]	6.19°	13.73°	4.42°	8.38°
	Gaze360 [13]	5.36°	11.04°	4.06°	7.06°
	CA-Net [3]	5.27°	11.20°	4.27°	8.27°
B	GazeTR-Pure [4]	5.72°	13.58°	4.74°	8.06°
	GazeTR-Hybrid [4]	5.33°	11.00°	4.18°	7.12°
	Proposed	**5.25°**	**10.70°**	**4.04°**	**7.00°**

comprehensive head pose and gaze range. They predefined the data set into 129K images for training, 17K for validation, and 26K for evaluation. We used this setup as is. The MPIIFaceGaze dataset is based on the MPIIGaze dataset and contains 45K images obtained from 15 subjects. We used the leave-one-person-out evaluation method to estimate the performance with this dataset. The RT-GENE dataset collected 123K images from 15 subjects and designated 13 subjects for training and two for verification. Although the resolution provided is high, the problem becomes difficult because the distance between the camera and the subject is far. Therefore, we used 3-fold cross-validation to evaluate the performance of this dataset. The EYEDIAP and MPIIFaceGaze datasets have the relatively limited head pose and gaze and assumed benchmarks in a controlled environment. Gaze360 and RT-GENE challenge their relatively wide head pose and gaze range and show performance in unrestrained environments.

Experimental settings The proposed method was implemented with PyTorch [21] [1]. All experiments were learned with the training set specified in each data set and evaluated by the test set for a fair experiment. The evaluation method remain the same. We used angular error as an evaluation metric that is commonly used in gaze inference. Since the error is the angle between the predicted gaze and the actual gaze, a model with a small error has better performance. Experiments were conducted with Intel Xeon CPU @ 3.40GHz, 256 GB RAM, and NVIDIA Tesla V100 GPU. For all training sequences, the batch size was set to 128. All networks (except for the pre-training module, facial landmark detection) were trained with RAdam [17], and the initial learning rate was set to 0.0001. The resolution of input image is 224×224. The sizes of the features that supplied to each stage are 112×112, 56×56 and 28×28, respectively. In the proposed model, each stage has 64, 192, and 384 channels, respectively, and the N_1, N_2, and N_3 are 1, 3, and 8, respectively. The kernel size and stride of the filter for convolution projection are 11 and 8, respectively, and the deconvolution is also the same. Dropout [24] with 0.2 was applied to the model.

[1]To ensure reproducibility, we will release the code when the paper is accepted.

4.1. Cross Dataset Evaluation

Cross-dataset evaluation on multiple datasets is frequently used to measure the generalization performance of gaze-inference models. In [29], they trained the off-the-shelf ResNet50 and conducted cross-dataset evaluation in various datasets. Following the same protocol, we trained the proposed model on the EYEDIAP dataset to compare cross-evaluation performance with the ResNet50. The model was also trained with the Gaze360 dataset and evaluated with three other datasets. The results are presented in Table 2. Our model obtained advanced results for all datasets. When trained on the EYEDIAP dataset, we found a performance improvement in the proposed model from 2.8 ° to 9.1 ° compared with ResNet50. In the case of the Gaze360 dataset, the performance improved from 1.69 ° to 8.9 ° than ResNet50. It shows higher accuracy when trained on the Gaze360 dataset than EYEDIAP, which seems to be

Table 5. Specification of the state-of-the-art models.

Method	# of Params.	# of FLOPs	Running Time(ms)
RT-GENE [8]	82.0M	30.81G	467
Gaze360 [13]	14.6M	12.78G	276
GazeTR-Pure [4]	227.3M	58.32G	1280
Proposed	74.8M	19.75G	379

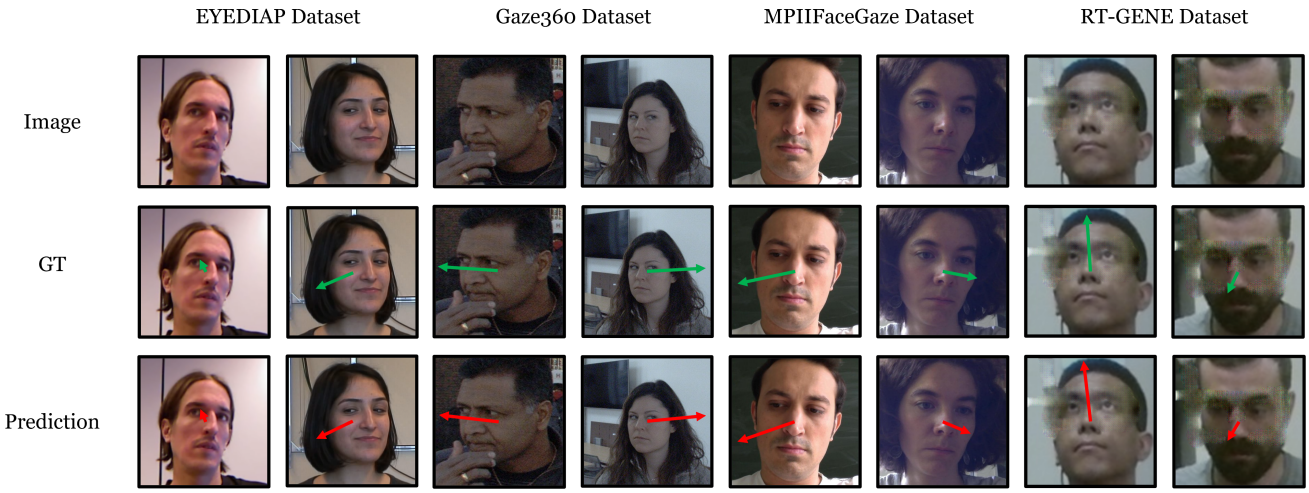

Figure 3. **Proposed method's Gaze estimation results on various dataset face images.** The first row images are input images, and the second and third rows are the ground truth for the gaze and the estimation result by the proposed network, respectively. We can see that the proposed method has good generalization performance, which can reliably estimate eye gaze from various kinds of dataset images.

because it has enough data to learn a wide gaze distribution. In other words, if we have enough data to train a transformer, self-attention can show higher generalization performance than convolution.

4.2. Ablation Study

To confirm the validity of our model design, we conducted the following ablation study on the elements while removing some components from the entire pipeline: a qualitative comparison of 'with convolution and deconvolution (proposed method)', 'without convolution projection', and 'without deconvolution' (See Table 3).

a) *w/o* convolution projection To investigate the effect of the convolution projection on the performance, we removed the convolution layer in the process of embedding the input to make the query, key, and value. Instead, we use a linear function on the non-overlapping patches for embedding. We conducted experiments on four datasets to ensure reliability, and the results are shown in Table 3. When the convolution projection technique was applied, the performance improved from 0.37 ° to 1.87 °, respectively, which shows that the proposed method is very effective compared to the alternative. It seems to be because convolution layer extracts information more efficiently in modeling the local context.

b) *w/o* deconvolution After calculating the attention score, we need to up-sample the output because our model down-samples queries, keys, and values. Otherwise, the resolution will remain small, and some spatial information will be lost. To check the effect of deconvolution, we compared the performance with the method [28] without up-sampling by using the query in its original size. This experiments were also carried out on four datasets, and, as shown in Table 3, the influence of deconvolution layer is significant. It can be seen that if we remove the deconvolution layer, the performance is worse than when we eliminate the convolution projection. We found degradation from 0.95 ° to 4.10 ° due to the removal of the deconvolution layer for each dataset. The performance degradation was particularly more significant in the Gaze360 and RT-GENE datasets than the other datasets because these datasets assume an unrestrained environment, making it difficult for the model to find information related to gaze inference. Additionally, the model is likely to try to fit the subject when self-attention preserves the spatial size of the query.

4.3. Comparison to the State-of-the-art Methods

We compared the performance of the proposed model and the state-of-the-art methods, which showed competitive performance in gaze estimation, with the EYEDIAP,

Gaze360, MPIIFaceGaze and RT-GENE datasets. The results are presented in Table 4. In the table, methods corresponding to category A (FullFace [33], RT-GENE [8], Dilated-Net [2], Gaze360 [13] and CA-Net [3]) are CNN- or RNN-based gaze estimation models, and the methods in category B are those that use a transformer.

Table 5 shows the number of parameters for each method and the flops required to derive the result. As shown in Table 4, the results of category B applying self-attention estimated gaze more accurately than those of category A. GazeTR-Pure [4] using ViT shows high accuracy compared to category A, but the requested computational cost increases rapidly as shown in Table 5. On the other hand, the proposed method requires smaller resource than ViT architecture and be estimated more accurately. In Table 5, GazeTR-Pure shows approximately three times the parameter and 2.95 times the FLOPs difference compared to our model. The running time also shows a proportional diversity. It was shown that our method using convolution effectively suppresses the cost increase that may occur when self-attention is applied with MLP embedding. The improvement is noticeable in the datasets in a more unconstrained environment. As shown in Table 1, the Gaze360 dataset contains a huge variance in head pose and gaze variance compared to other datasets. The performance of the proposed model in the Gaze360 and RT-GENE dataset increased from 0.30 ° to 4.29 ° and from 0.06 ° to 3.00 ° compared to other methods. The proposed model works well on datasets where it is difficult to learn the mapping between input and gaze. The results in the EYEDIAP and MPIIFaceGaze datasets also increased from 0.02 ° to 1.28 ° and from 0.02 ° to 0.89 °, respectively.

The proposed method's generalization performance can be seen in Figure 3. It visualizes some qualitative results of gaze estimation on various face images from different datasets.

5. Conclusion

This paper proposes a novel framework to solve the low generalization performance of gaze estimation networks with faces as input. We introduced self-attention with convolution and deconvolution that can handle global context and has better generalization performance. Convolution projection and deconvolution placed before and after self-attention effectively modeled the local context and reduced the amount of self-attention computation. Through rigorous experiments on four public datasets (EYEDIAP, Gaze360, MPIIFaceGaze, and RT-GENE), we validated that the proposed model outperforms other methods that are either CNN-based or transformer-based in terms of accuracy and computational cost. We adopted the self-attention structure to increase the model's generalization performance, but there is still scope for improvement, such as to identify a structure more suitable for removing individual characteristics.

Acknowledgment

This work was supported by the National Research Foundation of Korea through the Korean Government (MSIT) under Grant 2021R1A2B5B01001412, and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00034, Clustering technologies of fragmented data for time-based data analysis).

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 5

[2] Zhaokang Chen and Bertram E. Shi. Appearance-based gaze estimation using dilated-convolutions, 2019. 1, 6, 8

[3] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation, 2020. 1, 6, 8

[4] Yihua Cheng and Feng Lu. Gaze estimation using transformer. *ArXiv*, abs/2105.14424, 2021. 1, 3, 6, 7, 8

[5] Haoping Deng and Wangjiang Zhu. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3162–3171, 2017. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 2, 3

[7] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 3

[8] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *European Conference on Computer Vision*, pages 339–357, September 2018. 1, 3, 5, 6, 7, 8

[9] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap database: Data description and gaze tracking evaluation benchmarks. Idiap-RR Idiap-RR-08-2014, Idiap, 5 2014. 5

[10] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 4

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 2018. 4

[13] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE International*

Conference on Computer Vision (ICCV), October 2019. 3, 5, 6, 7, 8

[14] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3

[15] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2144–2151, 2011. 3

[16] Joseph Lemley, Anuradha Kar, and Peter Corcoran. Eye tracking in augmented spaces: A deep learning approach. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*, pages 1–6, 2018. 1

[17] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, April 2020. 6

[18] Sujitha Martin, Sourabh Vora, Kevan Yuen, and Mohan Manubhai Trivedi. Dynamics of driver's gaze: Explorations in behavior modeling and maneuver prediction. *IEEE Transactions on Intelligent Vehicles*, 3(2):141–150, 2018. 1

[19] Toan Q. Nguyen and Julian Salazar. Transformers without tears: Improving the normalization of self-attention. *CoRR*, abs/1910.05895, 2019. 5

[20] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. *Lecture Notes in Computer Science*, page 741–757, 2018. 1

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[22] Adrià Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1444–1452, 2017. 1

[23] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 3

[24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 6

[25] Julian Steil, Michael Xuelin Huang, and A. Bulling. Fixation detection for head-mounted eye tracking based on visual similarity of gaze targets. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018. 1

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 2, 3

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2, 3, 5

[28] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021. 4, 7

[29] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision (ECCV)*, 2020. 5, 6

[30] Xucong Zhang, Yusuke Sugano, A. Bulling, and Otmar Hilliges. Learning-based region selection for end-to-end gaze estimation. In *BMVC*, 2020. 1, 3

[31] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, 2015. 1

[32] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, June 2015. 1

[33] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2299–2308, 2017. 1, 3, 5, 6, 8