electronics

MDPI

Article

BoT2L-Net: Appearance-Based Gaze Estimation Using Bottleneck Transformer Block and Two Identical Losses in Unconstrained Environments

Xiaohan Wang [1], Jian Zhou [1,2,*], Lin Wang [1], Yong Yin [1], Yu Wang [1] and Zhongjun Ding [3]

[1] School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China; wangxiaohan@whut.edu.cn (X.W.)
[2] Hubei Key Laboratory of Broadband Wireless Communication and Sensor Networks, Wuhan 430070, China
[3] National Deep Sea Center, Qingdao 266237, China
* Correspondence: jianzhou@whut.edu.cn

Abstract: As a nonverbal cue, gaze plays a critical role in communication, expressing emotions and reflecting mental activity. It has widespread applications in various fields. Recently, the appearance-based gaze estimation method, which utilizes CNN (convolutional neural networks), has rapidly improved the accuracy and robustness of gaze estimation algorithms. Due to their insufficient ability to capture global relationships, the present accuracy of gaze estimation methods in unconstrained environments, has the potential for improvement. To address this challenge, the focus of this paper is to enhance the accuracy of gaze estimation, which is typically measured by mean angular error. In light of Transformer's breakthrough in image classification and target detection tasks, and the need for an efficient network, the Transformer-enhanced-CNN method is a suitable choice. This paper proposed a novel model for 3D gaze estimation in unconstrained environments, based on the Bottleneck Transformer block and multi-loss methods. Our designed network (BoT2L-Net), incorporates self-attention through the BoT block, utilizing two identical loss functions to predict the two gaze angles. Additionally, the back-propagation network was combined with classification and regression losses, to improve the network's accuracy and robustness. Our model was evaluated on two commonly used gaze datasets: Gaze360 and MPIIGaze, achieving mean angular errors of 11.53° and 9.59° for front 180° and front-facing gaze angles, respectively, on the Gaze360 testing set, and a mean angular error of 3.97° on the MPIIGaze testing set, outperforming the CNN-based gaze estimation method. The BoT2L-Net model proposed in this paper performs well on two publicly available datasets, demonstrating the effectiveness of our approach.

Keywords: unconstrained gaze estimation; Bottleneck transformer; combined loss function

Citation: Wang, X.; Zhou, J.; Wang, L.; Yin, Y.; Wang, Y.; Ding, Z. BoT2L-Net: Appearance-Based Gaze Estimation Using Bottleneck Transformer Block and Two Identical Losses in Unconstrained Environments. *Electronics* **2023**, *12*, 1704. https://doi.org/10.3390/electronics12071704

Academic Editors: Pratheepan Yogarajah, Muthu Subash Kavitha, Lamiaa Abdel-Hamid and Ananthakrishnan Balasundaram

Received: 9 March 2023
Revised: 28 March 2023
Accepted: 30 March 2023
Published: 4 April 2023

1. Introduction

Gaze estimation refers to the process of estimating the gaze direction of the eyes. This cue is crucial for nonverbal communication, as it provides insights into a person's level of engagement, interest, and attention during social interactions. Furthermore, gaze estimation is also one of the essential cues of many applications across a variety of fields, including saliency detection [1,2], virtual reality [3], first-person video analysis [4], human-computer interaction [5,6], affective computing [7], and medical diagnosis [8], etc.

There are two main categories of gaze estimation methods: model-based and appearance-based. Model-based gaze estimation methods utilize geometric models to calculate the gaze of human eyes [9–11]. However, these methods typically require specialized hardware, which limits their applicability in real-world environments. The appearance-based methods directly extract the gaze point from captured images, making it straightforward to estimate gaze in unconstrained environments. Recently, the field of gaze estimation has benefited greatly from the rapid development of deep learning techniques. As a result,

many researchers have applied deep learning algorithms to gaze estimation, resulting in the development of methods such as AGE-Net [12], CA-Net [13], and Dilated-Net [14]. Among these techniques, appearance-based gaze estimation has emerged as the most widely used method.

The earliest attempt to use neural networks for gaze estimation utilized monocular images as input [15]. Subsequent related research, typically employs popular backbone networks such as VGG [16], ResNet18 [17], and ResNet50 [18], to extract features, and then outputs the gaze direction after processing, and then processes them to output the gaze direction. These networks can take a single input [17,18], such as a face or eye image, to estimate gaze direction, or multiple inputs [12], such as face, eye images, and head pose, which are combined to estimate gaze direction.

The goal of gaze estimation, is to learn a mapping function that accurately predicts human eye gaze from facial appearance [19]. However, this is a complex problem, as factors such as illumination, environment, and personal head posture can all impact the accuracy of gaze point estimation. To address these challenges, the mapping function must be highly non-linear and capable of effectively integrating overall appearance information. Although numerous gaze estimation models based on convolutional neural networks (CNN) have been developed, they still lack the ability to capture the overall relationship between the facial appearance and the gaze accurately.

Transformer [20], proposed by Vaswani et al., has demonstrated exceptional performance in NLP (natural language processing) tasks. Recently, researchers have applied Transformer to computer vision tasks, with great success. In comparison to CNN, transformers excel at capturing global relations. One such approach is the Vision Transformer (ViT) [21], which employs a pure Transformer model for image classification and achieved superior performance compared to the state-of-the-art CNN models. However, since the self-attention storage and computation increase quadratically with the spatial dimension, it inevitably brings heavy computational costs. To mitigate this, some methods have attempted to incorporate transformers into the CNN backbone network or replace the convolutional block with an attention layer [22,23].

BoTNet [23] is an architecture that integrates the self-attention mechanism into various computer vision tasks. By utilizing Transformer's multi-head self-attention layer, instead of spatial convolution, in the last three Bottleneck blocks of ResNet50, BoTNet significantly enhances the baseline performance of instance segmentation and object detection. This is achieved while reducing the number of parameters needed and minimizing the delay overhead.

The task of gaze estimation is a type of numerical regression problem, commonly addressed through the use of the mean squared error (MSE) loss function for back-propagation. Petr et al. [17] proposed the pinball loss to estimate gaze direction and error boundaries together, which improves accuracy, particularly in unconstrained environments. However, its robustness and generalization ability still need to be improved.

This article proposes a novel 3D gaze estimation model based on Bottleneck Transformer and a multi-loss method in an unconstrained environment. The proposed model, called Bot2L-Net, incorporates self-attention using the Bottleneck Transformer block, which enhances the network's globality. Bot2L-Net is mainly composed of three Resnet Bottleneck blocks and three Bottleneck Transformer blocks. The model uses two identical loss functions, which are applied to classify and regress each gaze angle. These loss functions are then propagated back through the network to adjust the weight of the Bot2L-Net parameters, resulting in improved accuracy. The proposed model was evaluated on two commonly used gaze datasets: Gaze360 and MPIIGaze. On the Gaze360 testing set, the model achieved a mean angular error of 11.53° and 9.59° for front 180° and front-facing angles, respectively. On the MPIIGaze testing set, the mean angular error was 3.97°. Our model has demonstrated exceptional performance on two datasets, outperforming current CNN-based gaze estimation methods in terms of accuracy. The contributions of this paper can be summarized as follows:

- This paper proposes a gaze estimation network designed to operate in unconstrained environments. The network utilizes Bottleneck Transformer blocks to introduce self-attention, allowing it to be connected to Transformer. This design results in better overall capture capabilities while also requiring fewer parameters.
- We employ two identical loss functions to predict pitch and yaw angles. By combining the cross-entropy loss function with the MSE loss function, the resulting combined loss function achieves a lower angle error in the network.
- Further, we conduct a verification and comparison of the mean angular error of our model on the Gaze360 and MPIIGaze testing sets. Our results demonstrate that our model has a lower mean angular error and can accurately estimate gaze in unconstrained environments.

2. Related Work

2.1. Gaze Estimation Methods

2.1.1. Model-Based Gaze Estimation

Model-based gaze estimation methods typically use infrared light sources [9,10,24] to perform corneal reflections [11] on the eye to detect eyeball features, thereby estimating gaze. The accuracy of this gaze estimation method depends on the specific eye features of different testers, so additional calibration steps are often required. It is highly sensitive to input noise such as partial occlusion or illumination interference, and requires high-resolution images and fixed and uniform illumination. Therefore, this method cannot perform gaze estimation in an unconstrained environment.

2.1.2. Appearance-Based Gaze Estimation

Appearance-based gaze estimation techniques utilize extensive datasets of annotated eye or face images, to acquire a more direct mapping between images and gaze. Support vector regression [25], random forest [26], and deep learning-based gaze estimation methods, which are currently of great interest to researchers, are all applied in this way.

Zhang et al. pioneered the use of neural networks for gaze estimation [15], and contributed to the development of one of the most widely used gaze datasets in the field today: MPIIGaze. They also proposed a simple VGG-based architecture for predicting gaze using single-eye images [16]. In 2017, a full-face gaze estimation method was introduced, that leverages the attention mechanism [19]. This approach learns the weights of each position in the face region, to increase the importance of the eye region, and suppress the weights of irrelevant regions, resulting in higher accuracy. At the same time, they utilized gaze data from the MPIIGaze dataset and added full-face images, to propose the MPIIFaceGaze dataset. The method achieved an error of 4.8° on the MPIIFaceGaze dataset.

Cheng et al. [27] have proposed an innovative approach to asymmetric regression using two eyes. This method takes two eye inputs and assigns different weights based on the actual situation. Yu et al. [28] have developed a constraint model-based gaze estimation method, utilizing the concept of multi-task learning. In this approach, the eyes are detected while gaze estimation is carried out, and two learning tasks are performed simultaneously, to complement the information. Chen et al. [14] employed a null convolutional network to detect subtle alterations in eye images. Furthermore, they extended their research by introducing GEDD-Net [29], which employs both dilation convolution and gaze decomposition, resulting in better performance than using only dilation convolution. Wang et al. [30] utilized adversarial learning to align CNN-extracted features and improve gaze generalization performance.

Fischer et al. [31] developed a method to predict gaze angles by combining head pose vectors and VGG CNN features with eye crops. Kellnhofer et al. proposed Gaze360 [17], a large-scale gaze dataset and method for 3D gaze estimation in natural environments. They employed a temporal model with seven-frame sequences (LSTM) to predict gaze angle, and used pinball loss joint regression of gaze direction and error bounds to enhance gaze accuracy.

Cheng et al. [32] proposed the FAR-Net, which estimates the 3D gaze angle of both eyes using an asymmetric method, inspired by the asymmetric property of eyes. The model showed good performance on several publicly available datasets. In addition, Cheng et al. [13] proposed a coarse-to-fine adaptive network (CA-Net). The network leverages a face image to initially predict the primary gaze angles, which are further refined through the integration of residual estimates obtained from eye crops. Furthermore, they proposed a bi-gram model, to establish a link between the primary gaze angles and the residual estimates derived from the eyes.

Biswas et al. [12] proposed a gaze estimation network named AGE-Net, which incorporates the idea of attention mechanism. The network employs an attention branch to assign different weights to the features extracted from eye images, allowing for a more precise inference of the gaze point. Moreover, the network further refines the gaze output obtained from face images, resulting in improved prediction accuracy.

All of the aforementioned studies are based on CNN, which has a notable drawback: the convolution operation can only capture local information, and cannot establish long-distance connections for modeling the global image. This limitation weakens the ability of CNNs to capture global relationships, and makes it difficult to achieve further improvements in accuracy.

2.2. Transformer

Transformer [20], was proposed by Vaswani et al. and has been widely used in NLP (natural language processing), due to its excellent performance [33]. Recently, researchers have been exploring the use of Transformer in computer vision tasks, such as image classification and object detection.

Vision Transformer (ViT) [21] is a pioneering work, that introduced Transformer into image classification and achieved comparable or even better results than traditional convolutional neural networks (CNNs) on mainstream classification benchmarks. Other works have combined CNN and Transformer to achieve better performance in object detection tasks [34]. One such approach is Detection Transformer (DETR) [20], which treats object detection as a straightforward ensemble prediction problem and uses a Transformer encoder–decoder architecture as the detection head. DETR achieved competitive results on the quantitative evaluation of the COCO (common objects in context) dataset [35], a commonly used dataset for object detection tasks. Overall, the integration of Transformer into computer vision tasks has shown promising results and represents an exciting direction for future research.

The backbone network extended by Transformer can be divided into seven categories according to motivation and implementation [36], as shown in Figure 1. The original Visual Transformer (ViT) is a well-known method that has been widely adopted. Another method is the Transformer-enhanced CNN, which uses the strong global modeling ability of Transformer, to improve the long-distance dependence of the CNN backbone. In contrast, the CNN-enhanced Transformer introduces convolution-induced bias, to enhance Transformer's performance. Local attention-enhanced Transformer is another method that maintains a convolution-free architecture while enhancing the locality of Transformer, by redesigning block partitions and attention blocks. The Hierarchical Transformer, replaces the fixed-resolution columnar structure with a pyramid-shaped backbone. Additionally, the Deep Transformer prevents overly smooth attention maps, by increasing their diversity in the deep layers. Lastly, Transformer with self-supervised learning, is an approach that uses self-supervised learning techniques to enhance the efficiency of Transformer.

Transformer has a stronger modeling ability than CNN in theory. However, since the self-attention mechanism increases quadratically with the feature dimension, it inevitably brings heavy computational costs. To address this issue, Cordonnier et al. demonstrated that a convolutional layer can be approximated by a sufficient number of heads in MHSA (multi-headed self-attention) [37]. Some attempts have been made to insert Transformer into CNN or replace convolutional blocks with attention layers [22,23]. These methods aim

to enhance the modeling ability of CNN while maintaining its efficiency, by incorporating the advantages of Transformer.

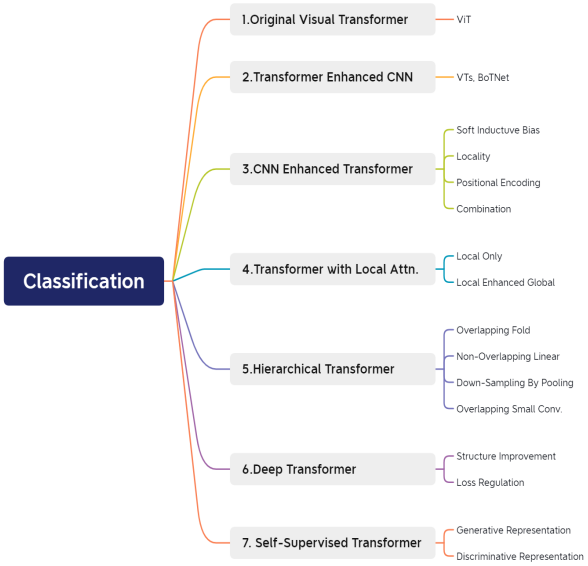

Figure 1. Vision Transformer for classification.

BoTNet [23] is a highly effective and versatile backbone architecture, that leverages self-attention mechanisms to support a range of computer vision tasks such as image classification, object detection, and instance segmentation. It proposes a novel approach, where consecutive Bottleneck blocks with self-attention can be treated as a Bottleneck Transformer block, by replacing the spatial convolution with global self-attention in the last three Bottleneck blocks of ResNet. Additionally, BoTNet achieves 84.7% top-1 accuracy with 75.1 M parameters on the ImageNet benchmark, surpassing most CNN models with similar parameter settings. These results demonstrate the effectiveness of the Transformer approach on standard convolutional models, and highlight the breakthrough achieved by BoTNet over the baseline.

3. Method

To extract image features and improve downstream tasks, many network architectures have been proposed, including various network structures with CNNs and Transformer as the backbone. The gaze estimation task is modeled as a regression from a normalized face image to a pitch-bias gaze direction vector. Our goal is to perform 3D gaze estimation in unconstrained environments, which requires training on a large dataset. Although the global integrity of Transformer is good, the self-attention mechanism leads to a quadratic increase in storage and computation, with respect to the spatial dimension. Training on a very large dataset, such as Gaze360, inevitably incurs heavy computational costs.

3.1. Bottleneck Transformer

In this section, we will introduce the basic structure and functions of the Bottleneck Transformer, which is a variation of the Transformer architecture widely used in NLP and CV (computer vision) tasks.

The ResNet Bottleneck block, with an MHSA layer, can be viewed as a Transformer block with a bottleneck structure and minor differences in modules. Thus, the ResNet Bottleneck block with the MHSA layer, is referred to as the Bottleneck Transformer block in [23]. Figure 2 [23] shows how to convert the Resnet Bottleneck block to the Bottleneck Transformer block. The key distinction lies in the replacement of the spatial 3×3 convolutional layer with MHSA. The symbols $\bigoplus$ represents element-wise addition.

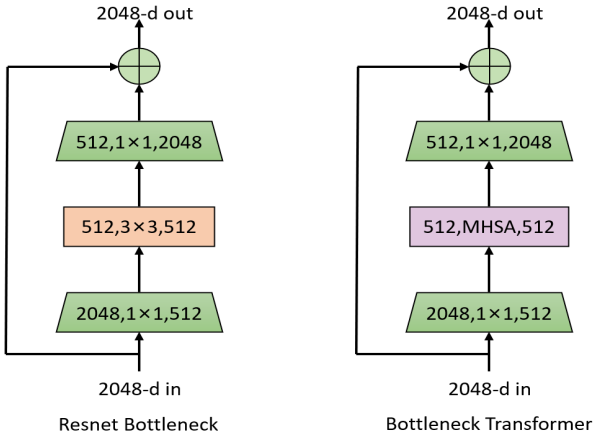

Figure 2. Comparison of Resnet Bottleneck and Bottleneck Transformer.

The structure of the self-attention layer is described in Figure 3 [23]. All-to-all attention is carried out on a 2D feature map, with split relative position encodings for both height and width (Rh and Rw). The attention logits are computed as qkT + qrT, where q, k, and r correspond to the query, key, and position encodings, respectively. In the BotNet architecture, relative distance encodings are utilized [38,39]. The symbols $\oplus$ and $\otimes$ represent element-wise addition and matrix multiplication, respectively, while 1×1 represents a pointwise convolution.

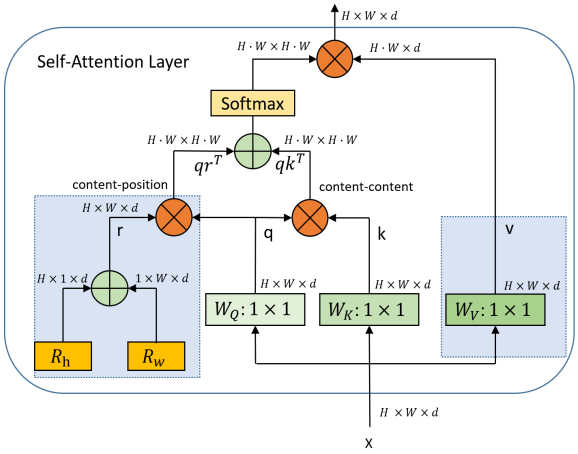

Figure 3. Multi-head self-attention (MHSA) layer used in the BoT block.

The paper [23], proposes a practical and straightforward example called BoTNet, which is connected to Transformer through BoT blocks. The approach involves replacing the last three Bottleneck blocks of ResNet50 with BoT blocks, specifically by substituting only the last three 3×3 convolutions. This replacement facilitates the network to learn the global features of the input. BoTNet achieves the highest top-1 accuracy of 84.7% on the ImageNet validation set. Additionally, the computation time on the TPU-v3 hardware is 1.64 times faster than the popular EfficientNet model.

BoTNet utilizes a hybrid design of convolution and global self-attention. Convolution efficiently learns abstract low-resolution feature maps from large images, while global self-attention processes and aggregates the information contained within these feature maps, so that convolution can be spatially downsampled, and attention can be applied to smaller resolutions, to effectively process large images. To enable attention to understand the relationship between objects and their positions, BoTNet employs 2D relative position self-attention, as described in [38,39].

Inspired by BoTNet, we used three Resnet Bottlenecks, three Bottleneck Transformers, and two fully connected layers to form a backbone network, and connect the network to Transformer through BoT block, to incorporate self-attention.

3.2. Loss Function

According to the data distribution characteristics of different machine learning tasks, it is often necessary to choose different loss functions, to achieve better results. The loss function commonly used in classification problems is cross-entropy loss. The L1 loss (mean absolute error, MAE) function and L2 loss (mean squared error, MSE) function are usually used in regression problems. Generally, when outliers are important to the task and cannot be discarded, the L2 loss function should be used. However, if the outliers only represent useless data, or data that is considered to be corrupted, the L1 loss function is more appropriate.

Three-dimensional gaze estimation is a numerical regression problem in machine learning. Most gaze estimation models based on neural networks use the L2 loss function for regression. In addition, for the gaze estimation problem, an angle error loss function can also be used.

We used both the cross-entropy and L2 loss functions in our approach, employing identical losses for each gaze angle. The cross-entropy loss function is utilized to forecast binned gaze classification. Additionally, we estimated the expectation of the binned gaze, to enhance the prediction accuracy. The L2 loss function is used to penalize the network, leading to a lower angle error. By combining both the loss functions, we can improve the performance of the network in terms of gaze estimation.

MSE is a popular loss function used in regression problems, including gaze estimation. It measures the average squared difference between the predicted values and the actual values of the target variable. Mathematically, the MSE is defined as the average of the squared differences between the predicted value, y^p, and the actual value y:

$$MSE(y, y^p) = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^p)^2 \tag{1}$$

The cross-entropy loss function is a commonly used loss function in classification problems. When used with the softmax layer, the cross-entropy loss function can be expressed as follows:

$$CE(y, y^n) = \sum_{i=1}^{n} y_i \log y_i^p \tag{2}$$

The proposed loss function for each gaze angle is a combination of mean squared error and cross-entropy loss. Specifically, it is defined as follows:

$$CL(y, y^p) = MSE(y, y^p) + CE(y, y^p) \tag{3}$$

As shown in Figure 4, where CL is combined with MSE and cross-entropy loss. Here, y_i represents the true values of the i-th sample, y_i^p represents the predicted values of the i-th sample, and n is the total number of samples.

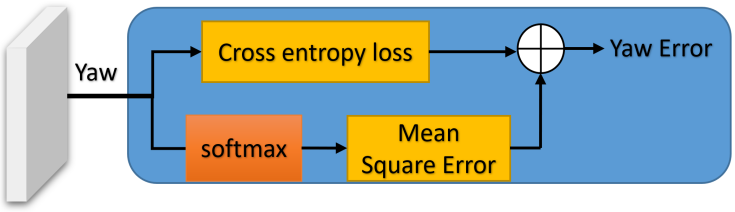

Figure 4. Combined loss. The symbols ⊕ represents element-wise addition.

In gaze estimation tasks, the input resolution is usually set to 224 × 224. However, to achieve higher accuracy in unconstrained gaze estimation, we need to utilize self-attention in more realistic scenarios. Considering that global self-attention across n entities requires $O(n^2d)$ memory and computation [20], we adopted the method of Transformer-enhanced CNN, which mainly used three Resnet Bottlenecks and three Bottleneck Transformers to form a network, and connected the network to Transformer through BoT block, to incorporate self-attention.

We designed a network based on BoTNet, as shown in Figure 5. We used stride 2 convolution and maximum pooling layer to downsample the input feature map, and then connected three Resnet Bottlenecks. In contrast to previous work [23], we did not change it on ResNet50, we only connected three Bottleneck Transformers with self-attention mechanism. The network was connected to Transformer through BoT block, to obtain a 2048-dimensional feature, and then the gaze estimation was performed through an average pooling layer and two fully connected layers. We used two identical loss functions, one for estimating yaw and one for estimating pitch. The two angles were back-propagated separately through two signals, which could adjust the network weight parameters more accurately and improve the accuracy. In addition, our number of network parameters was 41.2% less than ResNet50, as shown in Table 1.

Table 1. Comparison of BoT2L-Net and ResNet-50 network structures and parameters.

Stage	Output	ResNet-50	BoT2L-Net
C1	112×112	$7 \times 7, 64$, *stride* 2	$7 \times 7, 64$, *stride* 2
C2	56×56	3×3 *maxpool*, *stride* 2 $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	3×3 *maxpool*, *stride* 2 $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
C3	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$	
C4	14×14	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ MHSA, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
C5	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	
	1×1	Average pool 1000-d fc	Average pool 512-d fc 90-d fc ×2
params		23.5×10^6	13.8×10^6

3.3. BoT2L-Net

To process the output of the fully-connected layer in the gaze estimation model, we performed a series of steps. Firstly, we computed the binned classification loss, using cross-entropy between the output probability and the target label. In Gaze360, the gaze angles are divided into 90 classes, while in MPIIGaze, the gaze angles are divided into 28 classes. Secondly, we applied a softmax function to the network output logits, to obtain a probability distribution over the gaze angles. The expectation of this distribution was

then calculated, to obtain the predicted gaze direction. Finally, we calculated the MSE between the predicted gaze direction and the ground truth gaze direction, and added this MSE term to the classification loss to obtain the final loss.

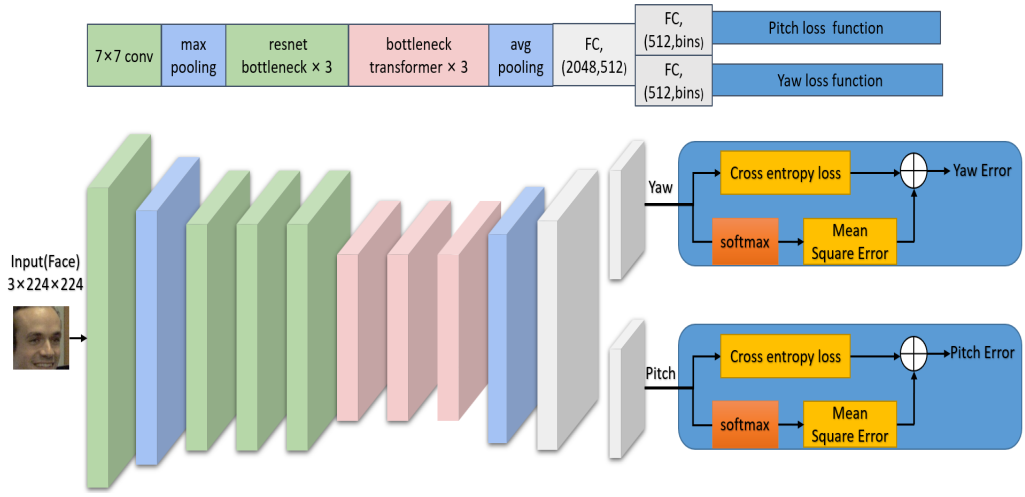

Figure 5. Bot2L-Net architecture.

3.4. Dataset

With the development of gaze estimation tasks, more and more gaze datasets have been proposed. Some datasets are captured using still recording setups [40,41], or cameras integrated into smartphones [3]. In order to enhance gaze estimation performance, large-scale datasets were proposed. To achieve our objective of unconstrained gaze estimation, we have opted to consider three widely used unconstrained datasets, namely, ETH-XGaze [18], Gaze360 [8], and MPIIGaze [19].

ETH-XGaze uses 4 light boxes and 18 high-definition cameras (6000 × 4000) to capture a wide range of head poses and gaze directions, under different lighting conditions. This dataset recorded nearly 600 sight directions of 110 subjects under 16 different lighting conditions, and collected a total of 1,083,492 images; the maximum yaw angle and pitch angle of the sight directions collected were ±120° and ±70°, respectively.

Gaze360 is a dataset that is designed for robust 3D gaze estimation in unconstrained images. It is one of the most extensive datasets available for gaze tracking and provides a wide range of 3D gaze annotations that cover a 360° range. Gaze360 requires the subjects to look at a moving target and uses multiple cameras to obtain the gaze direction of multiple subjects at the same time. The dataset collected 172,000 eye-sight data from 238 subjects in 5 indoor scenes and 2 outdoor scenes, including different backgrounds, time, and lighting.

MPIIFaceGaze is a dataset that contains facial images and corresponding gaze positions. MPIIFaceGaze is an extension of the MPIIGaze dataset, which matches gaze positions from the MPIIGaze dataset with full facial images, to more accurately reflect human gaze behavior. This dataset exhibits significant variations in appearance and lighting and is suitable for unconstrained gaze estimation.

However, since actual application images used for gaze estimation may not be ultra-high definition, the Gaze360 and MPIIFaceGaze datasets are suitable for model evaluation when using a training set. Figure 6 shows some sample pictures from the Gaze360 dataset and MPIIFaceGaze dataset.

Figure 6. Some pictures from the Gaze360 and MPIIFaceGaze datasets. (**Left**) Gaze360; (**Right**) MPIIFaceGaze.

4. Experiment

4.1. Setup

4.1.1. Data Preprocessing

In this paper, the normalization of images in MPIIFaceGaze and Gaze360 was performed using the processing method described in [15]. Specifically, a virtual camera was rotated and translated, to remove the head roll angle and maintain the same distance between the virtual camera and the reference point (i.e., the center of the face). In order to estimate gaze classification, the continuous gaze angles in each dataset were split into sets of gaze directions, with classification labels based on the range of gaze annotations, with 90 classes for Gaze360 and 28 classes for MPIIFaceGaze. Therefore, both datasets have continuous and binned labels, and multiple loss functions were used for classification and regression, to improve network accuracy.

The MPIIFaceGaze dataset is an extension of the MPIIGaze dataset, and it includes an evaluation set that consists of 15 subjects, with 3000 face images per subject. This dataset is specifically designed for gaze estimation research, and it provides a corresponding face image for each eye image in the MPIIGaze dataset. To evaluate the performance of gaze estimation algorithms on this dataset, the evaluation protocol used in [19] was followed. Normalization was performed on the face images, and leave-one-out cross-validation was used to evaluate the algorithms. This means that for each subject in the dataset, the algorithm is trained on all the other subjects and then tested on the remaining subject. This process is repeated for each subject in the dataset, and the results are averaged to provide an overall evaluation of the algorithm's performance.

For the Gaze360 dataset, the entire dataset was split into training, testing, and evaluation sets, using the method described in [17]. Based on the phi-ai laboratory's approach, the dataset was refined by removing images that did not have face detection results, as determined by the face detection annotations provided. Because some images in the Gaze360 dataset only show the subject's back, they are not suitable for appearance-based methods that rely on facial features. Therefore, these images were excluded from the dataset, to ensure the accuracy and effectiveness of appearance-based methods. The training set contains 80,942 images, the validation set contains 11,318 images, and the testing set contains 16,031 images. Examples of processed images in Gaze360 are shown in Figure 7.

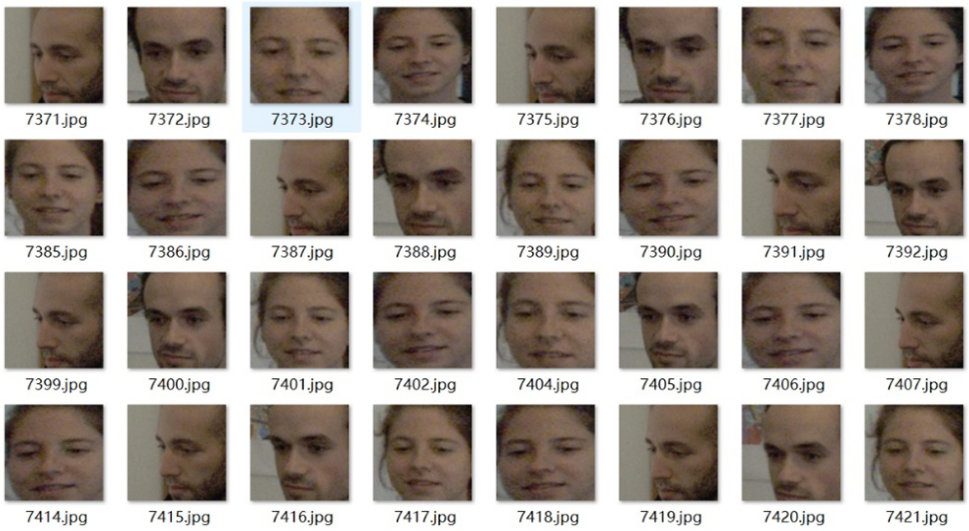

Figure 7. Processed images in Gaze360.

4.1.2. Training

Our proposed network (BoT2L-Net) is based on the BoT block in BotNet. There is currently no public pre-trained model, so we retrained the model ourselves. In order to meet the baseline [18] for gaze estimation, and to consider memory constraints, we uniformly resized the input data to 224 × 224 and normalized the data.

Training Gaze360 dataset: we trained the model using the AdamW optimizer, in the PyTorch 1.10 framework, with an initial learning rate of 0.0001 and exponential decay, with a decay rate of 0.97. The Gaze360 dataset has three evaluation ranges based on the range of gaze angles: full 360°, fronting 180°, and front-facing (within 20°). Based on the data processing and cleaning described earlier, we trained and evaluated the model on the front 180° and front-facing (within 20°) datasets. We trained the model for 150 epochs, and the mean angular error loss curve on the testing set is shown in Figure 8.

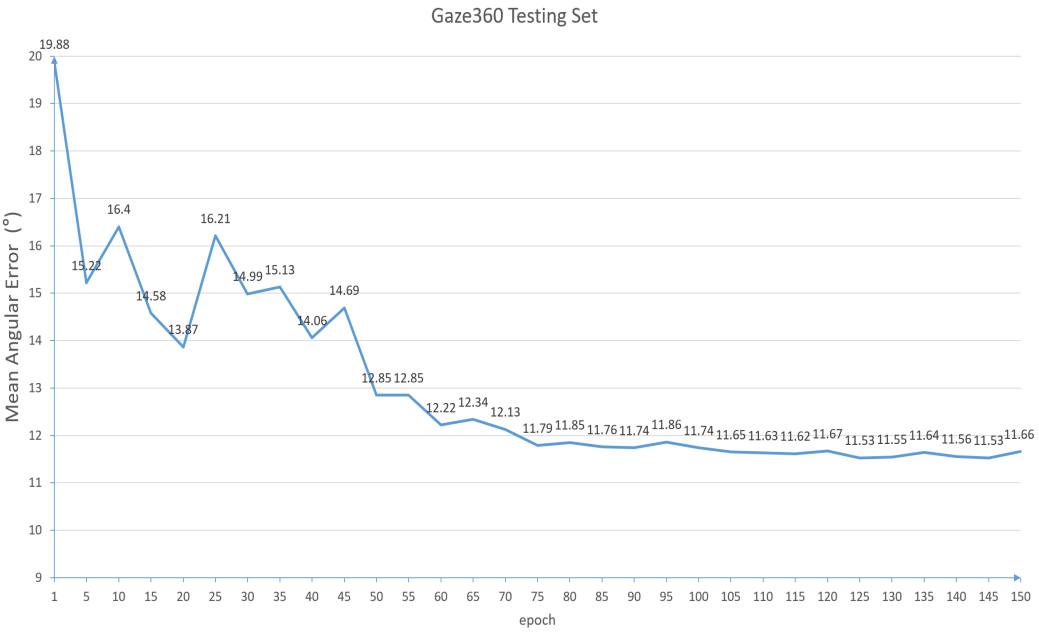

Figure 8. The mean angular error of the BoT2L-Net model trained for 150 epochs on the Gaze360 testing set.

Training MPIIGaze dataset: we employed leave-one-out cross-validation, as used in related work [12,13,19]. The dataset was split into 15 folds, and each time we trained the model using 14 folds as the training set and one fold as the testing set. For example, during the first training run, we used fold 0 as the testing set, and folds 1 to 14 as the training set. In the second run, we used fold 1 as the testing set and folds 0 and 2 to 14 as the training set. We continued this process until we had used all folds as the testing set once. The model was trained using the AdamW optimizer in the PyTorch 1.10 framework, with an initial learning rate of 0.0001 and exponential decay, with a decay rate of 0.97. We trained the model for 30 epochs and obtained 450 weight files.

4.2. Evaluation and Results

After estimating the pitch angle and yaw angle, the model can predict a 3D vector representing the gaze direction. The most commonly used evaluation index in the gaze field is the mean angular error (°), which is the angle between the predicted gaze direction and the ground truth gaze direction. The mean angular error (°) can be computed as follows:

$$E_{angular} = arccos \frac{g \cdot g^p}{|g||g^p|} \tag{4}$$

where g means ground truth gaze, g^p means predicted gaze, and $\cdot$ represents the dot product of two vectors. $|g|$ and $|g^p|$ represent the magnitudes of the two vectors, and *arccos* is the inverse cosine function.

We evaluated our proposed network on the Gaze360 dataset and compared its performance with that of state-of-the-art gaze estimation methods. We adhered to the original train–validation–test split of the datasets throughout the entire training and testing process. All the results presented below were obtained from the testing set. Table 2 shows the comparison of the mean angular errors between our proposed model and the state-of-the-art methods, on the testing set of the Gaze360 dataset. On the testing set, our proposed BoT2L-Net achieved average angular errors of 11.53° and 9.59° for the front 180° and front-facing (within 20°), respectively. The average angular error for the front 180° is 11.2% higher than that of the CNN-based gaze estimation method [13]. Since other methods are not evaluated in the front-facing, only the Gaze360 model could be compared. The mean angular error in the front-facing is 22.6% higher than that of the Gaze360 model [17].

Table 2. Comparison of mean angular error between our proposed model and state-of-the-art methods on the Gaze360 testing set.

Method	MPIIFaceGaze
MPIIGaze [16]	5.4°
AR-Net [27]	5.0°
Full-Face [19]	4.8°
Dilated-Net [14]	4.8°
GEDD-Net [29]	4.5°
FAR-Net [32]	4.3°
CA-Net [13]	4.1°
AGE-Net [12]	4.09°
Bot2L-Net (ours)	3.97°

We employed the leave-one-out cross-validation strategy, used in related works [12,18,31], to validate our model on the MPIIFaceGaze dataset. Table 3 shows a comparison between our proposed model and the state-of-the-art method, in terms of mean angular error, on the MPIIFaceGaze testing set. As previously mentioned, the MPIIFaceGaze dataset was tested using leave-one-out cross-validation, with different training

and testing sets in each round. The average of 15 rounds was taken for evaluation. BoT2L-Net achieved a mean angular error of 3.97°, which is 0.12° better than AGE-Net. We also provided the mean angular error of BoT2L-Net for each subject tested on the MPIIFaceGaze dataset and compared it with FAR-Net [32] and AGE-Net [12], as shown in Figure 9.

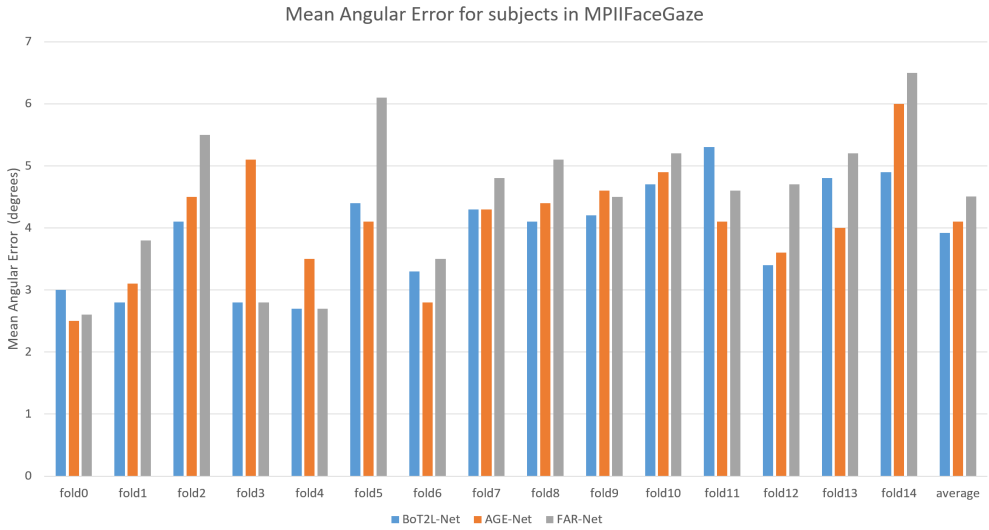

Figure 9. Mean Angular Error for subjects in the MPIIGaze testing set on BoT2L-Net, AGE-Net, and FAR-Net.

Table 3. Comparison of mean angular error between our proposed model and state-of-the-art methods, on the MPIIGaze testing set.

Method	Front 180°	Front Facing
Full-Face [19]	14.99°	N/A
Dilated-Net [14]	13.73°	N/A
RT-Gene [31]	12.26°	N/A
CA-Net [13]	12.20°	N/A
Gaze360 (LSTM) [17]	11.40°	11.10°
Bot2L-Net (ours)	11.53°	9.59°

5. Prediction in Unconstrained Environments

The BoT2L-Net model proposed in this paper, performs well on two unconstrained gaze estimation datasets and improves the previous methods, in terms of the evaluation metric of mean angular error. To verify the practical application of our model, we used RetinaFace [42] as the face detector and combined it with our proposed model. The results show that our model could achieve accurate gaze estimation in unconstrained environments, allowing for high-precision estimation of multiple gaze directions. Figure 10 shows that our results are comparable to subjective human judgments. This means that, based on the gaze results, we can use such nonverbal cues to gain insight into a person's level of interest and attention during social interactions.

Figure 10. Gaze estimation in the video. The green frames represent the result of face detection and the red arrows represent the result of gaze estimation.

6. Conclusions

In this work, we proposed a new model for 3D gaze estimation in unconstrained environments, which incorporates a Bottleneck Transformer block and a multi-loss method. The network architecture is designed to integrate self-attention through BoT block, and is trained with two identical loss functions, using back-propagation. We have evaluated the performance of the proposed model on two widely used 3D gaze datasets, Gaze360 and MPIIGaze, and have demonstrated its effectiveness on unconstrained images in video. However, the Transformer-enhanced-CNN method still has certain limitations. The network in this paper mainly consists of three ResNet Bottleneck blocks and three Bottleneck Transformer blocks, introducing a self-attention mechanism. The self-attention can result in a very large number of floating-point computations, so our model is only trained with a shallow network. If we continue to increase the network's depth, it requires devices with extremely high computing power for training and computation. In the future, when computing power permits, we can experiment with deeper networks. We hope that the results obtained from the gaze estimation task using our proposed model, will encourage future researchers to continue applying Transformer or Transformer-related methods to gaze estimation tasks, thereby improving the model's overall performance and robustness in unconstrained environments.

Author Contributions: Conceptualization, X.W., J.Z. and L.W.; methodology, X.W. and J.Z.; software, X.W., Y.Y. and Y.W.; validation, X.W.; resources, J.Z. and Z.D.; data curation, L.W. and Y.W.; writing—original draft preparation, X.W.; writing—review and editing, X.W. and J.Z.; visualization, X.W. and Y.Y.; supervision, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Key R&D plan of Shandong Province (2020JMRH0101), National Deep Sea Center.

Data Availability Statement: The publicly available datasets used in this research can be obtained through the following links: MPIIFaceGaze: https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/gaze-based-human-computer-interaction/its-written-all-over-your-face-full-face-appearance-based-gaze-estimation (accessed on 8 March 2023). Gaze360: http://gaze360.csail.mit.edu/download.php (accessed on 8 March 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, W.; Shen, J.; Dong, X.; Borji, A.; Yang, R. Inferring salient objects from human fixations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1913–1927. [CrossRef] [PubMed]
2. Wang, W.; Shen, J. Deep visual attention prediction. *IEEE Trans. Image Process.* **2017**, *27*, 2368–2378. [CrossRef] [PubMed]
3. Xu, Y.; Dong, Y.; Wu, J.; Sun, Z.; Shi, Z.; Yu, J.; Gao, S. Gaze prediction in dynamic 360 immersive videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5333–5342.
4. Yu, H.; Cai, M.; Liu, Y.; Lu, F. First-and third-person video co-analysis by learning spatial temporal joint attention. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef] [PubMed]
5. Hempel, T.; Al-Hamadi, A. Slam-based multistate tracking system for mobile human-robot interaction. In Proceedings of the Image Analysis and Recognition: 17th International Conference, ICIAR 2020, Póvoa de Varzim, Portugal, 24–26 June 2020; pp. 368–376.
6. Strazdas, D.; Hintz, J.; Khalifa, A.; Abdelrahman, A.A.; Hempel, T.; Al-Hamadi, A. Robot systemassistant (RoSA): Towards intuitive multi-modal and multi-device human-robot interaction. *Sensors* **2022**, *22*, 923. [CrossRef] [PubMed]
7. D'Mello, S.; Olney, A.; Williams, C.; Hays, P. Gaze tutor: A gaze-reactive intelligent tutoring system. *Int. J. Hum.-Comput. Stud.* **2012**, *70*, 377–398. [CrossRef]
8. Jiang, M.; Zhao, Q. Learning visual attention to identify people with autism spectrum disorder. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3267–3276.
9. Hennessey, C.; Noureddin, B.; Lawrence, P. A single camera eye-gaze tracking system with free head motion. In Proceedings of the 2006 Symposium on Eye Tracking Research & Applications, San Diego, CA, USA, 27–29 March 2006; pp. 87–94.
10. Yoo, D.H.; Chung, M.J. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Comput. Vis. Image Underst.* **2005**, *98*, 25–51. [CrossRef]
11. Huang, M.X.; Li, J.; Ngai, G.; Leong, H.V. Screenglint: Practical, in-situ gaze estimation on smartphones. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 2546–2557.
12. Biswas, P. Appearance-based gaze estimation using attention and difference mechanism. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA , 19–25 June 2021; pp. 3143–3152.
13. Cheng, Y.; Huang, S.; Wang, F.; Qian, C.; Lu, F. A coarse-to-fine adaptive network for appearance-based gaze estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 10623–10630.
14. Chen, Z.; Shi, B.E. Appearance-based gaze estimation using dilated-convolutions. In Proceedings of the Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 309–324.
15. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Appearance-based gaze estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, Boston, MA, USA, 7–12 June 2015; pp. 4511–4520.
16. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 162–175. [CrossRef] [PubMed]
17. Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; Torralba, A. Gaze360: Physically unconstrained gaze estimation in the wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6912–6921.
18. Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; Hilliges, O. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 365–381.
19. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. 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, Venice, Italy, 22–29 October 2017; pp. 51–60.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Singapore, 29–30 March 2021.
22. Wu, B.; Xu, C.; Dai, X.; Wan, A.; Zhang, P.; Yan, Z.; Tomizuka, M.; Gonzalez, J.; Keutzer, K.; Vajda, P. Visual transformers: Token-based image representation and processing for computer vision. In Proceedings of the International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2020; pp. 579–589.
23. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 16519–16529.
24. Zhu, Z.; Ji, Q. Eye gaze tracking under natural head movements. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; pp. 918–923.
25. Schneider, T.; Schauerte, B.; Stiefelhagen, R. Manifold alignment for person independent appearance-based gaze estimation. In Proceedings of the IEEE/CVF International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 1167–1172.

26. Huang, Q.; Veeraraghavan, A.; Sabharwal, A. Tabletgaze: Dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Mach. Vis. Appl.* **2017**, *28*, 445–461. [CrossRef]
27. Cheng, Y.; Lu, F.; Zhang, X. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 100–115.
28. Yu, Y.; Liu, G.; Odobez, J.M. Deep multitask gaze estimation with a constrained landmark-gaze model. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
29. Chen, Z.; Shi, B.E. Towards High Performance Low Complexity Calibration in Appearance Based Gaze Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1174–1188. [CrossRef] [PubMed]
30. Wang, K.; Zhao, R.; Su, H.; Ji, Q. Generalizing eye tracking with bayesian adversarial learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 11907–11916.
31. Fischer, T.; Chang, H.J.; Demiris, Y. Rt-gene: Real-time eye gaze estimation in natural environments. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 334–352.
32. Cheng, Y.; Zhang, X.; Lu, F.; Sato, Y. Gaze estimation by exploring two-eye asymmetry. *IEEE Trans. Image Process.* **2020**, *29*, 5259–5272. [CrossRef] [PubMed]
33. Radford, A.; Narasimhan, K.; Salimans, T. Improving Language Understanding by Generative Pre-Training. Open AI. Available online: https://openai.com/research/language-unsupervised (accessed on 11 June 2018).
34. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 213–229.
35. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
36. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A survey of visual transformers. *arXiv* **2021**, arXiv:2111.06091.
37. Cordonnier, J.B.; Loukas, A.; Jaggi, M. On the relationship between self-attention and convolutional layers. In Proceedings of the 8rd International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26 April–1 May 2020.
38. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Studying Stand Alone Self-Attention in Vision Models. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
39. Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; Le, Q.V. Attention augmented convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3286–3295.
40. Funes Mora, K.A.; Monay, F.; Odobez, J.M. EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras. In Proceedings of the Symposium on Eye Tracking Research and Applications, Safety Harbor, FL, USA, 22–31 March 2014; pp. 255–258.
41. Smith, B.A.; Yin, Q.; Feiner, S.K.; Nayar, S.K. Gaze locking: Passive eye contact detection for human-object interaction. In Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, St. Andrews, UK, 8–11 October 2013; pp. 271–280.
42. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-shot multi-level face localisation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5203–5212.