2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

A New Efficient Eye Gaze Tracker for Robotic Applications

Chaitanya Bandi[1] and Ulrike Thomas[2]

***Abstract*— Gaze estimation provides insight into a person's intent and engagement level, which is helpful in collaborative human-robot applications. With significant advancements in deep learning architectures, appearance-based gaze estimation has gained much attention. Appearance-based methods have shown significant improvement in gaze accuracy and, unlike traditional approaches, they function well in environments where there are no constraints. We present another convolution-based gaze estimation approach to further reduce the angular error. For estimating gaze under extreme conditions such as head variations and distances, full-face images have been shown to be efficient, so we rely on full-face and pay more attention to necessary features. With the proposed architecture, we achieve an accuracy of 3.75° on the MPIIFaceGaze dataset and 3.96° on the ETH-XGaze open-source dataset. In addition, we test eye gaze tracking in real-time robotic applications, such as attention detection, and pick-and-place.**

I. INTRODUCTION

The person's level of engagement is really helpful in human-robot interaction applications [18] like grabbing the attention of a robot by gazing at it. The willingness to interact can be measured using a non-verbal parameter known as the eye gaze. In addition to robotics, the gaze can be utilized in applications like human-computer interaction [46], [28], [38], virtual reality [33], [23], and behavioral analysis [17]. The estimation of eye gaze is categorized into model-based methods and appearance-based methods [14]. The traditional model-based eye gaze estimation methods [13], [31], [36], [12], [41] are accurate but the environment is highly controlled (i.e., fewer occlusions and stable laboratory settings). The stable laboratory settings also include calibrated cameras, fixed distances, or eye-tracking glasses that stabilize the eye to estimate the gaze. Such calibrated devices pre-assume the eye model and fail to provide results when there are drastic changes to the environment. The human-robot interactions cannot be held in such constrained environments which makes the traditional model-based methods obsolete. The aim of this work is to implement eye gaze tracking in real-world scenarios like Fig. 1 without using eye gaze glasses. The environment setup shown in the left image is for a pick and place task and the right image consists of a pepper robot that interacts with humans by determining the willingness to interact through the gaze.

Due to the limitations of constrained environment techniques, the focus of recent research has shifted toward appearance-based models. Application-specific devices like eye-tracking glasses or calibrated models are not necessary for convolutional neural networks-based (CNNs) techniques as on-the-shelf cameras are sufficient for image processing and gaze regression. Deep convolutional neural networks have been shown to be significant in extracting non-linear and high-level features in an image. Recent studies show that CNN-based architecture regresses the direction of gaze in eye images [47], [43], [10], [6], [30], face images [48], [42], [45], [32] or by combining both face and eye images [24], [3], [4]. As the face image contains all the information related to gaze and face angles, we directly regress the 2D gaze from the CNN architecture. The complete cascaded pipeline is depicted in Fig. 2. The contributions of this work are:

- We introduce a unique architecture that relies on a semantic segmentation from panoptic feature pyramid network [22], residual dilation blocks, and self-attention modules to regress the 2D gaze vector.
- The proposed unique network achieves state-of-the-art results on two large-scale open-source datasets known as the MPIIFaceGaze [48] dataset, and the ETH-XGaze [45] dataset.
- The proposed architecture is applied in real-time robotic applications like attention-grabbing, and pick-and-place.

Fig. 1. Human-Robot interaction environments. Left image: pick and place environment with Franka-Emika Panda robot. Right image: pepper robot interacting with a human.

II. RELATED WORK

Appearance-based methods have shown to be very efficient in cross-subject gaze estimation compared to model-based methods and we review recent deep-learning methods.

A. Appearance-Based Architectures

CNNs have been shown to be significant in many computer vision applications, including eye-gaze estimation that relies on input features. The previous works show evidence of estimating eye gaze from cropped eye region images or full facial images. One of the earliest works in [47] proposes to estimate gaze with just eye images with a simple multimodal

[1]Chaitanya Bandi is with Department of Robotics and Human Machine Interaction Lab, Technical university of Chemnitz, Germany chaitanya.bandi@etit.tu-chemnitz.de

[2]Ulrike Thomas is with Department of Robotics and Human Machine Interaction Lab, Technical university of Chemnitz, Germany ulrike.thomas@etit.tu-chemnitz.de

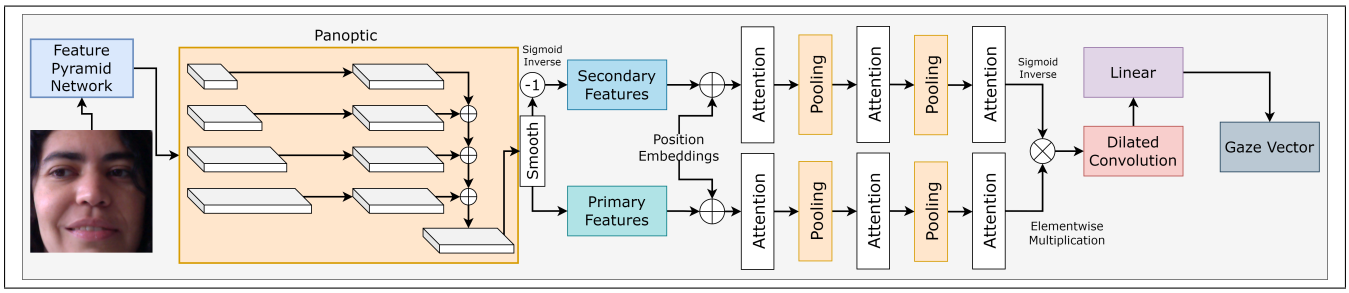

Fig. 2. The proposed architecture for eye gaze regression. The end-to-end architecture is a cascade of multiple mechanisms, the first cascade is a panoptic feature pyramid network for semantic segmentation. The smoothed panoptic features are inversed to obtain the primary and secondary features. Both features are passed through attention-pooling layers which are then element-wise multiplied. The features are then passed through the simple dilated residual block and linearized to obtain the gaze vector.

CNN inspired by LeNet architecture [26]. The authors also introduce the very first high-resolution eye dataset with ground truth gaze known as the MPIIGaze dataset. Later, the work in [43] considers using eye landmarks for further refinement of gaze on the UnityEyes [39] dataset. In natural environments, the distance between the camera and the subject is farther, and the resolution of the eye is quite low. RT-GENE [10] also introduces a large-scale dataset captured using eye tracking glasses which is then sematic inpainted using GAN architecture to remove the glasses from images. The architecture consists of three layered VGG-16 [34] based CNNs where two layers are for two eyes and one layer for the face for regression of eye gaze. With the asymmetry nature of the eyes, the work [6] proposes multiple streams for 3D gaze regression. The works [24], [3], [4], [25], [11], [35], [19], [8] did not only use eye images but also the face images in one way or other to estimate the eye gaze from deep learning architecture. The work in [7] reviews most of the appearance-based eye gaze estimation methods with existing benchmarks.

Recent works focus on regressing the eye gaze using full facial features for high-distortion applications. [48] propose a CNN-based architecture with spatial weights that outperforms eye image-based gaze estimation for regression and the dataset is a subset of the work [47] known as the MPIIFaceGaze dataset. The idea of mixed effects from statistics is implemented in a deep convolutional network to understand the hierarchical structure of repeated samples [42] to improve accuracy for real deployments. To avoid overfitting the small-scale dataset, a few calibration samples are used in the person-specific gaze estimation [32] architecture. Due to the lack of high-resolution gaze datasets, [15] introduces a large-scale dataset known as the ETHX-Gaze dataset and evaluates the dataset using existing ResNet-50 [15] architecture to regress the 2D gaze vector. The work [1] proposed L2CS-Net for estimation of gaze by using binned features from ResNet 50 [15] and proposes a new loss technique that uses classification and regression. The features from the ResNet are split into two fully connected layers for pitch and yaw angles. The very recent work [44] proposes a multiple-person gaze estimation in a single RGB image where the dataset is generated using a face-swapping technique from one dataset to multiple faces in the image. Although the technique is innovative, synthetic data is always a limitation compared to collecting and training with real data.

As our main architecture relies on the self-attention model we also review the effectiveness of transformer modules on eye gaze estimation. [5] proposes to study gaze estimation using vision-based transformers and hybrid transformers with CNNs. They show that the hybrid transformers achieve state-of-the-art performance and show the effectiveness of transformers in gaze estimation. [27] estimates gaze using a self-attention mechanism and augmented convolutions. The 14 layers are considered from ResNet architecture and the convolutions are replaced with the self-attention mechanism. Residual attention pooling network [2] where the features from the semantic network are forwarded to the attention mechanism connected to residuals and pooling is introduced to extract the eye gaze. The work [29] proposes to use a transformer model to further improve the eye gaze accuracy, the method uses a static transformer model and temporal difference module from recurrent neural networks. The authors propose to overcome the limitations of previous research such as disregarding the feature relations by just concatenation and complexity of dynamic feature extraction. For coarse and fine extraction of features, both eye images and face images are used by the authors.

III. METHODOLOGY

The architecture we propose consists of modules such as the panoptic feature pyramid network [22], residual blocks with dilation, attention mechanism, and pooling, and the architecture is known as a panoptic feature pyramid attention pooling network (Pan-AP network). The panoptic feature pyramid is a semantic segmentation branch which is an improved version of Mask RCNN with a feature pyramid network. The panoptic features are obtained by upsampling the top-down layers with a bilinear function. All the layers from the pyramid are summed together to form a semantic segmentation in pixel-wise. The building blocks of the feature pyramid network are residual networks [15]. The residual networks are originally introduced for image recognition, which is known to be efficient in feature extraction and avoids problems like vanishing gradients. The pooling

layer in deep learning techniques reduces the size of the features without losing the features. The overall flow of the pipeline is shown in Fig. 2. The process of self-attention applied in this work is explained in the next section.

A. Self-Attention Technique

The attention block takes n input features and returns n output features. The basic operation of attention is that it learns to pay more attention to the necessary features with an increased receptive field. The attention mechanism introduced in [37] works well for many applications such as natural language processing and computer vision [9], [16]. The attention mechanism is also known as scaled dot product attention, and it consists of queries (Q), keys (K), and values (V) as inputs. The same input features are copied to queries, keys, and values, and the attention from the work [37] is computed as

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

where $\sqrt{d_k}$ is a scaling factor. The attention mechanism is illustrated in Fig. 3 single head attention block which can be extended to n dimensional (D) space. The single-head attention mechanism is further extended to multi-head attention by combining multiple heads such as 2, 4, and 8 heads in parallel.

B. Pan-AP Architecture

The image of size $I \in R^{224\times224\times3}$ is passed as input to the Pan-AP architecture. The bottom-up and top-down layers of the panoptic network share lateral information between layers. As you can observe in Fig. 2, the features of each top-down layer are upsampled to obtain the size of $128\times56\times56$. Since all the layers are of similar size, the layers are then summed element-wise to obtain semantic features which are then smoothed and pooled to obtain features of size $128\times28\times28$. The features are then categorized into primary and secondary by inversing the features using sigmoid multiplication. The features are then added with the position embeddings and forwarded to the attention block. The input and output features from an attention block is of similar size, we apply convolution pooling to the features from the attention block to reduce the feature size. After the first attention block, the output size of the feature is $128\times28\times28$ which is then convolution pooled to obtain features of size $128\times14\times14$. The process of attention and pooling is repeated as follows: $attention_1 \rightarrow pooling_1 \rightarrow attention_2 \rightarrow pooling_2 \rightarrow attention_3$. Then, we obtain attended features of size $128\times7\times7$. The attention features from secondary features are mostly related to the background so we compute the inverse of the features for face feature enhancement. Finally, the attended features from the primary and secondary layers are multiplied and forwarded to a dilated convolution layer to increase the receptive field of features. The obtained 2048 features are then passed through a fully connected layer to regress the 2D gaze vector.

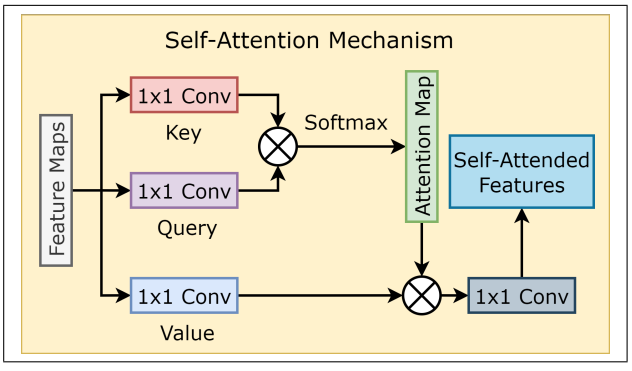

Fig. 3. Single-headed self-attention mechanism.

C. 3D To 2D Gaze Transformation

Optimizing the network for a 2D gaze vector is more feasible than a 3D gaze. We follow a similar process of 3D gaze to 2D gaze conversion and 2D gaze to 3D gaze conversion mentioned in [48], [15] as we use both of these datasets for experimentation. Given a 3D gaze vector, we compute 2D gaze yaw (θ) and pitch(ϕ) values as

$$\theta = \arcsin(v) \tag{2}$$

$$\phi = \arctan 2(u,w) \tag{3}$$

Similarly, given a 2D gaze vector we compute 3D unit gaze vector $[\text{u,v,w}]^T$ as

$$u = \cos(\theta)\cdot sin(\phi) \tag{4}$$

$$v = \sin(\theta) \tag{5}$$

$$w = \cos(\theta)\cdot\cos(\phi) \tag{6}$$

The conversion is specific to the ETHX-Gaze dataset [45] and MPIIFaceGaze [48].

IV. EXPERIMENTS

We evaluate the proposed Pan-AP architecture with two large-scale open-source gaze datasets and apply it in a real-time human-robot interaction environment.

A. Eye Gaze Datasets

To train and evaluate our proposed network strategy, we consider datasets with face images for gaze regression. To evaluate our architecture, we utilize two well-known large-scale open-source datasets. The first dataset is MPIIFaceGaze [48] dataset with good-resolution images where participants are using laptop cameras. The second dataset is ETH-XGaze [45] dataset with extremely high-resolution images.

MPIIFaceGaze dataset. MPIIFaceGaze [48] dataset is a part of the MPIIGaze [47] dataset. Originally, the MPIIGaze dataset consists of only eye images in the training and test set, and later full facial images are released as

TABLE I

COMPARISON TO THE STATE-OF-THE-ART ON FACE BASED GAZE ESTIMATION. THE GAZE ANGULAR ERRORS ARE IN DEGREES.

	Datasets	
Method	MPIIFaceGaze	ETH-XGaze
FewShotGaze [32]	5.2°	-
MPIIFaceGaze [48]	4.8°	-
ETH-XGaze [45]	4.8°	4.5°
RT-GENE [10]	4.3°	-
FARE-Net [8]	4.3°	-
ARes [27]	4.17°	-
CA-Net [4]	4.1°	-
AGE-Net [25]	4.09°	-
L2CS-Net [1]	3.92°	-
GazeTR [5]	3.88°	-
Pan-AP (Ours)	3.75°	3.96°

MPIIFaceGaze. The dataset is completely recorded using a laptop for daily activities over a long duration in an uncontrolled environment. Since the dataset consists of 15 subjects, equally sampled 45,000 full facial images are made available for experimentation.

ETH-XGaze. The ETH-XGaze [45] dataset is a very high-resolution dataset captured using 18 Canon 250D digital SLR cameras with a resolution of 6000×4000 pixels with extreme head variations. The setup for capturing the gaze is quite large, where the participants are instructed to focus on the shrinking circle, and when it becomes a dot mouse button is clicked. To obtain different illuminations on 110 participants, the lights are switched on and off.

B. Training Parameters

To backpropagate the weights of the proposed architectures, we use the L_1 loss function with regularization. The L_1 loss function is the difference between predicted gaze p_g and the actual gaze a_g. The loss function is

$$L_1 = |p_g - a_g| + \lambda \sum_{i=1}^{N} |w_i| \tag{7}$$

where λ is a regularization parameter and w_i are the network weights.

For a better understanding of the proposed architecture, we follow similar angular error computation as in [45], [48], [2]. The 3D angular error is calculated for both within-dataset and cross-dataset evaluations. The mathematical representation of 3D error between the ground truth gaze g_{gt} and the predicted gaze g_p is

$$\mathscr{L}_{angular} = \frac{\mathbf{g_{gt}} \cdot \mathbf{g_p}}{\|\mathbf{g_{gt}}\| \|\mathbf{g_p}\|} \tag{8}$$

There exist different optimization techniques in deep learning, we train the network with two widely used optimizers that is Adam [21] and RMSProp. From the initial experiments Adam performed better so we employ it with a learning rate of 1e-4 and to avoid fluctuations in loss we also introduce weight decay of 1e-6. We train the network on a 48GB Nvidia RTX Quadro graphical processing unit. To fully utilize the available GPU, we load 256 images for each batch. The proposed network is trained for 50 epochs for each evaluation.

TABLE II

CROSS-DATASET EVALUATION RESULTS IN DEGREES.

Train Test	MPIIFaceGaze	ETH-XGaze
MPIIFaceGaze [48]	3.75°	15.9°
ETH-XGaze [45]	6.2°	3.96°

C. Within Dataset Evaluation

We test the performance of the individual datasets where the model is trained and evaluated from similar subsets. Leave-one-person-out cross-validation measures the performance of the MPIIFaceGaze dataset. There exist 15 participants in MPIIFaceGaze dataset so the process is repeated 15 times leaving one person each time for validation. To compute the 3D angular gaze error, the 2D gaze vector is converted to 3D gaze as mentioned in section III.C. The mean angular gaze error of all subjects with average value is plotted in Fig. 4. The mean angular gaze error on the MPIIFaceGaze dataset is approximately 3.75°. The average error on all participants is approximately 3.75°. The figure shows that the angular error for the majority of the participants is less than 4.5°. The participant numbers P02 and P14 have the highest errors of 4.72° and 6.12° respectively.

The training samples and test samples are provided by default for the ETH-XGaze dataset. As the dataset images are very high quality, we apply different augmentation techniques like brightness, contrast, downsampling, and upsampling to reduce the quality of the image to match real-world applications. The training samples and their labels are made available, and the labels of the test samples are withheld. For the predicted gaze values for test samples, we obtain a 3D angular accuracy of 3.96°.

Finally, we compare the obtained results from the Pan-AP network with the state-of-the-art algorithms on face-based gaze estimation. The comparison results are mentioned in Table I. From the results, we can observe that our proposed model achieves state-of-the-art results on MPIIFaceGaze and ETH-XGaze datasets. A few results on the MPIIFaceGaze data with ground truth and predicted gaze is illustrated in Fig. 5.

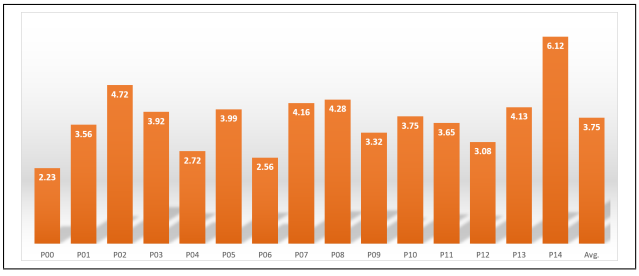

Fig. 4. Leave-one-person-out cross-validation errors of each subject in degrees on MPIIFaceGaze dataset trained on the Pan-AP architecture.

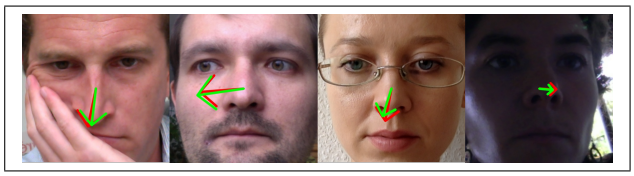

Fig. 5. Gaze output on 2D images. The output samples are from the MPIIFaceGaze dataset on different subjects with occlusions and brightness. The green arrow is the ground truth gaze and the red arrow is the predicted gaze.

D. Cross Dataset Evaluation

A model trained on one dataset is utilized for evaluating the other test set is known as cross-validation. The architecture is retrained with a complete MPIIFaceGaze dataset skipping leave-one-person-out validation. We do not repeat this step for the ETH-XGaze dataset as there is no such validation process. Table II consists of the cross-data validated gaze error for both datasets. From the results, we can observe that the model trained on the ETH-XGaze dataset performed well on the MPIIFaceGaze dataset. The same model is utilized for testing the real-world human-robot interaction scenario.

E. Ablation Study

In this section, we show the importance of each module in our architecture by experimenting with each module and their parameter. We mainly focus on the parameters of the self-attention module and the dilated convolution. We perform the ablation study with the MPIIFaceGaze dataset on the proposed Pan-AP architecture. The parameters in the convolution attention are the number of parallel heads and we experiment with 2, 4, and 8 heads. Dilated convolutions represent the receptive field. For the initial experiment, smoothed results from panoptic FPN are passed through 2 pooled convolution layers and fully connected layers to obtain the 2D gaze vector. We noticed an angular error of 4.3°. By adding the three attention layers with 2 parallel heads, the angular accuracy is reduced to 3.96°. By adding a dilation convolution and linear layers, we obtain an angular error of 3.8°. With 4 heads the angular error is reduced to 3.81° and for 8 heads the angular error increased to 3.93°. By adding a dilated convolution to 4 heads attention, we finally were able to achieve an angular error of 3.75°. Increasing the number of attention layers affects the real-time speed (i.e., fps) so we use only three convolution attention layers. On the ETH-XGaze dataset, we only notice a very minimal difference in angular error between 2 and 4 heads.

F. Human Robot Interaction Application

In this section, we apply the gaze estimation model in real-time human-robot interaction environments shown in Fig. 1. The goal of the first environment is to grab the attention of a robot in the environment and direct eye-gaze towards an object for picking. The distance between the camera and the subject ranged from 0.8 to 1.5 meters in the first experiment. For real-time application, the head pose is necessary, so we cascade the dlib [20] deep learning face detection with the proposed Pan-AP architecture for gaze estimation. From cross-dataset evaluation, the ETH-XGaze dataset works better compared to others from Table II so we use this trained model for interaction and application. We retrain the ETH-XGaze dataset by further adding augmentation techniques like brightness, noise, and contrast for real-time application. The Intel Realsense D435 camera is installed in the environment to obtain the RGB images and depth which in turn are calibrated to the robot workspace. Once the 2D gaze is estimated, the pupil center points from 2D head pose estimation are extended to 3D using depth information, and the gaze is converted to 3D as mentioned in Section III.C. The complete process runs at 16 frames per second on a 48GB Nvidia RTX GPU. We apply the proposed model in a real-time scenario and the YCB objects [40] used for experimentation can be observed in Fig. 6. In addition to that, we also used a wooden cube with a bullseye pattern for focusing. From the experiment, we were able to clearly distinguish the directions. In addition to that, we have conducted a few experiments to direct gaze toward objects for picking applications at varied distances. The process of picking an object is repeated for 20 iterations with 3 different subjects and varied distances between the camera and face and distance between two objects ranging from 0.3m to 0.01m. In all experiments, the successful picking accuracy is higher when the distance between the face and camera is closer and the distance between objects is higher as mentioned in Table III. The first column in the table is an approximate distance between the camera and face in meters, the second column is the distance between two objects placed on the table for picking experiments, and the last column is the success rate of picking an object based on two conditions.

Fig. 6. Objects used for pick and place in human-robot interaction environment.

TABLE III
EXPERIMENTS IN HUMAN-ROBOT INTERACTION ENVIRONMENT.

Camera-face distance	Object distance	Success rate
0.8m	0.3m	95%
1.0m	0.3m	95%
1.5m	0.3m	85%
0.8m	0.1m	90%
1.0m	0.1m	70%
1.5m	0.1m	65%
0.8m	0.01m	70%
1.0m	0.01m	55%
1.5m	0.01m	50%

The fluctuations of eye gaze make it harder to focus on certain regions or objects when the objects are placed close to each other.

In the second experiment, we worked with a pepper robot. In this scenario, the pepper robot will start interacting with humans by using eye gaze. If a person is gazing towards the robot for over a certain duration, the pepper robot automatically starts to interact with the human. In all experiments, the pepper robot was able to interact with a human by measuring attention. The future goal of this experiment is to include it in the library scenario for assistance in case of students need help with books.

Although the attention-grabbing of a robot with a gaze works well, picking up objects with a gaze needs further enhancements. For robotic applications, it is quite hard to focus on a single object due to various fluctuations, so for further development, we plan on including multi-modal communication by including the gesture to gaze to enhance the communication between humans and robots.

V. CONCLUSIONS

In this work, we introduced the Pan-AP network architecture for eye gaze estimation with a feature pyramid network, residual blocks, dilation, and attention mechanism. We performed the evaluation of the architecture with two large-scale open-source datasets and obtained state-of-the-art performance. We also performed experiments in human-robot interaction environments and concluded that the eye gaze works for certain applications. We further aim to include the human eye gaze in human-robot collaborations with multi-modal communications.

ACKNOWLEDGMENT

This work is Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 416228727 - SFB 1410.

REFERENCES

[1] Ahmed A. Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2cs-net: Fine-grained gaze estimation in unconstrained environments. *ArXiv*, abs/2203.03339, 2022.

[2] Bandi Chaitanya and Ulrike Thomas. Face-based gaze estimation using residual attention pooling network. *In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP*, 2023.

[3] Zhaokang Chen and Bertram E. Shi. Appearance-based gaze estimation using dilated-convolutions. *CoRR*, abs/1903.07296, 2019.

[4] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. *CoRR*, abs/2001.00187, 2020.

[5] Yihua Cheng and Feng Lu. Gaze estimation using transformer. 2022.

[6] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[7] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *CoRR*, abs/2104.12668, 2021.

[8] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020.

[9] 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. *ArXiv*, abs/2010.11929, 2021.

[10] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[11] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 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*, ETRA '14, page 255–258, New York, NY, USA, 2014. Association for Computing Machinery.

[12] Kenneth Alberto Funes Mora and Jean-Marc Odobez. Geometric generative gaze estimation (g¡sup¿3¡/sup¿e) for remote rgb-d cameras. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1773–1780, 2014.

[13] E.D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6):1124–1133, 2006.

[14] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, 2010.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[16] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

[17] Sabrina Hoppe, Tobias Loetscher, Stephanie A. Morey, and Andreas Bulling. Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience*, 12:105, 2018.

[18] Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 83–90, 2016.

[19] 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.

[20] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[22] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6392–6401, 2019.

[23] Robert Konrad, Anastasios Angelopoulos, and Gordon Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *CoRR*, abs/1906.09740, 2019.

[24] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra M. Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. *CoRR*, abs/1606.05814, 2016.

[25] Murthy L R D and Pradipta Biswas. Appearance-based gaze estimation using attention and difference mechanism. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3137–3146, 2021.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[27] Gabriel Lefundes and Luciano Oliveira. On estimating gaze by self-attention augmented convolutions. *ArXiv*, abs/2008.11055, 2020.

[28] Peng Li, Xuebin Hou, Xingguang Duan, Hiuman Yip, Guoli Song, and Yunhui Liu. Appearance-based gaze estimator for natural interaction control of surgical robots. *IEEE Access*, 7:25095–25110, 2019.

[29] Yujie Li, Longzhao Huang, Jiahui Chen, Xiwen Wang, and Benying Tan. Appearance-based gaze estimation method using static transformer temporal differential network. *Mathematics*, 11(3), 2023.

[30] Oliver Lorenz. and Ulrike Thomas. Real time eye gaze tracking system using cnn-based facial features for human attention measurement. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP*, pages 598–606. INSTICC, SciTePress, 2019.

[31] Atsushi Nakazawa and Christian Nitschke. Point of gaze estimation through corneal surface reflection in an active illumination environ-

ment. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 159–172, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[32] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation, 2019.

[33] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph.*, 35(6), Nov. 2016.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[35] B.A. Smith, Q. Yin, S.K. Feiner, and S.K. Nayar. Gaze Locking: Passive Eye Contact Detection for Human?Object Interaction. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 271–280, Oct 2013.

[36] Roberto Valenti, Nicu Sebe, and Theo Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2):802–815, 2012.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[38] Haofei Wang, Xujiong Dong, Zhaokang Chen, and Bertram E. Shi. Hybrid gaze/eeg brain computer interface for robot arm control on a pick and place task. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1476–1479, 2015.

[39] Erroll Wood, Tadas Baltrusaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications*, pages 131–138, 2016.

[40] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[41] Xuehan Xiong, Zicheng Liu, Qin Cai, and Zhengyou Zhang. Eye gaze tracking using an rgbd camera: A comparison with a rgb solution. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, page 1113–1121, New York, NY, USA, 2014. Association for Computing Machinery.

[42] Yunyang Xiong, Hyunwoo J. Kim, and Vikas Singh. Mixed effects neural networks (menets) with applications to gaze estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7735–7744, 2019.

[43] Yu Yu, Liu Gang, and Odobez Jean-Marc. Deep multitask gaze estimation with a constrained landmark-gaze model. In *ECCV 2018 Workshop*. Springer, 2018.

[44] Mingfang Zhang, Yunfei Liu, and Feng Lu. Gazeonce: Real-time multi-person gaze estimation. *ArXiv*, abs/2204.09480, 2022.

[45] 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, 2020.

[46] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Evaluation of appearance-based methods and implications for gaze-based applications. *CoRR*, abs/1901.10906, 2019.

[47] 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.

[48] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. *CoRR*, abs/1611.08860, 2016.