# Face-Based Gaze Estimation Using Residual Attention Pooling Network

Chaitanya Bandi[a] and Ulrike Thomas[b]
*Robotics and Human-Machine-Interaction Lab, Chemnitz University of Technology, Reichenhainer str. 70, Chemnitz, Germany*

Keywords: Gaze, Attention, Convolution, Face.

Abstract: Gaze estimation reveals a person's intent and willingness to interact, which is an important cue in human-robot interaction applications to gain a robot's attention. With tremendous developments in deep learning architectures and easily accessible cameras, human eye gaze estimation has received a lot of attention. Compared to traditional model-based gaze estimation methods, appearance-based methods have shown a substantial improvement in accuracy. In this work, we present an appearance-based gaze estimation architecture that adopts convolutions, residuals, and attention blocks to increase gaze accuracy further. Face and eye images are generally adopted separately or in combination for the estimation of eye gaze. In this work, we rely entirely on facial features, since the gaze can be tracked under extreme head pose variations. With the proposed architecture, we attain better than state-of-the-art accuracy on the MPIIFaceGaze dataset and the ETH-XGaze open-source benchmark.

## 1 INTRODUCTION

Eye gaze is a crucial nonverbal cue that determines a person's intent. The person's intent is extremely useful in human-robot interaction applications (Huang and Mutlu, 2016) such as attracting a robot's attention by glancing at it, and when combined with body motions, it is possible to strengthen communication between human and robot. Aside from robotics, the gaze can be used in human-computer interface (Zhang et al., 2019; Li et al., 2019; Wang et al., 2015), virtual reality (Patney et al., 2016; Konrad et al., 2019), and behavioral analysis (Hoppe et al., 2018). Model-based methods and appearance-based methods (Hansen and Ji, 2010) are used to estimate eye gaze. Although classic model-based eye gaze assessment approaches (Guestrin and Eizenman, 2006; Nakazawa and Nitschke, 2012; Valenti et al., 2012; Funes Mora and Odobez, 2014; Xiong et al., 2014) are accurate, the environment is extremely regulated (i.e., slight occlusions and static laboratory settings). Furthermore, the distance between the customized device or camera with RGB-D sensors and the eye is fixed (often to 60 cm) in order to estimate the gaze. The eye model is assumed to be constant across all participants, and without proper calibration, the system frequently fails to estimate the right gaze. Model-based solutions fail frequently in real-world applications that involve estimating the gaze in the wild (i.e., in an uncontrolled environment).

Because of the restrictions of the model-based techniques, recent research has switched to appearance-based models. Dedicated devices are not essential for appearance-based gaze estimating techniques because standard cameras are adequate for image processing and gaze regression. The appearance-based models are further classified into two types:

1) Feature-based methods and deep learning-based methods. The early works focus on effective feature extraction techniques like the histogram of oriented gradients (Martinez et al., 2012) to estimate gaze. The histogram of oriented gradients works well for low-level feature extraction but fails to effectively extract high-level features for gaze in images. One of the early efforts (Baluja and Pomerleau, 1994) tracks gaze using artificial neural networks using $15 \times 15$ retina input. Later appearance-based approaches (Tan et al., 2002) estimate eye gaze from images using non-linear mapping functions. Each calibrated subject has its mapping functions. The work (Williams et al., 2006) uses linear interpolation to do an appearance-based closest manifold point query. Training data is frequently used in appearance-based models. The paper introduces a semi-supervised Gaussian process with an uncertainty measure that learns map-

[a] https://orcid.org/0000-0001-7339-8425
[b] https://orcid.org/0000-0003-3211-4208



Bandi, C. and Thomas, U.
Face-Based Gaze Estimation Using Residual Attention Pooling Network.
DOI: 10.5220/0011789200003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 541-549
ISBN: 978-989-758-634-7; ISSN: 2184-4321

pings from partially labeled input. For unseen images, the sparse regression model infers mappings from processed pixel data in real time. The saliency gaze (Chang et al., 2019) estimates gaze on uncalibrated users to solve the problems of classic gaze estimation methods such as calibration, lighting, and position fluctuations. Using $l1$ optimization, the adaptive linear regression (Lu et al., 2014) approach handles calibration problems based on a large number of training samples, image resolution, and blinking. Although there is a slight improvement in accuracy, they are not reliable enough to apply in real-world scenarios.

2) Deep learning approaches such as convolutional neural networks (CNN) have been proved to be effective in extracting high-level image characteristics and learning non-linear information for regression applications. Recent research indicates that CNN-based design regresses the direction of human attention in eye images (Zhang et al., 2015; Yu et al., 2018; Fischer et al., 2018; Cheng et al., 2018; Lorenz. and Thomas., 2019a), face images (Zhang et al., 2016; Xiong et al., 2019; Zhang et al., 2020; Park et al., 2019), or from both face and eye images (Krafka et al., 2016; Chen and Shi, 2019; Cheng et al., 2020a).

We focus on regressing the 2D gaze vector from face images in this work because CNNs can regress outputs even with eye occlusions. To directly regress the 2D gaze vector, an image of the face is sent through the proposed network. Figure 1 depicts the basic architecture flow. This paper makes the following contributions:

[1] We provide a novel network design that regresses the 2D gaze vector using a Panoptic-feature pyramid network (PFPN) (Kirillov et al., 2019), residual blocks, pooling, and self-attention modules.

[2] Using the proposed technique, the network achieves cutting-edge performance on two separate datasets: the MPIIFaceGaze (Zhang et al., 2016) dataset and the ETH-XGaze (He et al., 2015) dataset.

# 2 RELATED WORK

Deep learning-based appearance methods have been found to be more efficient than model-based and feature-based learning methods in cross-subject gaze estimation.

## 2.1 Convolutional Neural Network Architectures

CNNs have proven to be useful in a variety of computer vision applications, including eye-gaze estimation. CNN-based gaze estimate is affected by the input features. CNNs can regress eye gaze utilizing features such as eyes and face either dependently or independently.

**Eye-based Methods.** The paper (Zhang et al., 2015) provides the first CNN-based gaze estimating methodology that works in real-world situations. LeNet architecture (Lecun et al., 1998) inspired the proposed multimodal CNN architecture. The multimodel CNN receives $60 \times 36$ pixel eye images as input and outputs a 2D gaze vector, providing an open-source unconstrained high-resolution dataset known as the MPIIGaze dataset. (Yu et al., 2018) present a multi-task framework that uses an end-to-end model known as the constrained landmark gaze model to localize eye landmarks and eye gaze. (Yu et al., 2018) use UnityEyes (Wood et al., 2016) and supplement data for training and evaluation to build an end-to-end model. In natural settings, the distance between the camera and subject is greater, and the resolution of the eye is fairly low. The work in (Lorenz. and Thomas., 2019b) present a multi-task CNN arhitecture that extracts the facial features at first and then eye features for gaze estimation using geometric method. (Fischer et al., 2018) present a CNN model for a new large-scale dataset known as the RT-GENE that feeds two eye regions to the VGG-16 (Simonyan and Zisserman, 2014) network individually. The characteristics are later concatenated with head posture information to regress the 2D eye gaze vector. (Cheng et al., 2018) proposes two networks for eye gaze regression that take advantage of eye asymmetry. The first network is an asymmetry regression network with four streams for 3D gaze regression and a two-stream assessment network for asymmetry correction.

**Face and Eye Combined Methods.** iTracker (Krafka et al., 2016) is one of the first attempts to use CNNs to forecast gaze based on the face, patches of both eye regions, and face grid. The iTracker is specifically built for commodity hardware such as mobile phones and tablets, and it makes use of a novel dataset known as GazeCapture. According to the study in (Chen and Shi, 2019), most CNN architectures use multi-layer downsampling, which degrades spatial resolution. Dilated convolutions are used to extract features to avoid this. The dilated convolutions are taken into account for both eye pictures but not for the facial region. A coarse to fine strategy (Cheng et al., 2020a) estimates coarse gaze direction from a facial image, fine gaze direction from an eye image, and final output gaze is refined. (L R D and Biswas, 2021) suggests the AGE-Net

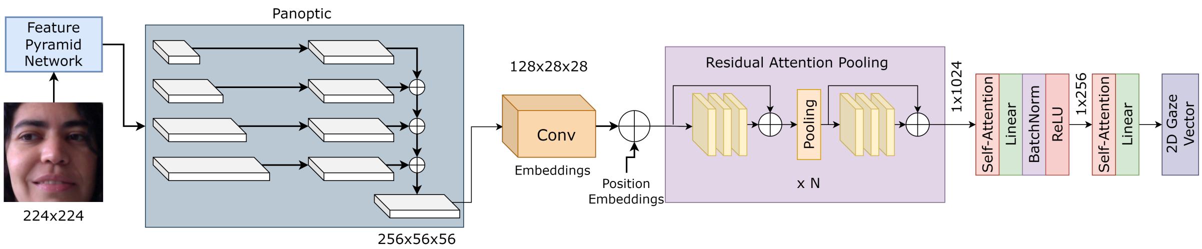


Figure 1: The overall architecture for eye gaze regression. The pipeline is an end-to-end network in which the first stage consists of a feature pyramid network with multiple output pyramids, which are then summed to form panoptic features. The panoptic features are smoothed and pooled using convolution. The convolution embeddings are then combined with position embeddings and forwarded to the residual attention pooling network. The output features are then self-attended, and linearized to obtain the gaze vector.

attention and difference mechanism, which consists of two networks: one that eliminates similarities in left and right eyes that are unimportant to eye gazing and the other that applies attention to eye image attributes. The work (Funes Mora et al., 2014) presents a new dataset known as the EYEDIAP dataset, while the work (Smith et al., 2013) introduces the Columbia gaze dataset, both of which are used by the majority of the works described above. (Cheng et al., 2021) compares most appearance-based eye gaze estimating algorithms to current standards, while (Kellnhofer et al., 2019) estimates gaze using temporal information. In response to the two-eye asymmetry characteristic, (Kellnhofer et al., 2019) introduces FARE-Net, which predicts 3D gaze angles for both eyes using an asymmetric approach. They assign uneven weights to each of the two eye losses and then sum these losses.

Recent research has demonstrated that eye gaze can be regressed utilizing face characteristics. (Zhang et al., 2016) includes a CNN-based framework with spatial weights for regression that surpasses eye image-based gaze estimation, and the dataset is a subgroup of the MPIIGaze dataset known as the MPIIFaceGaze dataset. To increase accuracy for real-world deployments, the principle of mixed effects from statistics is incorporated in a deep convolutional network to grasp the hierarchy system of repeated samples (Xiong et al., 2019). Person-specific gaze estimation (Park et al., 2019) increases the accuracy of eye gaze estimation datasets by employing a minimal number of calibration samples and avoiding overfitting for small-scale datasets. The work in (Zhang et al., 2020) regresses the 2D gaze vector and converts it to 3D for angular error calculation using a basic residual cnn architecture known as ResNet-50 (He et al., 2015). This paper also introduces the ETHX-Gaze dataset, which is a large-scale dataset. Recent work (Abdelrahman et al., 2022) introduced L2CS-Net for gaze estimation utilizing binned features from ResNet 50 (He et al., 2015) and provides a novel loss strategy that employs classification and regression. The ResNet features are divided into two fully connected layers for angle estimation in yaw and pitch directions. PureGaze (Cheng et al., 2022) uses adversarial training with a gaze estimation network and a reconstruction network to remove irrelevant features for gaze estimation.

# 3 METHODOLOGY

In this section, we present the **P**anoptic feature pyramid and **R**esidual **A**ttention **P**ooling (P-RAP) framework for eye gaze estimation. We investigate the essential building blocks of this architecture, such as panoptic-FPN, residual blocks, and attention processes, to gain a deeper understanding. Residual networks (He et al., 2015) were designed to increase the accuracy and performance of image recognition. Residual networks have been found to be more efficient in feature extraction and to optimize quicker with skip connections than networks such as VGG (Simonyan and Zisserman, 2014). The residual blocks are the fundamental backbone of the panoptic-FPN in our architecture. The panoptic-FPN (Kirillov et al., 2019) architecture is commonly used for object detection and semantic segmentation in order to acquire multi-scale characteristics for detecting smaller and larger objects. The network is made up of bottom-up and top-down layers containing lateral connections to increase object detection accuracy and image segmentation. We employ the panoptic-FPN architecture in this work to preserve the multi-scale aspects of the face and eyes. The basic architecture flow of panoptic features is shown in Figure 1.

## 3.1 Attention Mechanism

The attention mechanism accepts $n$ input features and returns $n$ output features. Attention's core operation

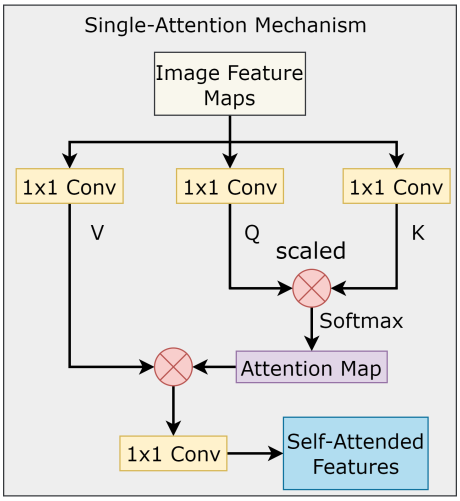


Figure 2: The self-attention mechanism.

is that it learns to pay greater attention to the required elements. The attention method proposed in (Vaswani et al., 2017) works well for a wide range of applications, including natural language processing (Vaswani et al., 2017) and computer vision (Dosovitskiy et al., 2021; Heo et al., 2021). The attention mechanism also referred to as scaled dot product attention, takes as inputs queries ($Q$), keys ($K$), and values ($V$). The identical characteristics from the input are replicated and passed as the queries, keys, values, and attention is calculated as

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

where $\sqrt{d_k}$ is a scaling factor. The attention mechanism is applicable to $n$-dimensional (D) space. Figure 2 depicts the single-head attention mechanism. By merging several heads simultaneously, the single-head attention process is expanded to multi-head attention. We experiment with 2, 4, and 8 heads for spatial or convolutional attention in this paper.

### 3.2 Panoptic and Residual Attention Pooling Network

We feed the input face image of size $I \in R^{224\times224\times3}$ to the P-RAP architecture. The panoptic-FPN consists of five bottom-up layers and four top-down layers. The last four top-down layers share the lateral feature information from the bottom-up layers. Each top-down layer from the FPN is then passed to a convolution layer to obtain a size of $256\times56\times56$. The four-layer outputs are then element-wise summed to obtain out features of size $256\times56\times56$ which are known as panoptic features. The panoptic features are then forwarded to a simple convolutional layer with pooling to obtain features of size $128\times28\times28$. The features are then combined with position embeddings similar to transformer encoder (Vaswani et al., 2017). We pass the features to the residual-attention-pooling (RAP) network as in Figure 1 purple block. The RAP network consists of residual blocks and attention convolution. Each residual-attention block consists of: $attention-convolution_1 \rightarrow batchNorm_1 \rightarrow ReLU \rightarrow attention-convolution_2 \rightarrow batchNorm_2 \rightarrow ReLU \rightarrow attention-convolution_3 \rightarrow batchNorm_3 \rightarrow skipconnection \rightarrow ReLU$. After the activation layer, the features are pooled and passed through the residual-attention block. The process is repeated 2 times and we linearize the output and pass through $1\times N$ self-attention layer. the process of linearized output is illustrated in Figure 1. Finally, the attended features are forwarded to a linear layer to regress the pitch and yaw angles of eye gaze (i.e., 2D gaze).

### 3.3 Gaze and Loss Function

Designing the network for a 2D gaze vector is more efficient than optimizing the network for a 3D gaze vector. We use the same nomenclature as (Zhang et al., 2016; He et al., 2015) to transform the regressed 2D gaze vector to a 3D gaze and vice versa as needed. We compute 2D gaze yaw ($theta$) and pitch ($phi$) values from a 3D gaze vector.

$$\theta = \arcsin(y) \quad (2)$$

$$\phi = \arctan 2(x,z) \quad (3)$$

Similarly, we compute 3D unit gaze vector $[x,y,z]^T$ given 2D gaze angles as

$$x = \cos(\theta)\cdot sin(\phi) \quad (4)$$

$$y = \sin(\theta) \quad (5)$$

$$z = \cos(\theta)\cdot\cos(\phi) \quad (6)$$

This conversion is unique to the ETHX-Gaze dataset (He et al., 2015) and the MPIIFaceGaze (Zhang et al., 2016). We employ the $L1$ loss function with regularization to backpropagate the weights of the proposed architecture. The loss function is

$$L_1 = |predicted_{gaze} - actual_{gaze}| + \lambda\sum_{i=1}^{N}|w_i| \quad (7)$$

where $\lambda$ is a regularization parameter and $w_i$ are the weights of the network.

Figure 3: MPIIFaceGaze (Zhang et al., 2015) sample images from dataset.

# 4 EXPERIMENTS

In this section, we evaluate the proposed P-RAP architecture and experiment with two open-source gaze datasets.

## 4.1 Gaze Datasets

We utilize datasets with face images to train and assess our proposed network technique because we intend to regress gaze purely on face images. We use two open-source datasets to test our architecture. 1) The MPIIFaceGaze (Zhang et al., 2016) dataset contains high-resolution photos of people who are closer to the camera. 2) The ETH-XGaze (He et al., 2015) collection contains incredibly high-resolution images.

**MPIIFaceGaze Dataset.** The MPIIFaceGaze (Zhang et al., 2016) dataset is a subset of the MPIIGaze (Zhang et al., 2015) dataset. The MPIIGaze dataset was originally composed of eye images for experimentation, but a subset of face images was eventually provided as MPIIFaceGaze. The MPIIGaze dataset is totally captured in an uncontrolled context, such as daily laptop usage over long periods of time. When the target was displayed, the individuals were prompted to hit a key. The dataset contains 213,659 photos from 15 distinct contributors. In the MPIIFaceGaze dataset, a subset of 45,000 samples with full facial images is released from this dataset. Each participant's dataset has 3,000 samples. A few samples are shown in Figure 3.

**ETH-XGaze.** The ETH-XGaze (He et al., 2015) dataset contains extremely high-resolution images acquired with 18 Canon 250D digital SLR cameras. The images captured have a resolution of $6000 \times 4000$ pixels. The ETH-XGaze dataset contains a large number of head position variations ranging from $\pm 80°, \pm 80°$. The ETH-XGaze dataset is a massive collection of over 1 million photos from 110 individuals. Figure 4 shows the facial cropped data images.

Figure 4: ETH-XGaze (Zhang et al., 2016) sample images from dataset.

## 4.2 Evaluation Metric and Training Parameters

The evaluation metric for measuring the performance is the 3D angular error. The angular error between the actual $g_{actual}$ and the predicted gaze $g_{pred}$ is computed as

$$\mathcal{L}_{angular} = \frac{\mathbf{g}_{\mathbf{actual}} \cdot \mathbf{g}_{\mathbf{predicted}}}{||\mathbf{g}_{\mathbf{actual}}|| \, ||\mathbf{g}_{\mathbf{predicted}}||} \tag{8}$$

The metric is utilized for both within and cross-dataset evaluation. To train the proposed architecture, we use the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 0.0001 and a weight decay of 1e-6. We use an Nvidia RTX Quadro with a 48GB graphical processing unit to train the networks. Each training batch consists of 256 - $224 \times 224 \times 3$ pixel images. For each evaluation, the architecture is trained for 50 epochs, and in most cases, the loss curves (training and validation loss) stabilize around $30^{th}$ epoch.

## 4.3 Within Dataset Evaluation

Within dataset evaluation evaluates the effectiveness of data from a similar subset on unknown subjects. The MPIIFaceGaze dataset started cross-validation with leave-one-person-out. The dataset contains 15 persons, 14 of which are used for training and one for testing. The procedure is evaluated 15 times, with the overall accuracy calculated by averaging the results. We use a similar cross-validation technique to evaluate the performance of the suggested architectures

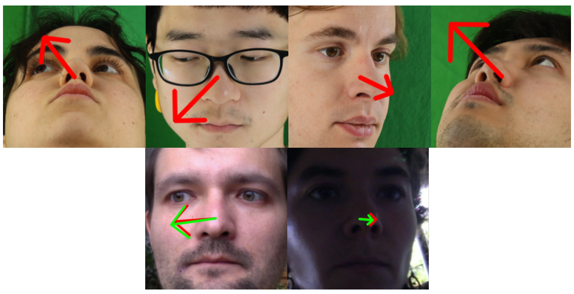

Figure 5: The output 2D gaze vector on the ETH-Xgaze and the MPIIFaceGaze test set. The red arrow is the predicted gaze and the green arrow is the ground truth gaze.

Table 1: Comparison of the proposed architecture results to the state-of-the-art. The values are within-dataset evaluation errors. The gaze angular errors are in degrees.

| | Datasets | |
|---|---|---|
| Method | MPIIFaceGaze | ETH-XGaze |
| FewShotGaze (Park et al., 2019) | 5.2° | - |
| MPIIFaceGaze (Zhang et al., 2016) | 4.8° | - |
| ETH-XGaze (He et al., 2015) | 4.8° | 4.5° |
| RT-GENE (Fischer et al., 2018) | 4.3° | - |
| FARE-Net (Cheng et al., 2020b) | 4.3° | - |
| CA-Net (Cheng et al., 2020a) | 4.1° | - |
| AGE-Net (L R D and Biswas, 2021) | 4.09° | - |
| L2CS-Net (Abdelrahman et al., 2022) | 3.92° | - |
| **P-RAP (Ours)** | 3.8° | 4.09° |

Table 2: Cross-dataset evaluation results in degrees.

| Train \ Test | MPIIFaceGaze | ETH-XGaze |
|---|---|---|
| MPIIFaceGaze (Zhang et al., 2016) | 3.8° | 27.95° |
| ETH-XGaze (He et al., 2015) | 6.79° | 4.09° |

on the MPIIFaceGaze dataset. The architecture regresses the 2D gaze vector, and to measure 3D angular error, we transform both the actual and predicted 2D gaze vectors to the previously specified 3D unit vector. Figure 7 depicts the mean 3D angular gaze error for 15 subjects in the MPIIFaceGaze dataset. The 3D angular error resulting from the proposed network trained on the MPIIFaceGaze dataset is represented in Figure 7. The average inaccuracy for all participants is about 3.8°. The chart demonstrates that the majority of the participants have an angle error of less than 4.5°. The most deviations are 4.8° and 6.27° for participants P02 and P14, respectively. The ETH-XGaze dataset has predefined training and test samples. We obtain a 3D angular accuracy of 4.09° on test set. The ground truth gaze of the ETH-XGaze test set is not available for direct evaluation. As the dataset is very recently released not many works are available for comparison.

Finally, we compare the P-RAP network results to the state-of-the-art methods for face-based gaze estimation. Table 1 contains the comparison findings. According to the results, our proposed model delivers state-of-the-art results on the MPIIFaceGaze and

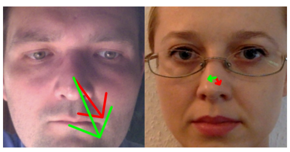

Figure 6: The high gaze angular error cases on the MPIIFaceGaze dataset. The red arrow is the predicted gaze and the green arrow is the ground truth gaze.

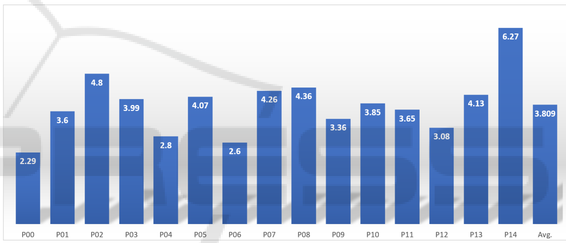


Figure 7: Participant-based mean angular gaze error in degrees on MPIIFaceGaze dataset trained on the P-RAP architecture.

ETH-XGaze datasets (as far as published work). Figure 5 and 6 depict a few output samples from both datasets with low and high precision.

## 4.4 Cross Dataset Evaluation

Cross dataset evaluation tests the performance of a model on a completely different dataset. For cross dataset evaluation, we retrained the architecture with complete MPIIFaceGaze dataset. We did not retrain the ETH-XGaze dataset as leave-one-out cross-validation is not performed during training. The cross dataset evaluation accuracy is mentioned in Table 2. The model trained on MPIIFaceGaze dataset is used for obtaining the gaze for the test set of ETH-XGaze dataset resulting in 27.9° angular error. Next, the model is trained on ETH-XGaze dataset and tested on MPIIFaceGaze dataset to obtain 3D angular error of 6.8°. From this, we can see that the model trained on the ETH-XGaze dataset performs better for the cross dataset evaluation.

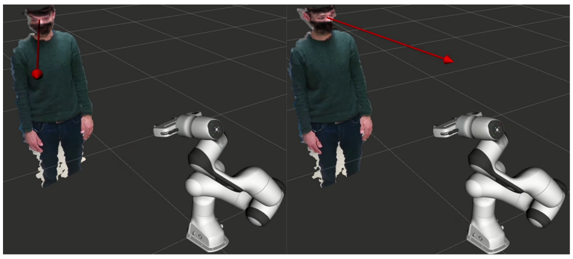

Figure 8: Gaze estimation in a human-robot interaction environment.

### 4.5 Human Robot Interaction Application

We use the gaze estimation method in a real-time human-robot interaction setting in this section. The purpose of this environment is to capture a robot's attention in a human-robot interaction scenario. The camera is approximately 1 to 2 meters away from the subject. According to cross-dataset examination, the ETH-XGaze dataset performs better than others from Table 2. We cascade the dlib (King, 2009) face detection and head posture estimation with the suggested FPN-AP architecture for gaze estimation for real-time applications. The suggested model is applied in a real-time situation, and the direction of gaze in the surroundings is shown in Figure 8. We can clearly discern the directions left, right, up, and down based on the experiments. In addition, we also tested the gaze in another human-robot environment for picking objects by gazing at them. From the experiments, we noticed that the distance between the camera and the human as well as the distance between objects are highly dependent. Although it worked for certain distances, it requires quite a huge improvement for real-time object-picking applications. As a further improvement, we are currently working on combining multi-modal communication information for the object-picking human-robot application.

## 5 CONCLUSION

We presented the P-RAP network design for eye gaze estimation with a panoptic feature pyramid network, residual blocks, and attention mechanism. We evaluated the framework using two large-scale open-source datasets. On both datasets, we conducted within-dataset and cross-dataset evaluations and obtained state-of-the-art performance. We aim to further improve the accuracy of gaze for real-time robotic applications in combination with multimodal communication.

## ACKNOWLEDGEMENTS

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

## REFERENCES

Abdelrahman, A. A., Hempel, T., Khalifa, A., and Al-Hamadi, A. (2022). L2cs-net: Fine-grained gaze estimation in unconstrained environments. *ArXiv*, abs/2203.03339.

Baluja, S. and Pomerleau, D. (1994). Non-intrusive gaze tracking using artificial neural networks. In Cowan, J., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann.

Chang, Z., Martino, M. D., Qiu, Q., Espinosa, S., and Sapiro, G. (2019). Salgaze: Personalizing gaze estimation using visual saliency.

Chen, Z. and Shi, B. E. (2019). Appearance-based gaze estimation using dilated-convolutions. *CoRR*, abs/1903.07296.

Cheng, Y., Bao, Y., and Lu, F. (2022). Puregaze: Purifying gaze feature for generalizable gaze estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Cheng, Y., Huang, S., Wang, F., Qian, C., and Lu, F. (2020a). A coarse-to-fine adaptive network for appearance-based gaze estimation. *CoRR*, abs/2001.00187.

Cheng, Y., Lu, F., and Zhang, X. (2018). Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Cheng, Y., Wang, H., Bao, Y., and Lu, F. (2021). Appearance-based gaze estimation with deep learning: A review and benchmark. *CoRR*, abs/2104.12668.

Cheng, Y., Zhang, X., Lu, F., and Sato, Y. (2020b). Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Fischer, T., Chang, H. J., and Demiris, Y. (2018). Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Funes Mora, K. A., Monay, F., and Odobez, J.-M. (2014). 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. Association for Computing Machinery.

Funes Mora, K. A. and Odobez, J.-M. (2014). 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.

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

Hansen, D. W. and Ji, Q. (2010). 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.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Heo, B., Yun, S., Han, D., Chun, S., Choe, J., and Oh, S. J. (2021). Rethinking spatial dimensions of vision transformers. In *International Conference on Computer Vision (ICCV)*.

Hoppe, S., Loetscher, T., Morey, S. A., and Bulling, A. (2018). Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience*, 12:105.

Huang, C.-M. and Mutlu, B. (2016). Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 83–90.

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., and Torralba, A. (2019). Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE International Conference on Computer Vision (ICCV)*.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Kirillov, A., Girshick, R. B., He, K., and Dollár, P. (2019). Panoptic feature pyramid networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6392–6401.

Konrad, R., Angelopoulos, A., and Wetzstein, G. (2019). Gaze-contingent ocular parallax rendering for virtual reality. *CoRR*, abs/1906.09740.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S. M., Matusik, W., and Torralba, A. (2016). Eye tracking for everyone. *CoRR*, abs/1606.05814.

L R D, M. and Biswas, P. (2021). 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.

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

Li, P., Hou, X., Duan, X., Yip, H., Song, G., and Liu, Y. (2019). Appearance-based gaze estimator for natural interaction control of surgical robots. *IEEE Access*, 7:25095–25110.

Lorenz., O. and Thomas., U. (2019a). 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.

Lorenz., O. and Thomas., U. (2019b). 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.

Lu, F., Sugano, Y., Okabe, T., and Sato, Y. (2014). Adaptive linear regression for appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Martinez, F., Carbone, A., and Pissaloux, E. (2012). Gaze estimation using local features and non-linear regression. In *2012 19th IEEE International Conference on Image Processing*, pages 1961–1964.

Nakazawa, A. and Nitschke, C. (2012). Point of gaze estimation through corneal surface reflection in an active illumination environment. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *Computer Vision – ECCV 2012*, pages 159–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Park, S., Mello, S. D., Molchanov, P., Iqbal, U., Hilliges, O., and Kautz, J. (2019). Few-shot adaptive gaze estimation.

Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., and Lefohn, A. (2016). Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph.*, 35(6).

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

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

Tan, K.-H., Kriegman, D. J., and Ahuja, N. (2002). Appearance-based eye gaze estimation. In *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, WACV '02, page 191, USA. IEEE Computer Society.

Valenti, R., Sebe, N., and Gevers, T. (2012). Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2):802–815.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Wang, H., Dong, X., Chen, Z., and Shi, B. E. (2015). 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.

Williams, O., Blake, A., and Cipolla, R. (2006). Sparse and semi-supervised visual mapping with the s3gp. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*.

Wood, E., Baltrusaitis, T., Morency, L.-P., Robinson, P., and Bulling, A. (2016). 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.

Xiong, X., Liu, Z., Cai, Q., and Zhang, Z. (2014). 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. Association for Computing Machinery.

Xiong, Y., Kim, H. J., and Singh, V. (2019). 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.

Yu, Y., Gang, L., and Jean-Marc, O. (2018). Deep multitask gaze estimation with a constrained landmark-gaze model. In *ECCV 2018 Workshop*. Springer.

Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., and Hilliges, O. (2020). Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation.

Zhang, X., Sugano, Y., and Bulling, A. (2019). Evaluation of appearance-based methods and implications for gaze-based applications. *CoRR*, abs/1901.10906.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. (2015). Appearance-based gaze estimation in the wild. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. (2016). It's written all over your face: Full-face appearance-based gaze estimation. *CoRR*, abs/1611.08860.