WACV #205

WACV #205

WACV 2025 Submission #205. Algorithms Track. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

EyeDentify: A Dataset for Pupil Diameter Estimation based on Webcam Images

Anonymous WACV Algorithms Track submission

Paper ID 205

Abstract

*In this work, we introduce EyeDentify, a dataset specifically designed for pupil diameter estimation based on webcam images. EyeDentify addresses the lack of available datasets for pupil diameter estimation, a crucial domain for understanding physiological and psychological states traditionally dominated by highly specialized sensor systems such as Tobii. Unlike these advanced sensor systems and associated costs, webcam images are more commonly found in practice. Yet, deep learning models that can estimate pupil diameters using standard webcam data are scarce. By providing a dataset of cropped eye images alongside corresponding pupil diameter information, EyeDentify enables the development and refinement of models designed specifically for less-equipped environments, democratizing pupil diameter estimation by making it more accessible and broadly applicable, which in turn contributes to multiple domains of understanding human activity and supporting healthcare. Our dataset is available at **LINK WILL BE IN THE FINAL VERSION**.*

1. Introduction

The cognitive state of humans is closely linked to features observable through their eyes. Fortunately, the accessibility of eye monitoring in everyday life is rapidly increasing, exemplified by recent advancements such as Apple's incorporation of camera-based eye tracking features [3, 14]. However, research in this domain primarily targets blink detection [17] and gaze estimation [36, 51], employing various methodologies, including the use of biomarkers [28], infrared spectrum reflected from the eyes [12], or image-based techniques [16]. In comparison, fewer explore pupil diameter estimation [5, 42], which also plays an undeniably crucial role in determining various physiological and psychological states. This oversight highlights a critical gap in the field, underscoring the need for more comprehensive approaches to fully leverage eye monitoring for cognitive state analysis for many reasons:

Previous studies show that the analysis of pupil diameter serves as an indicator of stress [37], attention [29, 45], or cognitive work loads [21, 25, 38]. In addition, the diameter of the pupil is also closely linked to the activity of the locus coeruleus [20, 34], a brain region critical for managing both short-term and long-term memory functions [21, 26]. Pupil diameter is also used for health check purposes, such as checking pupillary light reflex of patients with intracranial lesions in an intensive care unit [24]. Accurate pupil diameter estimation is thus fundamental to enhancing the capabilities of image-based eye tracking.

However, we identify three significant challenges in advancing the field of image-based pupil diameter estimation, which we want to address. The first challenge lies in collecting ground truth data. Previous works relied on capturing pupil images and subsequently measuring the diameter in pixels, a time-consuming process that is complicated by increasing participant numbers [5, 42]. We overcome this issue by applying a sensor substitution approach using Tobii eye-tracker with Tobii Pro SDK [1] as a reliable ground truth sensor. This approach allows for efficient data collection by acquiring ground truth diameter values from the eye-tracker and facial recordings via webcam.

The second challenge concerns data diversity. Previous studies have varied pupil diameter in participants by altering illumination displays [43]. We apply a similar approach, changing the computer display's color during our data collection. Unlike previous work [5, 42, 43], we impose fewer constraints, allowing them to choose their seating position and distance from the screen. This approach enables us to collect data under more natural, "in the wild" conditions, potentially enhancing the empirical validity of findings.

The third challenge is the prediction of pupil diameter itself. Previous studies [4, 22, 52] have highlighted that estimating gaze coordinates with a camera involves analyzing images of approximately 60×36 pixels [53]. The scale of our images will be smaller for pupil diameter estimation, necessitating analysis at an even finer resolution. This makes accurately predicting pupil diameters more complex than gaze estimation and presents a challenging task.

Table 1. Comparison of related datasets for eye monitoring. While most datasets have gaze coordinates [9, 17, 18, 22, 52, 53], there is a significant gap in pupil diameter informed [5, 40] datasets.

Dataset	Participants	Amount of data [frame]	Public	Gaze Coordinates	Pupil Diameter
MAEB [17]	20	1,440	✗	✓	✗
MPIIFaceGaze [53]	15	213,659	✓	✓	✗
Dembinsky et al. [9]	19	648,000	✓	✓	✗
Gaze360 [22]	238	172,000	✓	✓	✗
ETH-XGaze [52]	110	1,083,492	✓	✓	✗
VideoGazeSpeech [18]	*unknown*	35,231	✓	✓	✗
Ricciuti et al. [40]	17	20,400	✗	✓	✓
Caya et al. [5]	16	*unknown*	✗	✓	✓
EyeDentify (ours)	**51**	212,073	✓	✓	✓

In conclusion, we contribute to the image-based pupil diameter estimation field as follows:

1. **Creation of the Pupil Diameter Dataset EyeDentify**: We use the Tobii eye-tracker to collect accurate ground truth pupil diameter measurements. Concurrently, we captured facial video recordings with a built-in webcam. In total, data from 51 participants were gathered, comprising 212,073 images (i.e., blinking eyes removed from the initial 226,912 images).

2. **Data-Collecting Application ChameleonView**: Our own implemented web application called Chameleon-View captures various pupil diameters. It displays various screen colors during webcam recordings to ensure a diversity of pupil diameters. Moreover, it captures the timestamp and records videos while participants look at the display.

3. **Initial Pupil Diameter Estimations**: We evaluated several regression-based models to estimate pupil diameter sizes for both left and right eyes in a 5-fold cross-validation manner. For this, we tested ResNet-18 and ResNet-50, with ResNet-18 emerging as the most accurate, achieving a mean absolute error of 0.1340 ± 0.0196 and 0.1403 ± 0.0328 for the left and right eye, respectively.

2. Related Work

This section will examine existing datasets for eye monitoring, including gaze estimation and blink detection (Section 2.1). Furthermore, it surveys datasets for pupil diameter estimation (Section 2.2), as well as methods to estimate pupil diameters (Section 2.3). Table 1 compares all datasets appearing in eye monitoring publications, and we will compare our dataset EyeDentify with them.

2.1. Gaze Estimation and Blink Detection Datasets

Various gaze estimation, blink detection, and facial image data have been collected under different conditions, which we want to examine in the following. For instance, *MAEB* [17] has collected 20 subjects' images from nine different angles for robust image-based blink detection purposes. The study aims to complement the open-source dataset *HUST-LEBW* [19] with additional comparisons. Another example is *MPIIFaceGaze* [53], which has collected a daily life gaze coordinates dataset from 15 subjects. The strength of this dataset is that it has different illumination conditions and a collection of various day conditions, ranging from nine days to three months. Another dataset is presented by *Dembinsky et al.* [9], which collected a 360-minute dataset of 19 subjects with their corresponding gaze coordinates while speaking and/or listening. Thus, it incorporates different cognitive activities during recording. An alternative data collection setting is presented by *Gaze360* [22], which contains 238 subjects with 172,000 frame images. The distinguishing feature of this dataset is that it also contains outdoor data. Moreover, the camera used during data collection captured 360-degree view recordings, thus including the subjects' environment. Another varying axis during dataset collection is proposed by *ETH-XGaze* [52], which collected data from 110 subjects with 1,083,492 images under various hardware settings. More specifically, they used 18 high-resolution Canon 250D digital SLR cameras and four Walimex Daylight 250 light boxes. Lastly, *VideoGaze-Speech* [18] collected 35,231 frames from 29 videos. As a result, they extracted eye gaze data from existing videos; hence, the feature extraction approach varies from that of other studies.

Most mentioned datasets focus on enhancing gaze estimation and blink detection through diverse and complex recording conditions, but none incorporate pupil diameters.

2.2. Pupil Diameter Datasets

To the best of our knowledge, two data datasets have been used in publications for pupil diameter estimation that employ diverse approaches for data collection. For instance, *Ricciuti et al.* recruited 17 subjects ranging in age from 15 to 74 years and collected 20-second RGB video recordings under varying light intensities using a GoPro Hero 6 camera at a frame rate of 30fps [40]. Similarly, *Caya et al.* gathered data from 16 subjects aged 20 to 40 years, employing a camera mounted 10 cm from the subject's face in a controlled setting. Their study quantified a pupil diameter of 20 pixels as equivalent to 5 mm [5].

Despite these efforts, the mentioned works have not made their datasets publicly available. In contrast, our work and its dataset EyeDentify is not only inspired by their effort of collecting the necessary data for pupil diameter estimation but also presents the largest publicly accessible dataset for pupil diameter estimation from RGB images, significantly contributing to the broader field of eye monitoring.

2.3. Pupil Diameter Estimation Approaches

The estimation of pupil diameters has been done in various research studies. For instance, *Ni et al.* proposed a pupil diameter estimation method called BINOMAP [35]. Their approach uses master and slave cameras as a binocular geometric constraint for the input gaze images. The estimation model they implement is based on Zhang's algorithm [54], which achieved 0.022 ± 0.017mm mean absolute error. Similar to our work, *Caya et al.* uses their own data recording platform for data collection [5]. They captured the participant's facial images with a fixed camera distance of 10cm from the face. Then, the facial image is sent to Raspberry Pi for the pupil diameter estimation process. This estimation includes a RGB to grayscale conversion, contrast and brightness adjustment, image reshaping, and applying the Tiny-YOLO algorithm [23]. They achieved pupil measurements with a percent difference of 0.58% for the left eye and 0.48% for the right eye.

In summary, previous works have strong limitations in condition dependencies, such as having two cameras or keeping the face at a constant and fixed distance from the camera. Also, the data collection and evaluation must consider various pupil diameter sizes. We aim to address these limitations with the presented work.

3. Methodology

In this section, we describe the hardware (Section 3.1) and application we used for data collection (Section 3.2), the procedure for collecting the data (Section 3.3), as well as how we processed the raw recordings (Section 3.4).

3.1. Data Recording Hardware

The ground truth data for our pupil diameters is captured using the Tobii eye-tracker [1], a remote device capable of recording gaze coordinates and pupil diameters at 90Hz. It separately measures the diameters of the left and right pupils in millimeters with six-decimal precision. For camera-based recordings of the face and eyes, we utilize the built-in webcam of the Microsoft Surface Studio 1, which offers a screen resolution of 1280×720 pixels and a frame rate of 30.0 frames per second.

3.2. Data Collection Application

A distinctive aspect of our data collection methodology is using an application developed explicitly for EyeDentify. We implemented a web application called ChameleonView [1], which we used for our webcam data collection. The web application shows a button in the center of the screen that, when clicked by subjects, triggers a three-second webcam video recording. Additionally, we capture timestamps marking the start and end of each recording. These timestamps are crucial as they ensure synchronization between the webcam video and the Tobii eye-tracker data, as shown in Figure 1. During one data collection period, the subjects were asked to repeat this button-clicking process 50 times, resulting in 50 sessions for each participant.

To collect diverse pupil diameter sizes, we changed the computer screen background color [39, 49]. Thus, the 50 recordings had the following background color conditions (with corresponding HEX codes): the first ten recordings had a white display color (#ffffff), and the remaining 40 recordings were: black (#000000), red (#ff0000), blue (#0000ff), yellow (#ffff00), green (#008000), and gray (#808080), each in turn, five times displayed and a final display of white (#ffffff) for the last ten recordings again.

3.3. Data Collection Procedure

In this study, subjects began by completing a consent form and providing statistical information. Following this initial step, they proceeded to calibrate the Tobii eye-tracker. After calibration, the subjects started using the data collection application outlined in Section 3.2.

Upon completion of the experiment, the data from the Tobii, webcam video recordings, and corresponding timestamps were consolidated.

3.4. Data Preprocessing

After recording the raw webcam images alongside the measurements of the eye features, we merge them together for the final dataset EyeDentify. This involves two phases: Aligning the recordings (Section 3.4.1) and cropping the eyes (Section 3.4.2).

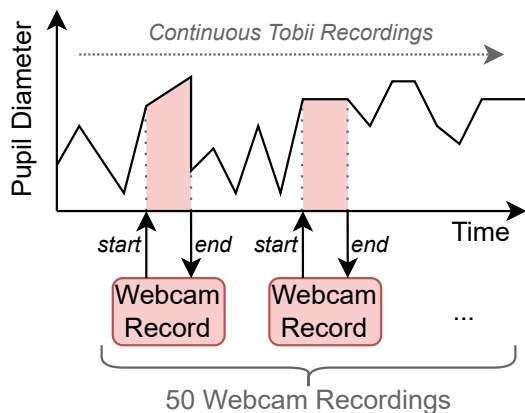[1]https://chameleon-view.netlify.app/

Figure 1. Data recording flow. Tobii eye-tracker records pupil diameter, and ChameleonView captures facial recordings using a webcam. Facial recordings start when the participant clicks on the button in the center. The start and end timestamp of the recording is collected in order to synchronize the data with an eye-tracker.

3.4.1 Aligning the Recordings

As mentioned in Section 3.2, we first synchronized the data collected from the Tobii eye-tracker with the webcam recordings for each participant, as shown in Figure 1.

We utilized a CSV file generated by the eye-tracker, which records various metrics, including the diameters of the right and left pupils, the average diameter of both pupils, and coordinates for both the left and right gaze positions. A separate timestamp CSV file also captured the start and end times for each participant's webcam recordings. The synchronization involves iterating through these timestamps and selecting all corresponding rows from the Tobii CSV file that match the start and end times specified for each recording segment. Each recording in our data collection app lasts three seconds, corresponding to three distinct timestamps per recording.

Since the video recording frame rate is 30 frames per second, this results in 90 frames per recording from our data collection application ChameleonView. Additionally, the Tobii eye-tracker operates at a frequency of 90Hz, providing 90 data points per second, resulting in 270 data points for each three-second recording.

To synchronize the 90 frames with the 270 Tobii-captured data points effectively, we concatenate each metric column separately across the 90 data points from the three unique timestamps (horizontally) in the Tobii-captured CSV file and compute a row-wise mean. This process is repeated for each recording. The first timestamp of the trio is designated as the primary timestamp for that recording, ensuring consistency in data alignment. Subsequently, we extract frames from the video recordings, yielding 90 images - aligned with 90 data points from the Tobii-eye tracker. This procedure is meticu-
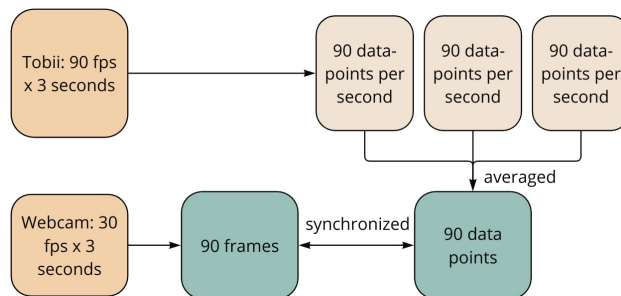

Figure 2. Data alignment flow of a single recording. To synchronize the 90 frames with the 270 Tobii-captured data points, each metric column is concatenated horizontally across the 90 data points from the three unique timestamps in the Tobii-captured CSV file, followed by computing a row-wise mean.

lously repeated for each participant to ensure uniformity and precision in the dataset preparation process. The entire data alignment flow is illustrated in Figure 2.

3.4.2 Cropping the Eyes

Once the frames are aligned with the Tobii eye-tracker data, the subsequent step involves cropping the eyes from each extracted frame. We used Mediapipe [30] for its robust face and eye detection capabilities. This tool not only identifies faces within the frames but also precisely locates facial landmarks, including the eyes, nose, iris, eyebrows, and mouth. For our purposes, we specifically utilize Mediapipe to crop the left and right eyes.

We utilize Mediapipe's face landmark detection bundle [2], which includes a series of models designed for face detection and the precise mapping of facial landmarks. The face mesh model identifies 478 distinct facial landmarks and accurately determines the coordinates for the left and right eyes and irises. Given the variability in image intensity and participants' distance from the webcam, the regions defined by these coordinates may vary in size, leading to inconsistencies in the dimensions of the cropped regions for each eye and iris.

To standardize our data, we crop a region around each eye by centering on the detected eye region and fixing the dimensions to 32 pixels in width and 16 in height. This method preserves the natural shape and scale of the eye, avoiding distortions that could result from resizing.

To improve the quality of the set of cropped images, we excluded frames where participants were blinking. For this reason, we first calculated the Eye Aspect Ratio (EAR), a concept adapted from *Soukupova et al.* [44] and tailored to suit the coordinates provided by Mediapipe's landmark detection outputs. This adaptation ensures the precise detection of blinks using the specific facial landmarks identified by
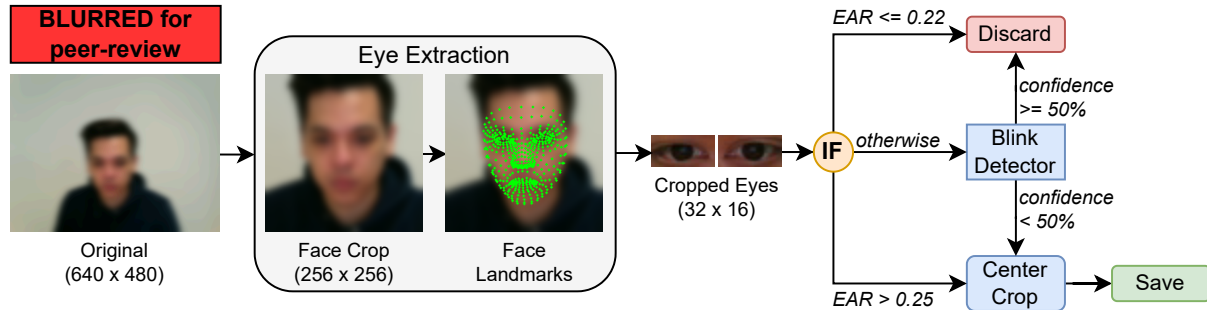
Figure 3. Pipeline of our data preprocessing. For face detection and landmark localization, we used Mediapipe to extract the respective cropped eye images (32x16), left and right, separately. Next, we applied blink detection on the cropped eyes using the Eye Aspect Ratio (EAR) and a pre-trained vision transformer for blink detection. Cropped eye images are then saved based on the EAR threshold and model confidence score.

Mediapipe and is described by

$$\text{eyes\_ratio} = \frac{\text{EAR}_{\text{left}} + \text{EAR}_{\text{right}}}{2}.$$

The EAR for the left and right eye is defined as

$$\text{EAR}_{\text{left}} = \frac{\|V_{l1} - V_{l2}\|}{\|H_{l1} - H_{l2}\|} \quad \text{and} \quad \text{EAR}_{\text{right}} = \frac{\|V_{r1} - V_{r2}\|}{\|H_{r1} - H_{r2}\|},$$

where $\|\cdot\|$ denotes the Euclidean distance between the specified points, V_{r1}, V_{r2} and V_{l1}, V_{l2} are the vertical landmarks of the right and left eyes, respectively, and H_{r1}, H_{r2} and H_{l1}, H_{l2} are the horizontal landmarks of the right and left eyes, respectively.

Using the EAR, we categorize frames as follows: frames with an EAR of 0.22 or lower are classified as 'blink detected' (indicating closed eyes), while those with an EAR above 0.25 are classified as 'blink not detected' (indicating open eyes). Frames with EAR values between 0.22 and 0.25 fall into a gray area, where classification is influenced by individual differences in eye structure. To enhance accuracy in this gray area and minimize both false positives and negatives in our blink detection, we use a pre-trained Vision Transformer (ViT) model [10] for open or closed-eye detection, trained on the MRLeyes dataset [13]. If the ViT model predicts closed eyes with a confidence score of 0.50 or higher, we classify the eyes as blinked and discard the frame; otherwise, we save it. This dual-measure approach ensures robust blink detection. Figure 3 shows the entire pipeline, including face detection, eye localization, cropping, and blink detection.

In humans, the complete blink process generally takes between 250 to 400 milliseconds [46], with the "blackout duration," when the upper lid covers the pupil, lasting from 40 to 200 milliseconds [11, 46, 47]. Given that our blink detection system identifies blinks based on the closing and opening of eyelids using EAR thresholds, we estimate a blink to last about 200 milliseconds. With a video capture rate of 30 frames per second, this translates to roughly 6 frames per blink. Detecting and excluding frames with blinks is crucial due to the frequency of blinks throughout the recording sessions. Figure 4 shows the pruning results for a particular participant across all sessions. The blue line indicates the total number of frames per session, while the orange line shows the frames retained after blink detection and removal.

4. Result

In this section, we describe the results of our dataset collection by describing the statistics of the participants (Section 4.1), the dataset distribution (Section 4.2), and first pupil diameter estimation results (Section 4.3). We use one common participant (number 6) to exemplify all visualizations in the following.

4.1. Dataset Statistics

In this study, we recruited 51 participants, including 39 males, 11 females, and 1 individual who preferred not to disclose their gender, all of whom consented to make their eye-cropped data public. The participants' ages ranged from 21 to 44 years (M=27.58). They were predominantly workers or university students residing in Germany with origin nationalities in Eastern Europe, South Asia, Southeast Asia, North Africa, and North America. We obtained informed consent from all participants prior to the experiment to ensure compliance with the General Data Protection Regulation (GDPR). Additionally, participants were informed that they could opt out of the experiment at any time.

From an initial total of 226,912 image frames, our final dataset (after preprocessing described in Section 3.4) consists of 212,073 images of left and right eyes, curated from 51 participants after removing frames where blinks were detected. We meticulously organize each participant's left and right eye images into separate directories. Additionally,

WACV
#205

WACV
#205

WACV 2025 Submission #205. Algorithms Track. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.
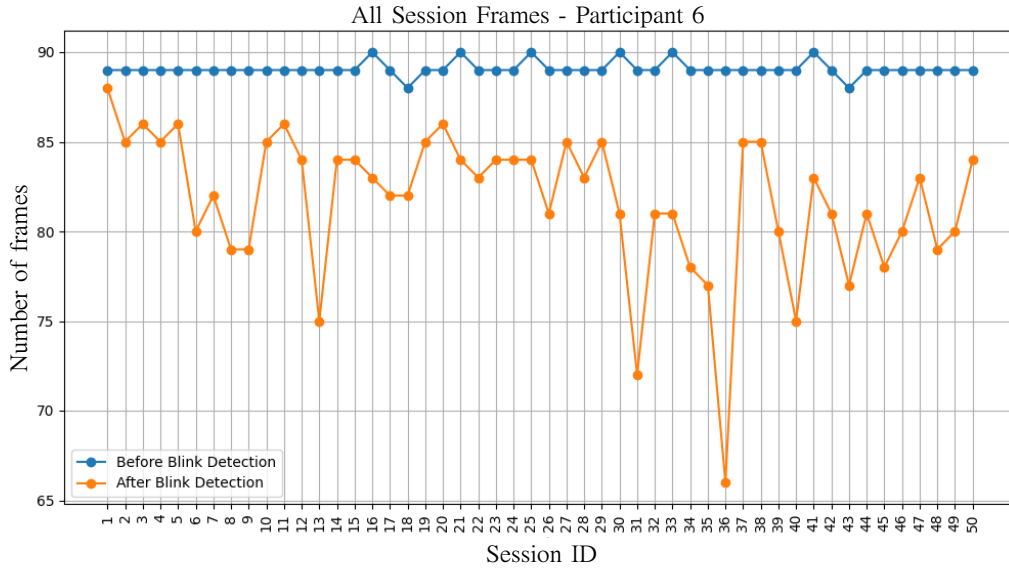

Figure 4. Visualization of total frames vs. frames after pruning due to blink detection for one participant in all recording sessions (50 in total) as outlined in Section 3.4. Note that each recording takes a total of three seconds, which is why the impact of a blink and the amount of blinks can vary significantly (around 40 - 200 ms per blink).
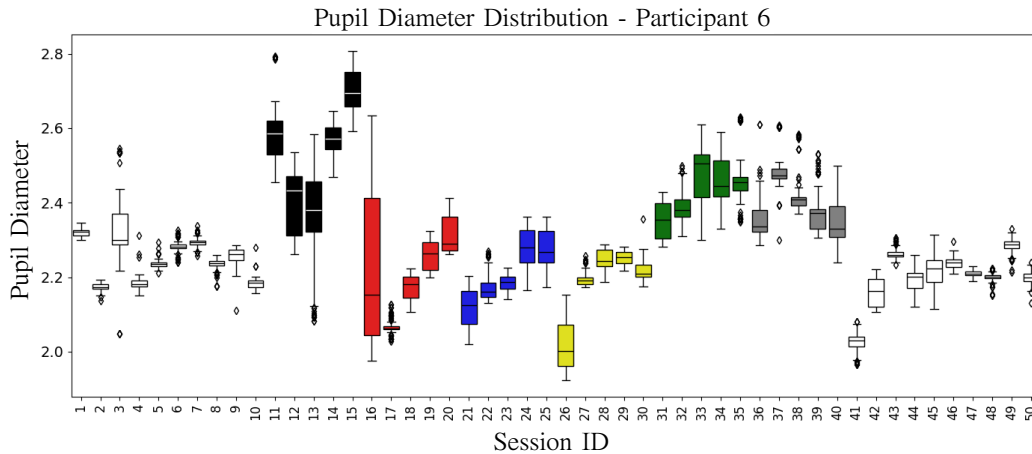

Figure 5. Pupil diameter distribution of one participant during the recordings. A set of different pupil diameter measurements and webcam images were captured during the different three-second long sessions (in total, 50 sessions). The colors of the boxes indicate the display color used during the recordings (white, black, red, blue, yellow, green, gray, and white again).

we include a CSV file for each session. This file precisely documents the timestamp, session ID, gaze data, pupil data, and the path for each frame retained in that session, ensuring a comprehensive record of the eye-tracking data post-blink-filtering. This structured approach facilitates easy access and analysis of the dataset.

4.2. Dataset Distribution

Figure 5 shows the distribution of pupil diameters using a boxplot visualization. An in-depth look at the individual pupil diameters within one recording is shown in Figure 6

for white and black display colors. For the same participant, Figure 7 shows exemplary the corresponding webcam recording images for both eyes (cropped). As explained in Section 3.2, each session shows varied display colors. The ChameleonView web application captures a three-second webcam video recording for each session, synchronized with the pupil diameter data collected by the eye-tracker over the same duration. Since the eye-tracker records data at 90Hz, each session generates 270 data points. The mean and variance depicted in the boxplots are computed from these recordings.
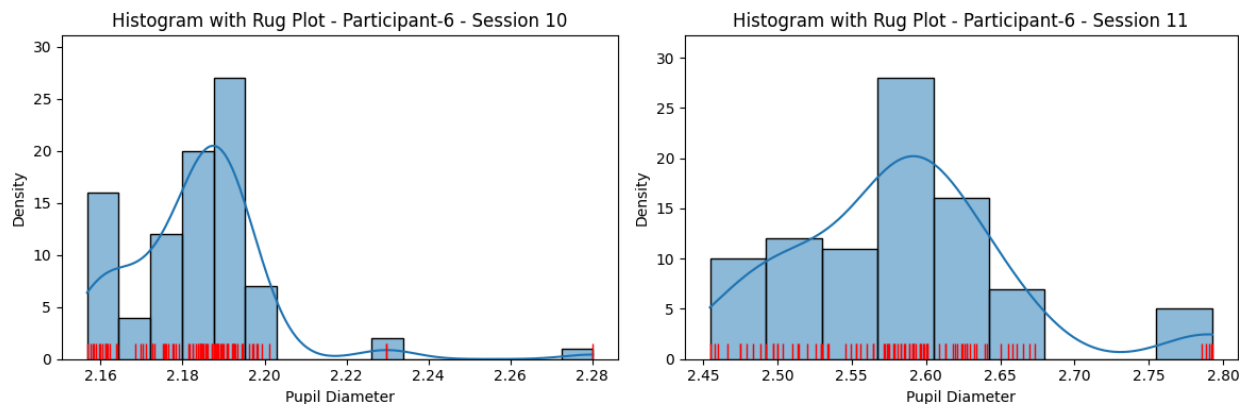
Figure 6. Two rug plot distributions of one participant with white display color (**left**) with pupil diameters ranging from 2.16mm to 2.28mm, and black display color (**right**) with pupil diameters ranging from 2.45mm to 2.80mm.

Figure 7. Example webcam recording images (first and last frame) of both eyes of one participant (same as in Figure 6) with white display color (**left**) and black display color (**right**).

Eye	Model	Validation MAE ↓	Test MAE ↓
Left	**ResNet-18**	0.0837 ± 0.0135	**0.1340** ± 0.0196
	ResNet-50	0.1001 ± 0.0197	0.1426 ± 0.0167
Right	**ResNet-18**	0.1054 ± 0.0173	**0.1403** ± 0.0328
	ResNet-50	0.1089 ± 0.0204	0.1588 ± 0.0203

Table 2. 5-fold cross-validation of ResNet-18 and ResNet-50 [15], evaluated separately for left and right eyes. Each group contains 10 randomly selected participants: 5 for validation and 5 for testing. The remaining participants were used to train the models. ResNet-18 performs the best for the pupil diameter estimation regarding mean values on the test partitions, whereas ResNet-50 shows a lower standard deviation, indicating more robustness for varied test partitions.

4.3. Pupil Diameter Estimation Result

Table 2 shows initial 5-fold cross-validation results on EyeDentify with the CNN models ResNet-18 and ResNet-50 [15]. For the left eye, the ResNet-18 model outperformed ResNet-50 with a lower MAE in both validation (0.0837 vs. 0.1001) and testing phases (0.1340 vs. 0.1426). Similarly, ResNet-18 again showed better performance for the right eye with a validation MAE of 0.1054 compared to ResNet-50's 0.1089 and a test MAE of 0.1403 versus 0.1588 for ResNet-50. The results indicate that ResNet-18 consistently yields lower error rates than ResNet-50 across both eyes and both evaluation phases. The uncertainties represented by the standard deviations suggest some variability in the model performances across different subsets of the data, where ResNet-50 seems to be more consistent on the respective test partitions.

Figure 8 shows the class activation map (CAM) [55] of the last convolution layer of ResNet18 and ResNet50 respectively. The figure illustrates that ResNet18 primarily focuses on the external regions of the eyes, capturing environmental and skin color variations, with minimal attention (left eye) to no attention (right eye) given to the iris. For the left eye, ResNet50 also tends to focus strongly on the outer regions of the eye and mostly outside the iris, attempting to learn

color intensities rather than concentrating on the iris or pupil. And, for the right eye, ResNet50 primarily focuses on the inner regions of the eye but neglects the surrounding color variations. These observations indicate that while focusing on the iris is crucial for accurately estimating pupil diameter, accounting for environmental variations can enhance performance and make the model more consistent.

5. Limitations

EyeDentify substantially contributes to the eye monitoring research community. Nonetheless, it is important to acknowledge some limitations of our work to guide future research and application.

Firstly, the data was recorded using a single camera model, limiting the diversity in recording devices. Given that webcams vary widely in resolution and optical characteristics, we recommend future studies to validate and possibly enhance our initial pupil diameter estimation models across
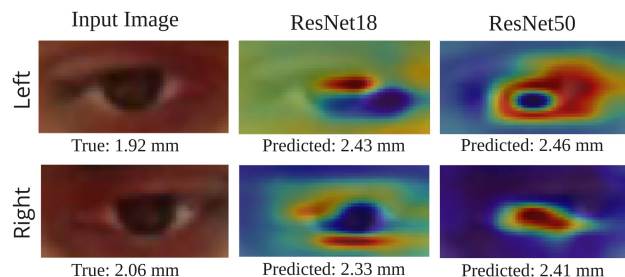
WACV
#205

WACV 2025 Submission #205. Algorithms Track. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

WACV
#205

Figure 8. ResNet18 ad ResNet50 test samples - CAM (class activation map) [55] visualizations with red display color. True and Predicted values show the original and estimated pupil diameter of the left and right eye in millimeters.

multiple camera types to ensure robustness.

Secondly, our EyeDentify primarily consists of participants without eyeglasses to maintain consistent conditions across all data. This exclusion limits the applicability of our findings to the general population, as eyeglass wearers comprise a significant demographic. Future work should consider this when they train models on EyeDentify for practical applications where users with eyeglasses are possible.

Lastly, EyeDentify does not account for personal eye characteristics that might influence measurements, such as eye color or health conditions. While we avoided collecting detailed personal meta eye data due to privacy concerns and the complexities of accurately self-reporting eye conditions, this omission can affect the applicability of our findings. Collaborating with ophthalmologists could enhance the collection process, ensuring accurate and privacy-sensitive inclusion of such data in future studies.

6. Future Work

Future work includes leveraging super-resolution models to enhance image quality, thereby providing more detailed features for deeper and more complex neural networks [32, 33]. For instance, face restoration or enhancement models such as GFPGAN [48] and CodeFormer [56], which have pre-learned facial features, can be particularly beneficial. Additionally, other state-of-the-art super-resolution models like regression-based transformed (SwinIR [27] or HAT [6]) or generative diffusion models (SR3 [41] or YODA [31]) can improve the resolution of webcam-captured eye images. These models can upscale images by factors of 2x, 3x, or 4x, facilitating the training of models on high-resolution images to estimate pupil diameter accurately, similar to gaze estimation using super-resolution [36, 51].

Furthermore, the features learned by models trained on high-resolution eye images for tasks such as gaze estimation (e.g., Appearance-based Gaze Estimation [7] and Super-Resolution for Appearance-Based Gaze Estimation [36]),

blink detection (e.g., Real-Time Eye Gaze and Blink Estimation in Natural Environments [8]), or segmentation of eye parts and regions (e.g., Iris Segmentation Model [50]) can be employed for transfer learning. This approach can be applied to the readily available low-resolution images or the high-resolution images generated by super-resolution models, enhancing pupil detection's overall performance and accuracy.

7. Societal Impact

We focus on publishing a dataset of everyday webcam images, which we hope will advance research about various fundamental physiological and psychological states and contribute to multiple human activity and healthcare domains. As such, these images and their associated pupil diameter information pose no immediate threat to individuals or organizations. The images were properly anonymized (by providing only cropped eyes). They do not depict any human or personal data that can be associated with any particular person due to the insufficient resolution of eye images for identification. As discussed in the limitations, EyeDentify may introduce biases inherent in the training data, i.e., missing nationalities or no eyeglasses. These biases could affect the quality and fairness of the pupil diameter estimations, particularly in diverse or underrepresented populations. Future work should aim to mitigate these biases and ensure that pupil diameter estimation technologies are equitable and inclusive.

8. Conclusion

We introduced EyeDentify, a novel dataset aimed at significantly enhancing eye monitoring research by enabling the development of models that estimate pupil diameter using standard webcam images. Such advancements are critical for understanding various physiological and psychological states through non-invasive methods, broadening the scope of applications in human-computer interaction, behavioral studies, and health diagnostics.

Moreover, EyeDentify represents a significant advancement by addressing the scarcity of publicly available datasets that combine eye images with precise pupil diameter annotations. By focusing on standard webcam recordings, EyeDentify democratizes access to pupil-related research, potentially expanding its applicability in low-resource settings and everyday computing environments. Furthermore, our findings demonstrate that models trained on EyeDentify, particularly the ResNet-18 architecture, achieve promising results in estimating pupil diameters with a high degree of accuracy.

Acknowledgements

WACV #205

WACV 2025 Submission #205. Algorithms Track. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

WACV #205

References

[1] Tobii AB. Specifications for the tobii eye tracker 4c. https://help.tobii.com/hc/en-us/articles/213414285-Specifications-for-the-Tobii-Eye-Tracker-4C, 2024. 1, 3

[2] Google AI. Face landmarker — mediapipe. https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker. 4

[3] Apple Inc. Apple announces new accessibility features, including eye tracking, music haptics, and vocal shortcuts, May 2024. Accessed: 2024-06-06. 1

[4] Ankur Bhatt, Ko Watanabe, Andreas Dengel, and Shoya Ishimaru. Appearance-based gaze estimation with deep neural networks: From data collection to evaluation. *International Journal of Activity and Behavior Computing*, 2024(1):1–15, 2024. 1

[5] Meo Vincent C. Caya, Christian Jorel P. Rapisura, and Renz Rienard B. Despabiladeras. Development of pupil diameter determination using tiny-yolo algorithm. In *2022 IEEE 14th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–6, 2022. 1, 2, 3

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pages 22367–22377, 2023. 8

[7] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 8

[8] Kevin Cortacero, Tobias Fischer, and Yiannis Demiris. Rt-bene: A dataset and baselines for real-time blink estimation in natural environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 8

[9] David Dembinsky, Ko Watanabe, Andreas Dengel, and Shoya Ishimaru. Eye movement in a controlled dialogue setting. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, ETRA '24, New York, NY, USA, 2024. Association for Computing Machinery. 2

[10] Dima806. Closed eye image detection using vision transformer (vit). https://www.kaggle.com/code/dima806/closed-eye-image-detection-vit. 5

[11] Marshall G Doane. Interactions of eyelids and tears in corneal wetting and the dynamics of the normal human eyeblink. *American journal of ophthalmology*, 89(4):507–516, 1980. 5

[12] Abdolhossein Fathi and Fardin Abdali-Mohammadi. Camera-based eye blinks pattern detection for intelligent mouse. *Signal, Image And Video Processing*, 9:1907–1916, 2015. 1

[13] R. Fusek. Pupil localization using geodesic distance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11241 LNCS:433–444, 2018. 5

[14] Robert Greinacher and Jan-Niklas Voigt-Antons. Accuracy assessment of arkit 2 based gaze estimation. In Masaaki Kurosu, editor, *Human-Computer Interaction. Design and User Experience*, pages 439–449, Cham, 2020. Springer International Publishing. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[16] Yoichiro Hisadome, Tianyi Wu, Jiawei Qin, and Yusuke Sugano. Rotation-constrained cross-view feature fusion for multi-view appearance-based gaze estimation. In *WACV*, pages 5985–5994, January 2024. 1

[17] Jeongmin Hong, Joseph Shin, Juhee Choi, and Minsam Ko. Robust eye blink detection using dual embedding video vision transformer. In *WACV*, pages 6374–6384, January 2024. 1, 2

[18] Yuqi Hou, Zhongqun Zhang, Nora Horanyi, Jaewon Moon, Yihua Cheng, and Hyung Jin Chang. Multi-modal gaze following in conversational scenarios. In *WACV*, pages 1186–1195, January 2024. 2

[19] Guilei Hu, Yang Xiao, Zhiguo Cao, Lubin Meng, Zhiwen Fang, Joey Tianyi Zhou, and Junsong Yuan. Towards real-time eyeblink detection in the wild: Dataset, theory and practices. *IEEE Transactions on Information Forensics and Security*, 15:2194–2208, 2020. 2

[20] Siddhartha Joshi, Yin Li, Rishi M Kalwani, and Joshua I Gold. Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, 89(1):221–234, 2016. 1

[21] Daniel Kahneman and Jackson Beatty. Pupil diameter and load on memory. *Science*, 154(3756):1583–1585, 1966. 1

[22] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, , and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, October 2019. 1, 2

[23] Ivan Khokhlov, Egor Davydenko, Ilya Osokin, Ilya Ryakin, Azer Babaev, Vladimir Litvinenko, and Roman Gorbachev. Tiny-yolo object detection supplemented with geometrical data. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pages 1–5. IEEE, 2020. 3

[24] Joji Kotani, Hiroyuki Nakao, Isamu Yamada, Atsushi Miyawaki, Naomi Mambo, and Yuko Ono. A novel method for measuring the pupil diameter and pupillary light reflex of healthy volunteers and patients with intracranial lesions using a newly developed pupilometer. *Frontiers in Medicine*, 8, 2021. 1

[25] Krzysztof Krejtz, Andrew T Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one*, 13(9):e0203629, 2018. 1

[26] Michal T Kucewicz, Jaromir Dolezal, Vaclav Kremen, Brent M Berry, Laura R Miller, Abigail L Magee, Vratislav Fabian, and Gregory A Worrell. Pupil size reflects successful encoding and recall of memory in humans. *Scientific reports*, 8(1):4949, 2018. 1

[27] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 8

[28] Mengxi Liu, Sizhen Bian, and Paul Lukowicz. Non-contact, real-time eye blink detection with capacitive sensing. In *Proceedings of the 2022 ACM International Symposium on*

WACV
#205

WACV 2025 Submission #205. Algorithms Track. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

WACV
#205

Wearable Computers, ISWC '22, page 49–53, New York, NY, USA, 2022. Association for Computing Machinery. 1

[29] Holger Lüdtke, Barbara Wilhelm, Martin Adler, Frank Schaeffel, and Helmut Wilhelm. Mathematical procedures in data recording and processing of pupillary fatigue waves. *Vision research*, 38(19):2889–2896, 1998. 1

[30] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 4

[31] Brian B Moser, Stanislav Frolov, Federico Raue, Sebastian Palacio, and Andreas Dengel. Yoda: You only diffuse areas. an area-masked diffusion approach for image super-resolution. *arXiv preprint arXiv:2308.07977*, 2023. 8

[32] Brian B Moser, Federico Raue, Stanislav Frolov, Sebastian Palacio, Jörn Hees, and Andreas Dengel. Hitchhiker's guide to super-resolution: Introduction and recent advances. *IEEE TPAMI*, 45(8):9862–9882, 2023. 8

[33] Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion models, image super-resolution and everything: A survey. *arXiv preprint arXiv:2401.00736*, 2024. 8

[34] Peter R Murphy, Redmond G O'connell, Michael O'sullivan, Ian H Robertson, and Joshua H Balsters. Pupil diameter covaries with bold activity in human locus coeruleus. *Human brain mapping*, 35(8):4140–4154, 2014. 1

[35] Yuming Ni and Biao Sun. A remote free-head pupillometry based on deep learning and binocular system. *IEEE Sensors Journal*, 19(6):2362–2369, 2019. 3

[36] Galen O'Shea and Majid Komeili. Toward super-resolution for appearance-based gaze estimation. *arXiv preprint arXiv:2303.10151*, 2023. 1, 8

[37] Marco Pedrotti, Mohammad Ali Mirzaei, Adrien Tedesco, Jean-Rémy Chardonnet, Frédéric Mérienne, Simone Benedetto, and Thierry Baccino. Automatic stress classification with pupil diameter analysis. *International Journal of Human-Computer Interaction*, 30(3):220–236, 2014. 1

[38] Bastian Pfleging, Drea K. Fekety, Albrecht Schmidt, and Andrew L. Kun. A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5776–5788, New York, NY, USA, 2016. Association for Computing Machinery. 1

[39] Brendan L Portengen, Giorgio L Porro, Saskia M Imhof, and Marnix Naber. The trade-off between luminance and color contrast assessed with pupil responses. *Translational Vision Science & Technology*, 12(1):15–15, 2023. 3

[40] Manola Ricciuti and Ennio Gambi. Pupil diameter estimation in visible light. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1244–1248, 2021. 2, 3

[41] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 45(4):4713–4726, 2022. 8

[42] Juni Nurma Sari, Adi N Hanung, Edi N Lukito, P. Insap Santosa, and Ridi Ferdiana. A study on algorithms of pupil diameter measurement. In *2016 2nd International Conference on Science and Technology-Computer (ICST)*, pages 188–193, 2016. 1

[43] K J Shanti, Bharatkumar Hegde, P R Shreya, and M M Chaitra. Automated system for determination of pupil size using multispectral imaging. In *2021 IEEE Bombay Section Signature Conference (IBSSC)*, pages 1–4, 2021. 1

[44] Tereza Soukupova and Jan Cech. Eye blink detection using facial landmarks. In *21st computer vision winter workshop, Rimske Toplice, Slovenia*, volume 2, 2016. 4

[45] Ruud L Van Den Brink, Peter R Murphy, and Sander Nieuwenhuis. Pupil diameter tracks lapses of attention. *PloS one*, 11(10):e0165274, 2016. 1

[46] Frances C Volkmann, Lorrin A Riggs, Aimee G Ellicott, and Robert K Moore. Measurements of visual suppression during opening, closing and blinking of the eyes. *Vision research*, 22(8):991–996, 1982. 5

[47] Frances C Volkmann, Lorrin A Riggs, and Robert K Moore. Eyeblinks and visual suppression. *Science*, 207(4433):900–902, 1980. 5

[48] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 8

[49] Barry Winn, David Whitaker, David B Elliott, and Nicholas J Phillips. Factors affecting light-adapted pupil size in normal human subjects. *Investigative ophthalmology & visual science*, 35(3):1132–1137, 1994. 3

[50] Worldcoin. Iris semantic segmentation. https://huggingface.co/Worldcoin/iris-semantic-segmentation. 8

[51] Jun-Seok Yun, Youngju Na, Hee Hyeon Kim, Hyung-Il Kim, and Seok Bong Yoo. Haze-net: High-frequency attentive super-resolved gaze estimation in low-resolution face images. In *ACCV*, pages 3361–3378, 2022. 1, 8

[52] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *ECCV*, pages 365–381. Springer, 2020. 1, 2

[53] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE TPAMI*, 41(1):162–175, 2019. 1, 2

[54] Zhengyou Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 666–673 vol.1, 1999. 3

[55] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 7, 8

[56] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *NeurIPS*, 35:30599–30611, 2022. 8