Gaze360: Physically Unconstrained Gaze Estimation in the Wild

Petr Kellnhofer*[1], Adrià Recasens*[1], Simon Stent[2], Wojciech Matusik[1], and Antonio Torralba[1]

[1] Massachusetts Institute of Technology, Cambridge MA 02139, USA
[2] Toyota Research Institute, Cambridge, MA, 02139, USA

{pkellnho,recasens,wojciech,torralba}@csail.mit.edu simon.stent@tri.global

∗ indicates equal contribution

Abstract

Understanding where people are looking is an informative social cue. In this work, we present Gaze360*, a large-scale gaze-tracking dataset and method for robust 3D gaze estimation in unconstrained images. Our dataset consists of 238 subjects in indoor and outdoor environments with labelled 3D gaze across a wide range of head poses and distances. It is the largest publicly available dataset of its kind by both subject and variety, made possible by a simple and efficient collection method. Our proposed 3D gaze model extends existing models to include temporal information and to directly output an estimate of gaze uncertainty. We demonstrate the benefits of our model via an ablation study, and show its generalization performance via a cross-dataset evaluation against other recent gaze benchmark datasets. We furthermore propose a simple self-supervised approach to improve cross-dataset domain adaptation. Finally, we demonstrate an application of our model for estimating customer attention in a supermarket setting. Our dataset and models are available at* http://gaze360.csail.mit.edu.

1. Introduction

In order to better understand humans – their desires, intents and states of mind – one must be able to observe and perceive certain behavioral cues. Eye gaze direction is one such cue: it is a strong form of non-verbal communication, signalling engagement, interest and attention during social interactions [1]. Detecting and following where another person is looking is a skill developed early on in a child's life – four-month-old infants are known to use eye gaze cuing to help visually process objects, for example [21]. Just as a parent's gaze can help to guide a child's attention, human gaze fixations have also been found to be useful in helping machines to learn or interact in various contexts [18, 22].

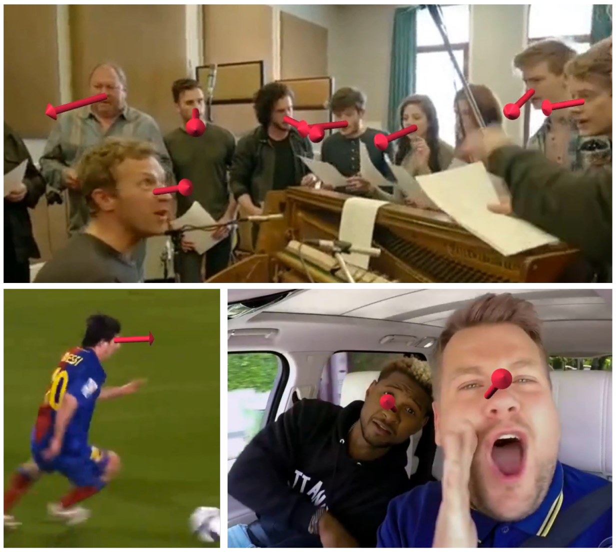

Figure 1. **Overview**: we introduce a novel dataset and method for estimating 3D gaze in-the-wild. This figure illustrates our model's output on unseen video gathered from YouTube, demonstrating its robustness to diverse, physically unconstrained scenes.

In recent years, while methods for related human modeling problems such as 2D body pose and face tracking have achieved impressive success by leveraging the representational power of deep convolutional neural networks along with very large annotated datasets [2, 6, 9, 14, 26], methods for gaze estimation have not yet reached such levels of performance. This is primarily due to the lack of sufficiently large and diverse annotated training data for the task. Collecting precise and highly varied gaze data with ground truth, particularly outside of the lab, is a challenging task.

In this work, we introduce an approach to help tackle this task and narrow the perceived performance gap:

- we first describe a methodology to efficiently collect annotated 3D gaze data in arbitrary environments;
- we use our method to acquire the largest 3D gaze

dataset in the literature by subject and variety, capturing video of 238 subjects in indoor and outdoor conditions, and we carefully evaluate the error and characteristics of the dataset;

- we train a variety of 3D gaze estimation models on the dataset before converging on a final model which uniquely takes a multi-frame input (to help resolve single frame ambiguities) and employs a pinball regression loss for error quantile regression to provide an estimate of gaze uncertainty;
- we demonstrate the usefulness of our dataset versus existing datasets by means of a cross-dataset model performance comparison (training on one dataset and testing on another), and introduce a simple method for self-supervised domain adaptation of gaze models;
- finally we demonstrate how our Gaze360 model can be applied to real-world use cases, such as estimating a customer's focus of attention in a supermarket.

2. Related Work

Gaze datasets. A summary of comparable gaze datasets is shown in Table 1. While many gaze-related datasets have been published in recent years [10, 12, 16, 17, 20, 23, 27, 30, 31], they are mostly geared towards physically constrained applications such as desktop or smartphone gaze tracking. Typically, these datasets are captured using a static recording setup [17, 20, 27, 34] or a camera integrated in a smartphone [10, 12, 28]. The static approach allows for more control and higher accuracy but can lack the diversity in illumination and motion blur useful for more general applications. Smartphone-based solutions overcome these flaws and have the advantage of straightforward scaling via crowd-sourcing to increase the subject variety. However, they lack head pose and gaze variability due to the collocation of the device's camera and screen, as well as the screen's relatively narrow area for projecting targets.

To try to capture the nature of human gaze in arbitrary natural scenes, it is important not to overly constrain the subject's pose, allowing for coverage over the full gamut of head and eyeball orientations in relation to the camera. While some existing datasets have relatively small head pose and gaze variation [16, 17, 20], others do provide a wider range [12, 27, 34] but are still restricted to primarily frontal rather than oblique views. While it is true that the eyes become increasingly occluded at larger angles of head yaw, we wish to capture such cases so that our model can be used in less constrained settings.

In one of the most comprehensive datasets from Zhu and Deng [34], the authors increased acquisition speed and viewpont variety by using an array of cameras in different poses. However, the setup was restricted to collecting data in the lab environment. While our approach also uses a multi-camera setup, our goal was to quickly acquire many subjects at once, using a free-moving rather than fixed target that allowed us to capture the full range of gaze directions, as described in Fig. 4 and Section 4. Moreover, as our capture setup is mobile, this allowed us to efficiently collect data from a broad demographic in more varied natural lighting environments, including a wider range of scale variation and image blur from subject motion during capture. This more closely approximates the domains of systems such as interactive robots or surveillance/monitoring cameras which might benefit from our gaze tracking model.

Table 1. **A comparison of popular gaze datasets.** The type and range of gaze labels, number of subjects and completeness of image data publicly available. *Full* stands for full face images, *Eyes* denotes crops of eye regions and *N/A* means that the dataset was not available for use. Asterisks indicate datasets containing partially occluded face images.

Dataset	Gaze	Range	# Subj.	Image	Outdoor
TabletGaze [11]	2D	∼80°	51	Eyes	No
iTracker [12]	2D	∼100°	1,450	Full	Partially
UT MV [23]	3D	∼50°	50	Eyes	No
Columbia [20]	3D	60°	56	Full*	No
RT-GENE [4]	3D	75°	15	Full*	No
MPIIFaceGaze [31]	3D	∼80°	15	Full	No
EYEDIAP [17]	3D	90°	16	Full	No
Weidenb. [27]	3D	180°	20	N/A	No
Zhu [34]	3D	180°	200	N/A	No
Gaze360 [ours]	**3D**	**360°**	**238**	**Full**	**Yes**

A recent work which also addresses gaze estimation in natural settings with larger camera-subject distances and less constrained subject motion, is that of [4]. Their approach to dataset generation was target-free, but required subjects to wear gaze-tracking glasses, used motion capture cameras to recover head pose, and needed a complicated semantic in-painting step to remove the gaze tracking glasses from the target image. In comparison, our approach is relatively simple, allowing us to scale to many more subjects (238 versus 15) and lighting conditions.

Geometric gaze models: Geometric models often use corneal reflections of near infra-red light sources [8, 29, 35] or other light sources with known geometry [10] to fit a model of the eyeball from which gaze can be inferred. Since these methods rely on a physical model, they generalize quite easily to new subjects with little or no training data, but at the cost of higher sensitivity to input noise such as partial occlusions or lighting interference. Since they also rely on a fixed light source, they are not feasible in unconstrained settings such as ours.

Appearance-based gaze models: Appearance-based methods learn a more direct image-to-gaze mapping, using large datasets of annotated eye or face images. Support vector regression [28], random forests [11] and most recently deep learning [4, 12, 30, 31, 34] have been applied in this way. A preprocessing step of eye or face detection is often required [12, 30]. Our model does not rely on eye or face

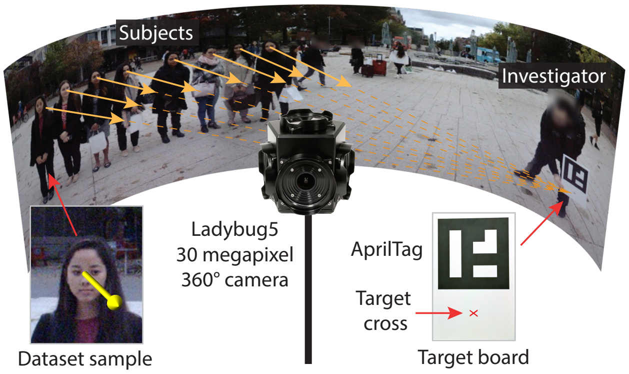

Figure 2. **Acquisition setup**. Our setup allows us to efficiently collect large volumes of diverse, annotated data for 3D gaze estimation. We create a dataset with 238 subjects in a wide range of lighting conditions (both indoor and outdoor) and distances and angles to subjects.

detectors, which enables it to achieve higher robustness in unconstrained settings when the required features become partially occluded. Dependency between gaze and head pose can either be handled by training implicitly [12, 30, 31] or modeled explicitly with separate branches [34].

Gaze estimation becomes more difficult under partial occlusion of eyes. Even at $90 - 135°$ head yaw a significant part of one eyeball is often still visible and informative for gaze estimation (see Supplemental). Existing methods [12, 32] do not deal with these cases and typically assume that the subject is facing the camera. However, such models do not generalize well to challenging applications such as in robotics or surveillance. Unlike previous approaches, our model is designed to cope with such situations by always providing best effort prediction along with an appropriate confidence measure. We learn to predict uncertainty via quantile regression [15] learned using a pinball loss. Our model outputs an estimated gaze direction even with fully occluded eyes by relying on visible head features, while at the same time informing about the limited accuracy of its prediction by outputting a correspondingly higher uncertainty value. In addition, unlike previous models, we investigate the use of additional frames to improve gaze estimates through the aggregation of image evidence over time. This increases the chance of capturing relevant features that may only be visible in few frames. We show how using motion significantly helps the system performance over a wide range of view angles.

3. Dataset collection method

There is currently no dataset suitable to learn a model capable of robustly estimating 3D gaze in-the-wild. Previous efforts to record large-scale datasets relied on careful acquisition setups with precisely measured subject and gaze target positioning [17, 23, 34]. Such setups are nearly impossible to move to different locations, can only record single subjects at a time and require constant verification of the desired gaze from the subject which makes the collection process inflexible and very slow. This is the reason why all existing datasets with 3D gaze labels are recorded in indoor environments and frequently use few subjects. As evidenced by the success of 2D body and face tracking models in the wild [2], to improve in-the-wild robustness it is important to collect data with a large number of different subjects, large variation in natural illumination and a wide range of head poses and gaze directions.

3.1. Setup

To tackle these issues we opted for a setup built around a Ladybug5 360° panoramic camera (Fig. 2) placed on a tripod in the center of the scene, and a large moving rigid target board marked with an AprilTag [25] and a cross on which subjects were instructed to continuously fixate. This allowed data from multiple subjects to be recorded simultaneously. The Ladybug5 consists of five synchronized and overlapping 5 megapixel camera units each with 120° horizontal field of view, plus one additional upward-facing camera which we do not use. We store each frame as 3382×4096 pixels image after fish-eye lens rectification. The face of a subject standing one meter away from the camera could be fully captured in at least one of the views. The camera is factory-calibrated and we rectified all images after capture to remove barrel distortion. The compactness of the setup, consisting of a single camera unit on a tripod together with a laptop and portable power source, allowed for easy portability and deployment for efficient data collection in many environments.

Subject positioning. To build the dataset, we use AlphaPose [3] to detect the position of head keypoints and feet of subjects in rectified frames from each camera unit independently. For very close subjects whose feet are beyond the camera field of view, we use the average body proportions of standing subjects to estimate their feet position from their hip position. The Ladybug camera provides a 3D ray in a global Ladybug Cartesian coordinate system $L = [\mathbf{L}_x, \mathbf{L}_y, \mathbf{L}_z]$ for every image pixel. We use it to derive the position of feet and eyes in spherical coordinates. The remaining unknown variable is the distance from Ladybug origin to eyes, d. We exploit a measured camera height above the horizontal ground plane that the camera and all subjects stand on. Although this limits our training data collection to flat surfaces, it is not restrictive at test-time. For further details on the trigonometry, please consult the supplementary materials.

Target positioning: Our target consists of a white board with a large AprilTag [25] on one side and a smaller cross beside it on both sides (Fig. 2). The cross serves as a gaze fixation target for the study subjects while the tag is used

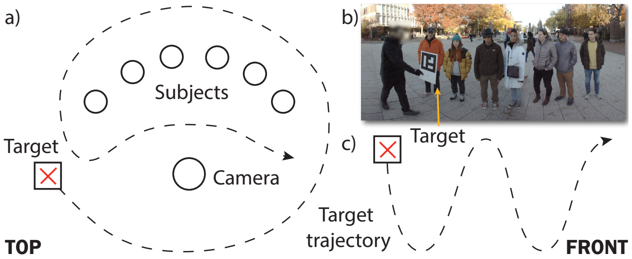

Figure 3. **Dataset collection protocol**: (a) the top view of the scene and target board trajectory showing full coverage around the subjects; (b) the image of the scene from the camera (stitched for illustration only); (c) the side view of the scene and target board trajectory showing large induced variation in pitch to the target.

for tracking of the board in 3D space. We use the original AprilTag library to detect the marker in each of the camera views and estimate its 3D pose using the known camera calibration parameters and marker size. We then use the pose and known board geometry to find the 3D location of the target cross $\mathbf{p}_t$.

Gaze direction: We compute the gaze vector in the Ladybug coordinate system as a simple difference $\mathbf{g}_L = \mathbf{p}_t - \mathbf{p}_e$. However, such a form would change with rotation of the camera and its coordinate system L. To remedy this, we express the gaze in the observing camera's Cartesian eye coordinate system $E = [\mathbf{E}_x, \mathbf{E}_y, \mathbf{E}_z]$. E is defined so that the origin is $\mathbf{p}_e$, $\mathbf{E}_z$ has the same direction as $\mathbf{g}_L$ and $\mathbf{E}_x$ lies in a plane defined by $\mathbf{L}_x$ and $\mathbf{L}_y$ (no roll). We can then convert the gaze vector to the eye coordinate system by:

$$\mathbf{g} = E \cdot \frac{\mathbf{g}_L}{||\mathbf{g}_L||_2}. \quad (1)$$

This definition of gaze direction guarantees that $\mathbf{g} = [0, 0, -1]$ when the subject looks directly at the camera, independently of the subject's position, and in general allows to express the gaze orientation from the local appearance of the head without the need for any global context.

3.2. Acquisition procedure

We acquired an institution review board approval for our dataset collection experiment. Subjects were instructed to stand around a camera at a distance of between $1 - 3$m (average 2.2m) and continuously track the target cross on the side of the marker board visible to them (Fig. 3). For safety, subjects were instructed to stay approximately in their starting locations as they would not be able to both track the target and see possible obstacles while moving.

The marker board was manipulated by one of the investigators who carried it once in a large loop around both the subjects and the camera ($2 - 5$m radius) and then in between the camera and subjects (Fig. 3a). While in motion, the target board was simultaneously moved up and down (Fig. 3c) to elicit gaze pitch variation. The loop part of the trajectory allowed to cover all possible gaze directions. The inner path was added to sample more extreme gaze pitch variation which can only be achieved from a closer distance due to limitations on the vertical position of the marker in the scene. We ensured that the marker board was always positioned to face the camera with the AprilTag as fronto-parallel as possible to reduce pose estimation error (Fig. 3b).

In order to capture a wide range of relative eyeball and head poses, we alternated between "move" and "freeze" instructions during each capture. While in the "move" state, subjects were allowed to naturally orient their head and body pose to help track the target. When the "freeze" instruction was issued, subjects were only allowed to move their eyes while maintaining a fixed head pose if possible.

4. Gaze360 dataset summary

Our dataset is unique for its combination of 3D gaze annotations, wide range of gaze and head poses, variety of indoor and outdoor capture environments and diversity of subjects. It is only surpassed in number of subjects by the GazeCapture [12] dataset (1,450 subjects), which is 2D and covers only a narrow gaze range for a limited use case. See Table 1 for a dataset comparison. Notably, our dataset is also the first to provide these qualities for short continuous videos (8 Hz).

Summary statistics. We collected 238 subjects in 5 indoor (53 subjects) and 2 outdoor (185 subjects) locations over 9 recording sessions. This is an acquisition speed that is unmatched by other on-site techniques and can only be compared to crowd-sourced approaches which, however, cannot compete in terms of experimental control. In total we acquired 129K training, 17K validation and 26K test images with gaze annotation. For privacy reasons we did not survey additional data about our subjects, but a visual inspection shows a wide distribution of subject ages, ethnicities and genders (58 % female, 42 % male). Please refer to Fig. 5 for examples.

Data distribution. We plot the angular distribution of the gaze labels covered by our and several other datasets using the Mollweide projection in Fig. 4. This illustrates how our dataset covers the entire horizontal range of 360°. While a portion of these gaze orientations correspond to fully occluded eyes (facing away from the camera), our dataset allows for gaze estimation up to the limit of eye visibility. This limit can, in certain cases, correspond to gaze yaws of approximately ±140° (where the head pose is at 90° such that one eye remains visible, and that eye is a further 50° rotated). The vertical range is limited by the achievable elevation of the marker. Sampling is less dense in the rear region (around the left and right borders of the map). This can be explained by occlusion of the target board by the subjects.

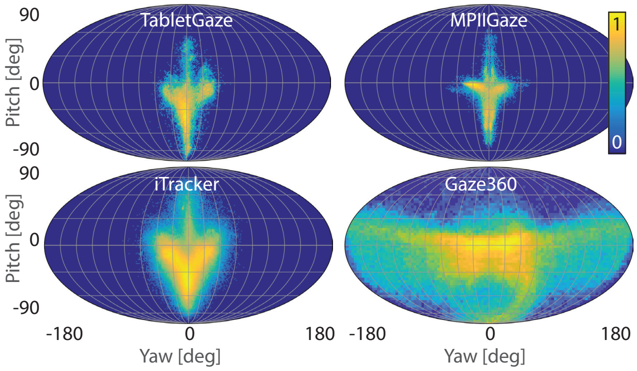

Figure 4. **Dataset statistics.** Joint distributions of the gaze yaw and pitch for TabletGaze [10], MPIIFaceGaze [31], iTracker [12] and our Gaze360 dataset. The Mollweide projection used to visualize the full unit sphere surface. All intensities are logarithmic.

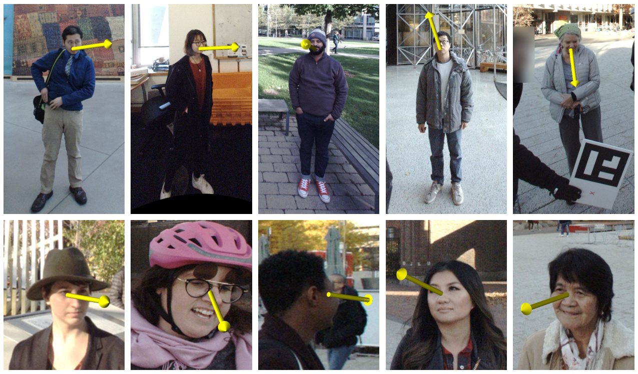

Figure 5. **Gaze360 dataset samples**: showing the diversity in environment, illumination, age, sex, ethnicity, head pose and gaze direction. Top: full body crops; bottom: closer-up head crops. Yellow arrows show measured ground-truth gaze.

Error characterization. In order to validate the accuracy of our gaze annotations we conducted a control experiment. We followed the standard acquisition procedure with our 360° camera and a single participant at a time wearing an additional front-facing test camera mounted above the right eye. We measured the 3D gaze in the test camera using the standard AprilTag based procedure and the known origin coinciding with the camera. Additional AprilTags in the background were used to register both cameras. We measured the mean difference between both gaze labels to be 2.9° over three recordings of two subjects. This is well within the error of appearance-based eye tracking at distance, validating our acquisition procedure as a means of collecting an annotated 3D gaze dataset.

5. Gaze360 model

Gaze is a naturally continuous signal. Gaze fixations and transitions yield a sequence of gaze directions. To exploit this, we propose a video-based gaze-tracking model using bidirectional Long Short-Term Memory capsules (LSTM) [5], which provide a means of modeling sequences where the output for one element is dependent on both past and future inputs. In this paper, we utilize sequences of 7 frames to predict the gaze of the central frame. Note that other sequence lengths including a single central frame alone are also possible.

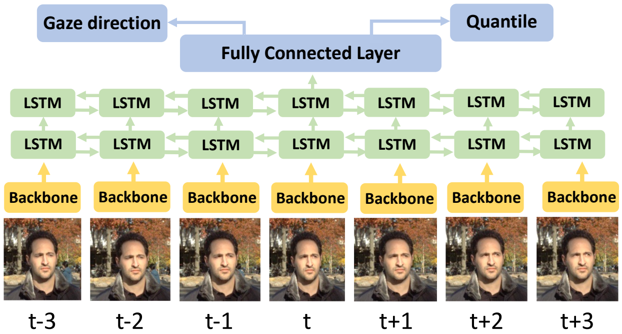

Figure 6. **Gaze360 model architecture**. The model receives multiple frames of input which are passed through a backbone network. The output for each frame is fed to a bidirectional LSTM to produce the compact representation which is used to make the final prediction of gaze direction and quantile regression. We use a 7-frame input window centered around the target frame.

Fig. 6 illustrates the architecture of the Gaze360 model. A head crop from each frame is individually processed by a convolutional neural network (backbone), which produces high-level features with dimensionality 256. These features are fed to bidirectional LSTMs with two layers which digest the sequence within forward and backward vectors. Finally, these vectors are concatenated and passed through a fully connected layer to produce two outputs: the gaze prediction and an error quantile estimation.

The gaze prediction output regresses the angle of the gaze relative to the camera view. In previous work, 3D gaze was predicted as a unit gaze vector [17, 34] or as its spherical coordinates [23, 31]. We use spherical coordinates which we believe to be more naturally interpretable in this context. We define the spherical coordinates such that the pole singularities correspond to strictly vertical gaze oriented either up or down, which are very rare directions.

We use an ImageNet-pretrained ResNet-18 [7] as the backbone network. All the models were trained in PyTorch using the Adam optimizer [13] with learning rate 10^{-4}.

5.1. Error quantile estimation

To the best of our knowledge, all existing research applying neural networks to the task of gaze estimation do not consider error bounds. Error bounds are useful when estimating gaze in unconstrained environments, because precision is likely to degrade when the eye is viewed from a sideways angle, or when one or more eyes are partially obscured (e.g. by glasses frames). In a classification setting, softmax outputs are often used as a proxy for confidence. However, for regression this is not possible, as the magnitude of the output corresponds directly to the predicted property.

To model error bounds, we use a pinball loss function [15] to predict error quantiles. We use one single network to predict both the mean value and the 10% and 90% quantile. The effect of this is that for a given image, we estimate through a single forward pass both the expected gaze direction and a cone of error within which the ground truth should lie 80% of the time. We assume that the distribution is isotropic in our spherical coordinate system. This assumption is not strictly true, especially for large pitch angles due to the space distortion around pole singularities. However, for most of the observed gaze directions (Fig. 4) it is a reasonable approximation to reduce dimensionality and simplify the interpretation of the result.

The output of our network is $f(I) = (\theta, \phi, \sigma)$, where (θ, ϕ) is the expected gaze direction in spherical coordinates, for which we already have a corresponding ground truth gaze vector in the eye coordinate system $\mathbf{g}$ (see Sec. 3.1) as $\theta = -\arctan \frac{g_x}{g_z}$ and $\phi = \arcsin g_y$. The third parameter, σ, corresponds to the offset from the expected gaze such that $\theta + \sigma$ and $\phi + \sigma$ are the 90% quantiles of their distributions while $\theta - \sigma$ and $\phi - \sigma$ are 10% quantiles.

Finally, we compute the pinball loss of this output. This will naturally force ϕ and θ to converge to their ground truth values and σ to the quantile threshold. If $y = (\theta_{\text{gt}}, \phi_{\text{gt}})$, the loss L_τ for the quantile τ and the angle θ can be written as:

$$\hat{q}_\tau = \begin{cases} \theta_{\text{gt}} - (\theta - \sigma), & \text{for } \tau \leq 0.5 \\ \theta_{\text{gt}} - (\theta + \sigma), & \text{otherwise} \end{cases} \qquad (2)$$

$$L_\tau(\theta, \sigma, \theta_{\text{gt}}) = \max(\tau \hat{q}_\tau, -(1-\tau)\hat{q}_\tau). \qquad (3)$$

A similar formulation is used for the angle ϕ. We average the losses for both angles and quantiles $\tau = 0.1$ and $\tau = 0.9$. Thus, σ is a measure of the difference between the 10% and 90% quantiles and the expected value.

5.2. Adapting to unseen domains

Despite the variety in the Gaze360 dataset, some real-world applications may benefit from a closer adaptation of the model to the target domain. For this reason, we introduce a self-supervised method for domain adaptation.

Our general model is fine-tuned using a mix of the labeled Gaze360 images and unlabeled images from the new domain. Inspired by [24], we introduce a discriminator which tries to identify the source domain of the image features as a binary classification task. The features are the output of the backbone network. The discriminator loss L_D is added to the original supervised loss L_τ for those images where ground truth is available.

In addition, we added a further loss to exploit the left-right symmetry of the gaze-estimation task as a means of encouraging model output consistency on unlabeled data. We use the model to compute the gaze of the original and horizontally flipped image, and the pinball loss L_S to minimize the angular difference between the prediction from the first input and horizontally mirrored prediction from the second input. While this loss by itself can lead to collapse to a gaze prediction along the line of symmetry, our observations in Sec. 6.2 show that this helps when used as a regularizer to improve performance in an unseen target domain. Altogether we minimize $L = \alpha \cdot L_\tau + L_D + \beta \cdot L_S$ where $\alpha = 60$ and $\beta = 3$ in our experiments.

6. Experimental Analysis

6.1. Model evaluation

In this section, we compare several approaches using the Gaze360 dataset. We compared the following methods: **Mean** - uses the mean gaze of the training set for all predictions; **Deep Head Pose** - a deep network based head pose estimator by Ruiz *et al.* [19]; **Static** - the backbone model, ResNet-18, and two final layers to compute the prediction; **TRN** - a version of Temporal Relation Network [33] where the features of frames at fixed windows around time t are concatenated before averaging the predictions of the temporal windows; **LSTM** - refers to the **Gaze360** architecture.

For each of the three architectures introduced above, we report accuracy of different baselines for uncertainty estimation: **MSE** - uses the mean squared error to regress only the spherical angles of gaze without uncertainty; **MSE+Drop** - using the MSE model, the uncertainty is estimated by 5 forward passes for each input while randomly dropping neurons in the last layer and computing the variance of the output; **Crop augmentation** - 5 random head crops are sequentially evaluated to estimate uncertainty using the variance of the 5 predictions of the MSE-trained model; and **Pinball Loss** - gaze direction and error bounds are jointly estimated using the pinball loss.

The angular errors in Table 2 are provided separately for the entire test set (*All 360°*) and for samples where the subject is looking within 90° (*Front 180°*) and 20° (*Front facing*) of the camera direction. We also report the Spearman's rank correlation between the error quantile estimate and the actual error, which is a metric for how well the predicted error bounds estimate the actual error.

The results confirm that eye-free **Mean** predictions as well as **Head pose** are insufficient to predict the rich variation of eye movement in our dataset. All of our gaze models outperform these simple baselines. We also observe that, under the same conditions, the error is generally lowest for the model using **Pinball** loss. The same trend can be seen for the correlation between the predicted uncertainty and actual prediction error. Additionally, only a single forward pass is required for the prediction. Hence, we chose the

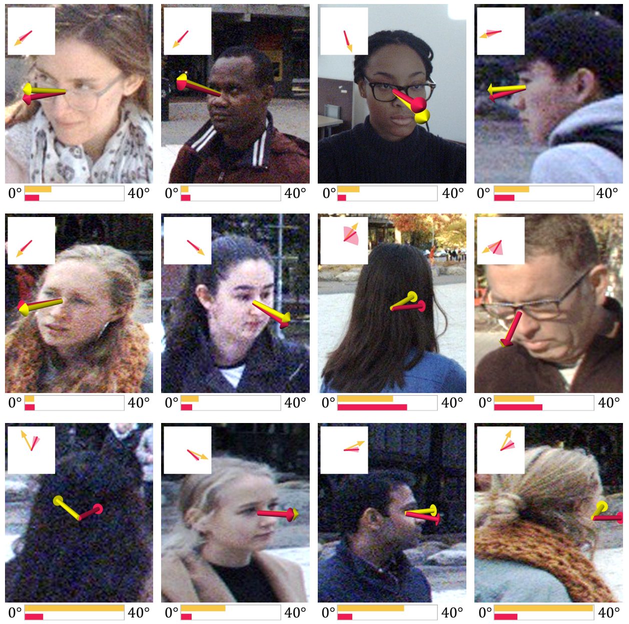

Figure 7. **Test set examples**: ground truth gaze (yellow) and Gaze360 predictions (red) are shown for unseen test subjects. The bars denote actual (yellow) and predicted (red) errors in degrees. The inset shows a top-down view of the gaze estimates and the predicted error versus ground truth. The bottom row shows sample failure cases where the model was overconfident.

Table 2. **Performance comparison on Gaze360 dataset.** The table below reports the mean angular errors for various models and benchmarks on the Gaze360 test data. The last column shows the correlation between the actual error and the predicted uncertainty.

Model	Uncert. Loss	*All* *360°*	*Front* *180°*	*Front* *Facing*	Uncert. Corr.
Mean	-	59.0	40.5	19.0	-
Deep HP	-	49.3	30.7	22.7	-
MSE Static	No	15.8	13.7	13.4	-
MSE TRN	No	14.3	11.8	11.8	-
MSE LSTM	No	14.1	12.1	11.6	-
MSE+Drop Static	No	15.8	13.7	13.4	0.24
MSE+Drop TRN	No	14.3	11.8	11.8	0.31
MSE+Drop LSTM	No	14.1	12.1	11.6	0.31
Crop Aug. Static	No	16.0	13.2	12.6	0.37
Crop Aug. TRN	No	14.2	11.5	11.4	0.39
Crop Aug. LSTM	No	14.1	11.6	11.2	0.37
PinBall Static	Yes	15.6	13.4	13.2	0.42
PinBall TRN	Yes	14.1	11.7	11.6	**0.46**
Pinball LSTM **(i.e., Gaze360)**	Yes	**13.5**	**11.4**	**11.1**	0.45

Pinball loss as our recommended approach.

Switching from a single-frame static model to a temporal model also benefits the gaze prediction accuracy substantially. We conclude that although the performance of **TRN** and **LSTM** is similar, we recommend the **Pinball LSTM** for its slightly better results in our metric and straightforward adaptation to use a different number of input frames.

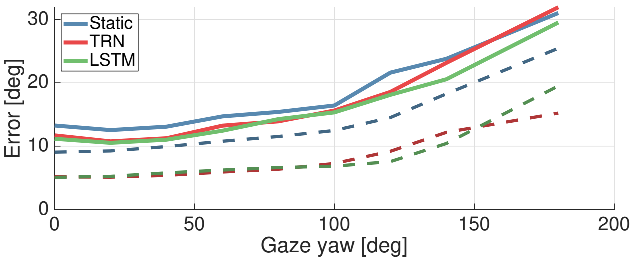

Figure 8. **Error measured on Gaze360 dataset using the Pinball models.** The full lines show prediction error, the dashed lines show predicted uncertainty.

In Fig. 8 we present the prediction error of the models using **Pinball** loss as a function of gaze yaw angle. As expected, accuracy falls with increasing gaze yaw angle. Unlike traditional eye trackers, our model smoothly transitions into head pose estimation (between head yaws of 90-150°) to provide a best guess of gaze even for rear views. This is accompanied by a higher associated uncertainty (dashed lines). Although the error for frontal views is generally larger than errors reported on existing high-resolution datasets, we next show that this is due to the challenging properties of Gaze360 which allow models trained on it to transfer better to physically unconstrained images.

In Fig. 7 we show sample results on our test data. The angular error denoted by the yellow bar intuitively grows as the eyes become smaller due to distance or occluded due to head pose variation. Although the prediction error for away-looking poses is on average large, the uncertainty measure provides a reasonable prediction of this behavior.

6.2. Cross-dataset evaluation

We evaluate the value of the Gaze360 dataset for gaze estimation in the wild by training the **Pinball Static** model using multiple pre-existing 3D gaze datasets and measuring cross-dataset test error. The comparison datasets we use are: **Columbia** [20] - high-resolution close-up faces; **MPIIFaceGaze** [31] - faces captured by webcams; **RT-GENE** [4] - low-resolution faces using in-painting to mask out eye-tracking glasses; **Gaze360** (Ours) - faces with varying resolution; For those datasets where no official splits were provided [20, 31] we use all available samples for training and do not measure the within-domain error.

Table 3 summarizes the results. This task is much more

Table 3. **Cross-dataset evaluation:** we report the mean angular errors for the Static model trained using different datasets.

Test **Train**	**Columbia**	**MPII** **FaceGaze**	**RT-GENE**	**Gaze360**
Columbia	-	12.3	32.8	57.9
MPIIFaceGaze	12.4	-	26.5	57.8
RT-GENE	24.2	18.9	-	56.6
Gaze360	9.0	12.1	23.4	-
Gaze360 + DA	**8.1**	**9.9**	**21.9**	-

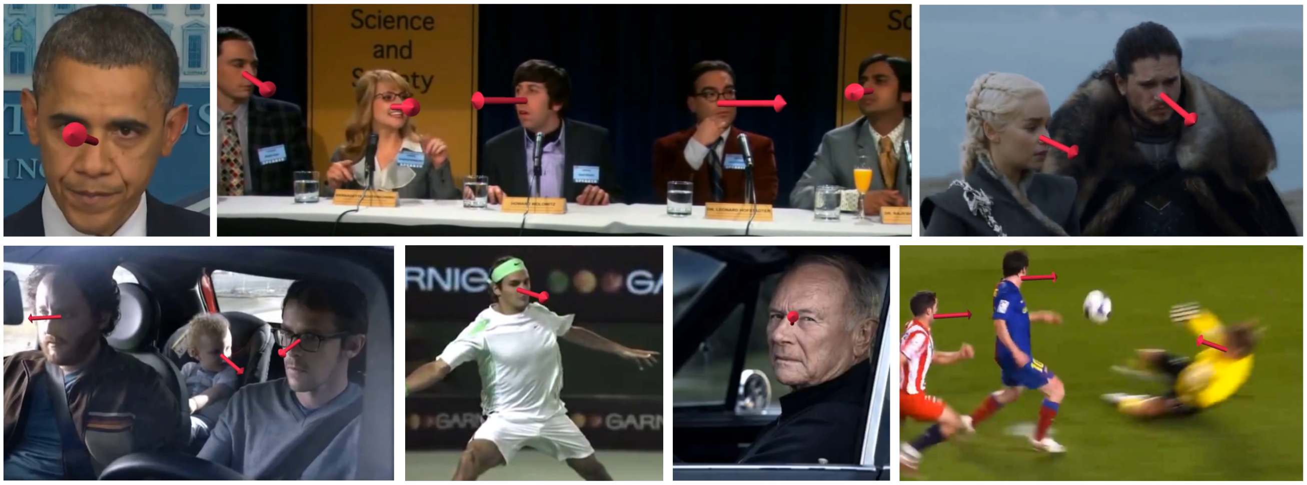

Figure 9. **Estimating 3D gaze in the wild**: further examples of our model's output on unseen video gathered from YouTube.

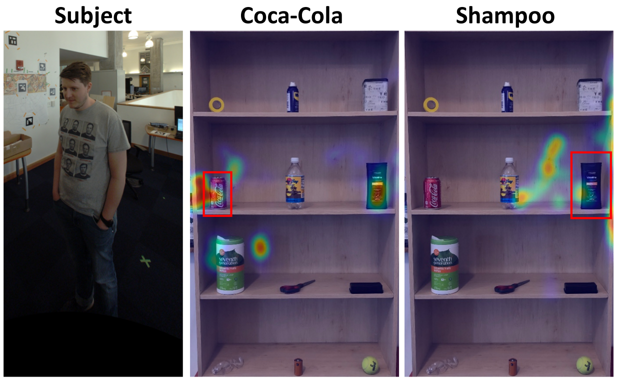

Figure 10. **An example application**: we use Gaze360 to passively infer the attention of a customer as they browse products on a shelf, using video (left) from a camera next to the shelf (right).

challenging than within-domain tests. The best results are consistently achieved when our dataset is used for training. In addition, we fine-tune our Gaze360-trained model on new domains (**Gaze360 + DA**) using the self-supervised approach described in Sec. 5.2, which does not utilize the ground truth labels in other datasets. Our domain adaption strategy improves performance further on all the datasets.

7. Tracking gaze in the wild

Prediction in unconstrained environments: The variation in appearance of subjects in the Gaze360 dataset allows our model to perform well without further training or fine-tuning on unseen image and video data from uncurated online sources. We demonstrate this visually on numerous examples in Figs. 1 and 9 and in our supplemental video.

Estimating attention in a supermarket: To illustrate one possible application of Gaze360, we apply it to the task of predicting which objects are being looked at on a supermarket shelf, which is relevant for product-placement in stores. We recreate a supermarket shelf and ask subjects to look at various objects while self-reporting those objects. We record them with a camera next to the shelf, as shown in Fig. 10. Despite a less than optimal view of the subject , we are able to predict which object is being looked at correctly 51% of the time. Using a smartphone camera embedded directly in the shelf (so that the view of subjects is closer to frontal), the accuracy increases to 68%. The objects along the bottom shelves have highest error rate, as the eyes become almost fully occluded when looking downwards. Finally, we are able to produce a heatmap of customer attention, shown in Fig. 10. While simple, this application demonstrates the flexibility of our system for use in a wide range of real-world applications.

8. Conclusion

In this work, we introduced a novel approach to efficiently collect annotated gaze data at scale and used it to generate a large and diverse dataset, suitable for deep learning of 3D gaze from images and video. We presented a new temporal appearance-based gaze model using a novel loss function to estimate error quantiles. Finally we demonstrated the value of (i) our dataset via careful cross-dataset performance comparison versus three existing 3D gaze datasets, and (ii) our model via application to unconstrained unseen imagery from YouTube videos. It is our hope that by using our dataset and model, researchers across a range of fields will be able to better leverage gaze as a cue to improve vision-based understanding of human behavior.

Acknowledgements. Toyota Research Institute provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

References

[1] Michael Argyle. Non-verbal communication in human social interaction. 1972. 1

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1, 3

[3] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 3

[4] 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)*, pages 334–352, 2018. 2, 7

[5] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional lstm networks for improved phoneme classification and recognition. In *International Conference on Artificial Neural Networks*, pages 799–804. Springer, 2005. 5

[6] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[8] Craig Hennessey, Borna Noureddin, and Peter Lawrence. A single camera eye-gaze tracking system with free head motion. In *ETRA*, 2006. 2

[9] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *CVPR*, 2017. 1

[10] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. Screenglint: Practical, in-situ gaze estimation on smartphones. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2546–2557, New York, NY, USA, 2017. ACM. 2, 5

[11] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 28(5):445–461, Aug 2017. 2

[12] Aditya Khosla*, Kyle Krafka*, Petr Kellnhofer, Harini Kannan, Suchi Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *CVPR*, 2016. 2, 3, 4, 5

[13] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[14] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. MultiPoseNet: Fast multi-person pose estimation using pose residual network. In *ECCV*, 2018. 1

[15] Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. 3, 6

[16] Christopher D McMurrough, Vangelis Metsis, Jonathan Rich, and Fillia Makedon. An eye tracking dataset for point of gaze detection. In *ETRA*, 2012. 2

[17] 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. *ETRA*, 2014. 2, 3, 5

[18] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *Intelligent Robots and Systems (IROS)*, pages 5048–5054. IEEE, 2016. 1

[19] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. *CoRR*, abs/1710.00925, 2017. 6

[20] Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. Gaze locking: Passive eye contact detection for human-object interaction. In *UIST*, 2013. 2, 7

[21] Tricia Striano and Vincent M Reid. Social cognition in the first year. *Trends in cognitive sciences*, 10(10):471–476, 2006. 1

[22] Yusuke Sugano and Andreas Bulling. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*, 2016. 1

[23] Yusuke Sugano, Yuki Matsushita, and Yuuki Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *CVPR*, 2014. 2, 3, 5

[24] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 6

[25] John Wang and Edwin Olson. AprilTag 2: Efficient and robust fiducial detection. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2016. 3

[26] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 1

[27] Ulrich Weidenbacher, Georg Layher, Petra-Maria Strauss, and Heiko Neumann. A comprehensive head pose and gaze database. *IET International Conference on Intelligent Environments*, 2007. 2

[28] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv:1504.06755*, 2015. 2

[29] Dong Hyun Yoo and Myung Jin Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *CVIU*, 2005. 2

[30] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, 2015. 2, 3

[31] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Honolulu, HI, USA, July 21-26, 2017*, pages 2299–2308, 2017. 2, 3, 5, 7

[32] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2019. 3

[33] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 6

[34] Wangjiang Zhu and Haoping Deng. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3, 5

[35] Zhiwei Zhu and Qiang Ji. Eye gaze tracking under natural head movements. In *CVPR*, 2005. 2