'Labelling the Gaps': A Weakly Supervised Automatic Eye Gaze Estimation

Shreya Ghosh[1], Abhinav Dhall[1,2], Munawar Hayat[1], and Jarrod Knibbe[3]

[1] Monash University, [2] IIT Ropar, [3] University of Melbourne
{shreya.ghosh,munawar.hayat}@monash.edu, abhinav@iitrpr.ac.in,
jarrod.knibbe@unimelb.edu.au

Abstract. Over the past few years, there has been an increasing interest to interpret gaze direction in an unconstrained environment with limited supervision. Owing to data curation and annotation issues, replicating gaze estimation method to other platforms, such as unconstrained outdoor or AR/VR, might lead to significant drop in performance due to insufficient availability of accurately annotated data for model training. In this paper, we explore an interesting yet challenging problem of gaze estimation method with a limited amount of labelled data. The proposed method utilize domain knowledge from the labelled subset with visual features; including identity-specific appearance, gaze trajectory consistency and motion features. Given a gaze trajectory, the method utilizes label information of only the start and the end frames of a gaze sequence. An extension of the proposed method further reduces the requirement of labelled frames to only the start frame with a minor drop in the generated label's quality. We evaluate the proposed method on four benchmark datasets (CAVE, TabletGaze, MPII and Gaze360) as well as web-crawled YouTube videos. Our proposed method reduces the annotation effort to as low as 2.67%, with minimal impact on performance; indicating the potential of our model enabling gaze estimation 'in-the-wild' setup[1].

Keywords: Gaze estimation · Weakly-supervised learning · Neural network.

1 Introduction

The 'language of the eyes' provides an insight into a complex mental state such as visual attention [29] and human cognition (emotions, beliefs and desires) [52]. Accurate gaze estimation has wide applications in computer vision-related assistive technologies [33,16] where the gaze is measured as a line of sight of the pupil in 3D/2D space or 2D screen location [24]. Recent advances in computer vision and deep learning have significantly enhanced the accuracy of gaze estimation [15]. Most promising eye gaze estimation techniques either require specialized hardware (for example Tobii [34]) or use supervised image processing solutions [62,46]. Device and sensor based gaze estimation methods are highly

---

[1] https://github.com/i-am-shreya/Labelling-the-Gaps

dependent on user assistance, illumination specificity, the high device failure rate in an uncontrolled environment and constraints on the device's working distance. On the other hand, supervised methods require a large amount of labelled data for training. Manual labelling of human gaze information is a complex, noisy, resource-expensive and time-consuming task. To overcome these limitations, weakly-supervised learning provides a promising paradigm since it enables learning from a large amount of readily available non-annotated data. Few works [26,17,57,10] explore in this direction to eliminate the data curation and annotation issue. However, these methods mostly investigate from a spatial analysis perspective. Human eye movement is a spatio-temporal, dynamic process which is either task-driven or involuntary action. Thus, it would be interesting to simplify the ballistic eye movement and curate large-scale training data for gaze representation learning. To this end, we propose a weakly supervised eye gaze estimation framework. Our proposed technique reduces the requirement of a large number of annotated training samples. We show that the technique can also be used to facilitate the annotation process and reduce the bias in the data annotations. The proposed method requires the ground truth labels of start and end frames in a pre-defined gaze trajectory. We further refine this strategy where only the start frame's gaze annotation is required. Our proposed method significantly reduces the annotation effort which could be beneficial for annotating large-scale gaze datasets quickly. Moreover, it can be used in several applications such as immersive, augmented and virtual reality [49,9] (especially in Foveated Rending (FR)), animation industry [18,38] and social robotics [2,58], where unsupervised or weakly supervised calibration is highly desirable. In Foveated Rending (FR), gaze based interaction demands low latency gaze estimation to reduce energy consumption. To achieve this, the virtual environment displays high-quality images only from the user's point of view and blurs the peripheral region. Due to the subsequent delays in the frame-wise gaze estimation pipeline, the usage of FR is quite limited and mostly headpose direction is used to approximate the field of view [9]. Our proposed method has the potential to bridge the gap and reduce energy consumption by interpolating the gaze trajectory of the user. Another potential application includes animation industry [18,38]. Given the start and end Point of Gaze (PoG) of a virtual avatar, our method can easily generate realistic labels for intermediate frames to display realistic facial gestures in the interaction environment [18,38]. Similarly, in social robotics, multi-modal gaze control strategies have been explored for guiding the robot's gaze. For example, an array of microphones has been utilized [2] to guide the gaze direction of a robot named Maggie. The other well-established methods include the usage of infrared laser and multimodal stimuli (e.g., visual, auditory and tactile) for modelling any known gaze trajectories [58]. Our proposed methods could eliminate the aforementioned requirement of specialised hardware or pre-defined heuristics to navigate the environment. The **main contributions** of the paper are as follows:

1. We propose two weakly supervised neural networks (2-labels, 1-label) for gaze estimation. '2-labels' and '1-label' require the labels of two and one frames in a gaze sequence, respectively.
2. We use task-specific information to bridge the gap between labelled and unlabelled samples. Our proposed method leverages facial appearance, relative motion, trajectory ordering and embedding consistency. This task-specific knowledge bridge the gap between labelled and unlabeled samples via learning.
3. We evaluate the performance of the proposed networks in two settings: 1) *On benchmark datasets* (CAVE, TabletGaze, MPII and Gaze360) and 2) *On unlabelled 'in the wild' YouTube data* where ground truth annotation is not available. Additionally, we perform cross-dataset experiments to validate the generalizability of the framework.
4. We also demonstrate the effectiveness of our proposed techniques by re-learning state-of-the-art eye gaze estimation methods with the labels generated by our method with very few prior annotations. The results indicate comparable performance for state-of-the-art frameworks (for example, 3.8 degrees by pictorial gaze [37] and 4 degrees with our 2-label technique).
5. We also validate our learning based interpolation method on unlabelled YouTube data where ground truth annotation is not available. Our experimental results suggest that this annotation method can be useful for extracting substantial training data for learning gaze estimation models.

2   Related Work

Gaze Estimation. Recent advances in computer vision and deep learning techniques have significantly enhanced the gaze estimation performance [37,24]. A thorough analysis of gaze estimation literature is mentioned in a recent survey [15]. Appearance based gaze estimation methods [30,63,31] learn image to gaze mapping either via support vector regression [46] or deep learning methods [27,63,59,60,23,12]. Among the deep learning based methods, supervised learning methods [63,23,60,37] mostly encode appearance based gaze which require a large amount of annotated data. To overcome the limitation, few works explore gaze estimation with limited supervision such as 'learning-by-synthesis' [48], hierarchical generative models [51], conditional random field [6], unsupervised gaze target discovery [62], unsupervised representation learning [10,57], weakly supervised learning [26], pseudo labelling [17] and few-shot learning [36,56]. Among these studies, the few shot learning approach required very few (≤ 9) calibration samples for gaze inference. Our method requires even less data annotation for gaze estimation (CAVE: 6.56%, TabletGaze: $< 1\%$, MPII: 4.67% and Gaze360: 2.38%).

Gaze Motion. Eye movements are divided into the following categories: *1) Saccade.* Saccades are voluntary eye movements to adjust the PoG in a visual field and it usually lasts for 10 to 100 ms. *2) Smooth Pursuit.* It is an involuntary eye movement that occurs while tracking a moving visual target. *3) Fixations.*
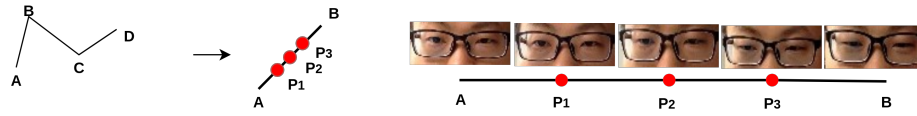
Fig. 1: Weakly supervised labelling approach illustration. The zig-zag path $\overrightarrow{AD}$ in the left is an example of a human gaze movement. This path can further be broken into several 'gaze trajectories' ($\overrightarrow{AB}, \overrightarrow{BC}$ and $\overrightarrow{CD}$). Given a gaze trajectory $\overrightarrow{AB}$ with gaze annotation for A and B, the objective of this work is to annotate all the intermediate unlabelled frames i.e. P_1, P_2 and P_3. The right side is an example of a gaze trajectory [63].

Fixations consist of three involuntary eye movements termed as *tremor, drift* and *microsaccades* [11]. The main objective is to stabilize the PoG on an object. Prior works along this line mainly used velocity based thresholding [25], BLSTM [47], Bayesian framework [42], and hierarchical HMM [67] for classification. Arabadzhiyska et al. [3] model saccade dynamics for gaze-contingent rendering. Our proposed method uses trajectory constrained gaze interpolation using temporal coherency with limited ground truth labels.

Gaze Datasets. In the past decade, several datasets [46,13,21,60,59,12,24,14,35] have been proposed to estimate gaze accurately. The dataset collection technique has evolved from constrained lab environments [46,61] to unconstrained indoor [21,60,59,12,61] and outdoor settings [24]. To consider both the constrained and unconstrained settings, we evaluate our weakly supervised framework in CAVE [46], TabletGaze [21], MPII [60] and Gaze360 [24] datasets.

Weakly Supervised Neural Networks. Over the past few years, several promising weakly supervised methods have been proposed which mainly infer on the basis of prior knowledge [8,28], task-specific domain knowledge [64,4,65], representation learning [53,66], loss-imposed learning paradigms [4,41] and combinations of the above [4]. Williams et al. [54] propose a semi-supervised Gaussian process model to predict the gaze. This method simplifies the data collection process as well. Bilen et al. [8] use pre-trained deep CNN for the object detection task. Arandjelovic et al. [4] propose a weakly supervised ranking loss for the place recognition task. Haeusser et al. [19] introduce 'associative learning' paradigm, which allows semi-supervised end-to-end training of any arbitrary network architecture. Unlike these studies, we explore loss-imposed domain knowledge for our framework.

3   Preliminaries

Gaze Trajectory. Human eye movement follows an arbitrary continuous path in three-dimensional space termed as the 'gaze trajectory' [40]. Gaze trajectories generally depend on the person, context and external factors [32]. Eye movements can be divided into three types: fixations, saccades and smooth pursuit [40]. Fixation occurs when the gaze may pause in a specific position voluntarily or

involuntarily. Conversely, gaze moves from one to another position for a saccade. The human gaze consists of a series of fixations and saccades in random order. In this work, we consider a small duration of eye movement from one position to another. Let us assume, the zig-zag path (left of Fig. 1) is an example of the human gaze movement path. This path can be divided into three small sub-paths $(\overrightarrow{AB}, \overrightarrow{BC}$ and $\overrightarrow{CD})$. We conduct our experiments on each such small sub-path. We use the term 'gaze trajectory' to refer to this simplified version of the gaze path (i.e. $\overrightarrow{AB}$). The red points $(P_1, P_2$ and $P_3)$ in $\overrightarrow{AB}$ are the frames in the A to B sequence. Considering these as discrete, they can be split into 3-point subsets with a constraint: each subset should contain the start and endpoints and the points should maintain the order of trajectory sequence. We term the 3 points set as *3-frame set*. For example, $\{A, P_1, B\}$, $\{A, P_2, B\}$ and $\{A, P_3, B\}$ are three 3-frame sets.

Problem Statement. In the context of a video, the points (A, P_1, P_2, P_3 and B) are frames in a specific trajectory order. Throughout this paper, we term A and B as start and end frames. Given the annotated start and end frames, the gaze trajectory sequence can be divided into small segments. There might be a high and insignificant temporal coherence if the segment duration is too small. On the other hand, a longer duration could affect the learning of meaningful representation due to diversity. We observed that the average difference in large time segment consisting of approx. ~ 80 frames is around $\sim 35°$. On the other hand, the smallest segment has angular difference of $< 1°$. On this front, from a long sequence, we mine 3-frame subsets of gaze trajectory for learning meaningful representation. We work on two experimental settings: (a) **2-labels**: *When the start and end frame annotations are available.* (b) **1-label**: *When only the start frame annotation is available.* Given a gaze trajectory similar to $\overrightarrow{AB}$ with labels for A and B, the objective of this work is to annotate the unlabelled frames i.e. $P_1, P_2, P_3, \ldots P_n$ where, n is the number of intermediate frames.

Notations. Suppose that we have a set of N '3-frame set' samples in a dataset $D = \{X_n, Y_{s_n}, Y_{e_n}\}_{n=1}^{N}$, where X_n is a nth 3-frame set consisting of start, middle/unlabelled and end frames $(f_s, f_{ul}, f_e)_n$, Y_{s_n} and Y_{e_n} are the nth start and end frame labels, respectively. Lets assume our model G with learnable parameter θ maps input $X_n \in \mathbb{R}^{3 \times 100 \times 50 \times 3}$ to the relevant label spaces i.e., $Y_{s_n} \in \mathbb{R}^3$, $Y_{ul_n} \in \mathbb{R}^3$ and $Y_{e_n} \in \mathbb{R}^3$. The mapping function is denoted as $G_\theta : X_n \to \{Y_{s_n}, Y_{ul_n}, Y_{e_n}\}$.

4 Gaze Labelling Framework

4.1 Architectural Overview of '2-labels'

The overview of the proposed framework is shown in Fig. 2. Given a 3-frame set $X_t : \{f_s, f_{ul}, f_e\}$, we define an encoder E which maps the input X_t to latent space $Z_t : \{Z_s, Z_{ul}, Z_e\}$, where $Z_s \in \mathbb{R}^{2048}, Z_{ul} \in \mathbb{R}^{2048}, Z_e \in \mathbb{R}^{2048}$ (Refer Fig. 2 Left). After $E : X_t \to Z_t$ mapping, two motion features M_{s_ul} and M_{ul_e} are extracted between start-middle frames and middle-end frames. These motion features are

concatenated with the latent embeddings. On top of it, Fully Connected (FC) layers having 512 and 1024 nodes are appended before prediction. Finally, the network predicts Y_s^p, Y_{ul}^p and Y_e^p, where, Y_s^p, Y_{ul}^p and Y_e^p are gaze information corresponding to the input frames $\{f_s, f_{ul}, f_e\}$. The backbone network is not architecture-specific, although we use VGG-16 [45] and Resnet-50 [20] for our experiments. The rationale behind the framework design is explained as follows:

Identity Adaptation The obvious usefulness of user adaptation of gaze calibration for AR and VR devices motivates us to design an identity specific gaze labelling framework [43]. Thus, we purposefully select identity specific 3-frame sets. At the same time, it is important for the framework to be able to learn the variations across a large number of subjects with a different head pose, gaze direction, appearance, illumination, image resolution and many other configurations. Additionally, the framework should encode rich features relevant to eye region appearance, which is the most important factor for weakly supervised gaze labelling. Similar to recent studies [59,36], we use eye region images as input for gaze inference. The 3-frame set is selected over a small duration temporal window in a video. As there is a subtle change in the appearance of eyes from one frame to another, the latent representation of the image also has minimal change. At a conceptual level, we are motivated by the smoothness constraint in optical flow algorithms. We choose to calculate cosine distance between start and middle, middle and end pair while calculating consistency loss. It helps in preserving the identity specific features across the 3-frame set.

Motion Feature. At first glance, the task of predicting gaze from sparsely labelled data may seem overly challenging. However, given a '3-frame set' sequence of a subject in very small time duration, there will be a high correspondence between frame A, P_i and B, where $i = \{1, 2, 3\}$. Given this constraint, the objective is reduced to modelling the head and eye motion information to bridge the gap. This motivates us to use motion features for sequence modelling. Similar to [7] in pose estimation domain, we encode a weak inter-frame motion by computing the ℓ_1 distance between two consecutive frames in the latent space. We define motion feature by $M_{s\_ul} \oplus M_{ul\_e}$ where $M_{s\_ul} = Z_s - Z_{ul}$ and $M_{ul\_e} = Z_{ul} - Z_e$. M_{s_ul} and M_{ul_e} represent motion feature between start-middle and middle-end frames, respectively. Further, we use this feature to estimate the gaze-direction in a given trajectory. To train the above-mentioned network, we use the following loss functions.

Regression Loss. Corresponding to each 3-frame set X_t, the start and end frames are annotated in the 2-labels setting. These annotations provide strong supervisory information to predict the gaze information of the middle unlabelled frame. It provides information regarding an arbitrary gaze trajectory. The unlabelled middle frame lies in between start and end frames in that specific trajectory. Thus, it belongs to the same distribution of the start and end frames. The regression loss is defined as: $l_{reg} = MSE(Y_s, Y_s^p) + MSE(Y_e, Y_e^p)$ Here, MSE is Mean Squared Error, Y_s, Y_s^p, Y_e and Y_e^p are start label, predicted start label, end label and predicted end label, respectively.
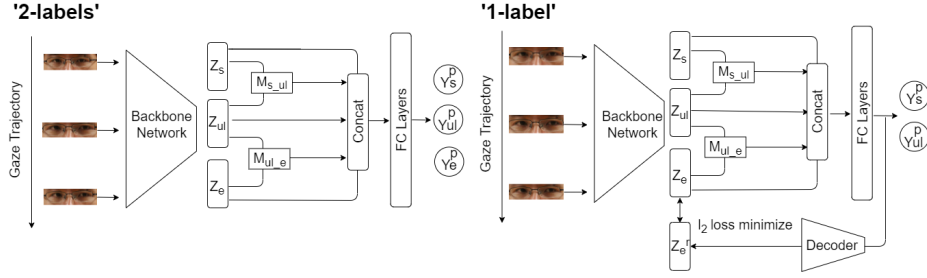
Fig. 2: Overview of '2-labels' and '1-label' network pipelines. The frameworks take 3-frame set as input and learn to interpolate intermediate labels via weak supervision. Refer Sec. 4 for more details.

Consistency Loss. The main aim of the consistency loss is to maintain consistency between latent and label space. Here, latent space is denoted as Z and label space is the output space of the '2-labels' framework. In the latent space, let the distance between Z_s and Z_{ul} be $d^z_{s\_ul}$ and the distance between Z_{ul} and Z_e be $d^z_{ul\_e}$. Similarly in the label space, let the distance between Y^p_s and Y^p_s be $d^y_{s\_ul}$ and the distance between Y^p_s and Y_e be $d^y_{ul\_e}$. According to our hypothesis, the distance from start to unlabelled frame and unlabelled to end frame remains consistent. The loss is defined as follows: $l_{consistency} = \{|d^y_{s\_ul} - d^z_{s\_ul}| + |d^y_{ul\_e} - d^z_{ul\_e}|\}$

In the equation, the cosine distance is considered. The rationale behind the choice is as follows: the cosine distance is applied pairwise to utilize the partial annotations and bridge the gap between labelled and unlabelled frames. The distance has following properties: it leverages the identity specific information across the 3-frame set and it captures the motion-similarity information which indirectly encodes relative ordering of the frames. The angular distance (d_{a_b}) between the frames a and b is defined as follows: $d_{a\_b} = \frac{a}{||a||_2} \cdot \frac{b}{||b||_2}$ where, a and b are latent or label space embeddings of the start, unlabelled and end frames. '(.)' denotes the dot product. According to [5], this term is calculated in two stages. First, the latent embeddings are L_2-normalized (i.e. $\frac{a}{||a||_2}$), which maps the d-dimensional latent embedding to a unit hyper-sphere, where the cosine similarity/distance and dot product are equivalent. As the human gaze follows a spatio-temporal trajectory [40], the 3-frame sets satisfy their relative trajectory ordering as well. The condition is satisfied by the consistency loss as it incorporates their relative position w.r.t. the middle unlabelled frame from label space to the latent space.

Overall Loss Function. The final loss function is defined below:

$$Loss = \lambda_1 l_{reg} + \lambda_2 l_{consistency} \tag{1}$$

It includes regression and consistency losses. Here, $\lambda_{1,2}$ are the regularization parameters. Further, we reduce the label requirement (by removing the requirement of the end frame's label) in the next framework with minimal impact in the performance.

4.2   Architectural Overview of '1-label'

The network architecture of 1-label network is shown in Fig. 2 Right. Similar to the 2-labels network, the motion feature and identity specific appearance are leveraged for gaze estimation. The main motivation for moving from 2-labels to 1-label architecture is to leverage rich features with even lesser number of total annotations in a dataset. In '1-label' architecture, we have additional decoder module defined as $D(Z'_e; \phi_D)$ consisting of an FC layer. The decoder D, parameterized by ϕ_D, maps the penultimate layer features (i.e. $Z'_e \in \mathbb{R}^{1024}$) to latent-embeddings $Z^r_e \in R^{2048}$. The rationale behind this remapping is adding constraints which can enhance the performance of the network by encoding meaningful representation from f_e. l_2 loss is computed between Z_e and Z^r_e. The backbone network is not architecture-specific similar to the 2-labels architecture, although we use VGG-16 [45] and Resnet-50 [20] as backbone networks for our experiments. To train this network, we use the following loss functions:

Regression Loss. Similar to 2-labels architecture, we compute the regression loss corresponding to the start label as follows: $l_{reg} = MSE(Y_s, Y^p_s)$ Here, MSE is mean squared error, Y_s and Y^p_s are start and predicted start label, respectively.

Consistency Loss. Similar to 2-labels, in the 1-label technique, we add the consistency loss $l_{consistency}$. The loss is defined as follows: $l_{consistency} = \{|d^y_{s\_ul} - d^z_{s\_ul}|\}$ Here, $d^y_{s\_ul}$ and $d^z_{s\_ul}$ are distance between start and unlabelled frame in label and latent space, respectively.

Similar Distribution. We leverage on the constraint that the gaze information belongs to similar distribution. Given a 3-frame set X_t, the output gaze should belong to a specific distribution and for that we compute Kullback-Leibler divergence (KL). To computes KL divergence loss between Y_{true} and Y_{pred}, the following equation is followed: $loss = Y_{true}(log\frac{Y_{true}}{Y_{pred}})$. The objective function $l_{divergence}$ is the loss term to minimize the divergence between the gaze information of start and unlabeled frames, defined as follows: $l_{divergence} = (KL(Y_s, Y_{ul}))$

Embedding Loss. It is the ℓ_2 loss computed between Z_e and Z^r_e. This pattern is predicting future embeddings from prior knowledge.

Overall Loss Function. The final loss function for '1-label' is an ensemble of regression, consistency, similar distribution and embedding losses, where, $\lambda_1 \cdots \lambda_4$ are the regularization parameters.

$$Loss = \lambda_1 l_{reg} + \lambda_2 l_{consistency} + \lambda_3 l_{divergence} + \lambda_4 l_{embedding} \qquad (2)$$

5   Experiments

On Benchmark Datasets. We validate the proposed methods on 4 benchmark datasets: CAVE, Tabletgaze, MPII and Gaze360. These datasets are collected in fully constraint (CAVE), less constrained (Tabletgaze and MPII) and unconstrained (Gaze360) environments.

Automatic 3-frame set mining: The first step to implement our proposed method is 3-frame set mining. For generating 3-frame sets, we perform dataset-specific

Table 1: Comparison of benchmark dataset statistics to show the amount of images in the original data and derived data (3-frames set mined).

Dataset	Original Dataset	Derived Dataset
CAVE [46]	5,880	3,024
MPII [63]	213,659	32,751
TabletGaze [21]	1,428 min	108,524
Gaze360 [24]	172,000	197,588

pre-processing as the datasets are collected in different setups. Please note that we require ground truth labels to define the gaze trajectories, especially for CAVE and MPII datasets as temporal information is not present. Moreover, the annotated subset required for weak supervision is a special subset which contains start and end frame of the trajectories. Additionally, there is no temporal overlap between 3-frame sets for any of the datasets. Table 1 shows the comparison among dataset statistics to show the amount of images in the original data and derived data (3-frames set mined used for weak-supervision). See supplementary material for more details.

On unlabelled 'in the wild' YouTube data. We evaluate our method on an 'in the wild' data i.e. when the expert/ground truth labels are not available. We leverage two eye symmetry property i.e. the change in relative position of the iris is symmetrical while scanning 3D space [10]. We collect approximately 400,000 frames from YouTube videos using this strategy. The details are mentioned in the supplementary material.

Experimental Settings. We define the following terminologies for easy navigation in the upcoming sections. *1) Original Data (OD):* It refers to benchmark dataset's unaltered data; *2) Derived Data (DD):* 3-frame set mined data derived from OD; *3) Original Labels (OL):* Original ground-truth labels provided with OD; and *4) Predicted Labels (PL):* Labels predicted from '2-labels' and '1-label' methods (i.e. Y_{ul}^p). We perform experiments with the following settings: **1) Validation w.r.t. Ground Truth Labels.** First, we applied '3-frame set mining' to obtain the DD (Refer Table 1) from OD. Further, we split the DD into 80%-20% train-test splits without any identity overlap. We evaluate our proposed method with OL. **2) Label Quality Assessment via State-of-the-art Methods' Performance.** We train the state-of-the-art methods [37,21,24] on PL and validate on OL for label quality assessment. **3) Experiments with Different Data Partitions.** We train state-of-the-art models [37,21,24] with different input data settings as follows: a) start and end frames of 3-frame sets, b) 50% of the whole data and c) newly labelled frames. **4) Ablation Studies.** We have conducted extensive ablation studies to show the importance of loss function, motion feature, sequential modelling and regularization parameters. **5) Gaze Labelling 'in the wild'.** We collect 'in the wild' gaze data from YouTube videos having *creative common licence* and compare the label quality with various model based techniques [50,55,22].

Table 2: Comparison of '2-labels' and '1-label' techniques using MAE, CC and Angular Error. Both frameworks with both VGG-16 and ResNet-50 backbone are trained on 80% of the Derived Data (DD, i.e. 3-frame set mined) and validated on 20% of the DD. This 80%-20% partition does not have any identity overlap. Here, MAE: Mean Absolute Error, CC: Correlation Coefficient, TG: TabletGaze.

| Dataset | VGG-16 | | | | ResNet-50 | | | | 2-labels | 1-label |
| | 2-labels | | 1-label | | 2-labels | | 1-label | | Angular Error (in °, TG: in cm) | |
	MAE	CC	MAE	CC	MAE	CC	MAE	CC		
CAVE	0.29	0.85	0.53	0.21	0.25	0.90	0.43	0.25	3.05	3.30
MPII	0.43	0.62	0.57	0.25	0.42	0.63	0.57	0.25	5.00	5.40
Gaze360	0.38	0.69	0.45	0.42	0.34	0.71	0.40	0.70	15.00	15.80
TabletGaze	0.49	0.54	0.50	0.58	0.47	0.55	0.49	0.55	2.27	2.61
YouTube 'in the wild'	0.27	0.90	0.36	0.81	0.24	0.92	0.34	0.86	9.41	12.07

Further, we did perform additional experiments to evaluate the generalizibity of the proposed method. See supplementary material for *cross dataset evaluation.* **Evaluation Metrics.** For quantitative evaluation, we use Mean Absolute Error (MAE), Correlation Coefficient (CC) and Angular Error (in °). Following each database's evaluation protocol, we follow 'leave-one-person-out' for MPII, cross-validation for CAVE and TabletGaze; and train-val-test partitions for Gaze360 dataset. We randomly split the 3-frame sets into 80%-20% training and testing sets. For a fair comparison between the datasets, we estimate the errors in normalized space i.e. we apply min-max normalization to the labels to map it to a [0-1] range. Additionally, we compute angular error (in °) except for the TabletGaze dataset, for which we compute the error in cm (similar to [21]). To compare with the state-of-the-art methods, we use similar evaluation protocols mentioned in the respective studies. See supplementary material for more details. **Training Details.** After the 3-frame set mining, we apply Dlib face detector [44] for eye detection. If face detection (dlib) fails especially for Gaze360 dataset, we use cropped headpose provided with the dataset[2]. Otherwise, we use the resized input image. For the backbone network, we choose VGG-16 [45] and ResNet-50 [20] architectures. For training, we use SGD optimizer with 0.001 learning rate with $1 \times e^{-6}$ decay per epoch. The values of $\lambda_1 \cdots \lambda_4$ are 1. In each case, the models are trained for 1,000 epochs with batch size 32 and early stopping.

6   Results

Gaze Labelling and Estimation Performance Comparison. To show the effectiveness of the proposed method, we evaluate on four benchmark datasets and YouTube data. First, we applied 3-frame set mining to get the derived data (Refer Table 1). Further, we split the derived data randomly into train and test sets (train set: 80% and test set: 20%). Please note that the test set does not

---

[2] https://github.com/erkil1452/gaze360/tree/master/dataset

Table 3: Results of the re-trained state-of-the-art methods on MPII and CAVE dataset in terms of angular error in °. OP=Original Predictions is the result as mentioned in the original papers [37].

Dataset	Method	Alexnet	VGG 16	Pictorial Gaze [37]
CAVE	OP	4.20	3.90	3.80
CAVE	2-labels	4.10	4.46	4.00
CAVE	1-label	4.55	4.84	4.36
MPII	OP	5.70	5.40	4.50
MPII	2-labels	5.90	5.79	4.70
MPII	1-label	6.30	6.20	4.90
Train:YouTube	CAVE	–	–	3.90
Train:YouTube	MPII	–	–	4.57

Table 4: Results of the re-trained methods on TabletGaze and Gaze360 dataset in terms of angular error in ° and cm. OP=Original Predictions (based on manual annotation) is the result as mentioned in the [24,21].

TabletGaze	mHOG+SVR [21]
OP	2.50
2-labels	2.70
1-label	3.10
Train:YouTube	2.30

Gaze360	Pinball LSTM [24]
OP	13.50
2-labels	14.40
1-label	17.20
Train:YouTube	12.80

have identity overlap with training partition. The results are mentioned in Table 2 in terms of MAE, CC and angular error. We use VGG-16 and Resnet-50 as backbone networks to show the impact of different network architectures. Quantitatively, ResNet-50 performs slightly better than the VGG-16. From Table 2, it is also observed that '2-labels' technique is closer to the original label distribution as compared to the '1-label' technique due to the absence of supervisory signal (i.e. absence of end frame label). Due to high-resolution images in CAVE dataset, generated labels' similarity as compared with original labels is high for '2-labels' setting. Instead of sequential modelling, our loss imposed gaze estimation method improves model performance.

Comparison with State-of-the-art Methods. We also evaluate the label generation quality of our proposed methods. For this purpose, we conduct experiments by training existing state-of-the-art methods [37,24,12,21] with the labels predicted from our method. The state-of-the-art network's performance is measured by comparing with the original ground truth labels. The performance comparison is mentioned in the Table 3 and 4. For [37,24], we use author's GitHub implementations [1]. It is observed that '2-labels' performs better than '1-label' for all the datasets. For CAVE dataset, 'gazemap' based method [37] for '2-labels' (i.e. 4.08°) performs better than other settings. It is to be noted that the results of re-trained methods are comparable to when they were trained with the original labels. By using less than 5% labelled data (For MPII 4.67% and

Table 5: Performance comparison for different input data settings with the state-of-the-art methods on MPII, CAVE, TabletGaze (TG) and Gaze360 dataset. NL: Newly Labelled, SF: Start Frame, EF: End Frame.

Dataset	CAVE	TG	Gaze360	MPII
Method	[37]	[21]	[24]	[37]
(SF+EF)	7.14	8.60	22.10	6.10
50% data	5.50	4.50	18.70	5.40
2-labels	4.00	2.70	14.40	4.70
NL frames	4.30	3.00	14.90	4.30
1-label	4.36	3.10	17.20	4.90

Table 6: Ablation study: effect of loss functions for '2-labels' (Eq. 1) & 1-label (Eq. 2). NA: Not Applicable

Loss	MAE (2-labels)	MAE (1-label)
l_{reg}	0.98	0.99
$l_{reg} + l_{consistency}$	0.42	0.76
$l_{reg} + l_{consistency} + l_{divergence}$ (for 1-label)	NA	0.58
$l_{reg} + l_{consistency} + l_{divergence} + l_{embedding}$ (for 1-label)	NA	0.57

Gaze360 2.38%), the labels generated by our weakly-supervised method perform favourably when evaluated on state-of-the-art methods [37,24]. This shows the usefulness of our weakly-supervised approach of label generation.

Re-train State-of-the-art Methods with Subset of Data. To further validate the generated labels, we perform following experiments: we re-train from scratch [37], [21] and [24] using the following labelled sets independently: a) start and end frames of 3-frame sets only, b) 50% of originally labelled training data, c) frames with labels generated with 2-labels method, d) frames with labels generated with 2-labels method apart from start and end frames, and e) frames with labels generated with 1-labels method. The results are shown in Table 5. When we train the networks with start and end frames, the error is high as the start and end frames consist of $< 10\%$ of the whole dataset. When we use 50% of the whole labelled data, the results significantly improve. Similarly, for newly labelled frames, the error is less as compared to above two settings. Please note that in the newly labelled case the training is performed on 90-95% of the training data. These results validate the quality of labels generated by our methods.

Ablation Studies.

Impact of Loss Function. We progressively integrate different parts of our method. We assess the impact of each loss term mentioned in Eq. 1 and Eq. 2 by considering them one at a time during learning on the MPII dataset. The results are shown in Table 6. For '2-labels' and '1-label', if only regression loss is considered, the error is high for the two techniques (0.98 and 0.99). We argue that this could be due to lack of domain knowledge. Further, consistency loss is added to the network, which reduces the error (0.42 and 0.76) significantly for both settings. On the other hand, KL divergence based loss is added to the '1-label' framework, which again reduces the error significantly from 0.76 to 0.58. Additionally, embedding loss is introduced to consider future frame consistency,

though we note that for MPII, the change in MAE is very less (i.e. from 0.58 to 0.57). These experiments clearly establish the individual importance of each of the proposed loss terms in our framework.

Impact of Motion Feature. To judge the impact of the motion feature, we evaluate the performance of '2-labels' on the CAVE dataset without using the motion feature i.e. the latent space features are directly concatenated (Refer Fig. 2). The results suggests that with the motion feature the MAE is reduced from 0.34 to 0.25. Thus, it is important to include this information in the framework.

Sequential Modelling Vs '2-labels'. We incorporate an LSTM module having 3 steps instead of the whole pipeline. The input to the LSTM module is 3-frame set and it estimate the gaze of the unlabelled frame. The MAE is quite high (i.e. 0.95) as compared to our ResNet-50 based '2-labels' framework (i.e. 0.25). The possible reason for this performance enhancement owes to task-relevant losses posed within very small temporal information.

Gaze Labelling 'in the wild'. The best results on the collected data in terms of MAE and CC are mentioned in the last row of Table 2. The angular errors w.r.t. the SLERP [39] for '2-labels' and '1-label' are 9.41° and 12.07°, although SLERP is a weak baseline to compare eye gaze in an unconstrained environment. When we use the generated labels on YouTube dataset to complement the other dataset, the performance improves for TabletGaze and Gaze360 (see Table 3

Table 7: Comparison of model based methods on YouTube data. MAE: Mean Absolute Error, AE: Angular Error.

Method	MAE	AE
2-labels	**0.24**	**9.41**
[50]	0.74	15.30
[55]	0.52	13.90
[22]	0.57	14.01

and 4). For adapting the SOTA methods (Table 3 and 4), we fine-tune the models following standard protocol [10,56]. The results suggest that the network learns a meaningful representation. Moreover, we compare our method with model based approaches [50,55,22]. The results are depicted in Table 7. From the table, it is observed that our proposed method outperforms model based methods.

Qualitative Analysis. Fig. 3 illustrates few examples of gaze trajectories along with the predictions of the proposed method. The trajectories consist of start, unlabelled and end frames and the trajectory length is not limited to 3. During training, the proposed method takes start, end and *one of the unlabelled frame* as input. Please note that the eye patch cropped from the facial images are used as input. In the '2-labels' case, the ground truth labels of both start and end frames are provided for weak supervision. In Fig. 3, the green and red arrow indicates ground truth and predicted gaze direction. For better understanding, we plot the gaze direction corresponding to each eye originated from detected pupil center. This qualitative analysis indicates that our weakly supervised method learns to interpolate gaze labels efficiently from terminal frames of a trajectory.
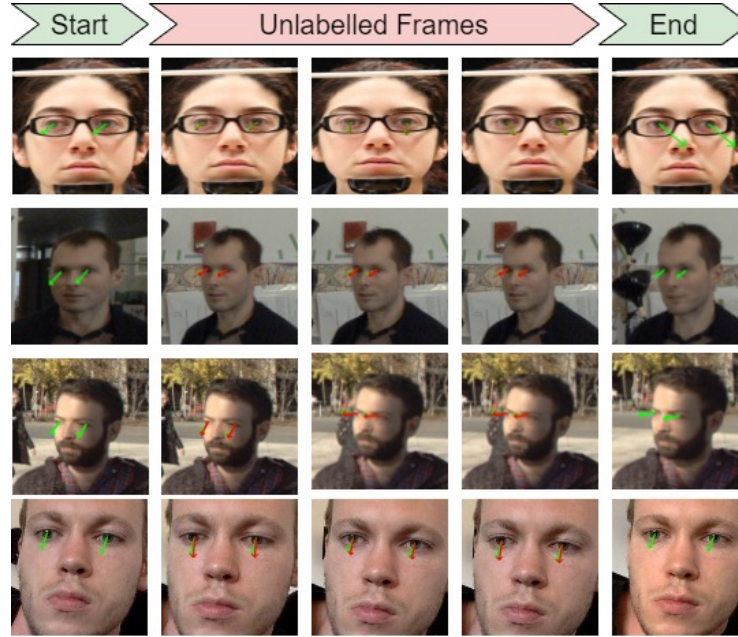
Fig. 3: Few examples of gaze trajectories along with qualitative prediction results. Here, red and green arrow represent predicted and ground truth gaze direction, respectively.

7   Limitations, Conclusion and Future Work

Our study introduces a weakly-supervised approach for generating labels for intermediate frames in a defined gaze trajectory. The proposed frameworks leverage task-specific domain knowledge i.e. trajectory ordering, motion and appearance features etc. With extensive experiments, we show that the labels generated by our methods are comparable to the ground truth labels. Further, we also show that the state-of-the-art existing techniques re-trained using the labels generated by our method give comparable performance. This applies that with just 1%-5.6% labelled data (dependent on the dataset) training can be performed with performance comparable to when 100% training data is available. Further, we also propose a technique to collect and label eye gaze 'in the wild'. The proposed method can be used for other computer vision based applications (e.g. gaze tracking devices for AR and VR) without the prior need of having to use the whole labelled dataset during training. Main limitations of our study are the fixed gaze trajectory requirement, inclusion of eye-blink and near frontal face. In the future, we will investigate subject specific gaze estimation in challenging situations such as low-resolution images, uneven illumination conditions and extreme head-poses. Although our methods consider gaze annotation in few of the aforementioned diverse conditions, it would be interesting to have more in-depth study in this domain.

References

1. gaze code:`https://github.com/swook/GazeML`,`https://github.com/Erkil1452/gaze360` (-)
2. Alonso-Martín, F., Gorostiza, J.F., Malfaz, M., Salichs, M.A.: User localization during human-robot interaction. Sensors **12**(7), 9913–9935 (Jul 2012). https://doi.org/10.3390/s120709913, `http://dx.doi.org/10.3390/s120709913`
3. Arabadzhiyska, E., Tursun, O.T., Myszkowski, K., Seidel, H.P., Didyk, P.: Saccade landing position prediction for gaze-contingent rendering. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
4. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: IEEE Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
5. Barz, B., Denzler, J.: Deep learning on small datasets without pre-training using cosine loss. In: IEEE Winter Conference on Applications of Computer Vision. pp. 1371–1380 (2020)
6. Benfold, B., Reid, I.: Unsupervised learning of a scene-specific coarse gaze estimator. In: IEEE International Conference on Computer Vision. pp. 2344–2351 (2011)
7. Bertasius, G., Feichtenhofer, C., Tran, D., Shi, J., Torresani, L.: Learning temporal pose estimation from sparsely-labeled videos. arXiv preprint arXiv:1906.04016 (2019)
8. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: IEEE Computer Vision and Pattern Recognition. pp. 2846–2854 (2016)
9. Blattgerste, J., Renner, P., Pfeiffer, T.: Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views. In: Proceedings of the Workshop on Communication by Gaze Interaction. pp. 1–9 (2018)
10. Dubey, N., Ghosh, S., Dhall, A.: Unsupervised learning of eye gaze representation from the web. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2019)
11. Duchowski, A.T., Duchowski, A.T.: Eye tracking methodology: Theory and practice. Springer (2017)
12. Fischer, T., Chang, H.J., Demiris, Y.: RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In: European Conference on Computer Vision. pp. 339–357 (2018)
13. Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In: ACM Symposium on Eye Tracking Research and Applications (2014)
14. Garbin, S.J., Shen, Y., Schuetz, I., Cavin, R., Hughes, G., Talathi, S.S.: Openeds: Open eye dataset. arXiv preprint arXiv:1905.03702 (2019)
15. Ghosh, S., Dhall, A., Hayat, M., Knibbe, J., Ji, Q.: Automatic gaze analysis: A survey of deep learning based approaches. arXiv preprint arXiv:2108.05479 (2021)
16. Ghosh, S., Dhall, A., Sharma, G., Gupta, S., Sebe, N.: Speak2label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset. arXiv preprint arXiv:2004.05973 (2020)
17. Ghosh, S., Hayat, M., Dhall, A., Knibbe, J.: Mtgls: Multi-task gaze estimation with limited supervision. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3223–3234 (2022)

18. Gumilar, I., Barde, A., Hayati, A.F., Billinghurst, M., Lee, G., Momin, A., Averill, C., Dey, A.: Connecting the brains via virtual eyes: Eye-gaze directions and inter-brain synchrony in vr. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–7 (2021)
19. Haeusser, P., Mordvintsev, A., Cremers, D.: Learning by association–a versatile semi-supervised training method for neural networks. In: IEEE Computer Vision and Pattern Recognition. pp. 89–98 (2017)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Computer Vision and Pattern Recognition. pp. 770–778 (2016)
21. Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision and Applications 28(5-6), 445–461 (2017)
22. Ishikawa, T.: Passive driver gaze tracking with active appearance models (2004)
23. Jyoti, S., Dhall, A.: Automatic eye gaze estimation using geometric & texture-based networks. In: International Conference on Pattern Recognition. pp. 2474–2479. IEEE (2018)
24. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: IEEE International Conference on Computer Vision (2019)
25. Komogortsev, O.V., Karpov, A.: Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades. Behavior research methods 45(1), 203–215 (2013)
26. Kothari, R., De Mello, S., Iqbal, U., Byeon, W., Park, S., Kautz, J.: Weakly-supervised physically unconstrained gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9980–9989 (2021)
27. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: IEEE Computer Vision and Pattern Recognition. pp. 2176–2184 (2016)
28. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: International Conference on Machine Learning Workshop. vol. 3, p. 2 (2013)
29. Liu, H., Heynderickx, I.: Visual attention in objective image quality assessment: Based on eye-tracking data. IEEE Transactions on Circuits and Systems for Video Technology 21(7), 971–982 (2011)
30. Lu, F., Chen, X., Sato, Y.: Appearance-based gaze estimation via uncalibrated gaze pattern recovery. IEEE Transactions on Image Processing 26(4), 1543–1553 (2017)
31. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Gaze estimation from eye appearance: A head pose-free method via eye image synthesis. IEEE Transactions on Image Processing 24(11), 3680–3693 (2015)
32. Majaranta, P.: Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies: Advances in Assistive Technologies. IGI Global (2011)
33. Mustafa, A., Kaur, A., Mehta, L., Dhall, A.: Prediction and localization of student engagement in the wild. arXiv preprint arXiv:1804.00858 (2018)
34. Niehorster, D.C., Hessels, R.S., Benjamins, J.S.: Glassesviewer: Open-source software for viewing and analyzing data from the tobii pro glasses 2 eye tracker. Behavior Research Methods pp. 1–10 (2020)
35. Palmero, C., Sharma, A., Behrendt, K., Krishnakumar, K., Komogortsev, O.V., Talathi, S.S.: Openeds2020: Open eyes dataset. arXiv preprint arXiv:2005.03876 (2020)

36. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: IEEE International Conference on Computer Vision. pp. 9368–9377 (2019)
37. Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation. In: European Conference on Computer Vision. pp. 721–738 (2018)
38. Park, W., Heo, J., Lee, J.: Talking through the eyes: User experience design for eye gaze redirection in live video conferencing. In: International Conference on Human-Computer Interaction. pp. 75–88. Springer (2021)
39. Peters, C., Qureshi, A.: A head movement propensity model for animating gaze shifts and blinks of virtual characters. Computers & Graphics **34**(6), 677–687 (2010)
40. Purves, D., Morgenstern, Y., Wojtach, W.T.: Perception and reality: why a wholly empirical paradigm is needed to understand vision. Frontiers in systems neuroscience **9**, 156 (2015)
41. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Mutual exclusivity loss for semi-supervised deep learning. In: IEEE International Conference on Image Processing. pp. 1908–1912 (2016)
42. Santini, T., Fuhl, W., Kübler, T., Kasneci, E.: Bayesian identification of fixations, saccades, and smooth pursuits. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications. pp. 163–170 (2016)
43. Schmidt, S., Bruder, G., Steinicke, F.: Depth perception and manipulation in projection-based spatial augmented reality. PRESENCE: Virtual and Augmented Reality **27**(2), 242–256 (2020)
44. Sharma, S., Shanmugasundaram, K., Ramasamy, S.K.: Farec—cnn based efficient face recognition technique using dlib. In: International Conference on Advanced Communication Control and Computing Technologies. pp. 192–195 (2016)
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (2014)
46. Smith, B., Yin, Q., Feiner, S., Nayar, S.: Gaze locking: passive eye contact detection for human-object interaction. In: ACM User Interface Software & Technology (2013)
47. Startsev, M., Agtzidis, I., Dorr, M.: 1d cnn with blstm for automated classification of fixations, saccades, and smooth pursuits. Behavior Research Methods **51**(2), 556–572 (2019)
48. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: IEEE Computer Vision and Pattern Recognition. pp. 1821–1828 (2014)
49. Swaminathan, A., Ramachandran, M.: Enabling augmented reality using eye gaze tracking (Jun 12 2018), uS Patent 9,996,150
50. Valenti, R., Sebe, N., Gevers, T.: Combining head pose and eye location information for gaze estimation. IEEE Transactions on Image Processing **21**(2), 802–815 (2011)
51. Wang, K., Zhao, R., Ji, Q.: A hierarchical generative model for eye image synthesis and eye gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 440–448 (2018)
52. Wang, W., Shen, J.: Deep visual attention prediction. IEEE Transactions on Image Processing **27**(5), 2368–2378 (2017)
53. Weston, J., Ratle, F., Mobahi, H., Collobert, R.: Deep learning via semi-supervised embedding. In: Neural networks: Tricks of the trade, pp. 639–655. Springer (2012)
54. Williams, O., Blake, A., Cipolla, R.: Sparse and semi-supervised visual mapping with the s3gp. In: IEEE Computer Vision and Pattern Recognition (2006)

55. Yamazoe, H., Utsumi, A., Yonezawa, T., Abe, S.: Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In: Proceedings of the 2008 symposium on Eye tracking research & applications. pp. 245–250 (2008)
56. Yu, Y., Liu, G., Odobez, J.: Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 11937–11946 (2019)
57. Yu, Y., Odobez, J.: Unsupervised representation learning for gaze estimation. IEEE Conference on Computer Vision and Pattern Recognition pp. 1–13 (2020)
58. Zabala, U., Rodriguez, I., Martínez-Otzeta, J.M., Lazkano, E.: Modeling and evaluating beat gestures for social robots. Multimedia Tools and Applications pp. 1–18 (2021)
59. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: IEEE Computer Vision and Pattern Recognition Workshop (2017)
60. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
61. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: European Conference on Computer Vision. pp. 365–381. Springer (2020)
62. Zhang, X., Sugano, Y., Bulling, A.: Everyday eye contact detection using unsupervised gaze target discovery. In: ACM User Interface Software and Technology. pp. 193–203 (2017)
63. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: IEEE Computer Vision and Pattern Recognition. pp. 4511–4520 (2015)
64. Zhang, Y., Dong, W., Hu, B.G., Ji, Q.: Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In: IEEE Computer Vision and Pattern Recognition. pp. 2314–2323 (2018)
65. Zhang, Y., Zhao, R., Dong, W., Hu, B.G., Ji, Q.: Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7034–7043 (2018)
66. Zhao, J., Mathieu, M., Goroshin, R., Lecun, Y.: Stacked what-where auto-encoders. International Conference on Learning Representations Workshop (2015)
67. Zhu, Y., Yan, Y., Komogortsev, O.: Hierarchical hmm for eye movement classification. In: European Conference on Computer Vision. pp. 544–554. Springer (2020)