

The Story in Your Eyes: An Individual-difference-aware Model for Cross-person Gaze Estimation

Jun Bao

Hangzhou Dianzi University
Hangzhou, China
baoj@hdu.edu.cn

Buyu Liu

NEC Laboratories America
San Jose, USA
buyu@nec-labs.com

Jun Yu

Hangzhou Dianzi University
Hangzhou, China
yujun@hdu.edu.cn

Abstract

We propose a novel method on refining cross-person gaze prediction task with eye/face images only by explicitly modelling the person-specific differences. Specifically, we first assume that we can obtain some initial gaze prediction results with existing method, which we refer to as *InitNet*, and then introduce three modules, the Validity Module (VM), Self-Calibration (SC) and Person-specific Transform (PT) Module. By predicting the reliability of current eye/face images, our VM is able to identify invalid samples, e.g. eye blinking images, and reduce their effects in our modelling process. Our SC and PT module then learn to compensate for the differences on valid samples only. The former models the translation offsets by bridging the gap between initial predictions and dataset-wise distribution. And the later learns more general person-specific transformation by incorporating the information from existing initial predictions of the same person. We validate our ideas on three publicly available datasets, EVE, XGaze and MPIIGaze and demonstrate that our proposed method outperforms the SOTA methods significantly on all of them, e.g. respectively 21.7%, 36.0% and 32.9% relative performance improvements. We won the GAZE 2021 Competition on the EVE dataset. Our code can be found here https://github.com/bjj9/EVE_SCPT.

1. Introduction

Gaze estimation from a single low-cost RGB sensor is an important topic in computer vision [20, 21]. Inputting face or eye images¹, gaze prediction aims to estimate the gaze direction or Point of Gaze (PoG) on screen. The task of cross-person gaze estimation is defined as one where a model is evaluated on a previously unseen set of participants, which is undoubtedly more challenging. Results from gaze prediction usually serves as input to tasks, e.g. human at-

¹Our method can handle both eye and face images and we refer to our input later in this paper as eye images for simplicity.

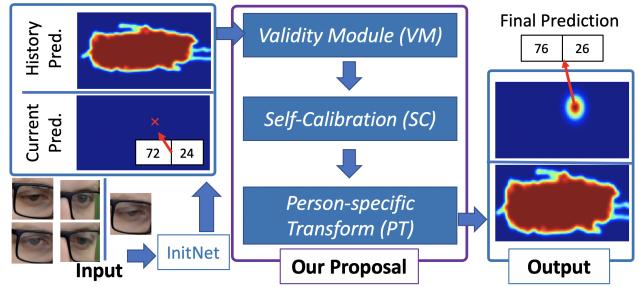


Figure 1: We propose to model the person-specific differences in cross-person gaze estimation problem. Specifically, our model takes the initial gaze predictions of current eye image and previous initial predictions of the same participant as input. After removing invalid samples via VM and explicitly modelling the person-specific offsets, it is able to improve the gaze prediction accuracy by a large margin.

tention estimation [6]. Therefore, having an accurate gaze estimation model can be essential for downstream tasks, e.g. intelligent user interfaces [8, 11].

Though there are many existing work aim to predict gaze location/direction w.r.t information from single images [21], video sequences and contents [20], the problem that there exist un-observable person-specific differences remains. To this end, many person-specific adaptation techniques are proposed, including exploiting networks to predict person-wise differences [17], directly estimating 6-degree calibration parameters of human eyes [16] or fine-tuning models with few labelled test samples [29].

We follow this adaptation line of work and propose a novel method to model person-specific differences. Unlike existing methods that require additional information, such as labelled samples from test participants to perform person-specific calibration or additional content information on screen, ours learns to compensate person-specific differences with eye images **only**. To achieve that, we propose three modules, one for modelling the reliability of samples

and two others for person-specific differences modelling. Specifically, we introduce a non-parametric module, or Validity Module (VM), that estimates the reliability of samples so that the invalid/unreliable samples would not affect our difference modelling procedure. We further propose two modules, Self-Calibration (SC) and Person-specific Transform (PT), to bridge the person-specific differences between human optical and visual axes. Our SC Module takes initial prediction from existing methods as input, models the gap between initial prediction w.r.t. dataset-wise gaze distribution and outputs refined prediction. And the PT Module explicitly learns a more general transformation on refined predictions by incorporating information from history refined predictions of the same participant. Please note that history predictions of the same participant can be interpreted in two ways. In the online setting where we are given consecutive frames as input, history predictions of the same participant mean all previous predictions from this particular participant. Given an offline scenario where all eye images, either sampled sparsely in video sequences or just some random images, of one participant are available, we then refer to all other images except the current one as its history. In general, our method is applicable to both video and image based dataset as long as multiple samples are available for one participant during test time.

It is well-acknowledged in vision literature that the angle kappa, the deviation between optical axis and visual axis, cannot be reflected by images taken with conventional cameras [2]². The standard deviation of angle kappa in normal population is around 1.8° in both horizontal and vertical axis [3, 1], leading to a random error of 2.0° - 2.3° when estimating gaze directions directly from images. Therefore, existing single-image-based methods have no chance to recover such individual differences and, in theory, their accuracy would ceil at 2.0° . Our goal is to model such differences thus effectively close the gap.

We validate our ideas on three publicly available datasets, EVE [20], MPIIGaze [33] and XGaze [30], and obtain the state-of-the-art (SOTA) performance on all three datasets, 27.6%, 36.0% and 32.9% relative performance improvement over existing SOTA methods. We further demonstrate the effectiveness of each module by performing ablation studies.

To summarize, our key contributions are:

- A **novel** gaze estimation method handles data with invalid samples and models person-specific differences under cross-person setting with only eye images.
- Our method includes (i) a Validity Module that estimates the reliability of samples, (ii) a Self-Calibration Module that models prediction offset w.r.t. dataset-wise distribution and (iii) a Person-specific Transform Module that explicitly compensates individual differences.

- Our model outperforms the SOTA by a large margin on three publicly available datasets. We won the GAZE 2021 Competition on the EVE dataset.

2. Related Work

Model-based Gaze Estimation Gaze estimation methods can generally be categorized into model-based or appearance-based. Model-based methods generally exploit geometric eye models and can be further distinguished into shape-based and corneal-reflection methods. Shape-based methods [12, 4] first detect eye shape and then infer gaze directions from detection results, e.g. the pupil centers. In contrast, corneal-reflection models rely on eye features by using reflections of an external infrared light source on the outermost layer of the eye, the cornea. Starting from limited stationary settings [19], corneal-reflection are capable of handling arbitrary head poses using multiple light sources or cameras [34] nowadays. Though more practical application scenarios [7, 27] have been observed with model-based methods, one main drawback of this line of work remains. That is, their gaze estimation accuracy is not satisfactory for real-world settings as they rely heavily on accurate eye feature detection results, which further requires high-resolution images and homogeneous illumination. Due to such requirements, these methods are not as desired as appearance-based methods in real-world settings or on commodity devices.

Appearance-based Gaze Estimation Appearance-based gaze estimation methods aim to map images directly to gaze. Compared to model-based approaches, appearance-based method achieves much better results for in-the-wild settings. Intuitively, these methods do not rely on any outputs from explicit shape extraction step thus are less constrained to image resolution or distances. Under restricted conditions, e.g. images are taken under limited and constrained laboratory conditions, regression techniques [18] or random forests ones [26] have been explored. With the benefits of large scale datasets, e.g. MPIIGaze [33], CNN-based methods [32, 33] further push this field fast forward. MPIIGaze therefore becomes a benchmark dataset for in-the-wild gaze estimation. Recently, larger datasets, such as EVE [20] and XGaze [30], provide more diverse data to evaluate gaze prediction performances with various experimental settings. To improve the prediction performance, researchers have explored model structures, e.g. introduce more complex or ensembles of CNNs [33, 9], model input, such as extend to multi-modal input [15, 28] or improve data normalization [31], and model representations, for instance, learning more informed intermediate representations [22]. However, in order to cover the significant variability in eye appearance caused by free head motion, these methods require more person-specific training data compared to model-based approaches, e.g. they generally require some specific domains

²It can be captured with dedicated equipment, e.g. Synoptophore.

or persons. Compared to existing methods, ours is more generic. Firstly, all the above mentioned methods, either appearance-based or model-based, are compatible with our model as long as they provide initial gaze estimations and our method can be applied to improve their performance. Secondly, our method explicitly models person-specific differences with eye images only.

Cross-person Gaze Estimation Modelling person-specific differences seems to be a natural way to perform cross-person estimation task. However, given restricted setting that the test participants are unseen during training time, incorporating personal modelling/calibration can be a challenging problem itself. Therefore, some existing methods try to explore additional data to tackle the cross-person gaze direction (and subsequent Point-of-Gaze) prediction problem. One line of work proposed to relax the restriction a bit by assuming that very few samples of a target test person’s data are available. Then they fine-tuned or adapted pre-trained model on this data and demonstrate advanced performance on the final test data from the same person [15, 23]. Building on top of this work, less samples, e.g. as few as 9 calibration samples for each test person, are required to achieve performance improvements [17]. Although the improvements seem to be promising, these methods all require labeled samples of test participants thus is not practical/generic. Another type of additional information comes from screen contents where the predicted visual saliency of the screen content is assumed to be available. With this assumption, researchers propose to align estimated PoG with an estimated visual saliency [25, 5]. To avoid the over-fitting problem of single saliency models on training data, multiple saliency models [24] are explored. More recent work [30] exploit both screen content and temporal cues. Unlike saliency-based method, screen content is incorporated in the form of region of interests on screen [30]. In contrast to existing methods that require additional information for person-specific modelling, such as test samples to be labelled, screen content or even consecutive frames from video sequences, our proposed method is more generic since we requires eye images **only**.

3. Our Framework

Our model includes four modules. The first module InitNet inputs eye images and outputs the initial Point of Gaze (PoG) on screen. The second Validity Module (VM) inputs the initial PoG predictions and outputs the reliability of each sample. The third module, or Self-Calibration (SC) Module, takes the validity predictions as well as the initial predictions as input, and learns to compensate the current prediction w.r.t. its history predictions and dataset-wise distribution. And we refer to the output from SC as refined PoG. Our last Person-specific Transform (PS) Module parses the refined

PoG and its refined history information outputs our final PoG by explicitly modelling person-specific differences. We provide more details for each module in Sec. 3.1 and then introduce learning process of each module in Sec. 3.2. Please note that our main contributions lie in the last three modules and we introduce the first module here for completeness.

3.1. Our Model

In this section, we focus on model structure and assume that supervisions are available for each module. Let’s first assume that we have an eye-tracking data set $\mathcal{D} = \{I_{t,j}, g_{t,j}, v_{t,j}\}_{t=1, j=1}^{N, M}$ of N samples for M participants. Specifically, $I_{t,j} = \{I_{t,j}^*\}_{*=l, r}$ representing the paired images for the t -th sample of j -th participants, where $* = l$ represents the left eye and $* = r$ is for the right. We define $g_{t,j}$ as a 2-dimensional vector denoting the ground truth Point of Gaze (PoG) location of the t -th sample of the j -th participant on screen. Finally, $v_{t,j}$ is binary indicating whether the current sample is valid or not. Fig. 2 gives an overview of the proposed method.

3.1.1 InitNet Module

The goal of InitNet module is to provide initial PoG estimation on the given images. Please note that the ultimate goal of this paper is to improve the gaze predictions from InitNet and it can be any existing gaze estimation methods. We denote the initial PoG prediction of the t -th paired images from participant j as $p_{t,j} = \frac{p_{t,j}^l + p_{t,j}^r}{2}$ where $p_{t,j}^l$ and $p_{t,j}^r$ denotes the PoG prediction on left eye image $I_{t,j}^l$ and right eye image $I_{t,j}^r$. Similarly, $p_{t,j}^l$ and $p_{t,j}^r$ are 2-dimensional vectors representing the gaze location on screen and $p_{t,j}$ averages the left and right positions.

Our InitNet follows the structure in [20]. Specially, InitNet takes either the left or right eye image, parses it to ResNet18 [10] and outputs $d_{t,j}^*$ for this particular image³. Mathematically, we denote InitNet as f_{init} and $d_{t,j}^* = f_{init}(I_{t,j}^*)$, where $* = \{l, r\}$ denotes left or right sample. In general, $d_{t,j}^*$ is represented by a 3-dimensional unit vector and can be converted to PoG. To achieve that, we first combine $d_{t,j}^*$ and with 3D gaze origin position, which is determined during data pre-processing. This would provide us a gaze ray with 6 degree of freedom. Given camera extrinsic as well as screen plane location, we can compute the intersection of this ray with the screen plane, which gives PoG $p_{t,j}^*$. Note that theoretically, $p_{t,j}$ can be represented by pixel dimensions or in centimeters⁴ and otherwise notified, we take the centimeter representation in our paper. Again,

³Note that compared to [20], we remove the pupil size prediction task as it deteriorates the gaze direction performance in practice.

⁴For instance, in EVE dataset [20], their screen is 1920 by 1080 with 553mm wide and 311mm tall.

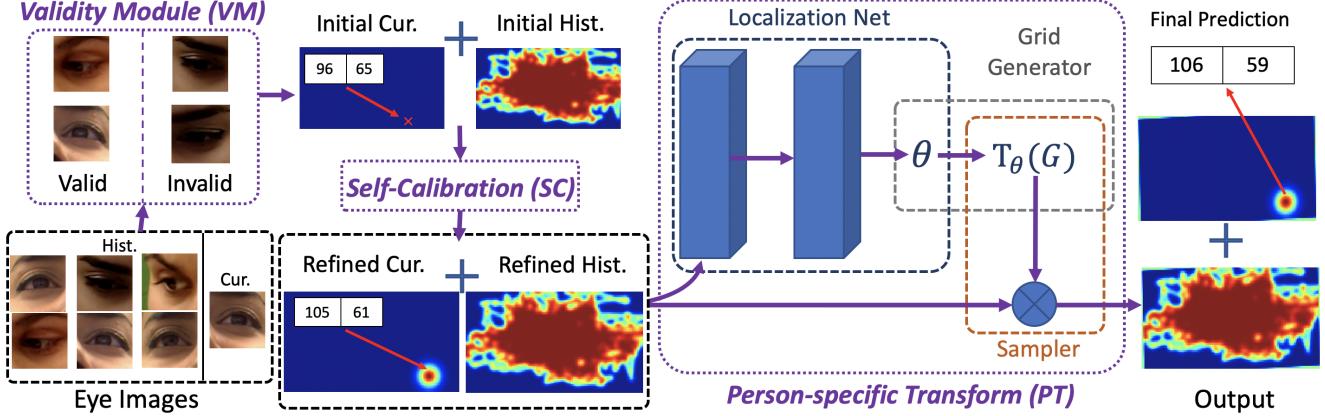


Figure 2: Assuming that we have current eye image and history eye images of the same person available, our method first models the validity of all images with the help of VM and then explicitly compensates the person-specific differences with SC and PT modules. We observe significant performance improvement over cross-person gaze estimation task.

InitNet is not restricted to ResNet18 but can be any State-Of-The-Art (SOTA) structure. We refer the readers to [20] and Sec. 4 for more details.

3.1.2 Validity Module

Predicting $p_{t,j}$ with $I_{t,j}$ can be noisy. For instance, participants may blink eyes and given partially visible pupil, the predicted PoG can be quite off. Errors from such noisy predictions can even propagate to future samples if patterns from previous predictions are modelled.

In view of this problem, we propose to identify how reliable $p_{t,j}$ would be and introduce our second module f_{vm} , or Validity Module, that takes $p_{t,j}^*$ as input and outputs $b_{t,j}^*$, where $b_{t,j}^*$ is binary denoting the reliability of current sample. Specifically, f_{vm} is an indicate function that considers both current prediction and history predictions of the same participant, if any. $b_{t,j}^*$ is one as long as it satisfies at least one of the following two requirements. 1. The initial prediction $p_{t,j}^*$ is inside the screen. 2. $p_{t,j}^*$ lies within 3 times of the standard deviation distance w.r.t. averaged history predictions, if any⁵. Finally, $b_{t,j} = b_{t,j}^l \times b_{t,j}^r$. Example VM results are shown in Fig. 3.

3.1.3 Self-Calibration Module

After obtaining the initial PoG $p_{t,j}$ from InitNet and reliability $b_{t,j}$ from Validity Module, our next step is to learn to compensate the prediction bias. Intuitively, every individual has a person-specific offset between his/her optical and visual axes in each eye. Although one can observe the

⁵We also tried to train a network to predict the $b_{t,j}^*$ with $I_{t,j}^*$ as input. In practise, the current formulation provides similar performance.

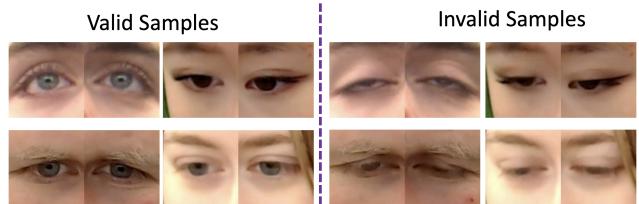


Figure 3: Examples for valid and invalid samples.

former by the appearance of iris in $I_{t,j}$, the later cannot be identified or detected easily as it is defined by the position of fovea at the back of eyeball. Since InitNet absorb the person-specific offset into its parameters, we further introduce Self-Calibration (SC) Module to model the offset thus to bridge the gap between visual and optical axes.

Specifically, we exploit dataset-wise distribution to model the person-specific offset. g_{tr} is the average PoG on the valid samples on training set and is defined as:

$$g_{tr} = \frac{\sum_{t=1, j=1}^{N, M} g_{t,j} \cdot v_{t,j}}{\sum_{t=1, j=1}^{N, M} v_{t,j}} \quad (1)$$

Then we model offset by measuring the gap between g_{tr} and averaged valid PoG locations and compensate our initial PoG w.r.t. learned offset to obtain the refined PoG, or $\hat{p}_{t,j}$. Mathematically, we define SC as:

$$\hat{p}_{t,j} = p_{t,j} - \left(\frac{\sum_{i \in \mathcal{H}_{t,j}} p_{i,j} \cdot b_{i,j}}{\sum_{i \in \mathcal{H}_{t,j}} b_{i,j}} - g_{tr} \right) \quad (2)$$

where $\mathcal{H}_{t,j}$ represents the set of history index. For online setting where we are given consecutive frames sequentially, $\mathcal{H}_{t,j} = \{1, \dots, t-1\}$. As for offline scenario where all samples of j -th participant are available, we define

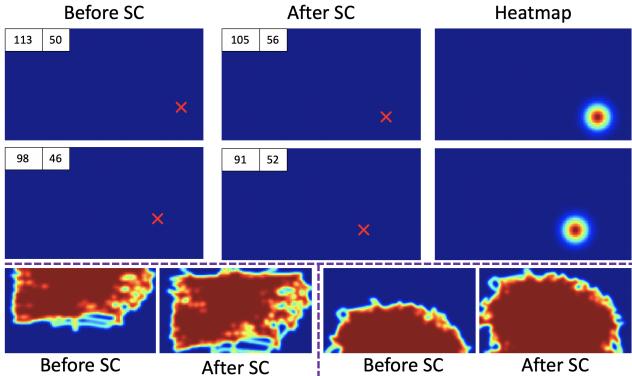


Figure 4: Example results of SC Module. SC is able to perform translation on initial predictions w.r.t. valid initial predictions and dataset-wise distribution.

$\mathcal{H}_{t,j} = \{1, \dots, t-1, t+1, \dots, N\}$. Please note that our SC Module is person-specific since we calibrate the predictions of $p_{t,j}$ based on the information of the same participant j . Theoretically, this module is able to perform self-calibration as long as there are multiple samples for the same participant. More details for the two settings can be found in Sec. 4.

We denote the refined PoG predictions for the t -th sample of the j -th participant obtained from SC Module as $\hat{p}_{t,j}$. The prediction history of this sample is defined as $\mathcal{P}_{t,j} = \{\hat{p}_{i,j}\}_{i \in \mathcal{H}_{t,j}}$. We visualize the input $p_{t,j}$ and output of SC Module $\hat{p}_{t,j}$ in Fig. 4 (See Before SC and After SC on the upper region). For clarity, the predicted PoG (in pixel space) are also reported by the 2D vector inside each image.

3.1.4 Person-specific Transform Module

Although SC Module is beneficial in terms of removing the person-specific difference, it only deals with cases where the difference is mainly translation. To this end, we further propose the Person-specific Transform (PT) Module to model and compensate more general differences.

Specifically, our PT Module takes the heatmaps of the refined PoG prediction $\hat{p}_{t,j}$ and its history $\mathcal{P}_{t,j}$ as input and outputs the transformed heatmaps for both. To obtain the heatmap for the t -th sample of j -th participant, or $r_{t,j} \in \mathbb{R}^{H \times W \times 1}$, we first map $\hat{p}_{t,j}$ to pixel space and then convolving a dirac delta function centered at $\hat{p}_{t,j}$ with a isotropic 2D Gaussian of fixed variance (See the middle and right column in Fig. 4 for $\hat{p}_{t,j}$ and $r_{t,j}$). We further denote the heatmap for $\mathcal{P}_{t,j}$ as $r_{t,j}^h \in \mathbb{R}^{H \times W \times 1}$. To obtain it, we repeat similar process described above on only valid samples. More specifically, $\hat{p}_{i,j}$ is valid as long as $\hat{p}_{i,j} \in \mathcal{P}_{t,j}$ and $b_{i,j} \neq 0$. Details of the process can be found in Alg. 1.

In practice, we find the order of samples does not affect the final performance. More importantly, instead of simply

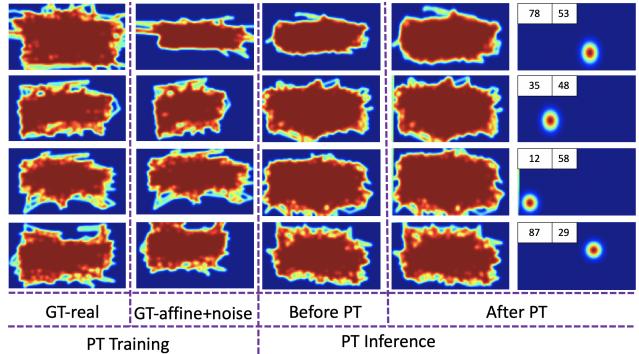


Figure 5: Examples for PT Module. We introduce diverse affine-like transformation with noise during training (left), which enables us to handle unseen cases thus we can effectively model cross-person differences during inference (right). We omit $r_{t,j}$ in Before PT.

combining the heatmap of all individual valid points in $\mathcal{P}_{t,j}$ to obtain the history heatmap, we propose to include the *trajectories* between two PoGs as well, as explained in the 11-th step in Alg. 1. Such design helps us to focus more on the PoGs that are distributed far away from center so that they can play more important role in our modelling process. Please note that we exclude invalid PoGs in this generation process in order to perform meaningful transformation, e.g. our PT Module will not be distracted by invalid predictions.

We concatenate $r_{t,j}$ and $r_{t,j}^h$ channel-wise to obtain our input $\mathbf{r}_{t,j} \in \mathbb{R}^{H \times W \times 2}$ for PT Module. Our PT Module consists of one localization net, one grid generator and one sampler [13]. The output of PT Module is denoted as $\hat{\mathbf{r}}_{t,j} \in \mathbb{R}^{H \times W \times 2}$, including both transformed sample heatmap $\hat{r}_{t,j}$ and transformed history heatmap $\hat{r}_{t,j}^h$. Both of them are in $\mathbb{R}^{H \times W \times 1}$ space. Mathematically, we denote PT Module as f_{pt} and $\hat{\mathbf{r}}_{t,j} = f_{pt}(\mathbf{r}_{t,j})$. To obtain the final PoG for $I_{t,j}$, one just need to apply softmax on $\hat{r}_{t,j}$, find the position with the maximum value and then re-scale the the position to screen size. We give some example results of PT in Fig. 5 (PT Inference). More details of our PT Module can be found in supplementary materials.

Note that SC Module is necessary yet important even SC seems to only consider translation-like offset while PT can compensate all affine-like transformations. For instance, PT is not able to handle scenarios where initial PoG predictions $p_{i,j}$ are mostly off screen. Or in another word, the translation is so large that some samples are out of scope. SC Module, in this case, is able to translate the initial predicted PoGs to some reasonable locations so that the later learning process of PT is meaningful and effective. In general, SC Module and PT Module are mutual beneficial and play different roles in our full model. We refer the readers to PT training in Fig. 5 and lower part of Fig. 4 for different roles that SC and

PT play in our model.

3.2. Model Training

We demonstrate our full model in previous section with the assumption that supervisions are all available for each module. In this section, we will describe how to train our model with given \mathcal{D} .

InitNet Module Given the structure defined in Sec. 3.1.1, we can prepare the ground-truth gaze direction $\tilde{d}_{t,j}$ with corresponding $g_{t,j}$ for each $I_{t,j}$. Then our loss function is defined as:

$$\mathcal{L}_{init}(\tilde{d}_{t,j}, d_{t,j}) = \frac{180}{\pi} \arccos\left(\frac{\tilde{d}_{t,j} \cdot d_{t,j}}{\|\tilde{d}_{t,j}\| \|d_{t,j}\|}\right) \quad (3)$$

Person-specific Transform Module The PT Module is designed to remove the remaining person-specific difference in $\mathbf{r}_{t,j}$ regardless of which network structure is used for the Initnet Module. As the characteristics of person-specific difference reflected in $\mathbf{r}_{t,j}$ varies across datasets and Initnet Modules, a generalised PT Module should be trained on samples that cover all these variations. In practice, we augment the ground-truth $\tilde{\mathbf{r}}_{t,j}$ to $\check{\mathbf{r}}_{t,j}$ in order to emulate all variations in person-specific difference. During training, $\check{\mathbf{r}}_{t,j}$ is served as input to the PT Module and $\tilde{\mathbf{r}}_{t,j}$ as ground truth; during inference, the PT Module takes in $\mathbf{r}_{t,j}$ and outputs $\hat{\mathbf{r}}_{t,j}$. $\tilde{\mathbf{r}}_{t,j}$ is obtained by concatenate $\tilde{\mathbf{r}}_{t,j}$ and $\tilde{r}_{t,j}^h$ channel-wise and similarly $\check{\mathbf{r}}_{t,j}$ is obtained by concatenate $\check{\mathbf{r}}_{t,j}$ and $\check{r}_{t,j}^h$ channel-wise. Similar to the process of obtaining $r_{t,j}$ and $r_{t,j}^h$ from $\hat{p}_{t,j}$, $b_{i,j}$ and $\mathcal{H}_{t,j}$ as shown in Alg. 1, we obtain $\tilde{r}_{t,j}$ and $\tilde{r}_{t,j}^h$ from $g_{i,j}$, $v_{i,j}$ and $\mathcal{H}_{t,j}$, and obtain $\check{r}_{t,j}$ and $\check{r}_{t,j}^h$ from $\check{g}_{i,j}$, $v_{i,j}$ and $\mathcal{H}_{t,j}$, where $\check{g}_{i,j}$ is augmented from $g_{i,j}$ as shown in Alg. 2. We refer the readers to supplementary for details on how we define the set for affine parameters \mathcal{A} and random noise \mathcal{N} during this augmentation process. We provide some visualization examples for $\tilde{r}_{t,j}^h$ and $\check{r}_{t,j}^h$ in PT training part of Fig. 5, where *GT-real* denotes the ground-truth heatmap and *GT-affine + noise* is the augmented one that emulates person-specific difference.

As long as we have training samples available for PT Module, we then define our loss function as:

$$\mathcal{L}_{pt}(\tilde{\mathbf{r}}_{t,j}, \check{\mathbf{r}}_{t,j}) = BCEloss(\tilde{r}_{t,j}^h, \check{r}_{t,j}^h) \quad (4)$$

where BCE is the binary cross-entropy loss. As can be seen in the equation above, we do not introduce any loss on transformed $\check{r}_{t,j}$. Our PT Module learns a transformation w.r.t. $\check{r}_{t,j}^h$ and applies the learned transformation directly on $\check{r}_{t,j}$. Although one can always introduce loss on $\check{r}_{t,j}$, in practice, we find that introducing either BCEloss on $\check{r}_{t,j}$ or

MSE loss on final numerical estimate of PoG deteriorates the model performance. This might because that learning transformation with per-sample gaze heatmap $\check{r}_{t,j}$ is less meaningful as there are only limited information on this heatmap. In addition, focusing on $\check{r}_{t,j}$ might further confuse the localization net in PT Module.

Algorithm 1 Framework for generating history heatmap $r_{t,j}^h$.

- 1: **Input:** Predicted history samples: $\mathcal{P}_{t,j} = \{\hat{p}_{i,j}\}_{i \in \mathcal{H}_{t,j}}$; Validity prediction: $\{b_{i,j}\}$ for all $i \in \mathcal{H}_{t,j}$
 - 2: **Output:** Heatmap $r_{t,j}^h$ for predicted history samples;
 - 3: **Initialize:** Set of visited samples: $\mathcal{S} = \mathcal{H}_{t,j}$; All zero heatmap: $r_{t,j}^h$
 - 4: Randomly select k from \mathcal{S} as long as $b_{k,j} \neq 0$;
 - 5: Map $\hat{p}_{k,j}$ to pixel space and plot in $r_{t,j}^h$ by setting the pixel value of corresponding location to 1.
 - 6: Update: $\mathcal{S} = \mathcal{S} \setminus k$
 - 7: **while** $\sum_{i \in \mathcal{S}} b_{i,j}$ greater than 0 **do**
 - 8: Randomly select m from \mathcal{S} as long as $b_{m,j} \neq 0$
 - 9: Map $\hat{p}_{m,j}$ to pixel space and plot it in $r_{t,j}^h$
 - 10: Update: $\mathcal{S} = \mathcal{S} \setminus m$
 - 11: In $r_{t,j}^h$, set the value of all pixels that lie in the segment bounded with two points, or $\hat{p}_{m,j}$ and $\hat{p}_{k,j}$, to 1
 - 12: Set k to m
 - 13: **end while**
 - 14: $r_{t,j}^h = g(r_{t,j}^h)$ where g denotes 2D Gaussian of fixed variance.
-

Algorithm 2 Framework for generating the augmented samples $\check{g}_{i,j}$ from $g_{i,j}$.

- 1: **Input:** Ground truth samples: $\{g_{i,j}\}$;
 - 2: **Output:** Augmented samples $\{\check{g}_{i,j}\}$
 - 3: **Initialize:** Affine transformation parameters: \mathcal{A} ; Noise parameter: \mathcal{N}
 - 4: Randomly sample one parameter set a from \mathcal{A}
 - 5: **for** $g_{m,j} \in \{g_{i,j}\}$ **do**
 - 6: Initialize: $\check{g}_{m,j} = g_{m,j}$
 - 7: Randomly sample one parameter set n_m from \mathcal{N}
 - 8: Map $\check{g}_{m,j}$ to pixel space
 - 9: Apply affine transformation a to $\check{g}_{m,j}$ in pixel space, add n_m on top of the transformed location
 - 10: Convert $\check{g}_{m,j}$ in pixel space back to direction space
 - 11: **end for**
-

4. Experiments

In this section, we demonstrate the effectiveness of our proposed model by conducting several experiments on three

Method	Requirements			Performance on EVE test set [20]		
	Static Image	Consecutive Frame	Screen Content	Gaze Dir. ($^{\circ}$) \downarrow	PoG (cm) \downarrow	PoG (px) \downarrow
EyeNet-Static [20]	✓			4.54	5.10	172.7
EyeNet-GRU [20]		✓		3.48	3.85	132.56
GazeRefineNet-Static [20]		✓	✓	2.87	3.16	109.85
GazeRefineNet-RNN [20]		✓	✓	2.57	2.83	98.38
GazeRefineNet-LSTM [20]		✓	✓	2.53	2.79	96.97
GazeRefineNet-GRU [20]		✓	✓	2.49	2.75	95.59
Ours-Online		✓		2.15	2.39	82.91
Ours-Offline	✓			1.95	2.17	75.19

Table 1: Performances on EVE test set. We outperform the SOTA significantly when having the same setting. And our method beats the SOTA with less information required. We rank the first in EVE leader-board upon submission.

publicly available datasets, EVE [20], MPIIGaze [33] and XGaze [30]. We demonstrate the state-of-the-art (SOTA) performance on these datasets and perform ablation study by validating the effectiveness of each module.

Datasets: **EVE** dataset is the main one that validate our ideas on. Specifically, 12,308,334 frames are provided in EVE. 54 participants are recorded with natural eye movements (as opposed to following specific instructions or smoothly moving targets). The gaze angles in EVE are in the range of [-60,60] degrees in the vertical and [-70, +70] degrees in the horizontal direction. It also covers a large set of head movement. We follow the standard split in [20] and report our final results on test sequence. We conduct the ablation study on its validation set. **MPIIGaze** consists of 213,659 images from 15 participants (six females, five with glasses), among which 14 of them are used for training and one for testing. In addition, 10 participants had brown, 4 green, and one grey iris colour. Participants collected the data over different time periods thus images are with high illumination diversity. The gaze angles in MPIIGaze are in the range of [-1.5,20] degrees in the vertical and [-18, +18] degrees in the horizontal direction. We follow the standard split in [33] and report our results. **XGaze** dataset consists of 1,083,492 images taken from 110 participants, including 47 female and 63 male. 17 of them wore contact lenses and 17 of them wore eyeglasses during recording. As for illumination condition, 16 controlled conditions are explored in XGaze. The gaze angles in XGaze are in the range of [-70,70] degrees in the vertical and [-120, +120] degrees in the horizontal direction. We follow the standard split as suggested in [30], e.g. with-in dataset setting where 80 participants are used for training and 15 are for testing, and report our performance on test set.

Evaluation Metrics: As for EVE, We follow [20] and report the predicting gaze direction and PoG. For the [30, 33], we report the gaze prediction error in degree.

Training Details: As for EVE dataset, we train InitNet Module from scratch for 8 epochs with ADAM [14], with

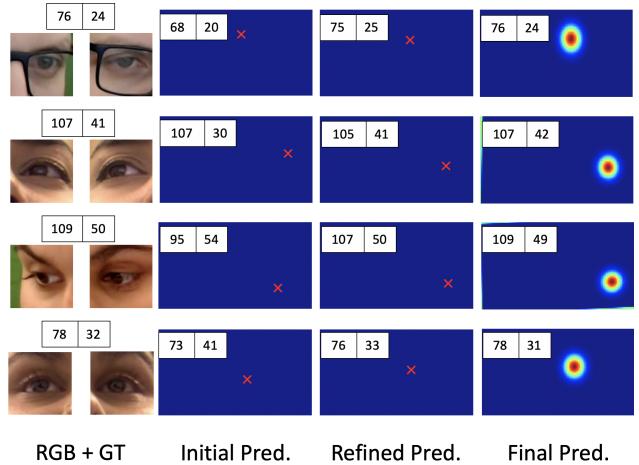


Figure 6: We demonstrate input RGB, GT, initial, refined and final predictions in this figure. Introducing each module can indeed boost performance.

learning rate set to $1e - 3$. Person-specific Transform (PT) Module is trained from scratch with SGD for 100 epochs and we set learning rate to 0.1 with momentum set to 0.9. Batch size is set to be 12 and 3200 for InitNet and PT respectively. As for MPIIGaze and XGaze datasets, we exploit [33] and [30] structure instead to demonstrate the generality of proposed method, e.g. our method is not restricted to any specific gaze prediction network but is able to improve the performance of existing methods in general. Please further note that we train our PT Module *only* on EVE and directly apply the pre-trained model on MPIIGaze and XGaze. Also, since both MPIIGaze and XGaze include only valid frames/images, we do not apply Validity Module to them. We have both online and offline version for EVE dataset as it provides video sequences. Only offline model is available on MPIIGaze and XGaze due to the lack of consecutive frames. We set H to 72 and W to 128 in experiment.

4.1. Results on EVE

In this section, we evaluate our method on EVE and demonstrate that our proposed method can aid in gaze estimation. Developed upon InitNet, we then evaluate the effects of individual modules in improving initial estimate of PoGs.

4.1.1 Main Results

Eye Gaze Estimation We first consider the task of gaze estimation purely from paired eye image patches. To obtain the ground-truth, we simply average the ground-truth of left and right eye and compared the averaged ground-truth with our prediction. Tab. 1 shows the performance of our full model on predicting gaze direction and PoG. We can see that our proposed method beats SOTA significantly on EVE, 27.6% relative improvement. Please note that the proposed method is directly comparable to SOTA [20] as all networks are trained on the training split of EVE. To have a fair comparison, our InitNet follows the same structure as EyeNet [20] (See Sec. 3 for details). More importantly, in contrast to incorporating contents in screen or requiring consecutive frames as input, our model relies only on eye image patches thus is more general and requires less information. Generally, we find that modelling person-specific differences is very important and our proposed modules, SC and PT, indeed benefits a lot in terms of compensate for the person-specific pattern.

Qualitative Results We visualize our results qualitatively in Fig. 6. We can see that when provided with initial estimates of PoG from InitNet, our Self-Calibration (SC) Module nicely recovers person-specific offsets at test time to yield improved estimates of PoG. Introducing PT Module further boosts the performance by compensating more general yet diverse offsets. By comparing the Ground Truth (GT) with our step-wise results, the mutual beneficial of SC and PT Modules are more observable. Compared to SOTA [20] that requires additional screen contents, our method is more general and is applicable to various datasets.

4.1.2 Ablation Study

To demonstrate the effectiveness of each module we introduced, we further conduct experiments on incrementally adding modules. We validate our ideas on EVE validation set and report the quantitative number in Tab. 2. As can be seen in this table, each module is indeed beneficial for PoG prediction task and combining all of them gives the best performance. Moreover, we are also able to beat the SOTA [20] method that requires both consecutive frames and screen contents as input on EVE validation set. We also perform ablation studies on the impact of history length, and we refer the readers to supplementary for more details.

Performance on EVE validation set [20]			
Method	Gaze Dir. ($^{\circ}$) \downarrow	PoG (cm) \downarrow	PoG (px) \downarrow
[20]	2.1	-	-
InitNet	2.41	2.72	94.33
+SC	2.32	2.66	92.19
+SC+VM	2.26	2.57	89.39
Full(online)	2.04	2.33	80.74
Full(offline)	1.89	2.16	74.85

Table 2: Ablation study on EVE validation set. Each component is beneficial for our task and our method outperforms the SOTA by a large margin.

Method	MPII test set [33]	XGaze test set [30]
	Gaze Dir. ($^{\circ}$) \downarrow	Gaze Dir. ($^{\circ}$) \downarrow
[30]	4.8	4.5
[21]	5.2	-
[9]	4.8	-
[22]	4.5	-
InitNet	5.73/4.83	4.50
Ours-Offline	4.14/3.02	2.88

Table 3: Performances on MPIIGaze and XGaze. Again, we outperform the SOTA significantly.

4.2. Results on MPIIGaze and XGaze

We conduct experiments on two more datasets, MPIIGaze and XGaze. Please note that to have fair comparisons as well as to demonstrate that our method is not restricted to specific network structure, we replace our InitNet structure with [33] and [30] for MPIIGaze and XGaze, respectively. Moreover, to showcase the generality of our proposed method, we train PT Module on neither of these two datasets, but directly applying PT Module trained on EVE and report the performance in Tab. 3. As can be seen in this table, the proposed method, again, outperform the SOTA by a margin, or relatively 32.9% improvement on MPIIGaze and 36.0% improvements on XGaze. Our results suggest that our proposed method can indeed model the person-specific difference very well, even on cross dataset evaluation. More importantly, our proposed method is general enough to improve various existing models and able to achieve best results.

We also visualize our results on MPIIGaze and XGaze in Fig. 7. As can be seen in these figure, our model can handle various type of input and can provide visually satisfactory gaze predictions in both datasets.

5. Conclusion

In this paper, we propose a novel method that aims to improve the cross-person gaze prediction task where only eye images are given as input and Point of Gaze (PoG) is the output. To this end, we introduce Validity Module (VM) to handle noisy data and two other modules, SC and PT,

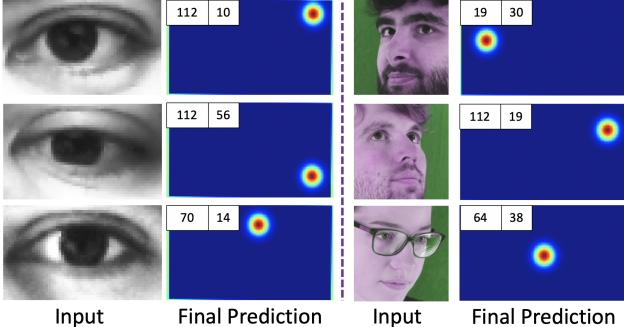


Figure 7: We visualize examples for MPIIGaze and XGaze on the left and right respectively.

to explicitly model person-specific differences in PoG prediction task. We estimate the reliability of each sample in VM and consider only valid ones for later difference modelling process. Our SC considers mainly the translation-like offsets and we further introduce PT to take into account more general and person-specific affine-like differences. We demonstrate the effectiveness of our proposed method on three publicly available datasets and report the SOTA performance. We also showcase the effectiveness and usefulness of each module in our ablation study.

References

- [1] David A Atchison, George Smith, and George Smith. *Optics of the human eye (pp. 30–38)*, volume 2. Butterworth-Heinemann Oxford, 2000. 2
- [2] Hikmet Basmak, Afsun Sahin, Nilgun Yildirim, Thanos D Papakostas, and A John Kanellopoulos. Measurement of angle kappa with synoptophore and orbscan ii in a normal population. *Journal of Refractive Surgery*, 23(5):456–460, 2007. 2
- [3] Esther Berrio, Juan Tabernero, and Pablo Artal. Optical aberrations and alignment of the eye with age. *Journal of vision*, 10(14):34–34, 2010. 2
- [4] Jixu Chen and Qiang Ji. 3d gaze estimation with a single camera without ir illumination. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008. 2
- [5] Jixu Chen and Qiang Ji. Probabilistic gaze estimation without active personal calibration. In *CVPR 2011*, pages 609–616. IEEE, 2011. 3
- [6] E. Chong, Nataniel Ruiz, Y. Wang, Y. Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *ECCV*, 2018. 1
- [7] Stefania Cristina and Kenneth P Camilleri. Model-based head pose-free gaze estimation for assistive communication. *Computer Vision and Image Understanding*, 149:157–170, 2016. 2
- [8] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017. 1
- [9] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gaze: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018. 2, 8
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 3
- [11] Michael Xuelin Huang, Tiffany C.K. Kwok, Grace Ngai, Stephen C.F. Chan, and Hong Va Leong. Building a personalized, auto-calibrating eye tracker from user interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016. 1
- [12] Takahiro Ishikawa. Passive driver gaze tracking with active appearance models. 2004. 2
- [13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koryay kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, 2015. 5
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7
- [15] Kyle Kraftka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016. 2, 3
- [16] Erik Lindén, Jonas Sjostrand, and Alexandre Proutiere. Learning to personalize in appearance-based gaze tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1
- [17] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 3
- [18] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Inferring human gaze from appearance via adaptive linear regression. In *2011 International Conference on Computer Vision*, pages 153–160. IEEE, 2011. 2
- [19] Carlos Hitoshi Morimoto, Arnon Amir, and Myron Flickner. Detecting eye position and gaze from a single camera and 2 light sources. In *Object recognition supported by user interaction for service robots*, volume 4, pages 314–317. IEEE, 2002. 2
- [20] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards end-to-end video-based eye-tracking. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 4, 7, 8
- [21] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377, 2019. 1, 8

- [22] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 721–738, 2018. [2](#), [8](#)
- [23] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–10, 2018. [3](#)
- [24] Yusuke Sugano and Andreas Bulling. Self-calibrating head-mounted eye trackers using egocentric visual saliency. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 363–372, 2015. [3](#)
- [25] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Calibration-free gaze sensing using saliency maps. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2667–2674. IEEE, 2010. [3](#)
- [26] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, 2014. [2](#)
- [27] Erroll Wood and Andreas Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 207–210, 2014. [2](#)
- [28] Yu Yu, Gang Liu, and Jean-Marc Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. [2](#)
- [29] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019. [1](#)
- [30] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020. [2](#), [3](#), [7](#), [8](#)
- [31] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2018. [2](#)
- [32] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015. [2](#)
- [33] 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*, 2017. [2](#), [7](#), [8](#)
- [34] Zhiwei Zhu and Qiang Ji. Eye gaze tracking under natural head movements. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 918–923. IEEE, 2005. [2](#)