Generalizing Eye Tracking with Bayesian Adversarial Learning

Kang Wang Rui Zhao
RPI
{kangwang.kw, zhaorui.zju}@gmail.com

Hui Su
RPI and IBM
huisuibmres@us.ibm.com

Qiang Ji
RPI
qji@ecse.rpi.edu

Abstract

Existing appearance-based gaze estimation approaches with CNN have poor generalization performance. By systematically studying this issue, we identify three major factors: 1) appearance variations; 2) head pose variations and 3) over-fitting issue with point estimation. To improve the generalization performance, we propose to incorporate adversarial learning and Bayesian inference into a unified framework. In particular, we first add an adversarial component into traditional CNN-based gaze estimator so that we can learn features that are gaze-responsive but can generalize to appearance and pose variations. Next, we extend the point-estimation based deterministic model to a Bayesian framework so that gaze estimation can be performed using all parameters instead of only one set of parameters. Besides improved performance on several benchmark datasets, the proposed method also enables online adaptation of the model to new subjects/environments, demonstrating the potential usage for practical real-time eye tracking applications.

1. Introduction

Eye gaze represents human's focus of attention or interests. The eye gaze for ourselves can help us better understand the visual world, and help us better interact with computers or large systems [1, 2, 3]. Furthermore, eye gaze also plays a crucial rule in understanding human's cognitive and emotional status, which have been used for marketing and advertising [4], social network [5, 7, 8, 6], web search [9, 11, 10], psychology study and medical research [12], etc.

Various techniques have been proposed to estimate eye gaze. Model-based methods [13, 14, 15, 16, 17, 18, 19, 20] rely on a geometric eye model to estimate eye gaze. The idea is to represent 3D eye gaze to two 3D points and their goal is to recover the 3D points. Despite their simplicity and good accuracy, the system is sensitive to key point detections and may not work in outdoor environments. Early appearance-based methods [21, 22, 23] try to extract hand-crafted features from eye images and map the features to eye gaze. However, they cannot handle large head poses and are restricted to controlled environments.

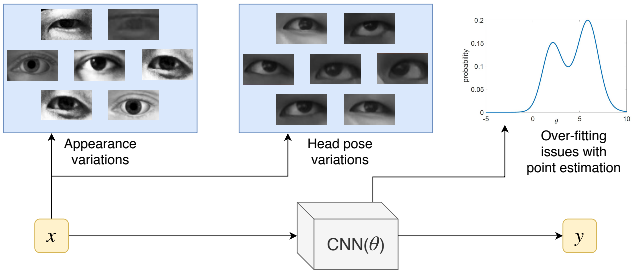

Figure 1. Three factors that affect the generalization performance of appearance-based gaze estimation methods.

More recently, appearance-based methods [24, 25, 26, 27, 28] with deep learning [29, 30, 31, 32] are the dominant approaches because of their improved performance over traditional model-based/appearance-based methods. However, researchers also begin challenging the generalization performance of deep learning-based approaches, since the trained model may totally fail for an unseen subject or in a new environment. This significantly limits the usage of appearance-based methods in practical eye tracking systems.

In this work, we study the following problem. Suppose we have a gaze estimator trained with data from a source domain, how can we generalize this gaze estimator to a target domain with few labeled data or no labeled data? we systematically study the factors that affect the generalization performance, and identify three major factors as in Fig. 1.

The first factor is the appearance variation, which is resulted from different combinations of illumination, skin color, eye texture, eye shape, imaging condition, glasses, etc. The example images in Fig. 1 come from different subjects/datasets with close to frontal eye gaze directions. It is difficult to model these individual factors separate, we therefore only model the coupled appearance variations.

The second factor is the head pose variation. Fig. 1 shows the images from the same subject looking at the same target but with different head poses. The head pose variations may not be obvious in the example images because we cut the eye images, however we can get a sense of head pose variations from the image brightness, shadows as well as the pupil positions. Although we can treat head pose variations as part of appearance variations, we would like to model them separately. The underlying reason is that head pose is resulted from geometric rotation and motion which have good analytical formulations. Compared to modeling appearance and pose variation together in a coupled way, we can benefit from a separate modeling.

The last factor is the over-fitting issue with point estimation. Traditional CNNs only estimate one optimal set of parameters, which work well for data with less variations. However, for practical environments with large variations as in Fig. 1, they may not work well since the parameter posterior is much more complex.

To deal with the three factors, we introduce a Bayesian adversarial learning approach. Our overall network is built on top of a traditional CNN that map eye image to eye gaze. Inspired by recent work on domain adaptation [33, 34], we first introduce an adversarial learning block, which is responsible for learning good features for eye tracking but can also generalize to appearance and head pose variations. The idea is to learn features that cannot discriminate the variations through a minmax objective. To handle the over-fitting issue resulted from point estimation, we extend the CNN to Bayesian Convolutional Neural Network (BCNN), where we can perform gaze estimation with multiple sets of parameters from the parameter posterior and hence improve the generalization. To summarize, we make the following novel contributions:

- We identify three major factors that affect the generalization performance of appearance-based gaze estimators and propose a Bayesian adversarial learning approach to deal with the three factors in a unified framework.

- We propose an adversarial learning approach which learns features that can handle appearance and head pose variations by combining appearance and model-based adversarial loss functions.

- We introduce a Bayesian framework that alleviates the over-fitting issues from point estimation and hence further improves the generalization.

2. Related work

2.1. gaze estimation

We focus on recent appearance-based methods with deep learning. In [24], the authors propose to map eye image to eye gaze with a LeNet architecture. To better handle head pose variations, they append the predicted head pose to the extracted feature vector to jointly estimate eye gaze. The authors in [25] propose a 4-pathway network to incorporate left, right eye images, face images and face location information to jointly estimate the eye gaze. In [26], the authors first decouple the eye gaze to eye pose and head pose. Then they use two CNN networks to estimate eye pose and head pose, which are then directly mapped to eye gaze with an analytical formulation. In [35], the authors propose to map the eye appearance to an intermediate gaze map and then map the gaze map to the final gaze. They argue that the two-step strategy is easier to learn than end-to-end models and therefore gives better accuracy. There are also hybrid-models [36, 37] that use CNN to map image to eye landmarks and then map eye landmarks to eye gaze. All these approaches implicitly or explicitly embed the head pose information to improve the generalization performance. However, their methods can only work in certain extent as the underlying CNN cannot capture all the variations in the image space, and their models only rely on one single set of parameters which are prone to over-fitting issues.

2.2. Domain adaptation

Because of dataset bias or domain shift, models trained on one dataset may fail on new datasets. Different domain adaptation techniques are proposed to reduce the effects of domain shift. Some of them learn the feature representations that can reduce domain shift in terms of maximum mean discrepancy [38], or correlation distance [39]. Recently, the adversarial learning [40] idea is employed to minimize the domain discrepancy through an adversarial objective [33, 34, 41]. By maximumly confusing the domain classifier, the learned feature representations can better generalize to both domains. Existing work on domain adaptation is designed to work for general tasks, and ignores domain knowledge for specific tasks. In this work, we incorporate the head pose knowledge for eye gaze and formulate them in a unified adversarial learning framework, and demonstrate better generalization.

2.3. Bayesian neural network

Bayesian neural network (BNN) [42] is a probabilistic interpretation of deep models by modeling the posterior distribution of the model parameters. BNNs avoid point estimation and provide robustness against over-fitting, which is crucial to generalize the learned model from the source domain to the target domain. However, inference in BNN is difficult because of the integration over the parameter space. Early attempts include the Laplace's method [43] and variational approaches [44], but the approximation error is large and the computational complexity remains large. Modern inference techniques include improved variational approaches [45, 46], Hamiltonian Monte Carlo based approaches and their variants [42, 47]. With these techniques, we can achieve better efficiency and scale up to large-scale datasets.

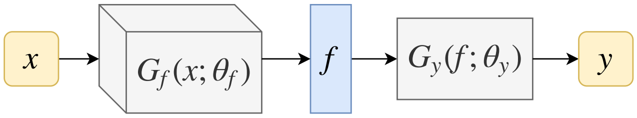

Figure 2. Illustration of a standard appearance-based gaze estimator.

3. Problem statement

Before discussing the proposed approach, we first introduce the baseline gaze estimator and our problem scenario.

Baseline gaze estimator.

We use a standard appearance-based gaze estimator (Fig. 2) as our baseline:

$$\mathbf{f} = G_f(\mathbf{x};\theta_f) \quad \text{and} \quad \mathbf{y} = G_y(\mathbf{f};\theta_y),$$

where $G_f(\cdot)$ is the feature extractor with parameter θ_f, $G_y(\cdot)$ is the gaze estimator with parameter θ_y, and $\mathbf{f}$ is the learned feature representations.

Problem scenario.

Suppose we have learned a baseline gaze estimator $\theta^s = \{\theta^s_f, \theta^s_y\}$ with data $\mathcal{D}_s = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n_s}$ from the source domain. This model can perform well on test data from a similar domain/distribution as $\mathcal{D}_s$, but may not generalize to a new domain/distribution. Formally, assume we have data $\mathcal{D}_t = \{\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n'_t}, \{\mathbf{x}_i\}_{i=1}^{n_t}\}$ ($n'_t \ll n_t$) from the target domain (Eg. new subjects, head poses or environments), we want to explore how we can adapt θ^s so that we can achieve good performance on data from $\mathcal{D}_t$. In this work, we are interested in both semi-supervised case and unsupervised case ($n'_t = 0$).

Next, we first discuss the proposed adversarial learning method in Sec. 4.1, then we introduce the Bayesian extension in Sec. 4.2.

4. Proposed approach

4.1. Adversarial learning

Our goal is to adapt the source model θ^s to a target model $\theta^t = \{\theta^t_f, \theta^t_y\}$ so that we can estimate gaze accurately on $\mathcal{D}_t$. To this end, we design a specific network as shown in Fig. 3. We introduce two additional classifiers compared to Fig. 2. The extracted features $\mathbf{f}$ are fed to three models:

- gaze estimator $G_y(\mathbf{f}, \theta_y)$: the output is the continuous eye gaze $\mathbf{y} \in \mathcal{R}^2$, $\mathbf{y}$ can represent the x and y coordinates on the screen or the pitch and yaw angles in 3D space.
- appearance classifier $G_a(\mathbf{f}, \theta_a)$: the output is a scalar probability $a \in [0, 1]$ indicating the probability of the input coming from the source domain $\mathcal{D}_s$.
- head pose classifier $G_h(\mathbf{f}, \theta_h)$: the output is a probability vector $\mathbf{h} = \{p_1, ..., p_k\}$ indicating the probability of each of the k head pose classes.

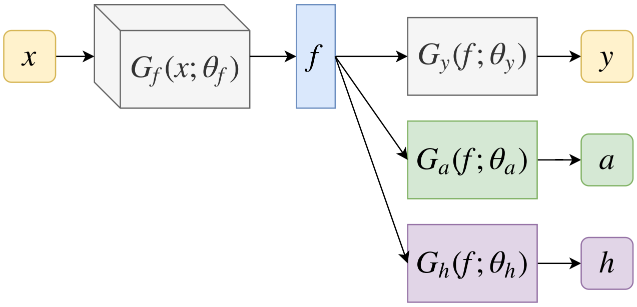

Figure 3. Illustration of the proposed adversarial learning method.

The loss function for the gaze estimator is defined as:

$$\mathcal{L}_y(\theta_f, \theta_y) = \frac{1}{n'_t} \sum_{i=1}^{n'_t} ||G_y(G_f(\mathbf{x}_i;\theta_f);\theta_y) - \mathbf{y}_i||^2 \tag{1}$$

For the appearance classifier, its goal is to differentiate images from source domain $\mathcal{D}_s$ or target domain $\mathcal{D}_t$, the loss function is defined as the binary cross-entropy:

$$\begin{aligned}\mathcal{L}_a(\theta_f, \theta_a) = &-\frac{1}{n_t}\sum_{i=1}^{n_t} \log(1 - G_a(G_f(\mathbf{x}_i;\theta_f);\theta_a)) \\ &-\frac{1}{n_s}\sum_{i=1}^{n_s} \log(G_a(G_f(\mathbf{x}_i;\theta_f);\theta_a))\end{aligned} \tag{2}$$

For the head pose classifier, its goal is to differentiate images with different head poses, the loss is defined as the multi-class cross-entropy:

$$\mathcal{L}_h(\theta_f, \theta_h) = -\frac{1}{n_t + n_s} \sum_{i=1}^{n_t+n_s} h_{i,j} \sum_{j=1}^{k} \log(G_h(G_f(\mathbf{x}_i;\theta_f);\theta_h) \tag{3}$$

where $h_{i,j}$ is the groundtruth probability for i-th image and j-th pose class.

There are 4 different sets of parameters, the learning of $\{\theta_y, \theta_a, \theta_h\}$ is easy because they only depend on θ_f. To this end, we can solve them given θ_f:

$$\hat{\theta_y} = \arg\min_{\theta_y} \mathcal{L}_y(\hat{\theta_f}, \theta_y) \tag{4}$$

$$\hat{\theta_a} = \arg\min_{\theta_a} \mathcal{L}_a(\hat{\theta_f}, \theta_a) \tag{5}$$

$$\hat{\theta_h} = \arg\min_{\theta_h} \mathcal{L}_h(\hat{\theta_f}, \theta_h) \tag{6}$$

The learning of θ_f is relatively difficult (depend on $\{\theta_y, \theta_a, \theta_h\}$) but is the key of our adversarial learning. Notice we want the learned features to produce small gaze estimation error but confuse appearance and pose classifiers. To this end, we have the following objective:

$$\hat{\theta_f} = \arg\min_{\theta_f} \mathcal{L}_y(\theta_f, \hat{\theta_y}) - \lambda_a \mathcal{L}_a(\theta_f, \hat{\theta_a}) - \lambda_h \mathcal{L}_h(\theta_f, \hat{\theta_h}) \tag{7}$$

where λ_a and λ_h are two positive balancing factors. The negative sign before the appearance and pose terms allows us to minimize them together with the gaze regression loss term.

Note the objective in Eq. (7) corresponds to the true minimax objective. Compared to Eq. (5) and (6), the only difference is the sign before the appearance and pose classifiers. We are actually optimize the same objective (different parameters) to opposite directions. However as [40, 34] point out, the $\log(1 - G_a(G_f(\mathbf{x}_i; \theta_f); \theta_a)$ term in Eq. (2) may be problematic and causes vanishing gradient when we minimize Eq. (7). We instead use the following new objective $\mathcal{L}_f(\theta_f, \hat{\theta_y}, \hat{\theta_a}, \hat{\theta_h})$ to solve θ_f:

$$\begin{aligned}\hat{\theta_f} &= \arg\min_{\theta_f} \mathcal{L}_f(\theta_f, \hat{\theta_y}, \hat{\theta_a}, \hat{\theta_h}) \\ &= \arg\min_{\theta_f} \mathcal{L}_y(\theta_f, \hat{\theta_y}) - \lambda_h \mathcal{L}_h(\theta_f, \hat{\theta_h}) \\ &+ \lambda_a \frac{1}{n_t} \sum_{i=1}^{n_t} \log G_a(G_f(\mathbf{x}_i; \theta_f); \hat{\theta_a})\end{aligned} \tag{8}$$

Eq. (8) and Eq. (7) has the same fixed-point properties but Eq. (8) can produce stronger gradients and improve the optimization.

Finally, we summarize the adversarial parameter learning algorithm in Alg. 1. After convergence, we discard the appearance and pose classifier parameters and only use θ_f^t and θ_y^t for our gaze estimation task.

4.1.1 Discussions

Motivation of head pose classifier and how to obtain head pose label. Existing domain adaptation approaches only consider the appearance adaptation. For our specific gaze estimation task, the target gaze label is a geometric entity and are strongly correlated with geometric features (Eg. facial/eye landmarks). In fact, there exists plenty of work on model-based / feature-based gaze estimation techniques. Inspired by this, we propose to explicitly embed the geometric dependence in the feature-learning process. However, it is difficult to analytically relate the eye gaze to geometric features (facial landmarks), we instead use head pose as an intermediate representation. For all training images, we perform offline detection of the landmarks $\mathbf{c}$ [48], then we can relate head $\{\mathbf{M}, \mathbf{t}\}$ pose with observed landmarks using a 3D shape model $\mathbf{S}$ [49, 50]:

$$\mathbf{c} = M\mathbf{S} + \mathbf{t} \tag{9}$$

By minimizing the projection error, we are able to recover the head pose, which is further quantized to k discrete pose classes. By using the head pose estimated from model-based methods, we implicitly encourage learning features that are not sensitive to geometric variations.

Algorithm 1: Adversarial parameter learning

```
1. Input: Source domain data $\mathcal{D}_s = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n_s}$, target
domain data $\mathcal{D}_t = \{\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n_t'}, \{\mathbf{x}_i\}_{i=1}^{n_t}\}$, source
model $\theta^s = \{\theta_f^s, \theta_y^s\}$.
2. Output: Target model $\theta^t = \{\theta_f^t, \theta_y^t\}$.
3. Initialization: $\theta_f^t = \theta_f^s$, $\theta_y^t = \theta_y^s$, $\theta_a^t = \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$,
$\theta_h^t = \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$, total iterations $T$.
4. for $iter \in \{1, ..., T\}$ do
    - Sample a batch of data from source and target:
      $\mathbf{x}^s \sim \mathcal{D}_s$, $\{\mathbf{x}^t, \{\mathbf{x}^{t\prime}, \mathbf{y}^{t\prime}\}\} \sim \mathcal{D}_t$.
    - Update $\theta_y^t$ with $\{\mathbf{x}^{t\prime}, \mathbf{y}^{t\prime}\}$ (Eq. (4)):
      $\theta_y^t \leftarrow \theta_y^t - \alpha\partial\mathcal{L}_y(\hat{\theta_f^t}, \theta_y^t)/\partial\theta_y^t$
    - Update $\theta_a^t$ with $\mathbf{x}^s$ and $\mathbf{x}^t$ (Eq. (5)):
      $\theta_a^t \leftarrow \theta_a^t - \alpha\partial\mathcal{L}_a(\hat{\theta_f^t}, \theta_a^t)/\partial\theta_a^t$
    - Update $\theta_h^t$ with $\mathbf{x}^s$ and $\mathbf{x}^t$ and their corresponding
      pose labels (Eq. (6)):
      $\theta_h^t \leftarrow \theta_h^t - \alpha\partial\mathcal{L}_h(\hat{\theta_f^t}, \theta_h^t)/\partial\theta_h^t$
    - Update $\theta_f^t$ with all data and other updated
      parameters (Eq. (8)):
      $\theta_f^t \leftarrow \theta_f^t - \alpha\partial\mathcal{L}_f(\theta_f, \hat{\theta_y}, \hat{\theta_a}, \hat{\theta_h})/\partial\theta_f^t$
      (Note for unsupervised learning, we discard the
      first $\mathcal{L}_y(\theta_f, \hat{\theta_y})$ term in Eq. (8) and optimize the
      rest two terms.)
```

4.2. Bayesian formulation

To alleviate the potential over-fitting issues with point estimation, we extend the deterministic model to a probabilistic Bayesian model. With Bayesian framework, gaze estimation for a new image $\mathbf{x}_t$ can be formulated as follows:

$$\begin{aligned}\mathbf{y}_t &= \arg\max_{\mathbf{y}_t} p(\mathbf{y}_t | \mathbf{x}_t, \mathcal{D}, \boldsymbol{\alpha}) \qquad (10)\\ &= \arg\max_{\mathbf{y}_t} \int_{\theta^t} p(\mathbf{y}_t|\theta^t) p(\theta^t|\mathcal{D}, \boldsymbol{\alpha}) d\theta^t \\ &\approx \arg\max_{\mathbf{y}_t} \sum_{i=1}^{m} p(\mathbf{y}_t|\theta^t[i]) \quad \text{where} \quad \theta^t[i] \sim p(\theta^t|\mathcal{D}, \boldsymbol{\alpha}) \\ & \qquad (11) \\ &\approx \frac{1}{m} \sum_{i=1}^{m} G_y(G_f(\mathbf{x}_t; \theta_f^t[i]); \theta_y^t[i])\end{aligned}$$

where $\mathcal{D} = \{\mathcal{D}_s, \mathcal{D}_t\}$, and $\boldsymbol{\alpha}$ is the prior for θ^t. Instead of performing a point estimation to estimate one optimal set of parameters, we perform Bayesian inference to obtain multiple sets of parameters drawn from its posterior. Gaze estimation is based on the average of multiple predictions and hence can improve the generalization. The extended Bayesian framework uses the same architecture as in Fig. 3, but now the network parameters $\{\theta_f^t, \theta_y^t, \theta_a^t, \theta_h^t\}$ are assumed to follow a probabilistic distribution. As in Eq. (11), the key to performing Bayesian inference is to effectively draw samples from the posterior distributions. It is difficult to draw $\{\theta_f^t, \theta_y^t, \theta_a^t, \theta_h^t\}$ all at once, we follow the idea in [51] to draw the 4 set of parameters alternately until final convergence. To draw samples alternately, we need to define the conditional posterior of the parameter given all other parameters, this will be discussed in Sec. 4.2.1. After that, we briefly introduce the algorithm to effectively draw samples from the posterior distributions (Sec. 4.2.2).

4.2.1 Construction of posterior distribution

We first assume the parameters follow a Gaussian prior distribution:

$$p(\theta_i^t|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{0}, \sigma\mathbf{I}), \quad \forall i \in \{f, y, a, t\} \tag{12}$$

where σ is the standard deviation. Next, we can construct the posterior by combining the likelihood models with the prior models. From the discussion in Sec. 4.1, we learn the 4 type of parameters alternatively, here we follow the same idea by constructing the conditional posterior given other parameters.

First, for the gaze branch, we assume the output eye gaze follows a Gaussian distribution:

$$p(\mathbf{y}|\mathbf{x}, \theta_f^t, \theta_y^t) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}(\mathbf{x}, \theta_f^t, \theta_y^t), \boldsymbol{\Sigma}(\mathbf{x}, \theta_f^t, \theta_y^t)) \tag{13}$$

where $\boldsymbol{\mu}(\mathbf{x}, \theta_f^t, \theta_y^t)$ represents the mean and $\boldsymbol{\Sigma}(\mathbf{x}, \theta_f^t, \theta_y^t)$ represents the covariance. In this work, covariance is assumed to be a diagonal matrix. To predict mean and covariance, we modify the gaze branch in Fig. 3 to output a 4-dimensional vector where the first 2 dimensions represent the mean and the last 2 dimensions represent the diagonal entries. The conditional posterior therefore follows:

$$\begin{aligned} & p(\theta_y^t|\theta_f^t, \theta_a^t, \theta_h^t) = p(\theta_y^t|\theta_f^t) \propto \\ & \prod_{i=1}^{n_t'} \mathcal{N}(\mathbf{y}_i; G_y^1(G_f(\mathbf{x}_i; \theta_f^t); \theta_y^t), G_y^2(G_f(\mathbf{x}_i; \theta_f^t); \theta_y^t)) p(\theta_y^t) \end{aligned} \tag{14}$$

where $G_y^1(\cdot)$ represents the first 2-dimension of the output (mean) and $G_y^2(\cdot)$ represents the last 2-dimension of the output (covariance). Intuitively, θ_y^t that yields good predictions (close to the groundtruth) should have larger probabilities.

Second, for the appearance branch, the conditional posterior follows:

$$\begin{aligned} & p(\theta_a^t|\theta_f^t, \theta_y^t, \theta_h^t) = p(\theta_a^t|\theta_f^t) \propto \\ & \prod_{i=1}^{n_t}(1 - G_a(G_f(\mathbf{x}_i; \theta_f^t); \theta_a^t)) \prod_{i=1}^{n_s} G_a(G_f(\mathbf{x}_i; \theta_f^t); \theta_a^t) p(\theta_a^t) \end{aligned} \tag{15}$$

If θ_a^t produces large probabilities (close to 1) for source data, while low probabilities (close to 0) for target data, then θ_a^t and its neighborhood should have large posterior probabilities.

```
Algorithm 2: Bayesian adversarial learning
1. Input: Source domain data $\mathcal{D}_s = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n_s}$, target
domain data $\mathcal{D}_t = \{\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n_t'}, \{\mathbf{x}_i\}_{i=1}^{n_t}\}$, source
model $\theta^s = \{\theta_f^s, \theta_y^s\}$.
2. Output: $m$ target model samples $\{\theta_i^t\}_{i=1}^m$.
3. Initialization: $\theta_f^t = \theta_f^s$, $\theta_y^t = \theta_y^s$, $\theta_a^t = \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$,
$\theta_h^t = \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$, burn in time $T$, collection interval $b$.
4. for $iter \in \{1, ..., T + m * b\}$ do
    - Sample a batch of data from source and target:
    $\mathbf{x}^s \sim \mathcal{D}_s$, $\{\mathbf{x}^t, \{\mathbf{x}^{t\prime}, \mathbf{y}^{t\prime}\}\} \sim \mathcal{D}_t$.
    - Sample $\theta_y^t$: $\theta_y^t \leftarrow \theta_y^t + \mathbf{v}_y$
    $\mathbf{v}_y \leftarrow (1 - \alpha)\mathbf{v}_y + \eta \frac{\partial \log p(\theta_y^t|\theta_f^t)}{\partial \theta_y^t} + \mathcal{N}(\mathbf{0}, 2\alpha\eta\mathbf{I})$
    - Sample $\theta_a^t$: $\theta_a^t \leftarrow \theta_a^t + \mathbf{v}_a$
    $\mathbf{v}_a \leftarrow (1 - \alpha)\mathbf{v}_a + \eta \frac{\partial \log p(\theta_a^t|\theta_f^t)}{\partial \theta_a^t} + \mathcal{N}(\mathbf{0}, 2\alpha\eta\mathbf{I})$
    - Sample $\theta_h^t$: $\theta_h^t \leftarrow \theta_h^t + \mathbf{v}_h$
    $\mathbf{v}_h \leftarrow (1 - \alpha)\mathbf{v}_h + \eta \frac{\partial \log p(\theta_h^t|\theta_f^t)}{\partial \theta_h^t} + \mathcal{N}(\mathbf{0}, 2\alpha\eta\mathbf{I})$
    - Sample $\theta_f^t$: $\theta_f^t \leftarrow \theta_f^t + \mathbf{v}_f$
    $\mathbf{v}_f \leftarrow$
    $(1 - \alpha)\mathbf{v}_f + \eta \frac{\partial \log p(\theta_f^t|\theta_y^t, \theta_a^t, \theta_h^t)}{\partial \theta_f^t} + \mathcal{N}(\mathbf{0}, 2\alpha\eta\mathbf{I})$
    - Collect sample $\{\theta_f^t, \theta_y^t\}$ every b iterations after
    burn in time.
```

Third, the conditional posterior for the head pose branch follows:

$$\begin{aligned} & p(\theta_h^t|\theta_f^t, \theta_y^t, \theta_h^t) = p(\theta_h^t|\theta_f^t) \propto \\ & \prod_{i=1}^{n_t+n_s} \prod_{j=1}^{k} G_h^j(G_f(\mathbf{x}_i; \theta_f^t); \theta_a^t)^{h_{i,j}} p(\theta_h^t) \end{aligned} \tag{16}$$

where $G_h^j(\cdot)$ represents the j-th element of the output of head pose branch. Similarly, θ_h^t should have large posterior probabilities if it produces correct pose classifications.

Finally, analogous to Eq. (8), we modify the appearance term to avoid vanishing gradient and the conditional posterior for θ_f^t follows:

$$\begin{aligned} & p(\theta_f^t|\theta_y^t, \theta_a^t, \theta_h^t) \propto \\ & \underbrace{p(\theta_y^t|\theta_f^t)}_{\text{gaze}} \underbrace{\prod_{i=1}^{n_t} G_a(G_f(\mathbf{x}_i; \theta_f^t); \theta_a^t)}_{\text{appearance}} \underbrace{(-p(\theta_h^t|\theta_f^t))}_{\text{head pose}} \underbrace{p(\theta_f^t)}_{\text{prior}} \end{aligned} \tag{17}$$

The conditional posterior in Eq. (17) tells when θ_f^t should have large probabilities: 1) the gaze term indicates θ_f^t should produce small gaze prediction error; 2) the appearance term regulates θ_f^t to produce large probability for data from target domain (confuse data from source and target domain); 3) head pose term, similarly maximumly confuse the pose classifier; 4) the prior term incorporates our prior knowledge about the parameter space. These four terms jointly contribute to the posterior distribution of θ_f^t, allowing us to obtain good samples that give good gaze estimation error while also improves the generalization performance.

4.2.2 Bayesian inference

Computing the posterior analytically is challenging, we instead employ the Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) [47, 51] to approximate the posterior. S-GHMC is an extension of HMC which supports mini-batch update. As a result, it can scale-up to large datasets and allow us to draw samples effectively. We leave the details of SGHMC for readers' own interest and only summarize the overall approximation algorithm in Alg. 2.

5. Experiments and Analysis

We evaluate the proposed method on four benchmark datasets: 1) MPIIGaze [24], which consists of data from 15 subjects in different environments; 2) UT [22], consists of 50 subjects, each with 8 head poses and 160 gaze directions; 3) Columbia [52], with 56 subjects and 5 head poses; and 4) EyeDiap [53], consists of data from HD/VGA camera, discrete and continuous targets and different head poses. Different approaches use different subsets of the data, we follow the same setting as [35] for the evaluation.

MPIIGaze and EyeDiap have continuous head pose angles, we follow the settings in MPIIGaze dataset to normalize head pose into 2 angles (2D region), then we manually set threshold of the two angles to divide the 2D region into 8 sub-regions (with approximately similar amount of data for each sub-region). UT and Columbia have different number of cameras with fixed head position, the number of head pose classes is equal to the number of cameras.

Our model input is eye image of size 36 × 60. Here is the summary of the architecture in Fig. 3: 1) $G_f(\mathbf{x}, \theta_f)$ (Conv(5, 5, 64), LeakyRelu(0.2), MaxPooling(2), Conv(5, 5, 32), LeakyRelu(0.2), MaxPooling(2), FC(128); 2) $G_y(\mathbf{f}, \theta_y)$ (FC(128), LeakyRelu(0.2), FC(2)); 3) $G_a(\mathbf{f}, \theta_a)$ (FC(500), LeakyRelu(0.2), FC(256), LeakyRelu(0.2), FC(1), Sigmoid); 4) $G_h(\mathbf{f}, \theta_h)$ (FC(500), LeakyRelu(0.2), FC(256), LeakyRelu(0.2), FC(k), Softmax). Notice we use a relative simple model compared to existing work with complex architectures.

For Bayesian inference, we need to modify the last layer of $G_y(\mathbf{f}, \theta_y)$ to output a 4-dimensional vector while other layers remain the same. The prior in Eq. (12) is set with $\sigma = 0.01$. For the inference in Alg. 2, we collect one sample every 64 iterations and use a total 100 samples to perform gaze estimation. With a Tesla M40 GPU, inference using one sample takes around 5ms.

5.1. Ablation study

We first perform a systematic study to evaluate different model components in Sec. 5.1.1 (unsupervised setting with no labeled data), then we study how number of annotated samples affect the model performance in Sec. 5.1.2.

5.1.1 Evaluation of different model components

We consider following 4 models:

- baseline: a standard CNN-based gaze estimator.
- baseline + appearance classifier: adding appearance classifier.
- baseline + appearance + pose classifiers: further incorporate head pose classifier.
- baseline + appearance + pose classifiers + Bayesian inference: perform Bayesian inference.

For each model, we consider 3 types of evaluations: 1) cross-subject; 2) cross-pose and 3) cross-dataset. For cross-subject evaluations, we perform 4-fold cross-validation by dividing all subjects into 4 clusters randomly. For cross-pose experiments, we perform 4-fold cross-validation for MPIIGaze, UT and EyeDiap. Their 8 head poses are divided into 4 clusters by grouping neighboring poses into the same cluster. For Columbia, we perform 5-fold cross-validation.

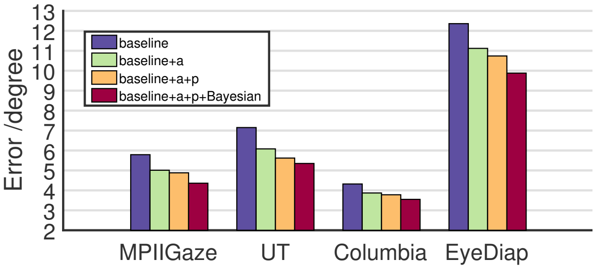

Figure 4. Cross-subject evaluations.

The cross-subject evaluations are shown in Fig. 4. We can see for the 4 datasets, adding appearance classifier shows a significant improvement over the baseline gaze estimator. First, the appearance variations are the most dominant variations, by learning features that cannot distinguish the variations, we can therefore achieve a large improvement. When we add the pose classifier, we can observe further improvement. The improvement is not as significant as appearance classifier, because head pose variations are also reflected by the underlying appearance change (handled by appearance classifier). And this is a cross-subject experiment, the head pose distributions for source and target domain appear similar. Explicitly using pose classifier is therefore less useful. Finally, when we perform Bayesian inference, we observe consistent improvement for the 4 datasets. Note the improvements for the 4 datasets are different. Bayesian inference gives a large improvement for EyeDiap (large variations and low-qualities), and a smaller improvement on Columbia because of its high-quality image conditions. The experiments demonstrate the contributions for each of the components in the proposed method.

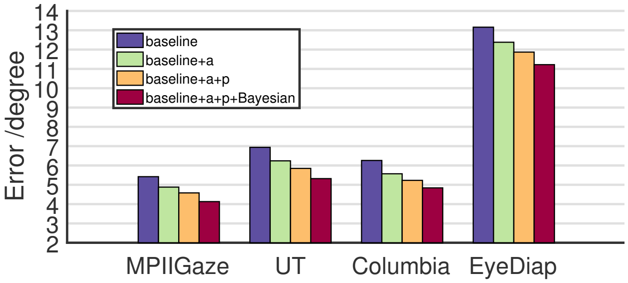

Figure 5. Cross-pose evaluations.

Next, we perform cross-pose evaluation as in Fig. 5. First, we observe consistent improvements for each model components on the 4 datasets. Second, compared to cross-subject experiments, the head pose classifier shows a more important role in cross-pose experiments, as the pose distributions for source and target are different. By explicitly force the model to learn features that are not pose-sensitive, we can obtain a larger improvement.

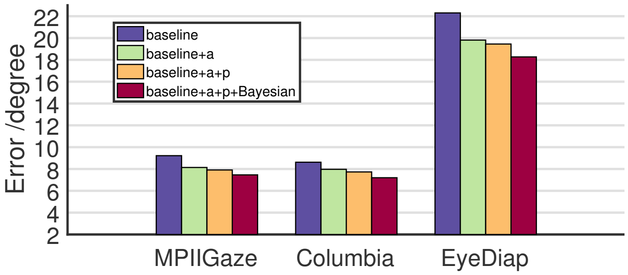

Figure 6. Cross-dataset evaluations with UT as source domain data.

The cross-dataset experiments are shown in Fig. 6. We observe similar patterns and all components contribute to the improved performance.

Table 1. Average improvement over baseline models.

	Cross-subject	Cross-pose	Cross-dataset
(a)	12.2%	9.3%	10.1%
(a, p)	15.6%	14.4%	12.4%
(a,p,B)	21.9%	21.1%	17.9%

Finally, we show the quantitative improvement over the baseline model in Tab. 1. The improvements are averaged over all datasets. From the results, we conclude that appearance classifier contributes most to the improvement, Bayesian inference demonstrates a mid-level role while pose classifier shows a relatively smaller improvement. But if source domain and target domain has different pose distributions, the pose classifier can play an important role since the basic appearance classifier cannot fully capture variations caused by geometric motions. In addition, different from appearance and pose classifiers, which address the generalization issue from data-variation perspective, the proposed Bayesian framework address the generalization issue from the model perspective. By introducing Bayesian inference instead of point-estimation, the underlying model yields better generalization capabilities.

5.1.2 Evaluation of number of labeled data

The previous study is conducted based on an unsupervised setting, we are also interested in a semi-supervised setting. We use UT dataset for our evaluation. We use $32,000$ images for source domain data and the rest $32,000$ images for target domain data. Next we random draw $k\%$ ($k \in \{0,1,2,5,10\}$) of the target domain data as labeled data, and the rest as the unlabeled (testing) data. We repeat the process 5 times and report the average performance. For a fair comparison, we also perform fine-tuning on the baseline models using the labeled data and compare with the proposed method.

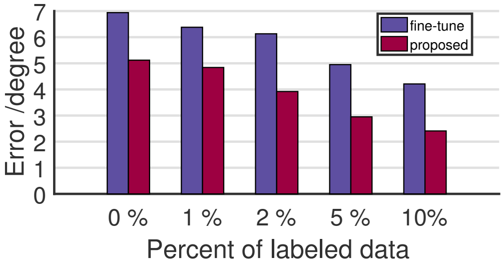

Figure 7. Cross-subject evaluation on UT.

The cross-subject evaluation is shown in Fig. 7, and the cross-pose evaluation is shown in Fig. 8. We can observe that with more labeled data, both the fine-tuned model and the proposed model can keep reducing the gaze estimation error, but the proposed method can always give better accuracy than the fine-tuned model. This demonstrate that the proposed approach can handle both unsupervised and semi-supervised scenarios.

5.1.3 Online eye tracking

The proposed method can serve as an online model adaptation technique for a real-time eye tracking system. Suppose

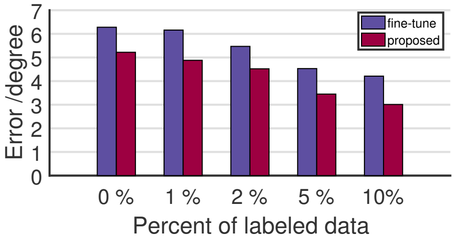

Figure 8. Cross-pose evaluation on UT.

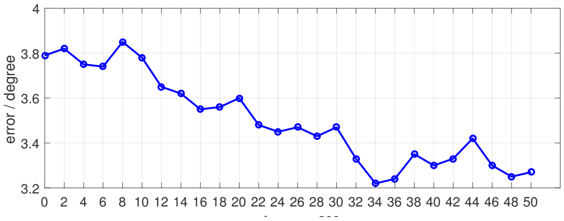

Figure 9. Online eye tracking performance with the proposed model adaptation approach.

we have a real-time eye tracking system trained with source domain data. A new subject tries to use the system but is not satisfied with the performance. In this case, we can use the proposed method to gradually adapt the model parameters so that it can produce good results for the new subject. In particular, we ask the new subject to use the system for some time and collect raw eye images. These raw eye images serve as unlabeled target domain data and can be used to adapt the model parameters with the proposed Bayesian adversarial learning. We conduct a simple experiment in the lab. The baseline gaze estimator is trained with data from 10 people, and we ask the new subject to use the system for some time. We also collect some labeled data for the new subject for testing. Fig. 9 shows the gaze estimation error as a function of time (we use the first T frames to adapt the baseline model). The results suggest that as we use the system and collect more data, we can gradually adapt to a new subject and improve the gaze estimation performance.

5.2. Comparison with State-of-the-art

Next we compare with state-of-the-art approaches on cross-subject and cross-dataset experiments.

Table 2. Cross-subject evaluation.

	[54]	[36]	[35]	Proposed
EyeDiap	-	11.9	10.3	9.9
UT	-	-	-	5.4
MPII	6.3	-	4.5	4.3

We first perform cross-subject experiments. In [54], the authors combined head pose feature and appearance feature to perform gaze estimation. They did not model the appearance variations and use a simple feature concatenation technique to incorporate head pose information. On the contrary, we use a similar baseline model, but explicitly consider appearance variations and incorporate pose information in a adversarial way. And combined with the Bayesian framework, we achieve an improvement around 2.0 degrees.

In [36], the authors first use a hourglass model to map eye appearance to eye landmarks and then use either feature-based or model-based method to map landmarks to eye gaze. In [35], the authors propose to map appearance to a gaze map then estimate gaze from gaze map. Both methods use a much more complex architectures than ours, but we still outperform them, demonstrating the effectiveness of the proposed Bayesian adversarial learning framework.

Table 3. Cross-dataset evaluation.

	[54]	[55]	[37]	[36]	Proposed
EyeDiap	-	-	-	26.6	18.3
MPII	13.9	8.9	7.7	8.7	7.4

We further perform cross-dataset experiments by using UT dataset as source domain data. Fig. 3 shows the results with unsupervised setting. We outperform all competing approaches on MPII dataset. Even with a relative small scale model, the proposed approach can still achieve better results. When evaluated on EyeDiap, we outperform [36] with a big margin. The reason is that the distribution of UT and EyeDiap differs significantly, minimizing the domain shift between them leads to a large improvement in the gaze estimation accuracy. In addition, EyeDiap has large variations which leads to complex parameter posterior distributions, using Bayesian inference is more effective in these cases which explains the large improvement.

6. Conclusion

In this paper, we systematically study the generalization issue of appearance-based gaze estimation methods. We identify three major factors: 1) appearance variations; 2) pose variations and 3) over-fitting issue with point estimation. By introducing an adversarial learning approach, we are able to learn better feature representations that can generalize to appearance and pose variations. With the extended Bayesian framework, we alleviate the over-fitting issue by using multiple sets of parameters to perform gaze estimation. Systematical experiments demonstrate the contributions from each model component, and the overall model also outperforms state-of-the-art on benchmark datasets.

Acknowledgments: The work described in this paper is supported in part by NSF award CNS 1629856, DoT DTRT5716C10037, and by RPI-IBM Cognitive Immersive Systems Laboratory (CISL).

References

[1] K. Wang, R. Zhao, and Q. Ji, "Human computer interaction with head pose, eye gaze and body gestures," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 789–789, IEEE, 2018. 1

[2] R. Zhao, K. Wang, R. Divekar, R. Rouhani, H. Su, and Q. Ji, "An immersive system with multi-modal human-computer interaction," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 517–524, IEEE, 2018. 1

[3] R. R. Divekar, M. Peveler, R. Rouhani, R. Zhao, J. O. Kephart, D. Allen, K. Wang, Q. Ji, and H. Su, "Cira: An architecture for building configurable immersive smart-rooms," in *Proceedings of SAI Intelligent Systems Conference*, pp. 76–95, Springer, 2018. 1

[4] C. H. Morimoto and M. R. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer vision and image understanding*, vol. 98, no. 1, pp. 4–24, 2005. 1

[5] W. A. W. Adnan, W. N. H. Hassan, N. Abdullah, and J. Taslim, "Eye tracking analysis of user behavior in online social networks," in *International Conference on Online Communities and Social Computing*, pp. 113–119, Springer, 2013. 1

[6] G.-J. Qi, C. C. Aggarwal, and T. S. Huang, "Online community detection in social sensing," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 617–626, ACM, 2013. 1

[7] J. Tang, X. Shu, G.-J. Qi, Z. Li, M. Wang, S. Yan, and R. Jain, "Tri-clustered tensor completion for social-aware image tag refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1662–1674, 2017. 1

[8] G.-J. Qi, C. C. Aggarwal, and T. Huang, "Link prediction across networks by biased cross-network sampling," in *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 793–804, IEEE, 2013. 1

[9] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky, "Eye tracking in web search tasks: design implications," in *Proceedings of the 2002 symposium on Eye tracking research & applications*, pp. 51–58, ACM, 2002. 1

[10] X. Wang, T. Zhang, G.-J. Qi, J. Tang, and J. Wang, "Supervised quantization for similarity search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2018–2026, 2016. 1

[11] S. Chang, G.-J. Qi, C. C. Aggarwal, J. Zhou, M. Wang, and T. S. Huang, "Factorized similarity learning in networks," in *2014 IEEE International Conference on Data Mining*, pp. 60–69, IEEE, 2014. 1

[12] W. A. Fletcher and J. A. Sharpe, "Saccadic eye movement dysfunction in alzheimer's disease," *Annals of neurology*, vol. 20, no. 4, pp. 464–471, 1986. 1

[13] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *TBE*, 2006. 1

[14] K. Wang and Q. Ji, "Hybrid model and appearance based eye tracking with kinect," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pp. 331–332, ACM, 2016. 1

[15] K. Wang, S. Wang, and Q. Ji, "Deep eye fixation map learning for calibration-free eye gaze tracking," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pp. 47–55, ACM, 2016. 1

[16] X. Xiong, Q. Cai, Z. Liu, and Z. Zhang, "Eye gaze tracking using an rgbd camera: A comparison with a rgb solution," *UBICOMP*, 2014. 1

[17] K. Wang and Q. Ji, "Real time eye gaze tracking with 3d deformable eye-face model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1003–1011, 2017. 1

[18] K. Wang and Q. Ji, "3d gaze estimation without explicit personal calibration," *Pattern Recognition*, 2018. 1

[19] L. Jianfeng and L. Shigang, "Eye-model-based gaze estimation by rgb-d camera," in *CVPR Workshops*, 2014. 1

[20] K. Wang and Q. Ji, "Real time eye gaze tracking with kinect," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 2752–2757, IEEE, 2016. 1

[21] L. Feng, Y. Sugano, T. Okabe, and Y. Sato, "Gaze Estimation From Eye Appearance: A Head Pose-Free Method via Eye Image Synthesis," *TIP*, 2015. 1

[22] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," *CVPR*, 2014. 1, 6

[23] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Inferring human gaze from appearance via adaptive linear regression," *ICCV*, 2011. 1

[24] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4511–4520, 2015. 1, 2, 6

[25] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2176–2184, 2016. 1, 2

[26] H. Deng and W. Zhu, "Monocular free-head 3d gaze tracking with deep learning and geometry constraints," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 3162–3171, IEEE, 2017. 1, 2

[27] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 100–115, 2018. 1

[28] K. Wang, H. Su, and Q. Ji, "Neuro-inspired eye tracking with eye movement dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012. 1

[30] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4373–4382, 2017. 1

[31] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010. 1

[32] H. Hu and G.-J. Qi, "State-frequency memory recurrent neural networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1568–1577, JMLR. org, 2017. 1

[33] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014. 2

[34] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, p. 4, 2017. 2, 4

[35] S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," *arXiv preprint arXiv:1807.10002*, 2018. 2, 6, 8

[36] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," *arXiv preprint arXiv:1805.04771*, 2018. 2, 8

[37] K. Wang, R. Zhao, and Q. Ji, "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. 2, 8

[38] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," *arXiv preprint arXiv:1502.02791*, 2015. 2

[39] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation.," in *AAAI*, vol. 6, p. 8, 2016. 2

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *NIPS*, 2014. 2, 4

[41] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017. 2

[42] R. M. Neal, *Bayesian learning for neural networks*, vol. 118. Springer Science & Business Media, 2012. 2

[43] J. S. Denker and Y. Lecun, "Transforming neural-net output levels to probability distributions," in *Advances in neural information processing systems*, pp. 853–859, 1991. 2

[44] G. E. Hinton and D. Van Camp, "Keeping the neural networks simple by minimizing the description length of the weights," in *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13, ACM, 1993. 2

[45] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight Uncertainty in Neural Networks," vol. 37, 2015. 2

[46] J. M. Hernández-Lobato, Y. Li, D. Hernández-Lobato, T. Bui, and R. Turner, "Black-box α divergence Minimization," *Black box learning & inference workshop (NIPS)*, pp. 1–5, 2015. 2

[47] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient hamiltonian monte carlo," in *International Conference on Machine Learning*, pp. 1683–1691, 2014. 2, 6

[48] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, vol. 1, p. 4, 2017. 4

[49] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 607–626, 2009. 4

[50] K. Wang, Y. Wu, and Q. Ji, "Head pose estimation on low-quality images," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 540–547, IEEE, 2018. 4

[51] Y. Saatci and A. G. Wilson, "Bayesian gan," in *Advances in neural information processing systems*, pp. 3622–3631, 2017. 5, 6

[52] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction," *ACM Symposium on User Interface Software and Technology*, 2013. 6

[53] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 255–258, ACM, 2014. 6

[54] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *CVPR*, 2015. 8

[55] T. Fischer, H. J. Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *European Conference on Computer Vision*, pp. 334–352, Springer, Cham, 2018. 8