Mixed Effects Neural Networks (MeNets) with Applications to Gaze Estimation*

Yunyang Xiong*
University of Wisconsin-Madison
yxiong43@wisc.edu

Hyunwoo J. Kim*
Korea University
hyunwoojkim@korea.ac.kr

Vikas Singh
University of Wisconsin-Madison
vsingh@biostat.wisc.edu

Abstract

There is much interest in computer vision to utilize commodity hardware for gaze estimation. A number of papers have shown that algorithms based on deep convolutional architectures are approaching accuracies where streaming data from mass-market devices can offer good gaze tracking performance, although a gap still remains between what is possible and the performance users will expect in real deployments. We observe that one obvious avenue for improvement relates to a gap between some basic technical assumptions behind most existing approaches and the statistical properties of the data used for training. Specifically, most training datasets involve tens of users with a few hundreds (or more) repeated acquisitions per user. The non i.i.d. nature of this data suggests better estimation may be possible if the model explicitly made use of such "repeated measurements" from each user as is commonly done in classical statistical analysis using so-called mixed effects models. The goal of this paper is to adapt these "mixed effects" ideas from statistics within a deep neural network architecture for gaze estimation, based on eye images. Such a formulation seeks to specifically *utilize information regarding the hierarchical structure of the training data — each node in the hierarchy is a user who provides tens or hundreds of repeated samples. This modification yields an architecture that offers state of the art performance on*various publicly available datasets *improving results by* 10-20%.

1. Introduction

Gaze serves as an important cue in understanding human attention, emotion and social interaction. Therefore, the ability to estimate and track gaze is important for various fields including psychology [19], neuroscience [13, 35, 8] and more recently, computer vision [45, 17, 21, 23]. While specialized eye gaze tracking hardware from several vendors have been available and used in research experiments for some time, in the last few years, many commodity products (such as [1]) can also be purchased that offer good real-time accuracy. Unfortunately, many of the high performing devices remain quite expensive and so, there is intensive work to come up with accurate computer-vision based gaze estimation techniques. One of the ideas in this line of work is appearance-based methods [48], using the appearance of the eye images to predict human gaze direction.

Generally speaking, appearance-based 3D gaze estimation can be formulated as a regression $f : \boldsymbol{x} \in \mathbf{R}^p \rightarrow \boldsymbol{y} \in \mathbf{R}^3$, where $\boldsymbol{x}$ is a set of features, e.g., image derived features and estimates of head pose from images and $\boldsymbol{y}$ is a gaze direction in 3D space. This problem can be approached in various ways. For example, we may use a standard k-NN regression estimator as in [46] or random forests for person-independent gaze estimation in [46]. The authors in [30] design adaptive linear regression whereas Schneider et al.[39] used support-vector regression with a polynomial kernel. More recently, deep neural networks have been successfully investigated for the problem in [60, 43]. For example, in [60], one provides an input eye image to a deep convolutional neural network and the last layer encodes the three-dimensional gaze vector. The parameters of the network can be trained with sufficient training data.

Are assumptions satisfied? Observe that independent of which scheme we use for inference, gaze estimation is a statistical fitting problem and understanding some of the basic assumptions and properties may suggest natural avenues of improvement. One of the basic assumptions most regression models make is that the samples are independent identically distributed (i.i.d.) [50]. It is meaningful to assess how (and whether) the common datasets used in many gaze estimation works satisfy (or violate) this property. **(A)** The Eyediap dataset [32] includes 94 video sequences of 16

*YX and HJK are corresponding authors. Work performed by HJK before joining Korea University.

subjects looking at three different targets with both static and free head motion under two different illumination conditions. **(B)** The UT Multiview dataset contains 160 gaze samples of 50 subjects recorded under controlled laboratory settings and 3D reconstructions of eye regions are used to generate synthetic images for arbitrary head poses. **(C)** A total of $214,000$ images from 15 subjects is collected in MPIIGaze dataset, which is considered to be a challenging dataset captured under extreme illumination conditions. The common feature in these (and other similar) datasets is that the data is "grouped" naturally which correspond to the number of participants in the acquisition study — these are repeated measurements and *not i.i.d*. If the statistical assumptions are violated, we can ask: **(1)** Is this just a theoretical issue or is this relevant in practice? **(2)** Are there simple fixes to this problem? For any gaze dataset, we cannot expect a researcher to collect i.i.d. data: which would mean spending effort into bringing in a participant and collect only one gaze sample.

Basic setup of this paper. The answer to question (1) is yes (we will experimentally show later). But first, we address (2) because it helps setup our formulation. Notice that the issue of non i.i.d. data is not restricted to computer vision and in fact, commonly occurs in social sciences, epidemiology and medicine. The de-facto recommendation when dealing with *multiple samples from each participant* is to utilize the so-called *mixed effects models* [22, 25] which are a special case of the more general hierarchical Bayesian model. The mixed effects models are composed of two parts: the *fixed effects* and the *random effects*. The fixed (global) effects are common in *all* samples and so the corresponding coefficients are called fixed. In contrast, the random (local) effects are specific to subjects (or groups). The random effects coefficients can vary depending on the subjects, assumed to be drawn from some unknown distribution. This approach is effective in many applications [22, 2, 16, 12, 20] and broadly used in standard statistical analysis. This raises the possibility of whether we can utilize the "mixed effects" idea in deep neural networks (especially CNNs) which offer state-of-the art performance in a range of problems in computer vision. Actually, using mixed effects is natural whenever images come clustered in groups/hierarchy (not necessarily individual subjects): this is common in fine-grained multi-label classification, object detection, medical imaging, and any longitudinal data. In general, other than participant specific random effects, we may even consider a separate "site" or "dataset" specific random effect. This is relevant because recent results have shown that even for linear regression, different datasets cannot be easily pooled in simple ways [62, 61].

Other models for repeated measurements. We should point out that neural networks for non i.i.d. data are not unique to this paper. In fact, recurrent neural networks (RNNs) have been studied in language modeling [31], speech recognition [37], image captioning generation [55], motion capture [47], and machine translation [5, 57] since the 1990s and recent work has substantially built upon the early formulations. However, RNNs are designed for **sequential (ordered) data** and do not directly fit our problem of gaze estimation based on a single input image. More importantly, RNNs do not explicitly exploit the group information (e.g., these 100 repeated measurements come from Alice), whereas mixed effects models explicitly use the subject (group) information by estimating the **random effects per subject** where the samples need not sequential. This is the main gap which is addressed by our work. The **contributions** of this paper are: **1)** We provide a mathematically sound neural network which includes the benefits of terms that model repeated measurements, arguably a better fit with the statistical properties of *most* available gaze datasets. **2)** Experimentally, we show that our formulation outperforms the state of the art by significant margins (10%-20%) on most available datasets.

1.1. Related work

Model-based methods for gaze estimation use a predefined geometric eye model and can be subdivided into feature-based methods and shape-based methods. The feature-based methods use predefined geometric eye features such as pupil center corneal reflection [14], iris contours [36], leverage infrared sensors [14], stereo cameras [42] and depth camera [18]. These approaches can offer high accuracy but their dependence on specialized hardware and calibration may not be suitable for more mass-market applications. On the other hand, shape-based methods [60, 49, 52, 3] extract shape parameters from observed eye images such as center of pupil, boundary of limbus and iris and seek to associate them with a geometric eye model to infer the gaze direction. These methods are quite successful but since summaries (optical axis, cornea radii, pupil radii) based on accurate registration are required, it is not clear whether they can yield high accuracy with low-resolution images from web cameras.

In contrast to the foregoing line of work, appearance-based methods, do not use an explicit geometric eye model and instead utilize eye images (or non-geometric features) as an input to directly learn the parameters of a mapping between eye images and gaze. While early works [4, 56, 48, 51, 41, 39] assumed fixed head poses for performing gaze estimation on eye images, more recent works [29, 6, 33, 27, 28, 54, 24] show promising results with arbitrary head poses, illumination and backgrounds. [4, 56] trained neural networks on eye images for gaze estimation. [48] utilized the local linearity of the eye appearance manifold and applied local interpolation to predict gaze. In [29], a calibrated approach, called adaptive linear regression, was

developed for gaze estimation that is robust to head movement. With the recent impact of deep neural network frameworks in computer vision, interest in appearance-based methods has been revived, in large part via the use of CNNs [60, 21, 58, 59, 10]. These works leverage relatively large scale datasets collected on participants' laptops and mobile devices in more general settings during daily life. These datasets aim to achieve appearance-based head-free and calibration-free gaze estimation in a wide range of scenarios involving significant variations in illumination, head pose, background and so on. We note that distinct from the approaches above, eye image synthesis has also been studied towards generating larger training data with bigger variations in head pose [28] and more recently, via generative adversarial networks (GANs) [43, 40]. [29, 26, 34] proposed personalized gaze estimation methods to handle variability across subjects using calibration samples.

2. Review of Mixed effects models

Many statistical models are *fixed effects models* that use a single "global" model and all parameters are associated with the the full set of samples without encoding information on which repeated samples came from which participants. In contrast, *random effects models* have a set of parameters for each subject (or group) and assume that the parameters are drawn from an unknown distribution. Unlike setting up a prior distribution as in Bayesian methods, chosen based on some domain knowledge, the unknown distribution in random effects models are also estimated [38].

How is a mixed effects model different? A model with both fixed effects and random effects is called a *mixed effects model* [22]. Mixed effects models describe relationships between a response variable and some covariates in data that are **grouped** based on some grouping criteria. If Alice and Bob provide 20 samples each, the model will know which samples are from whom. It is possible that the measurements from Alice have higher values and samples from Bob have a bigger variance. By associating such "random" effects to repeated measurements with the same "person" tag, mixed effects models flexibly represent the covariance structure of the grouping of the samples.

Recall that appearance based gaze estimation is a fitting problem $\boldsymbol{x} \rightarrow \boldsymbol{y}$ where $\boldsymbol{x} \in \mathbf{R}^p$ is an eye image or a feature vector (e.g., head pose, deep convolutional features) extracted from an eye image and $\boldsymbol{y} \in \mathbf{R}^3$ is a gaze direction. Our goal is to use fixed effects for a global model (as is the case in most existing schemes) but also include random effects to make use of information on which samples are from which participant — this yields subject-specific adjustments. We first introduce the linear and non-linear version of mixed effects model, covered in common textbooks. Then, with this concept in hand, we will add mixed effects terms within deep neural networks.

2.1. A Linear mixed effects model

We begin with a fixed effects model, i.e., a standard linear model, and then introduce a linear mixed effects models. Recall that a linear regression model is given as

$$y = \beta^0 + \beta^1 x^1 + \cdots + \beta^p x^p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \mathbf{\Sigma}_\epsilon) \quad (1)$$

where $\boldsymbol{x} = [x^1, \ldots, x^p] \in \mathbf{R}^p$, $\beta = [\beta^1, \ldots, \beta^p]^T \in \mathbf{R}^p$, $y \in \mathbf{R}$. We call this the "standard" linear model. When both $\boldsymbol{x}$ and $\boldsymbol{y}$ are multivariate measurements then we call it a *general linear model*. For simplicity of discussion, we introduce models with a univariate response variable (also called labels, dependent or target variable). Observe that in (1), all subjects have exactly the same function to map eye appearance to gaze directions; the noise permitted in the model estimation also comes from a distribution identical to everyone. But most gaze estimation datasets have repeated multiple measurements from a subject (see Fig. 1). These samples are not independent and each subject may have a slightly different mapping function. To address this issue, we may add random effects to (1) for participant. This yields a mixed effects model

$$y = \beta^0 + \beta^1 x^1 + \cdots + \beta^p x^p + u_i^1 z^1 + \cdots + u_i^q z^q + \epsilon_i,$$
$$u_i \sim \mathcal{N}(0, \mathbf{\Sigma}_{\boldsymbol{u}}) \text{ and } \epsilon_i \sim \mathcal{N}(0, \mathbf{\Sigma}_{\epsilon_i}) \quad (2)$$

where $\beta := [\beta^1, \ldots, \beta^p]^T$ are the fixed effects shared over the entire population, $u_i := [u_i^1, \ldots, u_i^p]^T$ are the random effects of the i_{th} subject (or group), and $z = [z^1, \ldots, z^q]$ is a design vector for q random effects. In the random effects part in (2), u_i allows subject-specific adjustment and ϵ_i is drawn from a *subject-specific* unknown distribution $\mathcal{N}(0, \mathbf{\Sigma}_{\epsilon_i})$ which enable precise handling of the non-i.i.d. nature of the data. Typically, the unknown distributions are assumed to be a zero mean Gaussian with an unknown covariance structure. Estimation involves estimating the parameters of a fixed effects model β, the random effects components $u_1, \ldots, u_N$ as well as $\mathbf{\Sigma}_{\epsilon_i}$ for all i, where N is the number of subjects. Since the linear mixed effects models have multiple random effects from unknown distributions, no closed form solution is available [53]. For estimation, EM algorithms and MCMC sampling are used [53, 15].

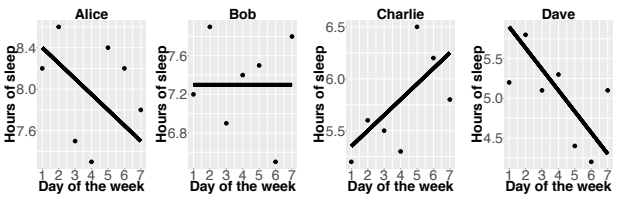

Figure 1: A simple example. Sleep measurements of Alice, Bob, Charlie, Dave, when pooled, do not correspond to an i.i.d. sample of the distribution, rather is hierarchically structured.

2.2. A Nonlinear mixed effects model

A nonlinear mixed effects model, as the name suggests, is an extension of linear mixed effects models but may use a nonlinear link function. Before describing the details, we setup useful notations. Let $\boldsymbol{y}_i = \left[y_{(ij)}\right]_{j=1}^{n_i}$ be a set of n_i repeated observations of a response variable from subject i: if Alice (indexed by i) provides 100 gaze measurements, then $n_i = 100$. Here $\boldsymbol{y}_i$ is a n_i dimensional vector, vertically stacked with $y_{(ij)}$ responses for subject i. We let $x_{(ij)}$ denote the predictor variables derived from the corresponding eye images for subject i and so, X_i is a $n_i \times p$ matrix if we assume a p-dimensional feature vector for the eye image. This can be setup as $\left[x^1_{(ij)}\ x^2_{(ij)}\ \ldots\ x^p_{(ij)}\right]_{j=1}^{n_i}$ where we collect for subject i, all p features for all n_i visits. The matrix Z_i of size $n_i \times q$ is a design matrix for q random effects (the random complement to the fixed X_i). In a simple case, Z_i can be $n_i \times 1$ matrix of ones to apply the subject-specific bias but Z_i can also include X_i or a subset of X_i features to denote subject-specific slopes as well. With this notation, a general, nonlinear mixed effects model [9] is

$$\begin{aligned} &\boldsymbol{y}_i = \nu(\Phi_i) + \boldsymbol{\epsilon}_i \text{ where } \Phi_i = X_i\beta + Z_i u_i \\ &u_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_u) \text{ and } \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_i}) \end{aligned} \tag{3}$$

where $\beta \in \mathbf{R}^p$ are the coefficients for the fixed effects, $u_i \in \mathbf{R}^q$ are the coefficients for the random effects, ν is a nonlinear function of the predictor variables (i.e., the feature vectors) and Φ_i collects for participant i the terms of fixed effects variable β plus the random effects variable u_i. Similar to the linear mixed effects model, the random effects $u_i \in \mathbf{R}^q$ and noise $\boldsymbol{\epsilon}_i = [\epsilon_{(i1)}, \ldots, \epsilon_{(in_i)}]^T$ are each a $n_i \times 1$ vector drawn from unknown normal distributions.

An EM algorithm is used to solve both the linear mixed effects model and nonlinear mixed effects model and can be described via standard E and M steps as described in [53] and implemented in many toolboxes.

3. Mixed effects in Deep Neural Networks

Our proposed model seeks to learn the mapping from the input features to gaze angles while being cognizant to individual level differences of the subjects. To do so, let us separate the nonlinear mixed effects model into fixed and random components. For simplicity, assume that the "X_i" in (3) for ν comes from any suitable neural network Γ. Since the input predictor is the eye appearance image, for this setting, it is valid to assume that the subject-specific random effects are also derived from X_i (Random Intercept and Random Slope Model, [11]). So, we may write

$$\boldsymbol{y}_i = \underbrace{\Gamma(X_i)\beta}_{\text{fixed effects}} + \underbrace{\Gamma(X_i)u_i}_{\text{random effects}} \tag{4}$$

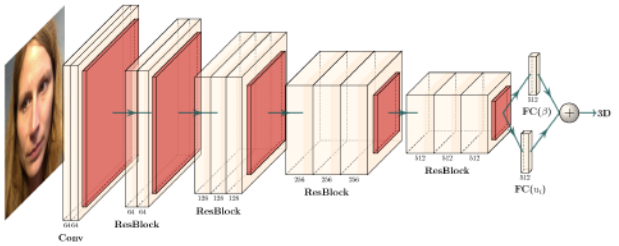

Figure 2: An overview of the ResNets structure incorporating linear mixed effects model, MeNets. The structure of ResBlock can be seen in Fig. 3 and the architecture details of MeNets are shown in the appendix. The hidden representation after the fully connected layer is $\Gamma(X_i)$, which is the input for the mixed effects models, $f(X_i) = \Gamma(X_i)\beta$ for fixed effects estimation and $\tilde{f}(X_i) = \Gamma(X_i)u_i$ is for random effects estimation. The output of the network is a 3D unit vector, the sum of fixed effects estimation and the random effects estimation.

Note that here we apply nonlinear transformation $\Gamma(\cdot)$ first and then learn fixed effects and random effects on top of the features whereas in (3), a nonlinear function $\nu(\cdot)$ is applied through $\Gamma(\cdot)$ and combines the fixed effects and random effects terms. This can be interpreted as using X_i for Z_i in (3) to learn subject-specific slopes instead of using entries in Z to indicate participant↔sample relationship as is commonly done. Since non-linearity is now subsumed within $\Gamma(\cdot)$, the form above suggests that we can use a deep neural network for $\Gamma(\cdot)$ and train a linear mixed effects model *after* the representations $\Gamma(X_i)$ have been learned. Our model uses the ResNets architecture for $\Gamma(\cdot)$ followed by a linear mixed regression loss after the fully connected (fc) layer: the loss is based on the gaze direction response variable. We perform end-to-end training.

We define β as the fixed effects coefficient on top of the final fc layer $\Gamma(x_{(ij)})$ and use u_i as the random effects coefficient on top of the final fc layer — these contribute to the linear mixed effects regression after the final fc layer $\Gamma(x_{(ij)})$. In other words, we define the fixed effects component of the regression model, $f(x_{(ij)}) = \Gamma(x_{(ij)})\beta$ and a separate random effects component of the model $\tilde{f}(x_{(ij)}) = \Gamma(x_{(ij)})u_i$. Then, the nonlinear mixed effects model can be written as an optimization problem by training the neural network in Fig. 2 with the loss function,

$$\min_{f,\tilde{f}} \sum_{ij} L(x_{(ij)}, y_{(ij)}) = \sum_{ij} ||y_{(ij)} - f(x_{(ij)}) - \tilde{f}(x_{(ij)})||^2, \tag{5}$$

where $f(\cdot)$ and $\tilde{f}(\cdot)$ are fully specified by $\{\beta, \Gamma(\cdot)\}$ and $\{u_i, \Gamma(\cdot)\}$ respectively.

3.1. Parameter estimation: Variational EM + SGD

To derive our algorithm, let us analyze the formulation in a little more detail. Expressed in terms of the response variable, for each subject i, the nonlinear mixed effects model

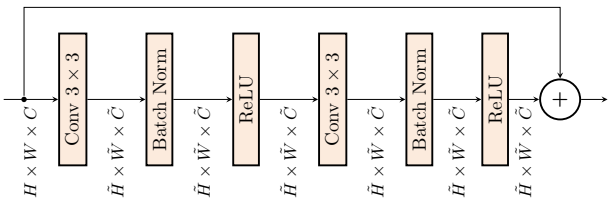

Figure 3: ResBlock structure.

can be formulated in design matrix form as,

$$\boldsymbol{y}_i = f(X_i) + \tilde{f}(X_i) + \boldsymbol{\epsilon}_i, i = 1, \ldots, N \tag{6}$$

where $\boldsymbol{y}_i = [y_{(i1)}, \ldots, y_{(in_i)}]^T$ is the $n_i \times 1$ vector of responses for the n_i observations for subject i, summing over all n_i's for N subjects gives us n observations, $X_i = [x_{(i1)}, \ldots, x_{(in_i)}]^T$ is the $n_i \times p$ design matrix, and $\boldsymbol{\epsilon}_i = [\epsilon_{(i1)}, \ldots, \epsilon_{(in_i)}]^T$ $n_i \times 1$ vector of errors, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_i})$. Recall that the hidden representation after the fully connected layer in Fig. 2 is $\Gamma(X_i)$ which is the input for the mixed effects models i.e., $f(X_i) = \Gamma(X_i)\beta$ and $\tilde{f}(X_i) = \Gamma(X_i)u_i$. We further assume that u_i and $\boldsymbol{\epsilon}_i$ are independent and normally distributed and that the between-subject observations are independent. We can verify that the covariance of observations $\boldsymbol{y}_i$ for subject i is

$$V_i = \text{COV}(\boldsymbol{y}_i) = \Gamma(X_i)\boldsymbol{\Sigma}_u\Gamma(X_i)^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_i}.$$

Estimation of a nonlinear mixed effects model can be performed by EM algorithm [53, 16] and our algorithm mimics this strategy for our formulation, see Algorithm 1. Our algorithm is a variational EM algorithm which involves an iterative optimization algorithm (SGD) within an EM procedure. In any EM procedure, data are assumed to be incomplete and the goal is to estimate the unobserved measurements and model parameters iteratively. In our problem, β, u_i and ϵ are unobserved measurements and they are estimated in the Expectation step. In the Maximization step, with the "complete" data (observed plus estimated measurements), the algorithm seeks to find the model parameters $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_i}$ simply by maximizing the likelihood. Consistent with convention, below, the estimate for any variable, say ρ, is given as $\hat{\rho}$.

What does the algorithm do? Algorithm 1 starts with initial values for $\hat{\beta}, \hat{u}, \hat{\sigma}^2, \hat{\boldsymbol{\Sigma}}_u$. Then in the **E step**, we calculate the fixed effects part of the response variable, $\boldsymbol{y}_i^{\text{fixed}}$, that is, the response variable from which we remove the current (i.e., estimated) value of the random effects term. For updating the fixed effects term's contribution to the response variable, we use SGD for fitting $(x_{(ij)}, y_{(ij)}^{\text{fixed}})$ using the convolutional neural network to obtain the $\hat{\beta}$ and $\Gamma(X)$. We then estimate the random effects part $\hat{u}_i$ based on removing the update fixed effects term. Lastly, we update the between-subject and within-subject variance based on the updated estimates of the residuals at **M step**. The algorithm keeps iterating by updating fixed effects and random effects component until convergence. The convergence of the algorithm can be captured by computing the following loss function at each iteration:

$$F(g, \beta, u_i | X, \boldsymbol{y}) = \sum_{i=1}^{N} [(\boldsymbol{y}_i - \Gamma(X_i)\beta - \Gamma(X_i)u_i)^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_i}^{-1} \tag{7}$$
$$(\boldsymbol{y}_i - \Gamma(X_i)\beta - \Gamma(X_i)u_i) + u_i^T \boldsymbol{\Sigma}_u^{-1} u_i + \log|\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_i}| + \log|\boldsymbol{\Sigma}_u|]$$

This is the negative log-likelihood function with a Gaussian assumption on noise (ϵ) and random effects u_i as in the classical mixed effects models (2). To predict the gaze estimate for a new observation j, we can encounter two cases. First, the subject was seen at training time and second, when the subject was not seen at training time. For the first case, we use both its corresponding population-level network regression term, $\hat{f}(x_{ij})(\cdot)$ and the predicted random effect term corresponding to subject i, $\hat{\tilde{f}}(\cdot)$ for prediction. For a subject *not* encountered at train time, next, we describe how the contribution from these terms can be approximated without knowing the subject "id" at test time.

Dealing with unseen subjects. A simple solution that works well is to concurrently learn to "predict" the random effects term $\tilde{f}(\cdot)$ based on the input eye images without knowing the subject id at test time. Assume we have two functions, a univariate function, $h(a)$ and a bivariate function, $l(a, b)$. By marginalizing over the variable b, we can find the best $h(a)$ such that $h(a) \approx \int_b l(a, b)db$. If $h(a)$ is learned correctly, it can act as a good proxy for $l(a, b)$ without access to information regarding the second variable b. Since our main model estimates a subject-specific $\tilde{f}(\cdot)$ we use a function $h(\cdot)$ to predict the random effect terms based only on the eye images. Notice that $h(\cdot)$ shares much of the same network architecture as $\tilde{f}(\cdot)$, so it is not necessary to have a separate "network", instead $h(\cdot)$ can simply be a fully connected layer at the end which predicts the random effects offset *using* $\Gamma(x_{ij})$ *as input*, i.e., $h(\cdot) : \Gamma(x_{ij}) \rightarrow \tilde{f}(x_{ij})$. In fact, training $h(\cdot)$ does not even need to happen concurrently with Alg. 1. Once the training of Alg. 1 has been completed, we can fix the weights of the convolutional layers, and learn a fully connected layer $h(\cdot)$ which will best predict $\tilde{f}(x_{ij})$ based only on the hidden representations $\Gamma()$ provided by the convolutional layers. At test time, for a new subject, we use the fixed effects terms from our model and add in the "offset" provided by $h(\Gamma(\cdot))$ using that specific participant's eye images (rather its hidden representations from $\Gamma(\cdot)$).

4. Experiments

In this section, we discuss the subject-independent gaze estimation task and validate the effectiveness of our MeNets

Algorithm 1 Variational EM + SGD for estimating the parameters of MeNet

0: **procedure** Training

1: $k = 0$, initialize random effects coefficients and between-subject covariance, $\hat{u}_{i(0)} = 0, \hat{\sigma}^2_{(0)} = 1, \hat{\mathbf{\Sigma}}_{i(0)} = I_p$, denote $X_i(0)$ the input images to the network.

E-Step :

2: $k = k + 1$, update the fixed part response $\boldsymbol{y}^{\text{fixed}}_{i(k)}$ by

$$\boldsymbol{y}^{\text{fixed}}_{i(k)} = \boldsymbol{y}_i - \Gamma(X_{i(k-1)})\hat{u}_{i(k-1)}, i = 1, \ldots, N. \quad (8)$$

3: Use the linear mixed deep net structure for training $(x_{ij}, y^{\text{fixed}}_{ij(k)})$ with stochastic gradient algorithm to obtain an estimate $\hat{\beta}_{(k)}, \Gamma(X_{i(k)})$.

4: Update random effects $\hat{u}_{i(k)}$ and $\hat{\boldsymbol{\epsilon}}_{i(k)}$ by

$$\hat{u}_{i(k)} = \hat{\mathbf{\Sigma}}_{u(k-1)}\Gamma(X_{i(k)})^T(\hat{V}_{i(k-1)})^{-1} \quad (9)$$
$$(\boldsymbol{y}_i - \Gamma(X_{i(k)})\hat{\beta}_{(k)}),$$
$$\text{where } \hat{V}_{i(k-1)} = \Gamma(X_{i(k)})\hat{\mathbf{\Sigma}}_{u(k-1)}\Gamma(X_{i(k)})^T$$
$$+ \hat{\sigma}^2_{(k-1)}I_{n_i}$$
$$\hat{\boldsymbol{\epsilon}}_{i(k)} = \boldsymbol{y}_i - \Gamma(X_{i(k)})\hat{\beta}_{(k)} - \Gamma(X_{i(k)})\hat{u}_{i(k)} \quad (10)$$

5: **M-Step :**Update between-subject variance $\hat{\sigma}^2_{(k)}$ and within-subject variance $\hat{\mathbf{\Sigma}}_{u(k)}$

$$\hat{\sigma}^2_{(k)} = \frac{1}{n}\sum_{i=1}^{N}[\hat{\boldsymbol{\epsilon}}^T_{i(k)}\hat{\boldsymbol{\epsilon}}_{i(k)} + \hat{\sigma}^2_{(k-1)} \quad (11)$$
$$(n_i - \hat{\sigma}^2_{(k-1)}\text{trace}(\hat{V}_{i(k-1)}))],$$
$$\hat{\mathbf{\Sigma}}_{u(k)} = \frac{1}{N}\sum_{i=1}^{N}[\hat{u}_{i(k)}\hat{u}^T_{i(k)} + (\hat{\mathbf{\Sigma}}_{u(k-1)} \quad (12)$$
$$- \hat{\mathbf{\Sigma}}_{u(k-1)}\Gamma(X_{i(k)})^T\hat{V}^{-1}_{i(k-1)}\Gamma(X_{i(k)})\hat{\mathbf{\Sigma}}_{u(k-1)})]$$

6: Keep iterating by repeating **E-step** and **M-step** until convergence.

6: **procedure** Prediction

7: For a new observation j for subject i, fixed component of deep net regression, $\Gamma(x_{(ij)})\hat{\beta}$ plus the predicted random part , $\Gamma(x_{(ij)})\hat{u}_i$.

8: For unknown observations, take the corresponding fixed component deep net prediction by $\Gamma(x_{(ij)})\hat{\beta}$ plus the predicted random part by $h(\Gamma(x_{(ij)}))$.

model. We compare the proposed method with various gaze estimation schemes. We conduct both within-dataset and cross-dataset evaluations to show the generalization power on unseen subjects (from the same and different datasets), which are expected in an unconstrained daily-life setting. We evaluate our method on three datasets, **MPIIGaze**, **UT-Multiview**, and **Real-video** dataset that we collected. Figure 4 (right) shows some examples of the actual test images. In contrast to gaze estimation schemes with calibrations, subject-independent gaze estimation algorithms are believed to require a large amount of training data. We briefly introduce each dataset.

MPIIGaze Dataset: MPIIGaze dataset [60] is a gaze dataset which contains 213,659 images collected from 15 laptop users over several months. It covers a large variability of head pose, appearance and illumination. The publicly available images include cropped eye images with some preprocessing such as registration from its corresponding full face dataset MPIIFace [58].

UTMultiview Dataset: UTMultiview dataset has a large number of 50 subjects with 64,000 (50 subjects $\times$ 8 views $\times$ 160 gaze directions) eye images in total. The authors further reconstruct 3D shapes of the eye regions and densely synthesize training eye images from dense viewing angles to provide 1,152,000 synthesized eye images.

Real-Video Dataset: Both of the datasets are well-organized and actively used for benchmarks [60, 46]. They are provided as high-resolution images, but cheap commodity devices in real life may only be able to offer lower resolution images. UTMultiview dataset even uses reconstructed 3D shapes to generate artificial eye images, which are different from real-life gaze estimation. The consequence of this gap can be observed in the poor gaze estimation performance [60] (Fig. 7) in cross-dataset evaluations, e.g., training on UTMultiview and testing on MPIIGaze dataset. In order to show the ability of our method to provide subject-independent gaze estimation for real data and complement the MPIIGaze dataset, we collect our own data from a commodity camera on a laptop. We developed an app to collect face image data and gaze direction concurrently. The collection procedure is simple. Participants were only asked to watch a gym video for several minutes. The app records subjects' face appearance and capture gaze directions using a Tobii eye tracker. The Tobii eye tracker is

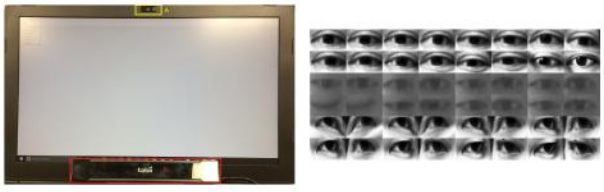

Figure 4: (Left) Face image data and gaze direction data collection set-up, the top yellow bounding box A refers to the webcam of the laptop, the bottom red bounding box B refers to Tobii X2-30 attached to the laptop. (right), Example of test images from 3 datasets, the first 2 rows from MPIIGaze dataset, the middle 2 rows from Real-Video dataset captured by our gaze data collection set-up, the last 2 rows from UT Multiview dataset.

specialized device to collect groud-truth gaze directions. It requires calibration for each session. The setup of our data collection system is shown in Fig. 4 (left) and the details of the data collection system are available in the appendix. We use the webcam of a MSI laptop to record eye images with resolution 848×480 and Tobii X-30 attached to the laptop to record corresponding gazes. The Tobii X-30 Compact can offer gaze direction estimation error less than 1° under ideal conditions including subject-specific accurate calibration [1, 7] and gives gaze direction estimation error 2.46° under non-ideal conditions. We collect this data from 7 volunteers: 1,711 to 7,605 images per participant.

Deciding the architecture: We implemented and experimented with using various state-of-the-art deep architectures to serve as the convolutional module in our formulation. We evaluated gaze estimation accuracy on the MPIIGaze dataset using leave-one-subject-out cross-validation. The two options we evaluated were the 18-layer ResNet and GoogLeNet. To setup the MeNets network, we change the last classification layer to two fully connected layers for gaze direction regression as shown in Fig. 2. The evaluations of MeNets with these two architectures versus the corresponding baselines (i.e., ResNet on its own, GoogLeNet on its own) is shown in Table 1, which also shows the performance of other state of the art approaches. From the results, our proposed MeNets outperforms the corresponding GoogLeNet and ResNet networks and all other contemporary approaches (a 10% improvement over GazeNet+, a paired Wilcoxon test gives p-value < 0.01). Based on these experiments, we use MeNets with the ResNets architecture in the remaining evaluations.

Convergence of Variational EM + SGD: Before additional accuracy plots, we present some results to evaluate the convergence of our estimation scheme. We conduct experiments under two settings: within-subjects (standard 10-fold cross-validation) and across-subjects (14 subjects for training and one subject for testing). Both settings show good convergence behavior. Evaluating the log-likelihood as a function of the number of iterations, we see that 7-8 iterations are enough (see the appendix for example).

Does the "mixed effects" terms in the MeNets model yield improvement? We evaluate whether the fixed effects terms alone in our architecture perform close to our MeNets model (with mixed effects terms). We used the strategy in [60] where a random subset for both training and testing is used and includes 1500 left + 1500 right eye samples for each person. Since eyes are not exactly symmetrical, we swap the right eye image horizontally and mirror the pose and gaze direction so that both eyes can be handled by a single regression function. Here, our MeNets model is trained with the random effects terms and then at test time, we use two options: use predicted random effects or not use predicted random effects. On MPIIGaze, even when

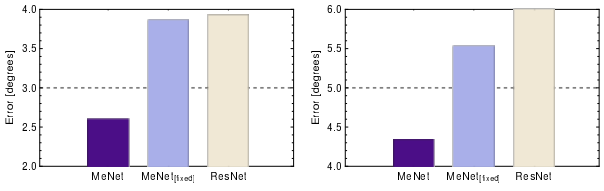

Figure 5: Comparison of MeNet, MeNet$_{[\text{fixed}]}$, and ResNet for gaze estimation under within-subjects and across-subjects settings. MeNet uses random effects for gaze estimation with ResNet architecture at test time, while MeNet$_{[\text{fixed}]}$ does not incorporate random effects at test time and ResNet does not consider any individual-level differences in training/testing. (left) MeNet offers the highest accuracy with 30% gain, 2.66 degrees vs 3.9 degrees by ResNet, even without incorporating random effects at test time, MeNet$_{[\text{fixed}]}$ outperforms ResNet. (right) Shows that MeNet still gives the highest accuracy for across-subjects gaze estimation, 4.34 degrees vs 6.0 degrees by ResNet.

the random effects are not used at test time (but the model was trained with the subject-specific random effects), we achieve better accuracy versus using the ResNets architecture on its own for gaze prediction. This is because incorporating subject specific random effects terms improves our estimating the fixed model in Fig 5. Using the full mixed effects, yields the best results (p-value < 0.01).

We also evaluate whether other gaze datasets demonstrate subject-specific random effects. By using subject-specific random effects, we see that linear mixed effects regression performs better than linear regression for gaze estimation on the three datasets in Table 1. Linear mixed effects model for gaze estimation is more accurate than just using linear effects model: 2° more accurate on Real-video and UT Multiview dataset. In summary, we find that the benefit of terms that account for repeated samples is clearly observed in our experiments.

Performance of MeNets on Within-Dataset evaluation: We compare our proposed method with other baseline methods and evaluate gaze prediction accuracy under leave-one-person-out setting. The model-based method EyeTab performs poorly on three datasets with mean error larger than 20 degrees. It means that appearance-based methods have an advantage over model-based methods for performing gaze estimation on real images. Since kNN, Random forests, ALR regression for gaze estimation underperform convolutional neural networks [54], in our experiments, we only report the performance of linear model, linear mixed effects model, support vector regression, several CNN based methods and MeNets. We perform leave-one-person-out gaze estimation on MPIIGaze and 3 fold cross validation on UT Multiview dataset(consistent with [46]). Table 1 show the mean estimation errors of within-dataset evaluation on MPIIGaze and Real-video dataset. Our MeNet obtains mean error 4.9 degrees on MPIIGaze under leave-one-person-out setting, while the state-of-the-art GazeNet+ can only offer mean error of 5.4 degrees on MPIIGaze (p-value < 0.01). Our gaze estimation also

Table 1: Comparison of our model with other baselines. Row 2 shows error with leave-one-person-out setting on MPIIGaze data and Row 3 shows error on Real-video. Row 4 shows error with 3-fold cross validation on UT Multiview. Row 5 shows error with cross-dataset evaluation with training data from the UT Multiview dataset and test error on MPIIGaze dataset. Our MeNet achieves large consistent accuracy improvement over baselines.

	MeNet	ResNet	GoogLeNet	GazeNet+[59]	iTracker[21]	MPIIGaze[54]	SVR	LR	LME
MPIIGaze	$\mathbf{4.90 \pm 0.59}$	6.04 ± 0.64	6.15 ± 0.81	5.40 ± 0.67	6.20 ± 0.85	6.59 ± 1.07	8.94 ± 3.20	7.44 ± 1.16	7.06 ± 1.07
RealVideo	$\mathbf{6.72 \pm 1.15}$	6.98 ± 1.63	7.13 ± 1.74	6.90 ± 1.34	7.65 ± 2.01	9.78 ± 2.85	12.67 ± 3.57	12.90 ± 2.71	10.14 ± 1.88
UT Multiview	$\mathbf{5.50 \pm 1.03}$	5.86 ± 1.10	5.97 ± 1.15	5.78 ± 1.04	N/A	5.98 ± 1.21	9.11 ± 2.27	9.07 ± 2.41	6.71 ± 1.41
Cross-dataset	$\mathbf{9.51 \pm 0.75}$	9.84 ± 1.73	9.97 ± 1.82	9.80 ± 1.83	N/A	13.30 ± 2.12	> 15	> 15	> 15

outperforms other CNN methods on UT Multiview. We significantly outperform all other baseline methods, which supports the advantage of incorporating the specific effects from each subject for gaze estimation.

How does this work for real video data? In order to show the ability of our method to provide subject-independent gaze estimation for real data, we perform gaze direction prediction for our Real-Video dataset. Since the resolution of the eye image in Real-Video dataset is lower than MPIIGaze dataset (and other real-world factors), it makes gaze prediction more challenging than on the MPIIGaze dataset. We compared our method with other gaze prediction methods. We outperform all other methods consistently. Moreover, when adding MPIIGaze dataset for gaze prediction on real video data, the accuracy increases to 6.11°. Some examples of gaze prediction can be seen in the appendix.

Cross-Dataset evaluation: We assess the effectiveness of our method for cross-dataset evaluations. We selected the UT Multiview dataset for training and perform estimation on MPIIGaze. Table 1 summarizes the mean angular errors and standard deviations of our method and other CNN based methods on MPIIGaze. Our method still outperforms the GazeNet+, but the improvement is not as large as in the within-dataset evaluation setting. Since UT Multiview dataset uses learning-by-synthesis approach for generating more eye images, the generated images are very different from images in MPIIGaze. In [43], by adding realism of generated images, the refined generated images offers better gaze estimation accuracy by more than 3°. The data shift problem is significant here, which domain adaptation techniques should be used to deal with, but were not utilized in this paper. Then, we add the MPIIGaze dataset to the training data and apply leave-one-person-out gaze estimation for real-video dataset, it improves the gaze estimation accuracy by more than 1° which partly supports the domain shift problem between UT Multiview and MPIIGaze.

Personalized gaze estimation: As our mixed effects model can learn the random effects associated with a specific person, our method is amenable to personalization via few calibration samples. In order to show our method can adapt to individual subjects, we perform personalized gaze estimation on MPIIGaze dataset. We pick calibration samples from MPIIGaze and use the remaining samples for evaluation. Given 200 calibration samples for each subject, our model achieves mean error of 3.8 degrees, comparable with state-of-the-art personalized gaze approaches [34].

Mixed-Dataset evaluation: In order to further show gaze estimation performance on multi-datasets, we assess the effectiveness of our method for mixed-dataset evaluations, where we pick samples from each dataset for training and pick other samples for testing. We pick 10 subjects from MPIIGaze dataset and the same number of eye images were generated from GAN [44] for training a ResNet model. Then, another 5 subjects from MPIIGaze dataset and the same number generated by GAN were used for testing. Although the actual gaze estimation task for both datasets is the same, the difference between the synthetic dataset and real dataset is large. Without considering the dataset difference, the trained ResNet model offers 16.36 degrees error, while our MeNet formulation using a random effects term for the dataset gives 11.2 degrees error and MeNet (without random effects) yields 11.7 error. It means that ***dataset specific random effects improve estimation over a model that is agnostic of this information***; suggesting the random effects at test time yields additional improvements.

5. Conclusion

Most researchers performing data analysis know that the choice of the correct model for the data at hand can lead to improvements in performance, and conversely a sub-optimal model can yield poor results. For appearance based gaze estimation, we explore how an appropriate statistical model that leverages information regarding repeated measurements from the same participant, a common feature of most if not all existing datasets, seems like a much better fit but has not been explored in computer vision much. To practicalize this observation within modern architectures, we propose a formulation that estimates a mixed effects model while leveraging the benefits of powerful deep neural networks. This conceptually simple idea leads to improvements (10-20% and more in some cases) over the state of the art on most gaze estimation datasets. Code and appendix are available at https://github.com/vsingh-group/MeNets.

Acknowledgments. This work was supported by UW CPCP AI117924 and NSF CAREER award RI 1252725, and partially supported by R01 EB022883, R01 AG062336, R01 AG040396 and UW ADRC (AG033514). We thank Karu Sankaralingam for discussions, and Mona Jalal, Ronak Mehta, Ligang Zheng, Brandon M. Smith, Sukanya Venkataraman, Haoliang Sun, Xiaoming Zhang, Sathya Narayanan Ravi and Seong Jae Hwang for helping with various aspects of the experiments.

References

[1] Accuracy and precision test report x2-30 fw 1.0.1.

[2] R Harald Baayen, Douglas J Davidson, and Douglas M Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412, 2008.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Winter Conference on Applications of Computer Vision*, pages 1–10, 2016.

[4] Shumeet Baluja and Dean Pomerleau. Non-intrusive gaze tracking using artificial neural networks. In *Advances in Neural Information Processing Systems*, pages 753–760, 1994.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[6] Jinsoo Choi, Byungtae Ahn, Jaesik Parl, et al. Appearance-based gaze estimation using Kinect. In *International Conference on Ubiquitous Robots and Ambient Intelligence*, pages 260–261, 2013.

[7] A Clemotte, M Velasco, D Torricelli, R Raya, and R Ceres. Accuracy and precision of the tobii x2-30 eye-tracking under non ideal conditions. *Eye*, 16(3):2, 2014.

[8] Kim M Dalton, Brendon M Nacewicz, Tom Johnstone, et al. Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience*, 8(4):519–526, 2005.

[9] Eugene Demidenko. *Mixed models: theory and applications with R*. John Wiley & Sons, 2013.

[10] Haoping Deng and Wangjiang Zhu. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 3162–3171. IEEE, 2017.

[11] Ludwig Fahrmeir and Gerhard Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media, 2013.

[12] Zixing Fang and Robert L Bailey. Nonlinear mixed effects modeling for slash pine dominant height growth following intensive silvicultural treatments. *Forest science*, 47(3):287–300, 2001.

[13] Nathalie George, Jon Driver, and Raymond J Dolan. Seen gaze-direction modulates fusiform activity and its coupling with other brain areas during face processing. *Neuroimage*, 13(6):1102–1112, 2001.

[14] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006.

[15] Jarrod D Hadfield et al. Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm r package. *Journal of Statistical Software*, 33(2):1–22, 2010.

[16] Ahlem Hajjem, François Bellavance, and Denis Larocque. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6):1313–1328, 2014.

[17] Eakta Jain, Yaser Sheikh, Ariel Shamir, and Jessica Hodgins. Gaze-driven video re-editing. *ACM Transactions on Graphics*, 34(2):21, 2015.

[18] Li Jianfeng and Li Shigang. Eye-model-based gaze estimation by rgb-d camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 592–596, 2014.

[19] Knut KW Kampe, Chris D Frith, Raymond J Dolan, et al. Psychology: Reward value of attractiveness and gaze. *Nature*, 413(6856):589–589, 2001.

[20] Hyunwoo J Kim, Nagesh Adluru, Heemanshu Suri, Baba C Vemuri, Sterling C Johnson, and Vikas Singh. Riemannian nonlinear mixed effects models: Analyzing longitudinal deformations in neuroimaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2540–2549, 2017.

[21] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, et al. Eye tracking for everyone. In *CVPR*, pages 2176–2184, 2016.

[22] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[23] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, pages 1346–1353, 2012.

[24] Yin Li, Alireza Fathi, and James M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013.

[25] Mary J Lindstrom and Douglas M Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687, 1990.

[26] Gang Liu, Yu Yu, Kenneth A Funes-Mora, Jean-Marc Odobez, and Eyeware Tech SA. A differential approach for gaze estimation with calibration. Technical report, 2018.

[27] Feng Lu, Takahiro Okabe, Yusuke Sugano, et al. Learning gaze biases with head motion for head pose-free gaze estimation. *Image and Vision Computing*, 32(3):169–179, 2014.

[28] Feng Lu, Yusuke Sugano, Takahiro Okabe, et al. Head pose-free appearance-based gaze sensing via eye image synthesis. In *International Conference on Pattern Recognition*, pages 1008–1011, 2012.

[29] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Inferring human gaze from appearance via adaptive linear regression. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 153–160. IEEE, 2011.

[30] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):2033–2046, 2014.

[31] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3, 2010.

[32] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258. ACM, 2014.

[33] Kenneth Alberto Funes Mora and Jean-Marc Odobez. Gaze estimation from multimodal kinect data. In *Computer Vision and Pattern Recognition Workshops*, pages 25–30, 2012.

[34] 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*, page 21. ACM, 2018.

[35] Kevin A Pelphrey, James P Morris, and Gregory McCarthy. Neural basis of eye gaze processing deficits in autism. *Brain*, 128(5):1038–1048, 2005.

[36] Michael J Reale, Shaun Canavan, Lijun Yin, Kaoning Hu, and Terry Hung. A multi-gesture interaction system using a 3-d iris disk model for gaze estimation and an active appearance model for 3-d hand pointing. *IEEE Transactions on multimedia*, 13(3):474–486, 2011.

[37] Anthony J Robinson. An application of recurrent nets to phone probability estimation. *IEEE transactions on Neural Networks*, 5(2):298–305, 1994.

[38] George K Robinson. That blup is a good thing: the estimation of random effects. *Statistical science*, pages 15–32, 1991.

[39] Timo Schneider, Boris Schauerte, and Rainer Stiefelhagen. Manifold alignment for person independent appearance-based gaze estimation. In *International Conference on Pattern Recognition*, pages 1167–1172, 2014.

[40] Matan Sela, Pingmei Xu, Junfeng He, Vidhya Navalpakkam, and Dmitry Lagun. Gazegan-unpaired adversarial image generation for gaze estimation. *arXiv preprint arXiv:1711.09767*, 2017.

[41] Weston Sewell and Oleg Komogortsev. Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 3739–3744. ACM, 2010.

[42] Sheng-Wen Shih and Jin Liu. A novel approach to 3-d gaze tracking using stereo cameras. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):234–245, 2004.

[43] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv:1612.07828*, 2016.

[44] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, page 6, 2017.

[45] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. Predicting primary gaze behavior using social saliency fields. In *ICCV*, pages 3503–3510, 2013.

[46] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *CVPR*, pages 1821–1828, 2014.

[47] Ilya Sutskever, Geoffrey E Hinton, and Graham W Taylor. The recurrent temporal restricted boltzmann machine. In *Advances in Neural Information Processing Systems*, pages 1601–1608, 2009.

[48] Kar-Han Tan, David J Kriegman, and Narendra Ahuja. Appearance-based eye gaze estimation. In *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*, pages 191–195. IEEE, 2002.

[49] Roberto Valenti, Nicu Sebe, and Theo Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2):802–815, 2012.

[50] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[51] Oliver Williams, Andrew Blake, and Roberto Cipolla. Sparse and semi-supervised visual mapping with the s^3gp. In *CVPR*, volume 1, pages 230–237, 2006.

[52] 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. ACM, 2014.

[53] Hulin Wu and Jin-Ting Zhang. *Nonparametric regression methods for longitudinal data analysis: mixed-effects modeling approaches*, volume 515. John Wiley & Sons, 2006.

[54] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, et al. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *CVPR*, pages 2235–2244, 2015.

[55] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[56] Li-Qun Xu, Dave Machin, and Phil Sheppard. A novel approach to real-time non-intrusive gaze finding. In *BMVC*, pages 1–10, 1998.

[57] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[58] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Its written all over your face: Full-face appearance-based gaze estimation. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[59] 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.

[60] Xucong Zhang, Yusuke Sugano, Mario Fritz, et al. Appearance-based gaze estimation in the wild. In *CVPR*, pages 4511–4520, 2015.

[61] Hao Henry Zhou, Vikas Singh, Sterling C Johnson, Grace Wahba, Alzheimers Disease Neuroimaging Initiative, et al. Statistical tests and identifiability conditions for pooling and analyzing multisite datasets. *Proceedings of the National Academy of Sciences*, 115(7):1481–1486, 2018.

[62] Hao Henry Zhou, Yilin Zhang, Vamsi K Ithapu, Sterling C Johnson, Vikas Singh, et al. When can multi-site datasets be pooled for regression? hypothesis tests, 2-consistency and neuroscience applications. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4170–4179. JMLR. org, 2017.