MTGLS: Multi-Task Gaze Estimation with Limited Supervision

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

[1]Monash University [2]Indian Institute of Technology Ropar [3]University of Melbourne

{shreya.ghosh, munawar.hayat, abhinav.dhall}@monash.edu jarrod.knibbe@unimelb.edu.au

Abstract

Robust gaze estimation is a challenging task, even for deep CNNs, due to the non-availability of large-scale labeled data. Moreover, gaze annotation is a time-consuming process and requires specialized hardware setups. We propose MTGLS: a Multi-Task Gaze estimation framework with Limited Supervision, which leverages abundantly available non-annotated facial image data. MTGLS distills knowledge from off-the-shelf facial image analysis models, and learns strong feature representations of human eyes, guided by three complementary auxiliary signals: (a) the line of sight of the pupil (i.e. pseudo-gaze) defined by the localized facial landmarks, (b) the head-pose given by Euler angles, and (c) the orientation of the eye patch (left/right eye). To overcome inherent noise in the supervisory signals, MTGLS further incorporates a noise distribution modelling approach. Our experimental results show that MTGLS learns highly generalized representations which consistently perform well on a range of datasets. Our proposed framework outperforms the unsupervised state-of-the-art on CAVE (by $\sim 6.43\%$) and even supervised state-of-the-art methods on Gaze360 (by $\sim 6.59\%$) datasets.

1. Introduction

Eye movement provides important cues for nonverbal behaviour analysis, revealing insights into visual attention [38] and human cognition (emotions, beliefs, and desires) [69]. Such insights have proved valuable across multiple application scenarios, including gaming platforms [10], measuring student's engagement [41], determining driver's attention [22], etc. Recently, significant research progress has been made across a broad range of gaze related tasks, including gaze tracking [66, 53, 46, 35], pupil detection [58], pupil and iris segmentation [61], and gaze estimation [46, 45, 35, 75, 84]. Automatic gaze estimation still remains challenging due to the individuality of eyes, eye-head interplay, occlusion, image quality, and illumination conditions. The current state-of-the-art gaze estimation methods [45, 76, 75, 46] employ deep Convolution Neural
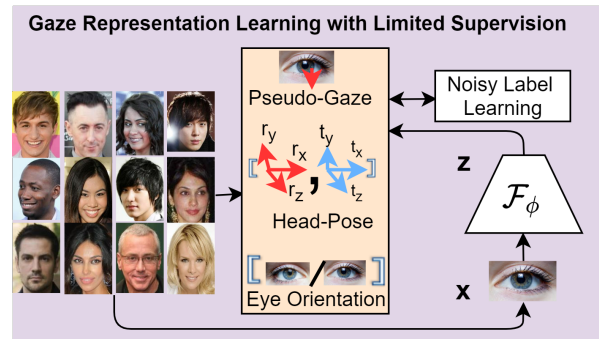

Figure 1. Overview of the proposed Multi-Task Gaze estimation framework with Limited Supervision (MTGLS). MTGLS learns a meaningful gaze embedding space guided by different auxiliary signals from 'in-the-wild' facial images curated from the web.

Networks (CNNs) to address these challenges. They generally require a lot of annotated data and person-specific calibration. Accurate annotation of gaze data is a complex, noisy, resource expensive, and time-consuming task. The annotation process can easily become noisy due to participants' non-cooperation (e.g., distraction), involuntary actions (e.g., eye blink), and measurement errors introduced by different data acquisition setups which limits merging multiple datasets. Research has sought to address these annotation challenges through data generation (e.g., synthesizing additional labelled data via gaze redirection [76] and GANs [67, 52]). However, these complex models are still constrained by the high computational complexity and quality of synthesized samples.

To this end, weakly/Semi/Self-Supervised Learning provides a promising paradigm as it enables learning from a large amount of readily available data without requiring any manual annotation efforts. To date, there are only a few existing works [76, 13] that learn low dimensional gaze representation guided by supervisory signals in the form of data augmentation, gaze redirection, and gaze zone. These existing works are either guided by weak auxiliary signals (e.g., gaze zone [13]), or are hard to scale since they require the generation of the eye or face image [76]. To learn a generalized and compact gaze representation, we propose an framework with a multi-task learning objective. Our rep-

resentation learning benefits from the progress made across other facial image analysis tasks (i.e. head-pose and facial landmarks), and is guided by three auxiliary gaze-relevant signals: (a) Line Of Sight (LOS) of the pupil (pseudo-gaze) determined by automatically localized facial landmarks [5], (b) head-pose in terms of Euler angles given by [1], and (c) orientation of eye patch (left/right eye). These auxiliary signals provide complementary supervision to MT-GLS. Our learned representations are highly generic and transferable. Unlike the existing methods [45, 75], which require person-specific calibration to adapt the label space, our approach achieves competitive results without such calibration, which is highly desirable in wearable eye devices, especially in augmented and virtual reality glasses [54, 7]. Moreover, our extensive empirical evaluations demonstrate that the representations learned by MTGLS generalize well across a range of datasets for different eye gaze estimation tasks. We further show that MTGLS can be used for real-time gaze inference applications, especially for driver's engagement. The **main contributions** of the paper are:

- MTGLS, a novel multi-task gaze estimation framework with limited supervision for rich visual representation learning, guided by complementary information in the form of line-of-sight of the pupil, head-pose, and orientation of eye patch.

- To overcome the impact of inherent noise in the auxiliary label space, we incorporate a probability distribution modelling based noise correction strategy which further boosts the performance of the model.

- We demonstrate the effectiveness of the proposed approach by extensive evaluations on downstream gaze estimation tasks on four benchmark datasets (i.e. CAVE, MPII, Gaze360, and DGW). The proposed approach not only outperforms unsupervised state-of-the-art methods on CAVE ($3.42° \rightarrow 3.20°$, $\sim 6.43\%$), but also performs favorably on the supervised state-of-the-art method on Gaze360 ($13.80° \rightarrow 12.89°$, $\sim 6.59\%$) datasets.

2. Related Work

Gaze Estimation. A thorough analysis of gaze estimation literature is mentioned in a recent survey [21]. Traditionally, gaze estimation methods are based on appearance features [40, 86, 29], or geometric attributes such as iris and corner detection [59]. Mostly different facial regions are widely used as input Region of Interest (RoI) in the literature such as whole face [11, 13, 83, 45], eye region [33, 82, 47, 46]. These features are then mapped to the screen coordinates using a learning algorithm including support vector regression [50] or deep neural nets (DNNs) models [37, 82, 83, 84, 33, 15]. Among the DNNs based methods, supervised learning methods [82, 33, 84]

have been most popular and successful, but they require a large amount of annotated training data. Some studies [52, 6, 80, 48, 45, 34, 25] explore gaze estimation in unsupervised and weakly/self/semi-supervised settings to reduce the data annotation burden. These approaches are mainly based on conditional random field [6], unsupervised gaze target discovery [80], few-shot learning [45, 75], weakly-supervised [36], and representation learning [76, 13]. Apart from these methods, Bayesian learning paradigms are also explored in the context of generalization [68, 67]. Nowadays there is an increasing trend to inject domain knowledge in terms of certain heuristics for gaze estimation such as the correlation between head-pose and gaze direction [45, 85], two eye symmetry [12], facial landmark based gaze modelling [68] and uncertainty in gaze modelling [36, 35].

Gaze Datasets. The performance of existing gaze estimation methods is generally dependent upon the amount of annotated data used during training. While several efforts have been made to collect and annotate gaze data in the previous decade [50, 17, 31, 84, 83, 15, 35, 19, 43, 79, 44], gaze data collection has multiple challenges, including the requirement of specialized hardware setup, inconsistency amongst different acquisition apparatus which limits merging multiple datasets, and resource-expensive data annotation process. Some approaches turn towards synthetic data generation to increment the size of the training data for gaze estimation task [18, 28, 70, 75, 49]. While these methods can enhance the training set size, the quality of their synthesized samples differ from their original counterparts. More specifically, these methods inherit several domain gaps, including content preservation (e.g., eye shape, eye appearance), network complexity, and resource expensiveness. Additionally, the gaze re-direction methods may have limited gaze direction samples while extrapolating on a new subject. These domain gaps are quite challenging and difficult to eliminate.

While gaze estimation still remains a challenging feat, limited by the capacity of available annotated data to train supervised models, significant research progress has been made across other facial analysis tasks, including face identification, classification, landmark localization, and head-pose estimation. In this work, we propose to leverage from existing facial image analysis models and distill their knowledge for gaze estimation, by making use of readily available facial image data. We demonstrate that the gaze representation learned by our method generalizes well for gaze estimation tasks across a range of datasets [84, 50, 35, 22], including those collected in controlled environments [50], and more challenging unconstrained datasets [84, 35]. Moreover, owing to the strong generalization capabilities of our learned representations, MTGLS can align the label space manifold of auxiliary and downstream gaze estimation label space which conse-
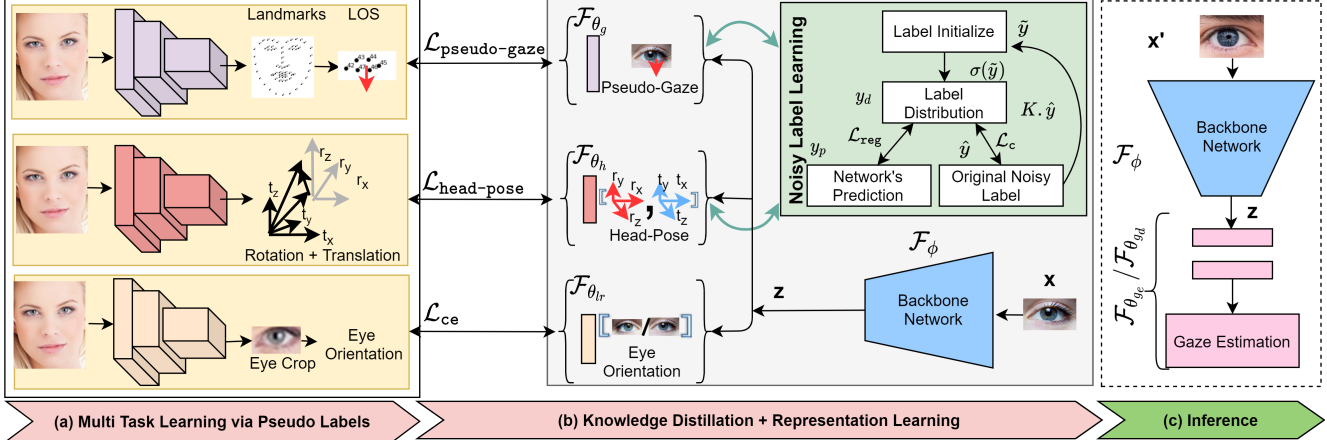
Figure 2. An overview of our proposed MTGLS framework. From left to right, we show *(a) Multi-task Learning via Pseudo Labels:* MTGLS learns gaze representations with limited supervision, by distilling knowledge from off-the-shelf deep models developed for facial image modelling. Here, MTGLS utilizes LOS, head-pose and eye orientation as auxiliary signals (Refer Sec. 3.1.1); *(b) Knowledge Distillation and Representation Learning:* Our gaze estimation framework consists of the backbone network and noisy label learning module (Refer Sec. 3.1.2). During the representation learning process, MTGLS considers the impact of inherent noise in the assigned soft label space; *(c) Inference:* For downstream adaptation, we use Linear Probing (LP), Fine-Tuning (FT) to show the generalization and transferability of the learnt features across different tasks.

quently helps in person-specific adaptation. Thus, it is quite promising for gaze calibration tasks having potential applications in AR/VR glasses where unsupervised gaze calibration is highly desirable. *Please note that the proposed method can be considered as self-supervised. We have used the term limited supervision as the pseudo labels are obtained from off-the-shelf models [65, 1].*

3. Method

Human eye gaze depends on several factors, including the relative location of the pupils, head-pose, and appearance. A strong feature representation of the eye is crucial for robust gaze modelling. Instead of limiting ourselves to scarcely available gaze annotated data, we distill knowledge from a vast amount of non-annotated data of human faces, which is easily available on the web. Our proposed framework, MTGLS is guided by three auxiliary signals: LOS of the pupil (pseudo-gaze), head-pose, and orientation of eye patch (i.e. if the eye patch belongs to the left or right eye). The proposed multi-objective learning formulation maps the input eye patch image to the optimal output representation manifold, which can further be utilized for other gaze estimation tasks. The visual illustration of MTGLS is shown in Fig. 2 and an algorithm is shown in Alg. 2.

3.1. Gaze Representation Learning

Preliminaries. Lets assume our MTGLS model $\mathcal{F}$ consists of feature embedding parameter ϕ and label space parameters θ. An input image $\mathbf{x}$ is mapped to the feature embedding space by $\mathcal{F}_\phi : \mathbf{x} \rightarrow \mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^d$. Consequently, the features $\mathbf{z}$ are mapped to three auxiliary output spaces $\mathbf{g} \in \mathbb{R}^3$, $\mathbf{h} \in \mathbb{R}^6$, and $\mathbf{lr} \in \mathbb{R}^2$, respectively by three

branches: $\mathcal{F}_{\theta_g} : \mathbf{z} \rightarrow \mathbf{g}$, $\mathcal{F}_{\theta_h} : \mathbf{z} \rightarrow \mathbf{h}$, and $\mathcal{F}_{\theta_{lr}} : \mathbf{z} \rightarrow \mathbf{lr}$. Here, $\mathbf{g}$, $\mathbf{h}$, and $\mathbf{lr}$ are the pseudo-gaze, head-pose, and eye orientation, respectively. The model $\mathcal{F}$ therefore consists of joint learning of three sub-networks: $(\mathcal{F}_\phi \circ \mathcal{F}_{\theta_g})$, $(\mathcal{F}_\phi \circ \mathcal{F}_{\theta_h})$, and $(\mathcal{F}_\phi \circ \mathcal{F}_{\theta_{lr}})$. MTGLS is trained on dataset $\mathcal{D}$ which consists of n image-label pairs denoted as: $\{\mathbf{x_i}, \mathbf{y_i}\}_n$, where $\mathbf{y_i} = \{\mathbf{g_i}, \mathbf{h_i}, \mathbf{lr_i}\}$ are the auxiliary task pseudo-labels, estimated by off-the-shelf facial image analysis models. The learned features by $\mathcal{F}$ are transferred and adapted for two gaze estimation tasks i.e. $\mathcal{F}_\phi \circ \mathcal{F}_{\theta_{g_e}}$ for gaze estimation and $\mathcal{F}_\phi \circ \mathcal{F}_{\theta_{g_d}}$ for driver's gaze zone estimation, using separate tasks specific modules i.e., $\mathcal{F}_{\theta_{g_e}}$ and $\mathcal{F}_{\theta_{g_d}}$.

3.1.1 Auxiliary Tasks

Self-training via knowledge distillation aims to learn rich representative features from a small initial annotated set to a large scale unlabeled instances with soft label assignment [72]. Knowledge distillation [30] is widely used in literature for representation learning [9, 20, 23]. MTGLS also adopt this soft label assignment strategy on large scale non-annotated data. Several prior works have reported the performance enhancement with the combination of representation learning and knowledge distillation [14, 42]. Along these lines, we propose the following auxiliary tasks aligned with gaze estimation for limited supervision.

a) Pseudo-Gaze: Prior literature on unsupervised and weakly supervised gaze representation learning utilize sparse gaze zone [13], gaze redirection [76] and looking at each other [36] soft labelling strategies for limited-supervision. Similarly, we define landmark-based LOS of the pupil as an auxiliary task and term it as *pseudo-gaze*.

Algorithm 1 Noisy Label Learning (NLL) for MTGLS

Require: $\mathcal{F}_\phi, \mathcal{F}_{\theta_g}, \mathcal{F}_{\theta_h}, \mathbf{Y_d}, \mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}}, \mathbf{y_d},$ and $\mathbf{y_p}$
1: **for e** epochs **do** ▷ Training
2: $\quad\hat{\mathbf{y}} \leftarrow$ noisy labels
3: $\quad\tilde{\mathbf{y}} \leftarrow K.\hat{\mathbf{y}}$
4: $\quad\mathbf{y_d} \leftarrow \sigma(\tilde{\mathbf{y}})$
5: $\quad\mathbf{y_p} \leftarrow \mathcal{F}_\phi \circ \mathcal{F}_\theta(\mathbf{x})$ ▷ $\mathcal{F}_\theta \in \{\mathcal{F}_{\theta_g}, \mathcal{F}_{\theta_h}\}$
6: $\quad\mathcal{L} = \mathcal{L}_{\text{reg}}(\mathbf{y_p}, \mathbf{y_d}) + \mathcal{L}_{\text{c}}(\mathbf{y_d}, \hat{\mathbf{y}}) + ||\omega||_2^2$
7: $\quad\{\phi, \theta, \mathbf{Y_d}\} \leftarrow \{\phi, \theta, \mathbf{Y_d}\} - \nabla_{\{\phi, \theta, \mathbf{Y_d}\}}\mathcal{L}$
8: **end for**

First, the eye region is registered via FaceX-Zoo's state-of-the-art landmark detection method [65]. We utilize the detected pupil center and eye landmarks to compute the gaze vector. The estimated gaze vector is from the 3-D eyeball center to the pupil location where the eyeball center is approximated by the eye landmarks. We follow the conventional details of this approximation technique from prior works [47, 3]. We use the cosine similarity based gaze direction loss (similar to the standard loss used in this domain [79]) to train the ($\mathcal{F}_\phi \circ \mathcal{F}_{\theta_g}$) sub-network on pseudo-gaze labels, defined by, $\mathcal{L}_{\text{pseudo-gaze}} = \frac{\mathbf{g}}{||\mathbf{g}||_2} \cdot \frac{\mathbf{g}'}{||\mathbf{g}'||_2}$, where, $\mathbf{g}$ and $\mathbf{g}'$ respectively denote the pseudo-gaze (ground truth) and predicted gaze in auxiliary label space.
b) Head-Pose: The head-pose also carries important information for gaze estimation, since there exists a strong correlation between eye and head once a person looks at something [82, 45, 85]. We use the state-of-the-art head-pose estimator 'img2pose' [1] to infer pseudo 6-D head-pose labels from the input face (Refer Fig. 2 (a)). Given a face image $\mathbf{x}$, we estimate 6-DoF pose from the img2pose model. We minimize the mean-squared error between predicted head-pose $\mathbf{h}'$, and pseudo head-pose (ground truth) $\mathbf{h}$ to train sub-network ($\mathcal{F}_\phi \circ \mathcal{F}_{\theta_h}$). The loss is defined as $\mathcal{L}_{\text{head-pose}} = MSE(\mathbf{h}, \mathbf{h}')$.
c) The Orientation of Eye Patch: We further take benefit from the eye orientation information to strengthen our representation learning. To this end, we define an auxiliary task in terms of eye orientation i.e., whether the eye patch belongs to the left or right eye of a subject. We minimize the binary cross-entropy loss ($\mathcal{L}_{\text{ce}}$) for left-right eye patch prediction corresponding to the ($\mathcal{F}_\phi \circ \mathcal{F}_{\theta_{lr}}$) sub-network.

3.1.2 Noisy Label Learning

Apart from the eye orientation labels, the other auxiliary signals (i.e. pseudo-gaze and headpose labels) are based upon off-the-shelf deep models [5, 1], and contain noise. To overcome inherent noise in auxiliary label space, we incorporate Noisy Label Learning (NLL). Our NLL module is inspired by [73] and is summarized in Alg. 1. Let's define $\tilde{\mathbf{y}} = K.\hat{\mathbf{y}}$, where, $\hat{\mathbf{y}}$ are the noisy labels and K is a

scalar. For our experiments, we empirically choose $K = 10$ similar to [73]. We further define the label distribution $\mathbf{y_d} = \sigma(\tilde{\mathbf{y}})$, where, σ is the sigmoid operator. Here, $\mathbf{y_d}$ attempts to model the unknown noise-free label distribution for inputs $\mathbf{X} : \{\mathbf{x_i}, \forall\, 1 \le i \le n\}$ via learning. Let, $\mathbf{Y_d}$ be the parameters of the label distribution $\mathbf{y_d}$ defined as $\mathbf{Y_d} = \{\mathbf{Y_d^i}, \forall\, 1 \le i \le n\}$. During each iteration of the training process, the parameters of $\mathbf{Y_d}$ (i.e., $\mathbf{Y_d^i}$) are updated along with the network parameters (i.e. ϕ and θ) via back-propagation process [56, 73]. We also hypothesize that $\tilde{\mathbf{y}}$ assists $\mathbf{y_d}$ to be a valid probability distribution. Thus, $\mathbf{y_d}$ is initialized in terms of $\tilde{\mathbf{y}}$ (i.e. $\sigma(K.\hat{\mathbf{y}})$) instead of random initialization. Hence, after initialization $\mathbf{y_d} \approx \hat{\mathbf{y}}$. The overall learning process is guided by three loss functions:
Regression Loss ($\mathcal{L}_{\text{reg}}$). is given by KL divergence between backbone network's prediction ($\mathbf{y_p}$) and label distribution ($\mathbf{y_d}$). i.e. $\mathcal{L}_{\text{reg}} = KL(\mathbf{y_p}||\mathbf{y_d})$
Compatibility Loss ($\mathcal{L}_{\text{c}}$). ensures that the label distribution ($\mathbf{y_d}$) are not divergent from the original noisy distribution ($\hat{\mathbf{y}}$), i.e. $\mathcal{L}_{\text{c}} = MSE(\mathbf{y_d}, \hat{\mathbf{y}})$
Regularization. ℓ_2 regularizer is employed to ensure that the network will not stop updating when the predictions become close to the label distribution ($\mathbf{y_d}$). It is defined as $||\omega||_2^2$, where, ω is the model parameter.
Overall NLL Loss ($\mathcal{L}_{\text{NLL}}$). The overall loss for NLL module is given by: $\mathcal{L}_{\text{NLL}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{c}} + ||\omega||_2^2$

3.1.3 Overall MTGLS Loss

Algorithm 2 summarizes the training procedure of the proposed MTGLS. Our overall learning is guided by the following objective functions:

$$\mathcal{L} = \mathcal{L}_{\text{pseudo-gaze}} + \mathcal{L}_{\text{head-pose}} + \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{NLL}}$$

Note that, we apply $\mathcal{L}_{\text{NLL}}$ on the auxiliary label space of pseudo-gaze and head-pose only as the eye patch is extracted from face using landmark detector is not prone to error (empirically, $\sim 100\%$ accurate). On the other hand, if the pre-trained models [65, 1] are more reliable, the dependency on NLL will reduce.

3.2. Evaluation Protocol of Limited Supervision

Our proposed MTGLS learns eye representation from the input eye patches and their pseudo-labels distilled from facial models in terms of pseudo-gaze, head-pose, and eye orientation. Following the standard evaluation protocols, we also adopt Linear Probing (LP) on frozen features [78, 26, 9] and Fine-Tuning (FT) the features on downstream tasks [9] (as illustrated in Fig. 2 Inference module). For LP, we use data augmentation in terms of random resize crops and horizontal flipping during training phase. Please note that during horizontal flipping the ground truth gaze

Algorithm 2 Training Procedure for MTGLS

Require: $\mathcal{F}_\phi, \mathcal{F}_{\theta_g}, \mathcal{F}_{\theta_h}, \mathcal{F}_{\theta_{lr}}$, and $\mathcal{D}$
1: **for e** epochs **do**   ▷ MTGLS Training
2:      $\mathbf{g} \leftarrow \text{LOS}(\mathbf{x})$ ▷ Localize Facial Landmarks
3:      $\mathbf{h} \leftarrow \text{img2pose}(\mathbf{x})$ ▷ Estimate head-pose
4:      $\mathbf{lr} \leftarrow \text{process}(\mathbf{x})$ ▷ Get left-right patch
5:      $\mathbf{z} \leftarrow \mathcal{F}_\phi(\mathbf{x})$
6:      $\mathbf{g}' \leftarrow \mathcal{F}_{\theta_g}(\mathbf{z}), h' \leftarrow \mathcal{F}_{\theta_h}(\mathbf{z})$, and $\mathbf{lr}' \leftarrow \mathcal{F}_{\theta_{lr}}(\mathbf{z})$
7:      $\mathcal{L}_0 = \mathcal{L}_{\text{pseudo-gaze}}(\mathbf{g}, \mathbf{g}') + \mathcal{L}_{\text{head-pose}}(\mathbf{h}, \mathbf{h}') +$
       $\mathcal{L}_{\text{ce}}(\mathbf{lr}, \mathbf{lr}') + \mathcal{L}_{\text{NLL}}(\mathbf{g}, \mathbf{g}') + \mathcal{L}_{\text{NLL}}(\mathbf{h}, \mathbf{h}')$
8:      $\{\phi, \theta_g, \theta_h, \theta_{lr}\} \leftarrow \nabla_{\{\phi, \theta_g, \theta_h, \theta_{lr}\}} \mathcal{L}_0$
9: **end for**
10: **for e** epochs **do**   ▷ Gaze Estimation
11:      $\mathbf{z} \leftarrow \mathcal{F}_\phi(\mathbf{x}')$ ▷ $x' \in \mathcal{D}$
12:      $\mathbf{y}' \leftarrow \mathcal{F}_{\theta_g}(\mathbf{z})$ ▷ $\theta_g \in \{\theta_{g_e}, \theta_{g_d}\}$
13:      $\mathcal{L}_1 = \mathcal{L}_{\text{LP}}(\mathbf{y}, \mathbf{y}')$ ▷ $\mathcal{L}_{\text{LP}} \in \{\mathcal{L}_{\text{gaze}}, \mathcal{L}_{\text{gaze-zone}}\}$
14:      $\{\theta_g\} \leftarrow \nabla_{\{\theta_g\}} \mathcal{L}_1$ ▷ $\theta_g \in \{\theta_{g_e}, \theta_{g_d}\}$
15: **end for**

labels change its sign. For this purpose, weights ϕ of the model are frozen and only the parameters of $\mathcal{F}_\theta$ are updated. During this gaze adaptation training process, the dataset and task specific samples $\mathbf{x}' \in \mathcal{D}$ are used to adjust the output manifold (i.e. respective label spaces). This adjustment process is enforced by applying the appropriate loss function, i.e., gaze direction loss ($\mathcal{L}_{\text{gaze}}$) when the output space is a 3-D gaze vector or cross-entropy loss ($\mathcal{L}_{\text{gaze-zone}}$) when the output space is a gaze zone, where $\mathcal{L}_{\text{gaze}} = \frac{\mathbf{y}}{||\mathbf{y}||_2} \cdot \frac{\mathbf{y}'}{||\mathbf{y}'||_2}$ and $\mathcal{L}_{\text{gaze-zone}} = CE(\mathbf{y}, \mathbf{y}')$ given $\mathbf{y}$ and $\mathbf{y}'$ are original and predicted labels. The overall linear probing process can be expressed as the following optimization problem, where, $\theta_g \in \{\theta_{g_e}, \theta_{g_d}\}$: $\min_{\phi, \theta_{\mathbf{g}}} \mathbf{E}_{\mathbf{x}', \mathbf{y} \sim \mathcal{D}}[\mathcal{L}_{\text{EP}}(\mathcal{F}_\phi \circ \mathcal{F}_{\theta_g}(\mathbf{x}'), \mathbf{y})]$

For FT, the backbone network is initialized with pre-trained weight and it is updated for the downstream task. The downstream task for drivers' gaze zone estimation is a 9-class classification problem. Thus, we additionally evaluate the representative features' quality with a weighted nearest neighbour classifier (k-NN) [9, 71]. We freeze the weights of the pre-trained model and extract the penultimate layer's feature for training. The weighted k-NN classifier then performs a similarity matching operation in the feature space to vote for the final label. Our empirical analysis suggests that it worked for 10 NN over several runs. Evaluation paradigm using weighted k-NN is simple and it does not require any complicated setup such as hyperparameter tuning, data augmentation etc.

4. Experiments and Results

We comprehensively compare our method on four publicly available benchmark datasets for different tasks from both quantitative (Sec. 4.2) and qualitative (Sec. 4.3) per-

spectives. Additionally, we perform extensive ablation studies to investigate the contributions of different individual components of the proposed pipeline (Sec. 4.2.3).

4.1. Experimental Protocols

Datasets. We evaluate MTGLS on four benchmark datasets: **CAVE** [50], **MPII** [84], **Gaze360** [35], and **DGW** [22]. CAVE [50] contains 5,880 images of 56 subjects collected in a constrained lab environment with 21 different gaze directions and head-poses for each person. MPII [84] gaze dataset has 213,659 images collected from 15 subjects during natural everyday activity in front of a laptop over a three-month duration. Gaze360 [35] is a large-scale dataset collected from 238 subjects in unconstrained indoor and outdoor settings with a wide range of head-poses. DGW [22] is a large scale driver gaze estimation dataset of 338 subjects collected in an unconstrained 'inside a car' scenario during different times of the day. For representation learning, we also used cleaned version of VG-GFace2 [8] containing 3M images of 8,631 identities and cleaned version of MS-Celeb-1M [24] dataset having 1M identities.

Settings. To be consistent with the previous methods [46, 45, 75, 76, 13], we follow the same experimental protocol as mentioned in the respective baseline papers for each dataset. The training set for these datasets is used for downstream adaptation and test set for inference. Specifically, we follow 'leave-one-person-out' for MPII, cross-validation for CAVE, and Train-Val-Test partitions for the Gaze360 and DGW datasets for our quantitative analysis. For validating few-shot learning, we report the person-specific gaze errors (in °) for a range of k calibration samples by following the evaluation protocol described in [45, 75].

Implementation Details. For images to be in common virtual camera points, we pre-process the input data based on the data-normalization procedure in [81]. For our experiments, we use ResNet-50 as a backbone containing 4 residual blocks with 64, 128, 256 and 512 filters. After training the model $\mathcal{F}$, we freeze the backbone network ($\mathcal{F}_\phi$) until the penultimate layer, i.e., the global average pooling layer. As shown in Fig. 2 (inference module), the feature embedding $\mathbf{z}$ is passed through two FC layers (512 and 256 neurons) for linear probing. The linear probing includes dataset-specific adaptation for gaze estimation and driver gaze zone estimation tasks. The hyperparameters are as follows: we use an SGD optimizer with an initial learning rate of 0.01 having momentum 0.9 and a weight decay of $1 \times e^{-4}$. The batch size is set to 32 with the number of epochs 500. For downstream tasks, we use the SGD optimizer for 20-25 epochs with a learning rate of 0.0001, momentum of 0.9 and weight decay of $1 \times e^{-4}$. Further, the value of the hyperparameter K (see Sec. 3.1.2) is tuned on the validation set during MTGLS training and is fixed to 10 as in [73].

Table 1. **Linear Probing (LP) and Fine-Tuning (FT) on CAVE, MPII and Gaze360.** Comparison of different pre-training methods on CAVE, MPII, and Gaze360 datasets. The error is reported in terms of error (in °) ± standard deviation.

Method	Pretrain	Dataset	Gaze Error (in °)	
			LP	FT
MTGLS	VGGFace2	CAVE	3.30 ± 0.97	3.16 ± 0.82
	MS-Celeb-1M	CAVE	3.24 ± 0.90	**3.00 ± 0.76**
	CAVE	CAVE	**3.20 ± 1.50**	3.14 ± 1.27
	VGGFace2	MPII	4.41 ± 0.98	4.23 ± 0.94
	MS-Celeb-1M	MPII	4.39 ± 0.89	**4.07 ± 0.95**
	MPII	MPII	**4.21 ± 1.90**	4.13 ± 1.31
	VGGFace2	Gaze360	12.90 ± 3.85	12.87 ± 3.69
	MS-Celeb-1M	Gaze360	13.00 ± 3.54	**12.83 ± 3.30**
	Gaze360	Gaze360	**12.89 ± 5.72**	12.88 ± 5.60

4.2. Quantitative Results

4.2.1 Comparison with Existing Methods

Following the standard evaluation protocols [50, 84, 35, 22], our empirical results on different datasets are presented in Table 1, 2 and 3. In **Table 1**, we compare the performance of MTGLS pre-trained on different datasets. Among the datasets, CAVE, MPII and Gaze360 have gaze labels; additionally, we explore representation learning from large scale facial datasets i.e. VGGFace2 and MS-Celeb-1M. Please note that the pre-training is performed without using any ground truth gaze labels. We report the results by following traditional evaluation protocol i.e., Linear Probing (LP) and Fine-tuning (FT). The gaze error is presented in error ± standard deviation format. For LP, pre-training on the same datasets without any label is performing better than the other settings as the downstream distribution is more aligned. However, the representation learned on large scale facial data (i.e. 5M images from MS-Celeb-1M) have less standard deviation in all cases. For FT, the model pre-trained on MS-Celeb-1M is more adaptive as compared to the other settings due to large scale data usage. Similarly in **Table 2**, the results suggest that our method consistently achieves competitive performance as compared to the state-of-the-art methods. *Please note that for a fair comparison with the previous works (Table 2), we only consider the models pre-trained on the same dataset's training partition without their gaze labels.* On CAVE dataset, MTGLS outperforms unsupervised gaze estimation [76] by ∼ 6.43% (3.42° →3.20°). Unlike [76], which adapts data augmentation via gaze redirection learning for unsupervised feature learning, our proposed MTGLS takes advantage of weak-supervision in the form of pseudo-labels obtained from off-the-shelf facial models [1, 65]. We further note from results in Table 2 that our method even performs better than the few of the supervised state-of-the-art methods on MPII [46, 68, 11, 15] when pre-trained on MPII dataset without labels and Gaze360 (13.80° →12.89°, ∼ 6.59%) datasets. We attribute our gain to the rich knowledge dis-

Table 2. **Gaze Estimation.** Comparison of different gaze estimation methods on CAVE, MPII, and Gaze360 datasets. * *denotes supervised methods.* MTGLS performs favorably against existing state-of-the-art, and achieves gains upto ∼ 6.43% against unsupervised SOTA on CAVE dataset, and ∼ 6.59% for supervised SOTA on Gaze360 dataset.

Dataset	Method (Pretrain on same data)	Gaze Error (in °)
CAVE	Jyoti et al. [33]*	**2.22**
	Park et al. [46]*	3.59
	Yu et al. [76]	3.42
	MTGLS	3.20
	MTGLS (MS-Celeb-1M)	3.00
MPII	Park et al. [46]*	4.50
	Wang et al. [68]*	4.30
	Cheng et al. [11]*	4.10
	Fischer et al. [15]*	4.30
	MTGLS	4.21
	MTGLS (MS-Celeb-1M)	**4.07**
Gaze360	Kellnhofer et al. [35]*	13.80
	MTGLS	12.89
	MTGLS (MS-Celeb-1M)	**12.83**

tilled from the models trained on generic face-relevant tasks (i.e. landmark localization [65], 6-D head-pose estimation [1]). Moreover, MTGLS learns highly generic, and transferable representation as it aligns the auxiliary output manifold with the actual gaze estimation task via multi-task learning. Further, we evaluate MTGLS on a different yet relevant task of driver gaze zone estimation on DGW dataset [22] to show the adaptability of MTGLS across

Table 3. **k-NN, Linear Probing (LP) and Fine-Tuning (FT) results on DGW dataset.** Performance comparison with state-of-the-art (SOTA) methods on driver gaze zone estimation on DGW dataset. * *denotes supervised methods.* Note that even after learning with limited supervision, MTGLS performs favorably against most of the supervised methods on DGW dataset.

| Method	Validation Accuracy (%)	Test Accuracy (%)
Vasli et al. [62]*	52.60	50.41
Tawari et al. [57]*	51.30	50.90
Fridman et al. [16]*	53.10	52.87
Vora et al. [63] (Alexnet face)*	56.25	57.98
Vora et al. [63] (VGG face)*	58.67	58.90
SqueezeNet [32]*	59.53	59.18
Ghosh et al. [22]*	60.10	60.98
Inception V3 [55]*	67.93	68.04
Vora et al. [64]*	67.31	68.12
ResNet-152 [27]*	68.94	69.01
Yoon et al. [74] (Face + Eyes)*	70.94	71.20
Stappen et al. [51]*	71.03	71.28
Lyu et al. [39]*	85.40	81.51
Yu et al. [77]*	80.29	82.52
MTGLS (k-NN)	63.90	65.35
MTGLS (LP)	73.83	75.10
MTGLS (FT)	79.63	80.05
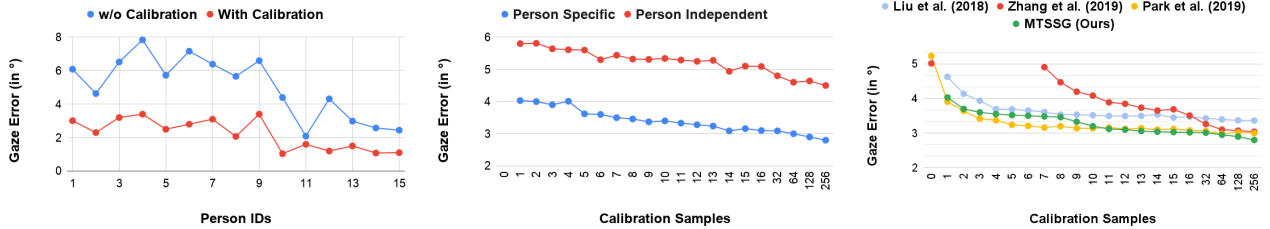
Figure 3. **Left**: Comparison of with and without person-specific calibration on MPII dataset. **Middle**: Comparison with person-specific and person-independent setting. **Right**: Comparison with state-of-the-art few-shot learning approaches. (Refer Sec. 4.2.2)

tasks. The results are shown in Table 3. Please note that the compared methods in Table 3 are supervised. MTGLS performs favorably against [62, 57, 16, 64, 63, 77, 39]. Some of the compared methods [77, 39] use additional information in the form of body pose and other cues for driver gaze estimation which are more aligned towards the driver monitoring task. Our simple eye patch-based representation learning (i.e. MTGLS (FT)) outperforms the baseline [22] by 19.53% and 19.08% margin on validation and test set, respectively. Even with weighted k-NN on the features of the penultimate layer, our method performs better than the supervised baseline [22] (3.80% on the validation set and 4.37% on the test set). These results indicate that MTGLS learns a highly generic and transferable eye representation using complementary auxiliary tasks.

4.2.2 Person Specific and Few-Shot Gaze Estimation

Person-specific adaptation can further improve gaze estimation performance and has therefore gained a lot of research attention [75, 45, 54, 7] due to its potential applications in AR/VR industry.

With and w/o Person Specific Calibration [Fig. 3 (Left)]. Similar to [45, 75], we conduct person-specific calibration on the MPII dataset having 15 identities. We randomly choose 200 calibration samples for each identity for adaptation and evaluate the performance on the remaining samples. The results in terms of gaze error (in °) corresponding to different person IDs are shown in Fig. 3 (Left). The results suggest that the angular error reduces significantly with the calibration. The average angular error reduces to 2.22° (with calibration) from 5.02° (without calibration). Please note that this average is not plotted in Fig. 3 (Left).

Person specific vs. Person Independent Calibration [Fig. 3 (Middle)]. Apart from person-specific adaptation, we perform person-independent adaptation to show the generalization capabilities of the MTGLS inspired by [45]. The underlying hypothesis behind this is that the latent representation is different across subjects due to the individuality of the eyes. For the person independent setting, we randomly choose 1-256 calibration samples for adaptation and evaluate the performance on the disjoint test set. For person dependent setting, we randomly choose 1-256 calibra-

tion samples for adaptation and evaluate the performance on the disjoint remaining person-specific test set. We report the average over the identities in the figure. The results are depicted in Fig. 3 (Middle), representing the number of calibration samples (x-axis) and corresponding gaze error ° (y-axis). From these results, we observe that for all of the calibration points varying from 1-256, the person-specific adaptation performs better than the person-independent setting. Another intuitive trade-off observed from the graph is that the difference between person-independent and person-specific adaptation decreases, as the number of samples increases. For example, between 1 and 256 samples the difference between two settings reduces by ~ 7% (1 sample: 37.77% [5.80° →4.03°], 256 samples: 30.51% [4.50° →2.8°]). This experiment further demonstrates the capabilities of MTGLS to learn generic representation.

Few-Shot Gaze Estimation [Fig. 3 (Right)]. Few-shot person-specific gaze estimation has recently gained popularity [45, 75], as it enables person-specific gaze calibration using only a few samples. Here, we investigate the transferable characteristics of our features for few-shot gaze estimation. We follow the same evaluation protocol as in [45, 75] for a fair and direct comparison with existing methods. Among 15 subjects of the MPII dataset, the last 500 images are reserved for testing, while k calibration samples are randomly selected for adaptation purposes. The results are shown in Fig. 3 (Right) indicate that our method performs favorably against the existing works (for ≥ 12), thus demonstrating the strength of our learned representations for low-annotation data regimes. The results for compared methods are adapted from [45].

4.2.3 Ablation Studies

Contribution of the Auxiliary Tasks. To study the impact of different auxiliary tasks on the overall performance of MTGLS, we progressively integrate them into our framework by keeping other parts fixed. We assess the impact of each auxiliary task and their different combinations during the representation learning on the CAVE dataset [50]. From the results in Table 4, it can be observed that the contribution of head-pose (3.52°) and pseudo eye gaze (3.40°) is greater as compared to the eye orientation (3.67°), in-

dicating that head-pose and pseudo-gaze provide more relevant information for gaze estimation as compared to eye orientation. Similarly, when the model is trained with both head-pose and pseudo-gaze, the effect is less as compared with the other combinations. Remarkably, the best performance is achieved when all of the three auxiliary tasks are combined i.e., by 2.12% ($3.30° \rightarrow 3.23°$). These empirical evaluations establish the complimentary strength and individual importance of each of the proposed auxiliary tasks in our framework.

Contribution of the Noise Removal in Label Space. Since two of our supervisory signals (pseudo-gaze and 6-D head-pose labels) contain inherent noise, MTGLS incorporates a noise correction strategy (see Sec. 3.1.2) based upon label distribution modelling. Our experiments reveal that when we integrate the NLL framework into MTGLS, the performance is significantly enhanced $\sim$ 7.12% ($3.23°$ $\rightarrow 3.00°$) thus suggesting the importance of this module.

Choice of Landmark Localization Model. Two of our auxiliary learning signals come from off-the-shelf face analysis models developed for landmark localization [65] and head-pose estimation [1]. Here, we evaluate the impact on the performance of MTGLS by considering other teacher models for localizing facial landmarks for gaze estimation. Our experiments on MPII dataset reveal that our landmark guided pseudo gaze method is more accurate than Open-Face 2.0 tracker [4] ($\sim$ 4.39% i.e. $9.10° \rightarrow 8.7°$) and Open-Face tracker [2] ($\sim$ 12.65% i.e. $9.96° \rightarrow 8.7°$). These empirical evaluations justify our choice of [65] for landmark localization.

4.3. Qualitative Results

Visualization of Gaze Representation. To further understand the effectiveness of learned latent representation qualitatively, we compute t-SNE [60] on the learned features space. After mapping the high dimensional feature to 2-D, DBSCAN clustering is performed (with the maximum distance = 1 and neighbourhood minimum sample = 12). The output is shown in Figure 4. We observe that the 2-D mapping of latent distributions is mapped to overlapping

Table 4. Contributions of different auxiliary tasks, their combinations, and noise correction towards overall MTGLS performance. Results are presented in terms of gaze error (in °) on CAVE dataset.

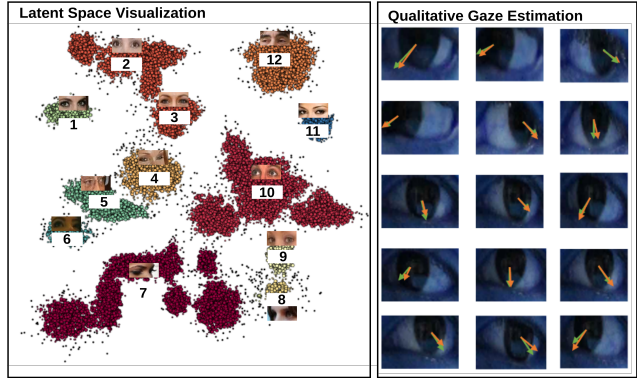| Pseudo-Gaze	Head-Pose	L/R Eye	NLL	Error (in °)
✓				3.40
	✓			3.52
		✓		3.67
✓	✓			3.30
	✓	✓		3.31
✓		✓		3.33
✓	✓	✓		3.23
✓	✓	✓	✓	3.00

Figure 4. **Left:** t-SNE visualization of the learned feature space by MTGLS. We observe well-formed clusters corresponding to different gaze directions and head-poses **Right:** Qualitative prediction result on CAVE dataset. Here, orange and green arrows represent predicted and ground-truth gaze direction, respectively.

clusters. We further overlay a representative eye patch corresponding to each of the formed clusters and observe that each cluster represents different head-pose, gaze, and their combination. We can notice clear patterns in learned space, i.e., the clusters within close proximity of each other have visual similarities in terms of eye gaze and head-pose (e.g., cluster pairs 2-3, 8-9, and 11-12).

Gaze Estimation. Fig. 4 (Right) illustrates the quality of our gaze estimation results. Here, the orange and green arrows indicate predicted and ground-truth gaze direction, respectively. It reflects the changes in the gaze estimation output manifold after representation learning via the proposed MTGLS. From the image, it is observed that our framework learns robust gaze estimation manifold.

5. Conclusion

We have proposed a multi-task gaze representation learning framework with limited supervision that exploits different auxiliary signals to minimize the misalignment between the output manifold of the actual and auxiliary tasks. Our method learns a robust feature embedding from abundantly available non-annotated facial images by distilling knowledge from deep CNN models developed for other facial analysis tasks. To counter the unavoidable noise in the auxiliary label space, we incorporate a probabilistic label distribution matching approach. Through a set of extensive empirical evaluations and ablations, we demonstrate that the gaze representations learned by our method are robust to several factors including appearance, head-poses, and illumination, and generalize well across different tasks and datasets. We further show that our generic representations transfer well for person-specific and few-shot calibrations. In future, it would be interesting to explore design choices across multi-task learning approaches, optimal pre-training strategies and combining the multi-task losses.

References

[1] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. *arXiv preprint arXiv:2012.07791*, 2020.

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

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018.

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[5] Bjorn Barz and Joachim Denzler. Deep learning on small datasets without pre-training using cosine loss. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1371–1380, 2020.

[6] Ben Benfold and Ian Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *IEEE International Conference on Computer Vision*, pages 2344–2351, 2011.

[7] Jonas Blattgerste, Patrick Renner, and Thies Pfeiffer. Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views. In *Proceedings of the Workshop on Communication by Gaze Interaction*, pages 1–9, 2018.

[8] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[10] Qin Cheng, Dimitris Agrafiotis, Alin M Achim, and David R Bull. Gaze location prediction for broadcast football video. *IEEE transactions on image processing*, 22(12):4918–4929, 2013.

[11] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10623–10630, 2020.

[12] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020.

[13] Neeru Dubey, Shreya Ghosh, and Abhinav Dhall. Unsupervised learning of eye gaze representation from the web. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2019.

[14] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021.

[15] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *European Conference on Computer Vision*, pages 339–357, 2018.

[16] L Fridman, P Langhans, J Lee, and B Reimer. Driver gaze estimation without using eye movement. *IEEE Intelligent Systems*, pages 49–56, 2015.

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

[18] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European conference on computer vision*, pages 311–326. Springer, 2016.

[19] Stephan J Garbin, Yiru Shen, Immo Schuetz, Robert Cavin, Gregory Hughes, and Sachin S Talathi. Openeds: Open eye dataset. *arXiv preprint arXiv:1905.03702*, 2019.

[20] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12742–12752, 2021.

[21] Shreya Ghosh, Abhinav Dhall, Munawar Hayat, Jarrod Knibbe, and Qiang Ji. Automatic gaze analysis: A survey of deep learning based approaches. *arXiv preprint arXiv:2108.05479*, 2021.

[22] Shreya Ghosh, Abhinav Dhall, Garima Sharma, Sarthak Gupta, and Nicu Sebe. Speak2label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset. *arXiv preprint arXiv:2004.05973*, 2020.

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[24] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[25] Junfeng He, Khoi Pham, Nachiappan Valliappan, Pingmei Xu, Chase Roberts, Dmitry Lagun, and Vidhya Navalpakkam. On-device few-shot personalization for real-time gaze estimation. In *IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[28] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6932–6941, 2019.

[29] Craig Hennessey, Borna Noureddin, and Peter Lawrence. A single camera eye-gaze tracking system with free head motion. In *Symposium on Eye tracking research & applications*, pages 87–94. ACM, 2006.

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[31] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 28(5-6):445–461, 2017.

[32] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size, 2016. cite arxiv:1602.07360Comment: In ICLR Format.

[33] Shreyank Jyoti and Abhinav Dhall. Automatic eye gaze estimation using geometric & texture-based networks. In *International Conference on Pattern Recognition*, pages 2474–2479. IEEE, 2018.

[34] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze embeddings for zero-shot image classification. In *IEEE Computer Vision and Pattern Recognition*, pages 4525–4534, 2017.

[35] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE International Conference on Computer Vision*, 2019.

[36] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. Weakly-supervised physically unconstrained gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9980–9989, 2021.

[37] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *IEEE Computer Vision and Pattern Recognition*, pages 2176–2184, 2016.

[38] Hantao Liu and Ingrid Heynderickx. Visual attention in objective image quality assessment: Based on eye-tracking data. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(7):971–982, 2011.

[39] Kui Lyu, Minghao Wang, and Liyu Meng. Extract the gaze multi-dimensional information analysis driver behavior. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 790–797, 2020.

[40] CH Morimoto, D Koons, A Amir, and M Flickner. Pupil detection and tracking using multiple light sources. *Image and Vision Computing*, 2000.

[41] Aamir Mustafa, Amanjot Kaur, Love Mehta, and Abhinav Dhall. Prediction and localization of student engagement in the wild. *arXiv preprint arXiv:1804.00858*, 2018.

[42] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9359–9367, 2018.

[43] Cristina Palmero, Abhishek Sharma, Karsten Behrendt, Kapil Krishnakumar, Oleg V Komogortsev, and Sachin S Talathi. Openeds2020: Open eyes dataset. *arXiv preprint arXiv:2005.03876*, 2020.

[44] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards end-to-end video-based eye-tracking. In *European Conference on Computer Vision (ECCV)*, 2020.

[45] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *IEEE International Conference on Computer Vision*, pages 9368–9377, 2019.

[46] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *European Conference on Computer Vision*, pages 721–738, 2018.

[47] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–10, 2018.

[48] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction. In *ACM Conference on Human Factors in Computing Systems*, pages 2594–2605, 2017.

[49] 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.

[50] BA Smith, Qi Yin, SK Feiner, and SK Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *ACM User Interface Software & Technology*, 2013.

[51] Lukas Stappen, Georgios Rizos, and Björn Schuller. X-aware: Context-aware human-environment attention fusion for driver gaze prediction in the wild. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 858–867, 2020.

[52] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *IEEE Computer Vision and Pattern Recognition*, pages 1821–1828, 2014.

[53] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, and Xian-Ming Liu. Toward statistical modeling of saccadic eye-movement and visual saliency. *IEEE Transactions on Image Processing*, 23(11):4649–4662, 2014.

[54] Ashwin Swaminathan and Mahesh Ramachandran. Enabling augmented reality using eye gaze tracking, June 12 2018. US Patent 9,996,150.

[55] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[56] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference*

on *Computer Vision and Pattern Recognition*, pages 5552–5560, 2018.

[57] A Tawari, K H Chen, and M M Trivedi. Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. In *IEEE Conference on Intelligent Transportation Systems*, pages 988–994, 2014.

[58] Marc Tonsen, Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Labelled pupils in the wild: a dataset for studying pupil detection in unconstrained environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 139–142, 2016.

[59] D Torricelli, S Conforto, M Schmid, and T D'Alessio. A neural-based remote eye gaze tracker under natural head motion. *Computer Methods and Programs in Biomedicine*, 2008.

[60] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[61] Viktor Varkarakis, Shabab Bazrafkan, and Peter Corcoran. Deep neural network and data augmentation methodology for off-axis iris segmentation in wearable headsets. *Neural Networks*, 121:101–121, 2020.

[62] B Vasli, S Martin, and M M Trivedi. On driver gaze estimation: Explorations and fusion of geometric and data driven approaches. In *IEEE Intelligent Transportation Systems*, pages 655–660, 2016.

[63] S Vora, A Rangesh, and M M Trivedi. On generalizing driver gaze zone estimation using convolutional neural networks. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 849–854. IEEE, 2017.

[64] S Vora, A Rangesh, and M M Trivedi. Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis. *IEEE Transactions on Intelligent Vehicles*, pages 254–265, 2018.

[65] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. Facex-zoo: A pytorh toolbox for face recognition. *arXiv preprint arXiv:2101.04407*, 2021.

[66] Kang Wang, Hui Su, and Qiang Ji. Neuro-inspired eye tracking with eye movement dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9831–9840, 2019.

[67] Kang Wang, Rui Zhao, and Qiang 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*, pages 440–448, 2018.

[68] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11907–11916, 2019.

[69] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017.

[70] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Gazedirector: Fully articulated eye gaze redirection in video. In *Computer Graphics Forum*, volume 37, pages 217–225. Wiley Online Library, 2018.

[71] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[72] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[73] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7025, 2019.

[74] Hyo Sik Yoon, Na Rae Baek, Noi Quang Truong, and Kang Ryoung Park. Driver gaze detection based on deep residual networks using the combined single image of dual near-infrared cameras. *IEEE Access*, 7:93448–93461, 2019.

[75] Y Yu, G Liu, and J Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019.

[76] Y Yu and J Odobez. Unsupervised representation learning for gaze estimation. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–13, 2020.

[77] Zehui Yu, Xiehe Huang, Xiubao Zhang, Haifeng Shen, Qun Li, Weihong Deng, Jian Tang, Yi Yang, and Jieping Ye. A multi-modal approach for driver gaze prediction to remove identity bias. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 768–776, 2020.

[78] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[79] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020.

[80] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Everyday eye contact detection using unsupervised gaze target discovery. In *ACM User Interface Software and Technology*, pages 193–203, 2017.

[81] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2018.

[82] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *IEEE Computer Vision and Pattern Recognition*, pages 4511–4520, 2015.

[83] X Zhang, Y Sugano, M Fritz, and A Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *IEEE Computer Vision and Pattern Recognition Workshop*, 2017.

[84] X Zhang, Y Sugano, M Fritz, and A Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[85] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. *NeurIPS*, 2020.

[86] J Zhu and J Yang. Subpixel eye gaze tracking. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.