2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI) | 979-8-3315-9944-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/IRI66576.2025.00070

Unsupervised Domain Adaptation for Appearance-based Gaze Estimation

Bhanuka Mahanama, Vikas Ashok, Sampath Jayarathna
Department of Computer Science, *Old Dominion University*, Norfolk, VA, USA
bhanuka@cs.odu.edu, vganjigu@cs.odu.edu, sampath@cs.odu.edu

***Abstract*—The human gaze provides informative cues on human behavior during day-to-day interactions, where extraction often requires specialized eye-tracking hardware, hindering the broad adoption of eye-tracking. Appearance-based eye tracking using commodity hardware offers an alternative approach for overcoming the problem by leveraging recent advancements in deep learning and large-scale datasets. However, due to the domain shift between the training and the testing environments, these approaches lose performance in practical applications where domain adaptation is required. We propose a novel unsupervised domain adaptation approach for appearance-based gaze estimation that utilizes adversarial training between domains. We train and test our approach against five publicly available datasets of varying domains, sizes, and tasks to illustrate the potential of our unsupervised domain adaptation approach, demonstrating promising improvements in gaze estimation and potential applications in a broader context.**

***Index Terms*—Gaze Estimation, Eye Tracking, Deep Learning, Domain Adaptation**

I. Introduction

The human gaze provides fascinating understandings of human behavior during our day-to-day interactions with insights into human behavior and psychology [1], [2]. The first step in quantifying the gaze to identify such relationships lies in estimating the gaze, often done using specialized computer hardware called eye trackers [3]. Despite a wide variety of specialized hardware, the high cost of hardware/ software and the invasive nature of the devices prevent the democratization of the technology among the masses [4], [5]. Making eye tracking available to the community at large requires us to address these challenges by developing novel methods for gaze estimation that use commodity hardware while being able to execute efficiently on different computational platforms.

Gaze estimation using commodity hardware eliminates the challenge by using RGB-only images, eliminating the need for specialized hardware. Furthermore, recent advancements in computer vision [3], [6] and the emergence of large-scale datasets [7]–[13] have contributed to commodity hardware-based gaze estimation becoming a subject of numerous studies. Despite the promising nature of these studies, the approaches proposed in the literature do not scale well for consumer-level applications.

These proposed gaze estimation models tend to be complex and bulky as studies focus mainly on improving the gaze estimation error, which often lacks parameter efficiency. As a result, these approaches require higher computational demands during training and gaze estimation. Moreover, these approaches also fail to scale across multiple input domains, such as input types, environments, and camera types. These methods often rely on the diversity of training data or annotated data from the target domain for domain adaptation. While both processes can improve the overall performance, they do not scale well due to the potential diversity of input domains and the expansive nature of annotated gaze data collection.

Unsupervised domain adaptation or transferring a model learned on a labeled domain to an unlabeled domain [14] provides an alternative approach to overcome the challenge due to the scalability of unlabeled target domain data collection. Further, employing the approach on a systematically scalable model such as EfficientNets [15], [16] provides scalability across multiple devices with differing computational capabilities. The contributions of our study are,

- Introduce a smaller and parameter-efficient gaze estimation model for appearance-based gaze estimation
- Improve the generalizability of the model through unsupervised domain adaptation
- Demonstrate the utility of the proposed model and approach through publicly available datasets

II. Related Work

Gaze estimation methodologies are broadly classified as model-based or appearance-based methods [17], [18]. Model-based approaches use ocular [19] or facial features [20] and employ a geometric model of the eye or face to estimate the gaze direction using landmarks such as the cornea [21], pupil center [22], and iris edges [23]. In contrast, appearance-based approaches utilize images to estimate the gaze directions using either ocular [6] or facial images [5], [7], [24], forming a mapping function between the image and the gaze directions. This eliminates the requirement for intermediate computation of facial landmarks. Based on the technique employed, these approaches can be further classified into conventional or deep learning approaches [17].

Conventional appearance-based approaches utilize image processing techniques combined with machine learning models (e.g., support vector machines [25], linear regression [26], or neural networks [27]) to estimate the gaze. Despite the simplicity, these models are often constrained by the capacity of the feature extractor and the complexity of the gaze estimation model [3], [23]. Instead of relying on generic features or dimensionality reduction techniques, deep learning methods approach this problem by detecting features and mapping

2835-5776/25/$31.00 ©2025 IEEE
DOI 10.1109/IRI66576.2025.00070

them to gaze estimation [17]. Recent studies in deep learning gaze estimation have shown Convolutional Neural Networks (CNNs) to be an excellent candidate for appearance-based gaze estimation [3], [6], [7], [11], [24], [28]–[30].

To achieve improved accuracy, CNNs typically scale up by adding more layers, which can often lead to bulkier, deeper CNN models [15], [16]. Despite the performance gain, these models tend to be computationally expensive due to their complexity. To derive an efficient gaze model, LiteGaze [24] derives models by sampling a super network trained for gaze estimation, followed by kernel shrinking. Despite the success, the approach offers limited scalability and accuracy due to the subnetwork selection process.

With the popularity of deep learning-based models in general computer vision applications, the wide adoption of mobile devices has led to the development of computationally efficient CNNs, such as Mobile-oriented models, such as MobileNet [31], ShuffleNet [32], and models with systematic scaling, such as EfficientNets [16]. Our approach uses EfficientNet-based models to achieve efficiency and scalability. Compared to LiteGaze, this eliminates the arbitrary nature in the development of the gaze model.

However, building an efficient feature extraction model may not guarantee effective domain adaptation for various gaze estimation environments. The traditional approach for domain adaptation requires costly data collection with various input and output domains in eye tracking, such as smartphones [11], laboratory settings [7], [8], and in-the-wild [10]. Unsupervised domain adaptation techniques offer an alternative approach by eliminating the requirement of target domain labels [14]. Recent studies on unsupervised gaze domain adaptations [33], [34] demonstrate the utility of the techniques in gaze estimation. Despite the success, these approaches often utilize complex training, domain adaptation techniques, and deep learning models. Thus, domain adaptation with simpler, parameter-efficient techniques remains largely unexplored.

III. Methodology

A. Architecture

Our proposed approach for gaze estimation starts with the intuition that a given facial image contains two features: gaze-defining and non-gaze-defining features, which we extend to any latent image representation. However, a typical autoencoder, which transforms an image into a latent representation, fails to capture this information as features get entangled between latent dimensions. As a result, a given autoencoder's latent dimension may contain both gaze-defining and non-defining latent features, producing a suboptimal representation for gaze estimation.

Our proposed model (see figure 1) uses an image encoder (E) that transforms a given image into a latent representation (z) and a decoder (D) that reconstructs the image from the latent representation. Based on our hypothesis, our latent space (z) comprises gaze-defining (z_g) and non-defining features (z_a) such that $z = \{z_a, z_g\}$. We define the latent space, z having N dimensions, with M dimensions corresponding to gaze-defining features. To enforce the constraint, we introduce a gaze decoder, G, that uses z_g to estimate a given facial image's gaze descriptor, presented by the gaze angles.

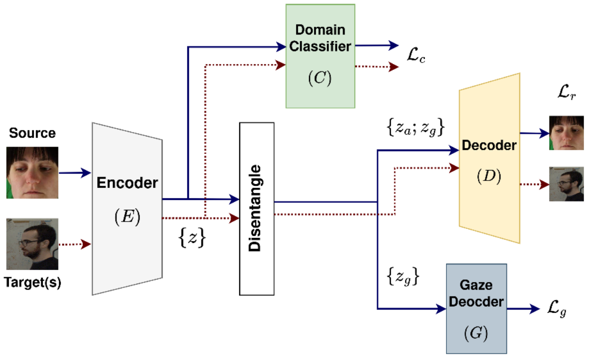

Fig. 1. Proposed architecture: Encoder (E), Decoder (D), and Domain Classifier (C) are trained on both target and source domains, while Gaze Decoder (G) is only trained on the source domain. D and C utilize the latent representation (z) from the E, whereas G uses only a part of the latent representation (z_g). Solid: Source domain ($\mathcal{S}$), Dotted: Target domain ($\mathcal{T}$).

Unlike supervised adaptation approaches that require labeled data from both domains, we employ unsupervised domain adaptation (UDA), where the target domain data remains unlabeled. The domain classifier (C) learns to distinguish between source and target domains, and the encoder (E) is trained adversarially to the classifier. This approach forces the encoder to learn domain-invariant gaze features, ensuring the model generalizes well to unseen target domains without requiring target labels.

We can represent a domain as a distribution over the input population X and the corresponding label space Y, which are image patches and gaze descriptors in appearance-based gaze estimation. In supervised gaze estimation, we find an objective function to approximate the relationship between the labeled samples $\{(x, y)\}$, where $y \in Y$ represents the ground truth gaze descriptors for the given image $x \in X$. In a naive domain adaptation, we train the model on source domain $\mathcal{S} = \{X^s, Y^s\}$ and fine-tune on a smaller target domain $\mathcal{T} = \{X^t, Y^t\}$. Despite the simplicity, this approach requires knowledge of the labels of the target domain (Y^t). However, our problem formulation considers the scenario of unknown or unavailable Y^t to address the problem in a broader context.

B. Losses

The primary loss of our proposed model is the loss of gaze estimation, where we minimize the error between the model estimates on gaze direction g and the gaze labels y. Here, we consider the gaze direction g relative to the camera represented using the *yaw* and *pitch* angles. We apply the loss only on the samples on source domain S containing the ground truth information (labeled source domain).

$$\mathcal{L}_g = \frac{1}{dim(y)} \sum_{x \in \{\mathcal{S}\}} |y - g| \text{ where } g = G(z_g) \qquad (1)$$

The domain classification loss is unsupervised and forces domain-invariant feature learning. Given a binary domain classification label c, indicating if $x \in \mathcal{S}$ or $x \in \mathcal{T}$, we define the classification loss of our model as $\mathcal{L}_c$ as,

$$\mathcal{L}_c = \sum_{x \in \{\mathcal{S}, \mathcal{T}\}} -c\, log(C(E(x))) - (1-c) log(1 - C(E(x))) \tag{2}$$

We apply the unsupervised reconstruction loss $\mathcal{L}_r$ to guide the model's reconstruction function, helping preserve visual consistency across domains. For any given image $x \in \{\mathcal{S}, \mathcal{T}\}$, we define $\mathcal{L}_r$ as,

$$\mathcal{L}_r = \frac{1}{dim(x)} \sum_{x \in \{\mathcal{S}, \mathcal{T}\}} |x - D(E(x))| \tag{3}$$

Finally, we combine the loss functions to form the multi-objective loss function $\mathcal{L}$ by introducing two hyperparameters λ_r and λ_c controlling the weights on $\mathcal{L}_r$ and $\mathcal{L}_c$ respectively (see equation 4). Joint optimization of both losses allows us to achieve domain alignment in addition to the main objective of gaze estimation. It is important to note that $\mathcal{L}_g$ is only optimized on the samples in the source domain $\mathcal{S}$.

$$\mathcal{L} = \mathcal{L}_g + \lambda_r \mathcal{L}_r + \lambda_c \mathcal{L}_c \tag{4}$$

C. Implementation

We use an encoder (E) built on the EfficientNetV2 [16] architecture with an empirically chosen latent dimension N = 128. The EfficientNet architecture provides parameter efficiency with a scaling mechanism using inverted residual blocks [15] and fused inverted residual blocks [16]. Our models use scaling parameters defined for the EfficientNetV2-B0 (*small*), EfficientNetV2-B2 (*medium*), and EfficientNetV2-B3 (*large*) for the encoder, followed by a series of dense layers forming the latent space. We use a series of convolution transpose blocks to form the decoder(D) of our model, with a scaling factor of two. The Gaze decoder (G) uses a fully connected layer with a hidden layer to predict the gaze directions as *pitch* (θ) and *yaw* (ϕ) angles. Similarly, the domain classifier (C) uses two fully connected layers and produces a sigmoid output on the domain label.

We train the model in two steps. First, we train the model for 50 epochs using the source domain with no domain adaptation for empirically chosen λ_r. Then, we evaluate the gaze estimation error of each model and choose the best-performing model in our next step of domain adaptation. Here, we continue the training process by including domain adaptation with the best performing λ_r and investigate the λ_c on target domain performance. Then, we train the model for another ten epochs, optimizing the aggregate loss function (see equation 4). To ensure faster convergence of the overall process, we use the ImageNet [35] pre-trained weights for the encoder at the initialization of training and optimize using the Adam optimizer with a linearly decaying learning rate.

Our source domain comprises the X-Gaze [7] dataset, which offers over one million high-resolution facial images of varying extreme head poses and gaze angles collected in a laboratory setting. The dataset comprises images of 110 participants captured under varying illumination conditions and gaze angles of range ($\pm 120, \pm 70$). We use the train split of the dataset with labels, with an 80-20 validation split formed by a random selection of participants from the dataset.

We use Gaze360 [10], RTGene [8], and Columbia Gaze [12] datasets as the target domain of the study. The Gaze360 dataset captures real-world images of users in uncontrolled environments. Since the dataset does not include gaze information as *pitch* and *yaw* angles, we convert the 3-D gaze ground truth using normalized vectors. Further, we use the facial bounding box information provided with the dataset and resize the image to (224×224) using linear interpolation. We use only the training and validation splits defined in the dataset during the training steps, and to enable comparisons, only use test splits for evaluation. The RTGene dataset comprises facial images of users captured in a laboratory setting while wearing a head-mounted eye tracker, used to generate ground truth labels. In addition to images of participants wearing the eye tracker, the dataset offers inpainted images, where the eye tracker is edited out with a Generative Adversarial Network. We use the inpainted images, with train, validation, and testing splits defined in the dataset. The Columbia Gaze dataset also comprises facial images captured in a laboratory setting, totaling 5880 images. Our experiments use the Columbia dataset only for evaluation due to its relatively small size. Here, we use a crop of the central region of the image containing the participant's face downsampled to (224×224) as the input, with gaze directions computed from the experiment metadata.

In addition to the above datasets, we use RGBD Gaze [11] dataset to demonstrate the utility of our approach in two real-world applications. The RGBD dataset comprises facial images captured from smartphones under four contexts (standing, walking, sitting, and lying) with gaze location on the smartphone screen of 45 users.

We use the baselines provided in each dataset, where applicable, as the primary baseline of the experiments. However, due to the variability in architectures, experimental differences, and model complexities, we use a secondary benchmark developed using an EfficientNetV2-B0 network (Baseline) trained on the source domain with no domain adaptation. Similar to our proposed model implementation, we perform global pooling on the final fused-MBConv block and pass through a series of fully connected layers to obtain gaze angles.

IV. Experiments and Results

A. Gaze Estimation

We first compare our proposed model's gaze estimation error after completing the first step of the training process, before domain adaptation against the ETH-XGaze test set (see Table I). Our proposed models use significantly fewer parameters during inference (E and G) than the ResNet-50-based ETH X-Gaze benchmark, with our *small* model using 76% lesser parameters with a 10% higher gaze error, while

TABLE I

Comparison of gaze estimation model mean angular error in degrees on ETH X-Gaze dataset [7] with the number of parameters in each model.

Model	Gaze Error	#Params
ETH X-Gaze Baseline [7]	**4.50**	23M
ResNet18	7.87	11M
EfficientNetV1-B1 [3]	5.12	6.9M
LiteGaze-XS (Finetuned) [24]	6.93	2.5M
LiteGaze-L (Finetuned) [24]	7.63	5.5M
Baseline model (Gaze loss only)	7.38	6.2M
Small ($\lambda_r = 0.01$)	4.98	6.2M
Small ($\lambda_r = 0.10$)	**4.95**	6.2M
Small ($\lambda_r = 0.25$)	4.99	6.2M
Small ($\lambda_r = 0.50$)	5.09	6.2M
Small ($\lambda_r = 1.0$)	5.10	6.2M
Small ($\lambda_r = 0.10$)	4.95	6.2M
Medium ($\lambda_r = 0.10$)	4.76	9.2M
Large ($\lambda_r = 0.10$)	**4.60**	13.4M

large uses 40% lesser parameters with comparable error. Further compared studies on parameter-efficient smaller models such as LiteGaze [24], ResNet-18, and our baseline model (a standard gaze model), the proposed architecture achieves a superior performance. This indicates that our modifications yield better in-domain generalizability with upwards of 49% error improvement (Baseline vs. *small* $\lambda_r = 0.10$), in addition to the efficiency gain achieved by the EfficientNetV2 architecture.

B. Cross Domain Evaluation

We further evaluate the cross-domain dataset performance of our model against similar cross-domain benchmarks in Gaze360 [10], RTGene [8], and Columbia [12] Gaze datasets before domain adaptation. For cross-domain evaluation with Gaze360, we report the performance for test samples where the subject is looking within 90° (front 180°) of the camera direction to maintain the same output domain as the ETH X-Gaze dataset. Our proposed models perform poorly compared to benchmarks found in the literature that utilize the entire test set and the Gaze360 benchmark on the front-facing images (see table II). Our models on the RT-Gene test set yield better gaze error than the X-Gaze and Gaze360 benchmarks, trained only on the corresponding datasets, indicating resiliency to domain shift. For the smaller Columbia dataset, most configurations of the proposed model outperform the Gaze360 benchmark and the baseline model (see table II), with the error improvement of 6.6% for *small* and 18.9% for *large*.

Considering the overall performance of $\lambda_r = 0.1$, we evaluate and report their gaze estimation errors after domain adaptation in Table III. The *small* models improve performance in Gaze360 and RTGene estimation errors in 9 out of 12 instances when compared to source-only models irrespective of the target domain while maintaining a mean gaze error at 6.15° (increase of 6.3%). Similarly, domain adaptation on *medium* yields better performance, with a slight degradation for *large*. Moreover, results indicate that Gaze360 is a better candidate for more generalized domain adaptation, as evidenced by improvements in gaze error attributable to the extensiveness and diversity compared to RTGene.

We also report the contribution of the discriminative domain classification on domain adaptation of *Small* in Table III. For Gaze360, discriminative ($\lambda_c = 0.1, 0.2$) and non-discriminative ($\lambda_c = 0$) models yield similar performance levels irrespective of the target domain. However, the $\lambda_c = 0.1, 0.20$ offers a considerable gain for RTGene, whereas $\lambda_c = 0$ results in a higher estimation error.

C. Qualitative Study

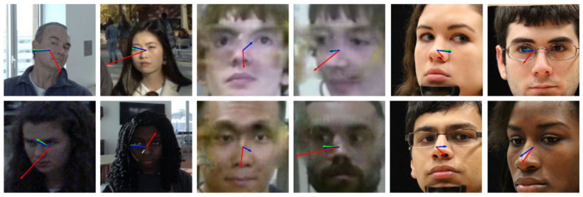

Fig. 2. Examples of gaze estimate correction post-domain adaptation with *Small* with $\lambda_r = \lambda_c = 0.1$ using Gaze360 as target domain. Green: ground truth, Red: before target training, Blue: gaze estimation after target training. Left: Gaze360 [10], Center: RTGene [8], Right: Columbia [12]

We qualitatively examine the gaze error improvement due to domain adaptation on Gaze360 using the *small* ($\lambda_c = 0.1$), with figure 2 illustrating gaze corrections upon domain adaptation for three datasets in consideration. The alignment of post-domain adaptation estimates with ground truth indicates our approach improves the estimates with **unlabelled target domain** combined with the labeled source domain. Further, our domain adaptation improves estimates on non-target and non-source domains, as evidenced by corrections on RT-Gene and Columbia gaze datasets by domain adaptation on Gaze360.

D. Utility Study: Smartphone Gaze Estimation

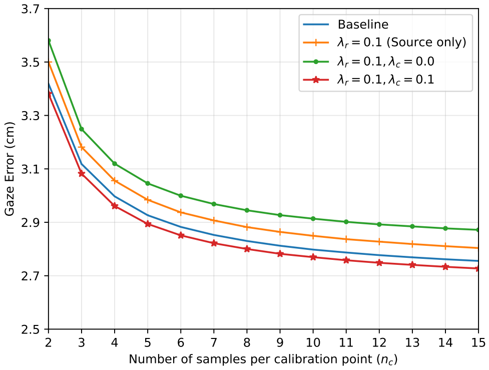

Fig. 3. Smartphone gaze estimation error on RGBD dataset with 5-point calibration using *small* ($\lambda_r = \lambda_c = 0.1$)

Next, we demonstrate the utility of our proposed approach through the utility of gaze estimation on smartphones by

TABLE II

CROSS DATASET GAZE ESTIMATION ERROR IN DEGREES BEFORE UNSUPERVISED DOMAIN ADAPTATION COMPARED TO SIMILAR STUDIES.

Model	Train Description	Gaze360	RTGene	Columbia
ETH X-Gaze Baseline [7]	ETH X-Gaze only	27.3	31.2	-
GazeCaps [29]	ETH X-Gaze + fine-tuning	10.04	6.92	-
GazeTR [36]	ETH X-Gaze + fine-tuning	10.62	6.55	-
Gaze360 Baseline [10]	Gaze360	13.5 (Front: 11.4)	23.4	9.0
Baseline model	ETH X-Gaze only	29.96	20.76	6.20
Small ($\lambda_r = 0.01$)		28.91	22.44	6.20
Small ($\lambda_r = 0.1$)		31.06	**17.48**	**5.79**
Small ($\lambda_r = 0.25$)		30.18	19.77	5.79
Small ($\lambda_r = 0.5$)		27.78	21.21	9.18
Small ($\lambda_r = 1.0$)		**27.10**	23.05	6.39
Small ($\lambda_r = 0.1$)	ETH X-Gaze only	31.06	**17.48**	5.79
Medium ($\lambda_r = 0.1$)		32.54	19.70	5.83
Large ($\lambda_r = 0.1$)		**29.07**	20.55	**5.03**

TABLE III

CROSS DATASET GAZE ESTIMATION ERROR IN DEGREES POST-DOMAIN ADAPTATION USING GAZE360 [10] AND RTGENE [8].

	λ_r	λ_c	Target	Gaze360	RTGene	Columbia
Small (Source only)	0.1	-	-	31.06	17.48	5.79
Small	0.1	0.0	Gaze360	27.67	25.12	6.23
Small	0.1	0.1	Gaze360	**27.30**	12.26	**5.77**
Small	0.1	0.2	Gaze360	27.99	13.16	5.99
Small	0.1	0.0	RTGene	30.71	24.08	5.82
Small	0.1	0.1	RTGene	33.24	**11.53**	6.14
Small	0.1	0.2	RTGene	30.78	13.02	6.92
Medium (Source only)	0.1	-	-	32.54	19.70	5.83
Medium	0.1	0.1	Gaze360	**26.71**	**9.88**	**5.30**
Medium	0.1	0.1	RTGene	33.13	15.03	5.68
Large (Source only)	0.1	-	-	29.07	**20.55**	**5.03**
Large	0.1	0.1	Gaze360	29.92	21.44	5.86
Large	0.1	0.1	RTGene	**28.57**	31.42	5.48

evaluating the *small* ($\lambda_c = 0.1$) on the RGBD gaze dataset [11], only using the RGB modality. The dataset contains the smartphone's on-screen gaze location information, the face's location, the device model, and the orientation, along with facial images captured under four contextual conditions: standing, walking, sitting, and lying. In our experiment, we crop and transform the facial images to match the specifications of the gaze model and transform the on-screen gaze location to front camera coordinates. We combine the face image patch with the image patch and image frame ratio, indicating the user's distance from the screen, to estimate the gaze location through linear regression. We train the regression model by simulating a five-point calibration process (four corners and center). For a selected participant, we randomly sample n_c training samples for each calibration point, followed by evaluation using non-calibration points. We repeat this process for arbitrarily chosen 1000 iterations for each user in the dataset for $n_c \in [2, 15]$ and report the Euclidean mean gaze error in cm (see Figure 3).

Across all values of n_c, the post-domain adaptation model with discriminative classification outperforms other models (non-discriminative domain adaptation and source-only training), indicating better generalizability achieved through our approach and utility in real-world applications. With three calibration samples for each point, our model achieves a better performance than TabletGaze [37], which reports an estimation error of $3.17\,cm$. Even though EyeTab [22] and RGBDGaze [11], with estimation errors of $2.58\,cm$ and $2.26\,cm$ (RGB only) report better accuracies, it is essential to note that our approach does not require a large-scale labeled smartphone gaze dataset. Moreover, we achieve personalization through $5n_c$ samples captured with dynamic world conditions, such as users holding devices at varying angles and distances.

V. CONCLUSION AND FUTURE WORK

In this study, we proposed an unsupervised domain adaptation approach for gaze estimation, which has demonstrated utility in real-world applications. Our approach stems from the intuition that gaze-defining and non-defining features in a facial image can be extended to latent representations combined with domain adaptation using discriminative classification. Compared to models of similar complexities, we achieve improved generalizability, domain adaptability, scalability, and parameter efficiency, as demonstrated through tests against five publicly available datasets with different input and output domains. Compared with traditional gaze estimation approaches that heavily rely on supervised training for domain adaptation, our approach combines a **labeled source** and an **unlabeled target** domain through unsupervised domain adaptation. Considering the lower complexity of unlabeled data, our approach offers low-cost, efficient, broader domain adaptation, such as in smartphone gaze estimation with potential applications in multi-user eye-tracking [3].

Further, scalable gaze models also offer applicability in low-power interfaces and on-device gaze estimation, enabling widespread adoption in consumer and research settings. Future work will focus on alternative gaze estimation approaches, such as CNN and attention-driven hybrid approaches, and potential alternative approaches for domain alignment.

REFERENCES

[1] B. Mahanama, Y. Jayawardana, S. Rengarajan, G. Jayawardena, L. Chukoskie, J. Snider, and S. Jayarathna, "Eye movement and pupil measures: A review," *frontiers in Computer Science*, vol. 3, p. 733531, 2022.

[2] B. Mahanama, M. Sunkara, V. Ashok, and S. Jayarathna, "Disetrac: Distributed eye-tracking for online collaboration," in *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, 2023, pp. 427–431.

[3] B. Mahanama, V. Ashok, and S. Jayarathna, "Multi-eyes: A framework for multi-user eye-tracking using webcameras," in *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2024, pp. 308–313.

[4] B. Mahanama, "Multi-user eye-tracking," in *2022 Symposium on Eye Tracking Research and Applications*, 2022, pp. 1–3.

[5] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2176–2184.

[6] B. Mahanama, Y. Jayawardana, and S. Jayarathna, "Gaze-net: appearance-based gaze estimation using capsule networks," in *Proceedings of the 11th Augmented Human International Conference*, 2020, pp. 1–4.

[7] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *European Conference on Computer Vision*. Springer, 2020, pp. 365–381.

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

[9] 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*, vol. 41, no. 1, pp. 162–175, 2017.

[10] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6912–6921.

[11] R. Arakawa, M. Goel, C. Harrison, and K. Ahuja, "Rgbdgaze: Gaze tracking on smartphones with rgb and depth data," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 329–336.

[12] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: passive eye contact detection for human-object interaction," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 2013, pp. 271–280.

[13] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *Proceedings of the symposium on eye tracking research and applications*, 2014, pp. 255–258.

[14] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–46, 2020.

[15] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[16] ——, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[17] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Eye gaze estimation: A survey on deep learning-based approaches," *Expert Systems with Applications*, vol. 199, p. 116894, 2022.

[18] Y. Cheng, H. Wang, Y. Bao, and F. Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[19] M. Kassner, W. Patera, and A. Bulling, "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, 2014, pp. 1151–1160.

[20] M. X. Huang, T. C. Kwok, G. Ngai, H. V. Leong, and S. C. Chan, "Building a self-learning eye gaze model from user interaction data," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1017–1020.

[21] M. Kassner, W. Patera, and A. Bulling, "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, 2014, pp. 1151–1160.

[22] E. Wood and A. Bulling, "Eyetab: Model-based gaze estimation on unmodified tablet computers," in *Proceedings of the symposium on eye tracking research and applications*, 2014, pp. 207–210.

[23] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Eye gaze estimation: A survey on deep learning-based approaches," *Expert Systems with Applications*, vol. 199, p. 116894, 2022.

[24] X. Guo, Y. Wu, J. Miao, and Y. Chen, "Litegaze: Neural architecture search for efficient gaze estimation," *Plos one*, vol. 18, no. 5, p. e0284814, 2023.

[25] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: passive eye contact detection for human-object interaction," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 2013, pp. 271–280.

[26] A. Papoutsaki, J. Laskey, and J. Huang, "Searchgazer: Webcam eye tracking for remote studies of web search," in *Proceedings of the 2017 conference on conference human information interaction and retrieval*, 2017, pp. 17–26.

[27] W. Sewell and O. 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*, 2010, pp. 3739–3744.

[28] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4511–4520.

[29] H. Wang, J. O. Oh, H. J. Chang, J. H. Na, M. Tae, Z. Zhang, and S.-I. Choi, "Gazecaps: Gaze estimation with self-attention-routed capsules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2668–2676.

[30] V. Nagpure and K. Okuma, "Searching efficient neural architecture with multi-resolution fusion transformer for appearance-based gaze estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 890–899.

[31] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[32] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[33] Y. Liu, R. Liu, H. Wang, and F. Lu, "Generalizing gaze estimation with outlier-guided collaborative adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3835–3844.

[34] X. Cai, J. Zeng, S. Shan, and X. Chen, "Source-free adaptive gaze estimation by uncertainty reduction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 035–22 045.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[36] Y. Cheng and F. Lu, "Gaze estimation using transformer," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 3341–3347.

[37] Q. Huang, A. Veeraraghavan, and A. Sabharwal, "Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets," *Machine Vision and Applications*, vol. 28, pp. 445–461, 2017.