IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 September 2025, accepted 29 September 2025, date of publication 3 October 2025, date of current version 9 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3617076

RESEARCH ARTICLE

Gaze Data Imbalance: An Overlooked Challenge in Appearance-Based Gaze Estimation

JAN GLINKO
Department of Decision Systems and Robotics, Gdańsk University of Technology, 80-222 Gdańsk, Poland
e-mail: jan.glinko@pg.edu.pl

This work was supported by the Polish High-Performance Computing Infrastructure PLGrid [High-Performance Computing (HPC) Centers: [Academic Computer Center (ACK) Akademickie Centrum Komputerowe] Cyfronet AGH University of Krakow (AGH)] under Grant PLG/2024/017808.

ABSTRACT Data imbalance exists in appearance-based gaze estimation datasets, hurting the model's generalizability and fairness. Unfortunately, this aspect is usually overlooked, and researchers focus mainly on developing more and more sophisticated gaze estimation neural networks. In this work, we identify two types of imbalance in gaze estimation data. The first is related to the uneven distribution of ground truth gaze vectors, and the second one comes from the uneven ethnicity distribution of dataset participants. We prove the negative impact of both of them on the model's generalizability. Therefore, we propose Uniform Gaze Sampling and Uniform Ethnicity Sampling, simple yet effective re-sampling techniques tailored for gaze estimation. Moreover, we introduce balanced metrics, i.e., Balanced Gaze Error and Balanced Ethnicity Error, for a fair performance evaluation. Finally, we demonstrate the usefulness of the proposed methods and metrics on four benchmarks. To the best of our knowledge, we are the first to address data imbalance in gaze estimation comprehensively.

INDEX TERMS Appearance-based gaze estimation, deep learning, neural networks, imbalanced learning.

I. INTRODUCTION

Eye-tracking has become increasingly prevalent in various fields, such as psychology [1], [2], human-computer interaction [3], [4], and market research [5], [6]. This technology provides valuable insights into human behavior and cognition by tracking and recording eye movements. Eye tracking can be accomplished through different imaging devices, such as infrared [7], standard [8], or 3D cameras [9] and with different methods, for instance appearance-based [10], model-based [9], [11], or feature-based [12].

With the widespread availability of high-quality cameras in everyday devices like smartphones and laptops, there has been a notable surge of interest in appearance-based gaze estimation among eye-tracking researchers. This method analyzes facial images to estimate where a person looks. The recent advancements in deep learning and convolutional neural networks have significantly improved the accuracy and robustness of gaze estimation systems in the wild, making them more practical for real-world deployment. For a detailed review of appearance-based gaze estimation methods, please refer to works such as [10], [13], and [14].

However, due to the complexity of learning individual gaze patterns and the increasing size of deep neural networks, a substantial amount of data is required to train appearance-based gaze estimation models efficiently. An often overlooked issue is the imbalance present in state-of-the-art benchmark datasets.

While imbalanced learning has been widely recognized and addressed in other areas of machine learning and data analysis [15], [16], [17], its impact in the context of gaze estimation has received relatively little attention. Moreover, unlike gaze estimation, which is a regression problem, most methods addressing data imbalance focus on classification. Additionally, gaze data imbalance is a complex issue and comprises three main factors: uneven gaze points distribution, uneven head pose distribution, and uneven ethnicity distribution. All of those constitute challenges in constructing precise and resilient gaze estimation models. While the imbalance in ground truth gaze points and head

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Huang.

pose distribution results from the limited size of gaze estimation datasets caused by the difficulties of acquiring gaze data, the imbalance in ethnicity arises from the uneven distribution of human races among volunteers participating in creating such datasets. This is essential because of variations in anatomical eye structure across human races [18], [19]. As presented in Figure 1, the structure of Asian eyes is distinct, characterized by a covered inner corner, single or double eyelids, downward-facing eyelashes, and occasionally a fold at the corner. In contrast, European or Indian eyes typically feature an exposed inner corner and an external fold at the outer edge.

The head pose distribution imbalance is efficiently addressed by the normalization techniques [20], [21]. On the other hand, imbalances in distributions of ground truth gaze points and participants' ethnicity remain outside the area of gaze estimation research. Additionally, it is worth emphasizing that all three types of imbalance can be found in state-of-the-art benchmarks, such as MPIIGaze [13], ETHXGaze [22], EYEDIAP [23], and GazeCapture [24].

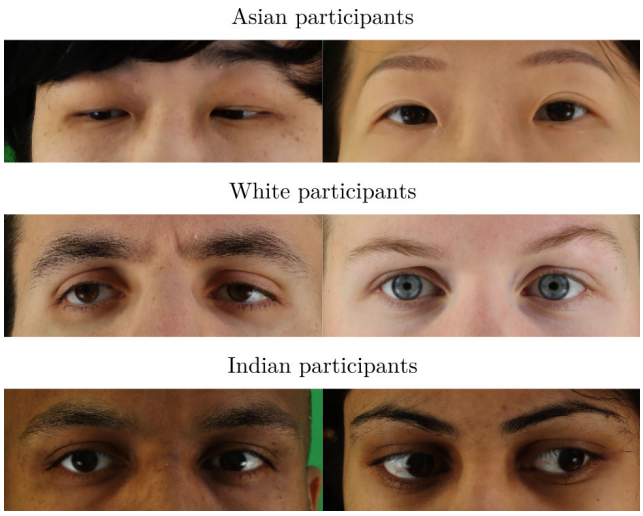

FIGURE 1. Comparison of eye appearance among participants from different ethnicity groups. The structure of Asian eyes is distinct, characterized by a covered inner corner, single or double eyelids, downward-facing eyelashes, and occasionally a fold at the corner. In contrast, White and Indian eyes typically feature an exposed inner corner and an external fold at the outer edge.

This study aims to tackle the issues of imbalances in distributions of ground truth gaze points and participants' ethnicity in gaze estimation benchmarks. We introduce new data sampling techniques called Uniform Gaze Sampling and Uniform Ethnicity Sampling to address these imbalances during the training of gaze estimation models. Additionally, we compare the performance of proposed samplers with existing methods addressing data imbalance in the regression domain, such as Feature Distribution Smoothing (FDS) [25], re-weighting techniques like Dense Weight [26] and Balanced Mean Squared Error [26], and the focal loss adapted for regression [25], [27]. Moreover, we compare our sampling strategies with robust, appearance-based gaze estimation networks that quantize gaze space, transforming gaze estimation into a classification task [28]. On top of that, we propose two balanced evaluation metrics, Balanced Ethnicity Error, and Balanced Gaze Error, designed explicitly for imbalanced gaze estimation benchmarks. Finally, we comprehensively evaluate the proposed sampling methods with the existing methods on the four most popular appearance-based gaze estimation benchmark datasets.

To the best of our knowledge, we are the first to try to investigate imbalances in appearance-based gaze estimation datasets comprehensively. By raising awareness of this issue and providing potential solutions and metrics, we hope to encourage researchers and practitioners in the eye-tracking community to consider and address gaze data imbalance in their work, ultimately advancing the reliability and validity of findings derived from eye-tracking research.

II. RELATED WORK

A. IMBALANCED REGRESSION

Numerous methods have been investigated for dealing with imbalanced classification, such as resampling [29], [30], [31] and reweighting [32], [33], [34], [35]. At the same time, imbalanced regression is relatively under-explored.

Earlier works adopt methods designed for classification, like synthetic resampling [36], [37]. However, creating meaningful synthetic samples for image data through simple interpolation is challenging. This is because such techniques often produce blurry, unrealistic images and fail to capture the highly non-linear relationship between subtle changes in eye appearance and the corresponding gaze direction. For this reason, we focus on re-sampling and re-weighting techniques. Moreover, in appearance-based gaze estimation, slight differences in the eye image result in a significant change in ground truth values. On the other hand, recent methods focus more on weighting samples [25], [26], [27], [38], assigning higher weights to underrepresented areas. Those methods can be easily adapted to appearance-based gaze estimation, and we compare their efficiency with our methods.

Imbalanced regression is addressed by ensembling approaches in specific domains, like predicting sleep apnea severity [39], river discharge [40] or fuel properties based on their molecular structures [41]. Such an approach is built upon specialized domain knowledge and is hard to generalize to other regression tasks. It is similar to the ethnicity imbalance problem in the appearance-based gaze estimation domain.

B. IMBALANCED GAZE ESTIMATION

A limited amount of attention is paid to imbalanced learning in eye-tracking research.

In a feature-based gaze estimation, [42] recognizes the imbalance class distribution in eye images, especially in the pupil area. They introduce a combined loss function consisting of Generalized Dice Loss, Boundary Aware Loss, Surface Loss, and Cross Entropy Loss for semantic

segmentation of eye region. This ensures sharp semantic boundaries and improves accuracy. Similarly, in a model-based gaze estimation, [43] indicates that iris occlusion due to eyelid shape or eyelashes frequently breaks ellipse fitting algorithms that rely on well-defined pupil or iris edge segments. They propose a convolutional neural network-based framework that is robust to occlusions. However, these works do not present the direct impact of the proposed methods on the results achieved for the underrepresented ground truth gaze angle areas.

In an appearance-based gaze estimation, a primary method for addressing the imbalance of the data and achieving a robust predictor is to quantize the gaze space and convert the regression task into a multi-class classification [28], [44], [45]. Additionally, [44] proposes a tolerant and talented training scheme, an iterative random knowledge distillation framework enhanced with cosine similarity pruning and aligned orthogonal initialization to address overfitting caused by data imbalance. Moreover, they propose an adversarial training-based Disturbance with Ordinal loss to improve robustness. However, this work is limited only to single benchmark datasets and 2D points of gaze instead of 3D gaze vectors. Significantly, information about errors is limited to the overall distance score. Similarly, [28] proposes a two-factor loss function including Categorical Cross Entropy (CCE) and Mean Squared Error (MSE). A continuous value of gaze angle is restored based on predicted gaze bin probabilities. MSE loss is calculated from the continuous gaze angle and is summed with CCE to compute the ultimate loss value. Work [45] builds upon that and adds more classification heads with coarser levels of gaze space quantization. Similar to [28], a continuous gaze angle value is restored from the most granular classification head. To calculate the value of the loss function, CCE from each classification head is summed with MSE. However, [28] and [45] do not identify the imbalance of the used datasets or demonstrate the usefulness of the proposed methods in improving the generalizability of gaze estimation models.

In addition, in the eye-tracking research with dedicated hardware, [46] reports that Tobii TX300 eye-tracker accuracy and precision were worse for Asian participants than for African and White participants. Moreover, large gaze angles proved detrimental to trackability for all African, Asian, and White participants. This is similar to our observations of data imbalance effects in appearance-based gaze estimation methods.

To the best of our knowledge, no previous studies have addressed the imbalance in appearance-based gaze estimation data. This represents a significant gap, particularly given recent developments in appearance-based gaze estimation neural networks.

III. METHODOLOGY

A. PROBLEM SETTING

Evaluation of appearance-based gaze estimation methods with standard, un-balanced metrics — henceforth referred to as overall metrics — lead to incorrect conclusions about their performance. These metrics, such as Mean Gaze Angle Error (MGAE) or Mean Absolute Error (MAE), are typically calculated across the entire dataset without accounting for the underlying data imbalance, which can be misleading. For instance, predictions for individual users may fall into a gaze range that is underrepresented in the dataset, or the anatomical structure of the user's eye may differ from that of most of the participants in the dataset. Both factors can significantly disrupt the accuracy of the gaze estimation model, bringing it below the value derived from the general metrics.

To address the issue, balanced evaluation metrics are employed to fairly evaluate a model's performance and generalizability within different gaze ranges and participant ethnicities. Such metrics can not only help evaluate the model reliably but also identify the operating point for which the model will give the best results, which can help design a further eye-tracking experiment. Moreover, imbalance-aware sampling techniques are used during the training of gaze estimation neural networks to mitigate the impact of data imbalance on the final model performance.

B. UNIFORM GAZE SAMPLING

To motivate the need for our approach, we first analyze the relationship between data distribution and model error on four popular benchmarks: MPIIFaceGaze, EYEDIAP, GazeCapture, and ETHXGaze. Figure 2 visualizes the ground truth distributions for both pitch and yaw angles alongside the normalized error of our baseline model. A critical observation across all datasets is that the gaze vector distributions are heavily imbalanced, often exhibiting a long-tail shape where most samples are concentrated in the center of the gaze range.

More importantly, the figure reveals a strong correlation between this data sparsity and model performance. The normalized error is consistently higher in the underrepresented "tail" regions of the distribution for both pitch and yaw angles. This demonstrates that even a well-trained model struggles in areas with fewer training examples, leading to a loss of generalizability and fairness. This direct link between data imbalance and increased prediction error is the primary motivation for our proposed Uniform Gaze Sampling (UGS) method.

UGS is based on resampling techniques and aims to equalize the distribution of gaze vectors. To implement UGS, we first quantize ground truth space into N equal bins of width w. With bins labeled from $i = 1$ to N, the boundaries of each bin i are then

$$Bin_i = [(i-1) \cdot w, i \cdot w) \tag{1}$$

where Bin_i represents the range of ground truth values in the i-th bin. For each ground truth value x, we determine the bin i to which it belongs using the formula

$$i = \lfloor \frac{x}{w} \rfloor + 1 \tag{2}$$

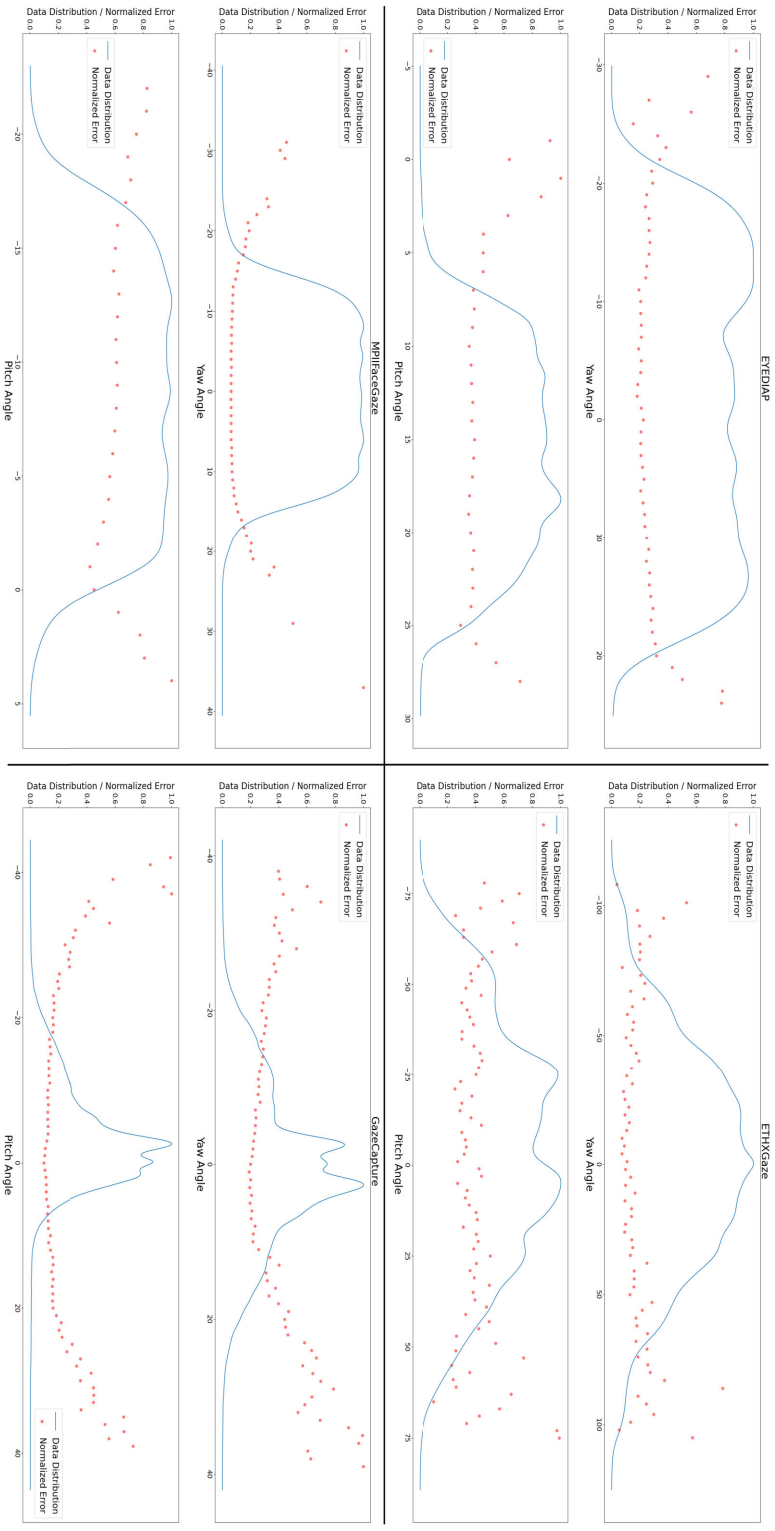

FIGURE 2. Comparison of scaled ground labels distribution and normalized quantized error of baseline model for pitch and yaw angles across appearance-based gaze estimation datasets: MPIIFaceGaze, EYEDIAP, GazeCapture, and ETHXGaze. The quantization range is 1 degree. It can be observed that all dataset distributions are long-tail or imbalanced, which poses challenges in training fair gaze estimation models. Significantly, the normalized error value is higher in the underrepresented areas across all datasets for both pitch and yaw. Notably, for the MPIIFaceGaze and EYEDIAP pitch plots, the error is shown to increase in regions of low data density, regardless of the absolute gaze angle value.

Once all ground truth values have been assigned to their respective bins, we proceed with uniform sampling within bins. Finally, we uniformly sample continuous ground truth values from the selected bin. In summary, UGS ensures approximately uniform data distribution in the whole range.

In appearance-based gaze estimation, a single image is associated with a two-dimensional ground truth gaze vector, consisting of pitch and yaw angles. To apply Uniform Gaze Sampling (UGS) to this 2D space, we adopt a strategy that balances both components within each training batch.

For a given batch of size B, we construct it by sampling in two halves. First, $B/2$ samples are selected using the UGS procedure on the pitch dimension. This involves uniformly sampling from the quantized pitch bins and then randomly selecting a data point from that bin, regardless of its yaw value. Next, the remaining $B/2$ samples are selected by applying the same UGS procedure to the yaw dimension, this time ignoring the pitch value during sampling. These two sets of samples are then combined to form the final training batch of size B. This approach ensures that every batch actively contributes to mitigating imbalances across both the pitch and yaw distributions. The precise procedure is detailed in Algorithm 1.

Algorithm 1 Batch Construction With UGS for 2D Gaze

Require: Full training dataset D, Batch size B
Ensure: A single training batch T of size B

$T \leftarrow []$
$N_{samples} \leftarrow B/2$
{Sample based on Pitch distribution}
for $i = 1$ to $N_{samples}$ **do**
  Uniformly select a random pitch bin p_{bin}
  Randomly select a sample s from D such that $s_{pitch} \in p_{bin}$
  Append s to T
end for
{Sample based on Yaw distribution}
for $i = 1$ to $N_{samples}$ **do**
  Uniformly select a random yaw bin y_{bin}
  Randomly select a sample s from D such that $s_{yaw} \in y_{bin}$
  Append s to T
end for
return T

C. UNIFORM ETHNICITY SAMPLING

We manually labeled participants in all four datasets with one ethnicity label: Asian, Black, Indian, Latino, or White. This revealed a second type of imbalance stemming from the uneven ethnic distribution of participants. As noted earlier, anatomical variations in eye structure across different ethnicities can significantly impact a model's performance. Figure 3 illustrates the extent of this issue by presenting the participant ethnicity distribution across the four benchmark datasets. A clear pattern emerges: all datasets are dominated by participants of White ethnicity, whose representation ranges from 45% to 77%. In stark contrast, other groups are significantly underrepresented. For example, the proportion of Asian participants is much smaller, varying from just 6% to 20%. This severe imbalance leads to skewed performance, as the models are better optimized for the majority group. For instance, on the GazeCapture dataset, the baseline model's Mean Gaze Angle Error (MGAE) is 3.21° for White participants but 5.34° for Asian participants.

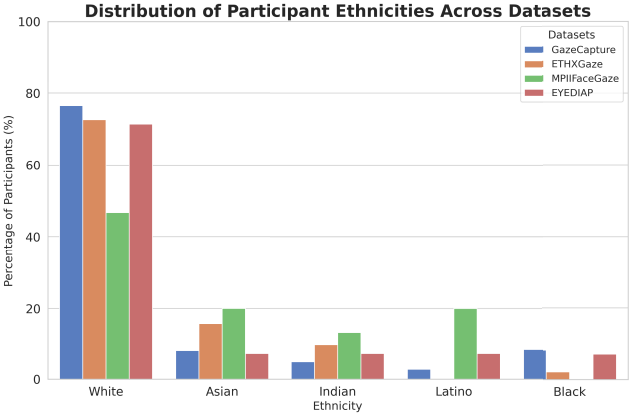

FIGURE 3. Distribution of individual ethnicities across appearance-based gaze estimation datasets: GazeCapture, ETHXGaze, MPIIFaceGaze, and EYEDIAP. Imbalance in favor of White ethnicity can be observed in all datasets. White ethnicity occupancy varies from 45% to 77%. On the other hand, the occupancy of Asian ethnicity varies from 6% to 20%. Notably, ETHXGaze does not include participants of Latino ethnicity, and MPIIFaceGaze does not include participants of Black ethnicity.

To overcome this issue, we propose Uniform Ethnicity Sampling (UES). By uniformly sampling ethnicity labels during training, UES aims to mitigate the impact of this uneven distribution. After an ethnicity is sampled, a data point is chosen randomly from the corresponding participants. This approach reduces the MGAE gap among different ethnicities by ensuring equal participation of all ethnicities in the training process.

D. BALANCED GAZE ERROR

Standard overall metrics are often insufficient for imbalanced regression problems as they can mask poor performance in underrepresented regions of the data distribution. A model may achieve a low overall error by performing well on the high-density center of the gaze range while failing on less frequent, extreme gaze angles, as shown in Figure 3. To provide a more honest assessment of a model's generalizability across the entire gaze space, we propose the Balanced Gaze Error (BGE), a balanced evaluation metric that ensures a fair evaluation of a gaze estimation model's performance.

To compute BGE, we first quantize gaze space into N bins of width w using Equation 1 and 2. Next, for each bin, we calculate the error using all values assigned to that bin with the formula

$$E_i = \frac{1}{n_i} \sum_{j=1}^{n_i} L(\hat{y}_j, y_j) \tag{3}$$

where n_i is the number of samples falling in bin i, $\hat{y}_j$ is predicted value, y_j is ground truth value, and $L()$ is a loss function. Finally, we compute BGE with the equation

$$BGE = \frac{1}{N}\sum_{i=1}^{N} E_i \tag{4}$$

It is worth noting that the greater the w, the closer the BGE values approach the value of the overall metric. A particular case is the w equal to the range of the entire set. Therefore, the general metric can be seen as a special case of BGE. Significantly, BGE is not limited to gaze estimation and can be used to fairly evaluate the performance of a model in every regression task. The choice of bin width w is a tunable hyperparameter. To balance resolution with statistical robustness, we empirically set the bin width hyperparameter to $w = 1°$. A more exhaustive sensitivity analysis is an important direction for future work.

E. BALANCED ETHNICITY ERROR

Given the significant performance disparities between ethnic groups due to data imbalance and anatomical variations, a standard overall metric is inadequate for assessing model fairness. To explicitly measure and address these performance gaps, we propose the Balanced Ethnicity Error (BEE), a metric designed to evaluate a model's performance equitably across all participant ethnicities.

The structure of Asian eyes is distinct. They are characterized by a covered inner corner, single or double eyelids, downward-facing eyelashes, and occasionally a fold at the corner. In contrast, European eyes typically feature an exposed inner corner and an external fold at the outer edge. Considering variations in anatomical eye structure across human races is crucial for a fair assessment of the appearance-based gaze estimation model, especially given the significant discrepancies in results for different ethnicities.

To evaluate appearance-based gaze estimation model performance considering the ethnicity imbalance, we propose Balanced Ethnicity Error (BEE). To compute BEE value, given M ethnicities in the dataset, we first calculate the error for all ethnicities separately using the formula

$$Et_k = \frac{1}{l_k}\sum_{j=1}^{l_k} L(\hat{y}_j, y_j) \tag{5}$$

where l_k is the number of samples corresponding to a given ethnicity label, and L is a loss function. Finally, we compute BEE with the equation

$$BEE = \frac{1}{M}\sum_{k=1}^{M} Et_k \tag{6}$$

where M is the number of ethnicities in a dataset. Notably, BEE cannot be evaluated given only ground truth labels and requires additional data labeling, which can be performed manually or by specialized software, e.g., DeepFace [47]. We note that this introduces potential biases, both from subjective labeling and from model-assisted prediction. We stress that these labels are used here solely for research evaluation, and that ethical safeguards and responsible use are critical in real-world applications. It is also limited to the gaze estimation domain. Still, it reveals how important it is to have an in-depth understanding of the particular problem before applying neural networks to solve it.

F. NEURAL NETWORK

We establish the baseline gaze estimation model to further evaluate the impact of proposed and state-of-the-art methods. We chose a convolutional neural network as this architecture is overwhelmingly prevalent in the field of appearance-based gaze estimation, a fact well-established by numerous foundational works and comprehensive surveys on the topic [10], [14]. In detail, two separate, one-layer dense networks with 512 inputs are applied on top of the ResNet18 [48] backbone to predict final gaze angle values. Pretrained ImageNet [49] weights were used for a ResNet18 backbone initialization, while the dense layers were initialized randomly. In summary, the baseline network is a multi-head CNN with two heads and a shared feature extractor.

IV. IMPLEMENTATION DETAILS

In this section, we depict the characteristics and the evaluation scenario for each of the used benchmark datasets. Moreover, we detail the hyperparameters used for training and the state-of-the-art methods.

A. DATASETS

The datasets for our experiments—GazeCapture, MPIIGaze, ETHXGaze, and EYEDIAP—were chosen for several key reasons. First, they are widely recognized as standard benchmarks in the appearance-based gaze estimation community, ensuring our results are comparable to a broad range of existing and future work. Second, they collectively represent a diverse set of real-world challenges, including the largest in-the-wild dataset (GazeCapture), significant within-person variations (MPIIGaze), and extreme head poses (ETHXGaze). Most importantly, these datasets were selected because they all prominently feature the gaze and ethnicity imbalances that are central to our investigation, making them ideal for evaluating the methods we propose. **GazeCapture** [24] is the largest available in-the-wild gaze dataset. It consists of almost 2.5 million images, which were collected using Apple devices, both phones and tablets. The images vary in camera location, including landscape and portrait modes. We followed the train-test split from the original work [24]. Validation data were omitted in both the training and test phases. The same subject filtration as in [50] was applied, and the same head pose data was used in image normalization.

MPIIGaze [13] is a well-established dataset for benchmarking in-the-wild gaze estimation methods. It comprises images from 15 participants. Images were collected over

several dozen days. Thanks to this, significant within-person variation, including illumination, make-up, and facial hair changes, makes it more challenging. Data was collected with the user's laptop camera. For the evaluation, images from MPIIFaceGaze [20], a subset of MPIIGaze, were used. The MPIIFaceGaze consists of about 2900 images for each participant. A one-person-out evaluation was performed in this work, as is done in [51].

ETHXGaze [22] covers a wide range of head poses, gaze directions, and lighting conditions. It consists of over 1 million high-resolution images collected from 110 participants. 95 participants were used for training. ETHXGaze does not provide annotations for the test set. The training part of person-specific evaluation data was used to compute Balanced Gaze Error and Balanced Ethnicity Error. It contains 15 participants and 200 images per participant.

EYEDIAP [23] comprises images from 16 participants recorded under two different target scenarios: floating target and screen target, and two different head pose scenarios: static and dynamic. This work used only the screen target scenario with static and dynamic head pose. No frames were skipped during extraction from videos. This resulted in 100 000 images in total. The one-person-out evaluation was performed to assess the performance of a trained neural network.

B. DATA NORMALIZATION

To ensure that the neural network learns the mapping from eye appearance to gaze direction, rather than relying on head orientation, we used a standard data normalization procedure. The primary goal of this process is to disentangle gaze and head pose. This is achieved by creating a virtual camera in a canonical 3D space. For each input image, the 3D head pose is estimated, and the image is then warped to match the viewpoint of this virtual camera. This crucial step unifies the input data by normalizing the head's scale and orientation, making it appear as if all images were captured from a consistent frontal perspective. The corresponding gaze vectors are also transformed into this new canonical coordinate system.

For reproducibility and a fair comparison with prior work, we adopted the specific normalization procedures established for each benchmark. For GazeCapture, MPIIFaceGaze, and EYEDIAP, the standard method described in [20] was applied. The virtual camera parameters were set as follows: distance from a face to 600 mm, focal length to 960 pixels, and resolution to 244 × 244 pixels. For ETHXGaze, the normalization procedure used in the original work [22] was followed, and the virtual camera resolution was set to 448 × 448 pixels.

C. NEURAL NETWORK TRAINING

In each experiment, the neural network was trained for ten epochs. RAdam [52] was used as an optimizer to reduce the impact of the learning rate on final results. A learning rate

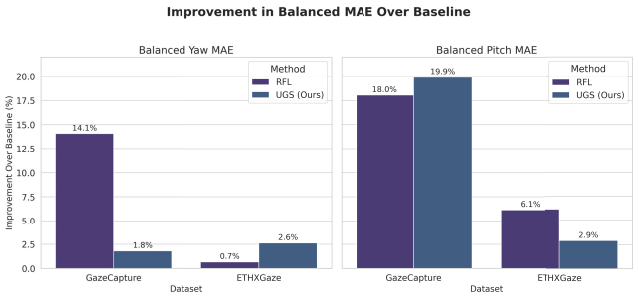

FIGURE 4. Visualization of the percentage improvement in Balanced Yaw and Pitch MAE over the baseline for the RFL and UGS methods on the GazeCapture and ETHXGaze datasets. Data is derived from Table 1.

was set to 10^{-5}, and a weight decay to 0.0005. The batch size was set to 256. Mean Squared Error (MSE) was used as a baseline loss function. Before feeding to a neural network, images were resized to 448×448 pixels and normalized using mean and standard deviation.

D. EVALUATION METRICS

To report evaluation results, we introduce proper metrics. For gaze angle, we use Mean Gaze Angle Error (MGAE), given by the equation

$$MGAE = \frac{180}{\pi G} \sum_{g=1}^{G} arccos\left(\frac{\hat{v_g} \cdot v_g}{||\hat{v_g}||||v_g||}\right) \tag{7}$$

where $\hat{v}$ is predicted gaze vector, v is ground truth gaze vector and G is number of predictions. Cosine similarity of vectors is an adequate metric for expressing the accuracy of gaze vector estimation. Please note, that using $MGAE$ as L function in BEE results in Balanced Ethnicity Mean Gaze Angle Error (Balanced Ethnicity MGAE). In the rest of the paper, we will use Balanced MGAE as equivalent to BEE.

To evaluate the accuracy of pitch and yaw separately, we use Mean Absolute Error (MAE). Thus, BGE becomes Balanced Yaw MAE for yaw gaze vector component, and Balanced Pitch MAE for pitch gaze vector component.

E. STATE-OF-THE-ART COMPARISON

We employed state-of-the-art imbalanced regression and imbalanced gaze estimation methods to compare against the baseline and proposed uniform samplers. Hyperparameters of them are detailed in this section.

Regression Focal Loss (RFL) [25], [27] is given by the equation

$$L_f = L + [\sigma(\beta L)]^{\gamma} \tag{8}$$

where L is the MAE, σ is sigmoid function, and β and γ are hyperparameters controlling shape of the final loss. β and γ were set to 3 and 10, respectively.

Feature Distribution Smoothing (FDS) [25] kernel was set to Gaussian.

Dense Weight (DW) [26] α was set to 1, and the bandwidth selection method was the Silverman algorithm.

TABLE 1. Comparison of imbalanced gaze experiments on GazeCapture, MPIIFaceGaze, ETHXGaze, and EYEDIAP. Lower MAE is better. Best results per column are bolded. Green color indicates improvement over the baseline.

Dataset	Method	Balanced Yaw MAE↓	Balanced Pitch MAE↓	Yaw MAE↓	Pitch MAE↓
GazeCapture	Baseline	3.91 ± 0.15	4.82 ± 0.46	1.92 ± 0.11	2.47 ± 0.06
	RFL	**3.36 ± 0.06**	3.95 ± 0.14	1.75 ± 0.07	2.48 ± 0.10
	FDS	9.73 ± 0.68	8.94 ± 0.38	3.93 ± 0.40	3.17 ± 0.07
	BMC	3.74 ± 0.19	5.13 ± 0.23	1.91 ± 0.11	2.52 ± 0.15
	DW	3.70 ± 0.14	4.36 ± 0.20	1.97 ± 0.04	2.53 ± 0.05
	L2CS	3.74 ± 0.11	4.02 ± 0.48	1.72 ± 0.03	2.34 ± 0.06
	UGS (Ours)	3.84 ± 0.25	**3.86 ± 0.21**	2.00 ± 0.10	2.56 ± 0.12
MPIIFaceGaze	Baseline	5.09 ± 0.17	**2.94 ± 0.41**	2.45 ± 0.29	2.75 ± 0.29
	RFL	5.00 ± 0.08	3.01 ± 0.50	2.43 ± 0.36	2.84 ± 0.40
	FDS	5.15 ± 0.12	3.07 ± 0.56	2.47 ± 0.35	2.78 ± 0.36
	BMC	**4.99 ± 0.12**	2.98 ± 0.51	2.51 ± 0.26	2.91 ± 0.47
	DW	5.62 ± 0.18	3.10 ± 0.26	2.75 ± 0.02	2.86 ± 0.20
	L2CS	5.57 ± 0.12	3.25 ± 0.40	2.70 ± 0.15	2.96 ± 0.20
	UGS (Ours)	5.11 ± 0.06	3.08 ± 0.43	2.45 ± 0.26	2.89 ± 0.31
ETHXGaze	Baseline	4.58 ± 0.03	3.42 ± 0.14	3.46 ± 0.04	3.20 ± 0.21
	RFL	4.55 ± 0.10	**3.21 ± 0.12**	3.31 ± 0.05	2.97 ± 0.14
	FDS	42.86 ± 16.21	31.59 ± 11.98	27.34 ± 9.93	21.69 ± 8.88
	BMC	4.68 ± 0.04	3.59 ± 0.22	3.55 ± 0.13	3.31 ± 0.23
	DW	4.62 ± 0.06	3.58 ± 0.20	3.79 ± 0.09	3.50 ± 0.19
	L2CS	4.71 ± 0.35	3.60 ± 0.68	3.56 ± 0.28	3.76 ± 0.82
	UGS (Ours)	**4.46 ± 0.09**	3.32 ± 0.06	3.55 ± 0.04	3.20 ± 0.05
EYEDIAP	Baseline	3.52 ± 0.09	3.98 ± 0.20	2.75 ± 0.09	3.18 ± 0.14
	RFL	**3.42 ± 0.10**	4.10 ± 0.09	2.79 ± 0.11	3.11 ± 0.03
	FDS	10.97 ± 0.92	7.44 ± 0.35	8.29 ± 0.57	5.70 ± 0.54
	BMC	3.48 ± 0.15	4.00 ± 0.16	2.82 ± 0.16	3.32 ± 0.18
	DW	3.80 ± 0.06	4.09 ± 0.09	3.06 ± 0.06	3.28 ± 0.08
	L2CS	3.93 ± 0.08	4.14 ± 0.09	3.24 ± 0.11	2.95 ± 0.16
	UGS (Ours)	3.87 ± 0.05	**3.96 ± 0.19**	3.02 ± 0.02	3.07 ± 0.15

Balance Mean Squared Error (BMC) [38] noise was set as a learnable parameter with a learning rate 10^{-5} and an initial value of 1.

L2CS-Net (L2CS) [28] bin size was 3 degrees for pitch and yaw. The number of bins was 30 for EYEDIAP, GazeCapture, and MPIIFaceGaze and 80 for ETHXGaze.

V. RESULTS

We report the results for four appearance-based gaze estimation datasets. Because the appearance-based gaze estimation neural networks are sensitive to randomness, we repeated the training and evaluation process four times using different random seeds. The results presented are the average and standard deviation of all runs.

We evaluate proposed and state-of-the-art methods against the baseline using both the standard (overall) and proposed balanced metrics. EYEDIAP dataset was excluded from the ethnicity experiments because it contains only one participant from each non-white ethnicity. This makes the one-person-out evaluation procedure ill-conditioned because, for the non-white participants, the ethnicity label in a test set is not present in a train set.

A. IMBALANCED GAZE

We trained the baseline network standalone and with UGS and compared its performance with state-of-the-art imbalanced regression methods, namely RFL, FDS, DW, BMC, and L2CS, applied to the baseline. Each approach was evaluated using overall and Balanced Gaze Error metrics. Results are presented in Table 1.

GazeCapture: RFL achieves the best results for balanced yaw. It outperforms the baseline by 0.55 and has the lowest standard deviation. It is worth noting that all methods except FDS gave better results than baseline. UGS improved balanced yaw by 0.07. For balanced pitch, the best-performing method is UGS. It outperforms the baseline by 0.96 and reduces the standard deviation by 0.25. RFL, L2CS, and DW improved the balanced pitch by 0.87, 0.80, and 0.49, respectively. For imbalanced metrics, only L2CS and RFL improve the result with respect to the baseline.

MPIIFaceGaze: Only BMC and RFL improve over the baseline for balanced yaw by 0.10 and 0.09, respectively. Notably, none of the methods outperforms the baseline in balanced pitch. Similarly, the baseline achieves the best results in overall metrics. We hypothesize that this lack of improvement is a direct consequence of the dataset's small size, which comprises only 15 participants. When combined with the rigorous one-person-out evaluation protocol, the training set can become extremely sparse for certain gaze angles. Forcing the model to focus on these few, hard-to-learn examples with imbalanced learning techniques appears to destabilize the training process rather than improve generalization. This suggests that the inherent data scarcity creates a condition where these methods are ineffective. As in GazeCapture, significant discrepancies can be observed between balanced and overall yaw. On the other hand, balanced and overall pitch scores are similar, which indicates that the pitch distribution is balanced within MPIIFaceGaze.

ETHXGaze: UGS achieves the best results in balanced yaw, improving over the baseline by 0.12. Notably, this is the only method that consistently improved balanced yaw across all experimental runs. For a balanced pitch, only RFL and UGS perform better than the baseline by 0.21 and 0.1, respectively, while the other methods worsen the

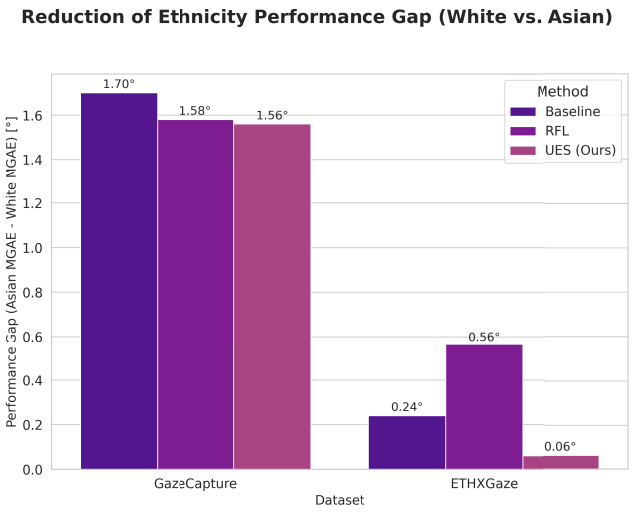

FIGURE 5. Visualization of the MGAE performance gap (Asian MGAE - White MGAE) for the Baseline, RFL, and UES methods. A smaller bar indicates a fairer model. Our proposed UES method shows a substantial reduction in the gap on the ETHXGaze dataset. Data is derived from Table 2.

metric. For imbalanced metrics, RFL outperforms baseline by 0.15 and 0.23 for yaw and pitch, respectively. Similar to MPIIFaceGaze and GazeCapture, significant discrepancies can be observed between balanced and overall yaw, and balanced and overall pitch scores are comparable.

EYEDIAP: RFL and BMC outperform the baseline for balanced yaw by 0.10 and 0.04, respectively. In a balanced pitch, only UGS improves over the baseline by a slight margin, namely 0.02. For imbalanced yaw, no method improves over the baseline. For imbalanced pitch, L2CS, UGS, and RFL reduce error by 0.23, 0.11, and 0.07, respectively. Notably, of all the datasets, EYEDIAP has the most similar balanced and overall metrics. This confirms that the EYEDIAP ground truth distribution is the closest to the uniform, which can also be observed in Figure 2.

Summary: Imbalanced regression methods improve the performance of the baseline for balanced metrics among all datasets except MPIIFaceGaze. RFL and UGS give the best results among tested approaches. Moreover, RFL often improves baseline results for imbalanced metrics, too. On the other hand, FDS led to a substantial degradation in performance across all datasets except MPIIFaceGaze. To visually summarize the impact of the most effective methods, Figure 4 illustrates the percentage improvement in both Balanced Yaw and Pitch MAE. The visualization clearly shows that UGS and RFL provide significant gains on the GazeCapture and ETHXGaze datasets, with our proposed UGS method yielding the highest improvement of nearly 20% in Balanced Pitch MAE on GazeCapture.

B. IMBALANCED ETHNICITY

Although, unlike UES, the state-of-the-art methods are unaware of the ethnicity imbalance in the data, and their foundations are mainly built upon an uneven ground truth distribution, we decided to test their effect on the BEE metric. This is motivated by our observation that the discrepancies in accuracy for different ethnicity groups are also apparent in the training phase of the appearance-based gaze estimation neural network. Therefore, our goal was to investigate whether a greater emphasis on hard examples during training could reduce the difference in performance between different ethnicities.

We trained the baseline network standalone and with UES and compared its performance with state-of-the-art imbalanced regression methods, namely RFL, FDS, DW, BMC, and L2CS, applied to the baseline. Each approach was evaluated using overall and Balanced Ethnicity Error metrics. Results are presented in Table 2.

GazeCapture: L2CS achieves the best results in Balanced Ethnicity Mean Gaze Angle Error (Balanced MGAE), outperforming the baseline by 0.36°. Moreover, RFL and UES improve over the baseline by 0.29° and 0.17°, respectively. Additionally, L2CS is the best-performing method for overall (imbalanced) MGAE and for White, Asian, Black, and Latino ethnicities. RFL achieves the lowest MGAE for Indian ethnicity. Notably, UES improves over the baseline for all ethnicities except White. Using the UES as an example, it can be observed that although the MGAE for each ethnic group, except for White, is lower than the baseline, the overall MGAE remains almost unchanged. This demonstrates the high prevalence of White ethnicity over others in the overall metrics.

MPIIFaceGaze: The baseline achieves the best results in all metrics. The ineffectiveness of the imbalanced learning methods on this dataset is likely attributable to its structural limitations. With only 15 total participants and a strict one-person-out evaluation protocol, the training data for minority ethnicities becomes critically sparse. For instance, when one of the Indian participants is held out for testing, only a single participant of that ethnicity remains for training. Under these conditions of extreme data scarcity, methods designed to up-weight or re-sample minority groups lack sufficient data to learn meaningful features and instead introduce noise, failing to improve the model's performance. Similar to GazeCapture, a significant discrepancy can be observed between the results of different ethnicities, in particular, White and Asian.

ETHXGaze: RFL achieves the best results in balanced and overall MGAE, outperforming the baseline by 0.25° and 0.23°, respectively. It should be emphasized that the only method, besides FRL, that outperforms baseline results is UES. Namely, UES improves balanced and overall MGAE by 0.21° and 0.13°, respectively. Moreover, UES achieves the best results for Asian ethnicity, while RFL performs better for White and Indian ethnicities.

Summary: UGS, RFL, and L2CS are efficient methods to deal with ethnicity imbalance in appearance-based gaze estimation datasets. Within all benchmarks, one can observe worse results for Asians than for other ethnicities. To better illustrate the impact on model fairness, Figure 5 visualizes

TABLE 2. Comparison of ethnicity-related experiments on GazeCapture, MPIIFaceGaze, and ETHXGaze. MPIIFaceGaze does not include Black ethnicity, and the ETHXGaze test set does not include both Black and Latino ethnicities. Best results are bolded. Green color indicates improvement over the baseline.

Dataset	Method	Balanced Ethnicity MGAE↓	MGAE↓ (Overall)	MGAE↓ (White)	MGAE↓ (Asian)	MGAE↓ (Black)	MGAE↓ (Indian)	MGAE↓ (Latino)
GazeCapture	Baseline	3.83 ± 0.07	3.46 ± 0.12	3.40 ± 0.09	5.10 ± 0.17	3.50 ± 0.17	3.26 ± 0.17	3.92 ± 0.03
	RFL	3.54 ± 0.14	3.31 ± 0.12	3.28 ± 0.08	4.86 ± 0.04	3.26 ± 0.26	2.73 ± 0.20	3.59 ± 0.17
	FDS	6.25 ± 0.34	5.71 ± 0.34	5.59 ± 0.28	7.62 ± 0.30	5.73 ± 0.34	5.69 ± 0.43	6.59 ± 0.32
	BMC	3.82 ± 0.21	3.48 ± 0.18	3.42 ± 0.14	5.02 ± 0.15	3.58 ± 0.33	3.20 ± 0.18	3.92 ± 0.27
	DW	3.81 ± 0.05	3.52 ± 0.05	3.48 ± 0.03	4.98 ± 0.15	3.56 ± 0.14	3.13 ± 0.09	3.92 ± 0.16
	L2CS	**3.47 ± 0.09**	3.18 ± 0.07	3.13 ± 0.05	4.73 ± 0.10	3.18 ± 0.27	2.79 ± 0.02	3.51 ± 0.18
	UES (Ours)	3.66 ± 0.15	3.50 ± 0.23	3.51 ± 0.23	5.07 ± 0.07	3.16 ± 0.16	2.93 ± 0.19	3.67 ± 0.20
MPIIFaceGaze	Baseline	**4.22 ± 0.13**	4.08 ± 0.07	4.00 ± 0.19	4.35 ± 0.43	–	5.02 ± 0.26	3.56 ± 0.20
	RFL	4.26 ± 0.07	4.12 ± 0.03	4.03 ± 0.28	4.40 ± 0.37	–	5.09 ± 0.12	3.65 ± 0.19
	FDS	4.27 ± 0.05	4.12 ± 0.03	3.98 ± 0.17	4.54 ± 0.28	–	5.00 ± 0.09	3.64 ± 0.16
	BMC	4.34 ± 0.17	4.21 ± 0.17	4.08 ± 0.30	4.51 ± 0.33	–	5.10 ± 0.11	3.68 ± 0.23
	DW	4.39 ± 0.18	4.26 ± 0.18	4.13 ± 0.33	4.67 ± 0.36	–	5.10 ± 0.12	3.70 ± 0.28
	L2CS	4.44 ± 0.16	4.27 ± 0.12	4.11 ± 0.25	4.76 ± 0.28	–	5.18 ± 0.19	3.81 ± 0.27
	UES (Ours)	4.38 ± 0.06	4.21 ± 0.04	4.03 ± 0.11	4.73 ± 0.19	–	5.12 ± 0.10	3.72 ± 0.06
ETHXGaze	Baseline	4.80 ± 0.20	4.84 ± 0.13	4.81 ± 0.04	5.05 ± 0.17	–	4.54 ± 0.42	–
	RFL	**4.55 ± 0.12**	4.61 ± 0.08	4.51 ± 0.05	5.07 ± 0.32	–	4.06 ± 0.14	–
	FDS	37.45 ± 13.77	37.52 ± 13.76	36.82 ± 13.59	39.85 ± 14.38	–	35.66 ± 13.44	–
	BMC	5.09 ± 0.23	5.06 ± 0.15	4.90 ± 0.13	5.43 ± 0.13	–	4.93 ± 0.59	–
	DW	5.26 ± 0.32	5.28 ± 0.22	5.33 ± 0.11	5.17 ± 0.39	–	5.27 ± 0.54	–
	L2CS	5.07 ± 0.58	5.22 ± 0.67	5.49 ± 0.86	4.83 ± 0.38	–	4.90 ± 0.50	–
	UES (Ours)	4.59 ± 0.15	4.71 ± 0.10	4.79 ± 0.10	4.85 ± 0.15	–	4.14 ± 0.36	–

the MGAE performance gap between White and Asian participants. The chart provides strong evidence for the effectiveness of our proposed method; while the baseline model exhibits a notable performance gap on the ETHXGaze dataset, UES reduces this gap by 75% to just 0.06°. This demonstrates a significant step towards creating more equitable and fair gaze estimation systems. Notably, the predominant ethnicity has the most significant impact on overall metrics, demonstrating the weakness of such metrics in a fair evaluation.

VI. LIMITATIONS

A key limitation is the reduced efficacy of imbalanced learning methods on smaller datasets. As demonstrated in our analysis of the MPIIFaceGaze results, when the training data is inherently sparse due to a low number of participants and challenging evaluation protocols, these techniques can struggle to improve generalization. This highlights a boundary condition for their effectiveness, suggesting that a minimum threshold of diverse data is necessary for these balancing methods to be beneficial.

Moreover, imbalanced learning methods also cause a drop in overall metrics. This can be a drawback when the distribution of the test set can be estimated and is not uniform.

Both of these limitations are interesting issues for future research.

VII. CONCLUSION

We identified two types of imbalance in appearance-based gaze estimation, namely gaze imbalance and ethnicity imbalance, and we proved the negative impact of these phenomena on the performance and fair evaluation of the appearance-based gaze estimation neural networks. We, therefore, proposed Uniform Gaze Sampling and Uniform Ethnicity Sampling to mitigate this issue. Further, we compared the proposed methods with state-of-the-art imbalanced learning techniques and gaze space quantization. Moreover, we introduced two balance metrics, namely, Balanced Gaze Error (BGE) and Balanced Ethnicity Error (BEE), for a fair assessment of appearance-based gaze estimation model performance. Finally, we demonstrated the positive impact of our re-sampling and state-of-the-art imbalanced regression methods on balanced and overall metrics within the four most popular benchmark datasets.

Importantly, the evaluation framework centered on BGE has implications that extend well beyond gaze estimation. BGE provides a generalizable template for rigorously assessing model performance in any domain grappling with long-tail regression challenges, such as in medical diagnostics or financial forecasting, where extreme values are rare yet critical. By offering a clear methodology to quantify performance across all regions of a continuous target variable, not just the dense center, we believe this approach can foster the broader adoption of imbalance-aware evaluation for regression tasks throughout the machine learning landscape.

By raising awareness of gaze and ethnicity imbalances and providing potential solutions and metrics, we hope to encourage researchers and practitioners in the eye-tracking community to consider and address gaze data imbalance in their work. We point out the possibility of combining all types of imbalanced regression techniques as a future extension of our work.

Furthermore, while our work focuses on addressing data imbalance, we recognize that the computational cost of deep learning models is a critical factor for real-world deployment. For appearance-based gaze estimation to be practical on resource-constrained devices like smartphones or AR/VR headsets, models must be both accurate and efficient. Future

research could therefore explore the synergy between our data-balancing techniques and established model compression methods. Promising directions include investigating the impact of model pruning to remove redundant parameters [53], quantization to reduce memory footprint and leverage faster integer arithmetic [54], and knowledge distillation to transfer knowledge from a large model to a more compact one [55]. Pursuing these optimizations would be a valuable step towards developing gaze estimation systems that are not only robust and fair but also computationally efficient.

Finally, our findings offer a key takeaway for practitioners: there is an explicit trade-off between optimizing for overall performance versus ensuring fairness and robust worst-case performance. The choice of model should be guided by the specific application and its expected test distribution. For systems where the real-world data is known to mirror the imbalanced training set, a standard model might be preferred. However, for applications where reliability and equity across all scenarios and user groups are paramount, such as assistive technologies or medical devices, a model trained with balancing techniques is the superior choice, even if it results in a minor drop in the overall metric.

VIII. ACKNOWLEDGMENT

The author gratefully acknowledges the Polish High-Performance Computing Infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities.

REFERENCES

[1] R.-M. Rahal and S. Fiedler, "Understanding cognitive and affective mechanisms in social psychology through eye-tracking," *J. Experim. Social Psychol.*, vol. 85, Nov. 2019, Art. no. 103842. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022103119300782

[2] M. L. Mele and S. Federici, "Gaze and eye-tracking solutions for psychological research," *Cognit. Process.*, vol. 13, no. S1, pp. 261–265, Aug. 2012.

[3] P. Majaranta and A. Bulling, "Eye tracking and eye-based human–computer interaction," in *Advances in Physiological Computing*. Cham, Switzerland: Springer, 2014, pp. 39–65.

[4] R. J. K. Jacob and K. S. Karn, "Eye tracking in human–computer interaction and usability research," in *The Mind's Eye*. Amsterdam, The Netherlands: Elsevier, 2003, pp. 573–605.

[5] S. Białowąs and A. Szyszka, "Eye-tracking in marketing research," *Manag. Econ. Innov.-Methods Instrum.*, vol. 1, no. 69, pp. 91–104, 2019.

[6] M. Wedel and R. Pieters, "Eye tracking for visual marketing," *Found. Trends Marketing*, vol. 1, no. 4, pp. 231–320, 2006.

[7] M. Mokatren, T. Kuflik, and I. Shimshoni, "3D gaze estimation using RGB-IR cameras," *Sensors*, vol. 23, no. 1, p. 381, Dec. 2022.

[8] J. Chen and Q. Ji, "3D gaze estimation with a single camera without IR illumination," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[9] X. Zhou, H. Cai, Y. Li, and H. Liu, "Two-eye model-based gaze estimation from a Kinect sensor," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1646–1653.

[10] Y. Cheng, H. Wang, Y. Bao, and F. Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 7509–7528, Dec. 2024.

[11] E. Wood and A. Bulling, "EyeTab: Model-based gaze estimation on unmodified tablet computers," in *Proc. Symp. Eye Tracking Res. Appl.*, Mar. 2014, pp. 207–210.

[12] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, Jun. 2018, pp. 1–10.

[13] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.

[14] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Eye gaze estimation: A survey on deep learning-based approaches," *Expert Syst. Appl.*, vol. 199, Aug. 2022, Art. no. 116894.

[15] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci.*, vol. 513, pp. 429–441, Mar. 2020.

[16] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," 2013, *arXiv:1305.1707*.

[17] H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, "Imbalance class problems in data mining: A review," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, p. 1552, Jun. 2019.

[18] T. D. Brown Jr., F. C. Dane, and M. D. Durham, "Perception of race and ethnicity," *J. Social Behav. Personality*, vol. 13, no. 2, pp. 295–306, 1998.

[19] S. McClure, M. Poole, and E. P. Anderson-Fye, "Race, ethnicity, and human appearance," in *Encyclopedia of Body Image and Human Appearance*. Oxford, U.K.: Academic, 2012, pp. 707–710.

[20] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2299–2308.

[21] X. Zhang, Y. Sugano, and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, Jun. 2018, pp. 1–9.

[22] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 365–381.

[23] K. A. F. 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 *Proc. Symp. Eye Tracking Res. Appl.* New York, NY, USA: ACM, Mar. 2014, pp. 255–258, doi: 10.1145/2578153.2578190.

[24] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," 2016, *arXiv:1606.05814*.

[25] Y. Yang, K. Zha, Y. Chen, H. Wang, and D. Katabi, "Delving into deep imbalanced regression," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11842–11851.

[26] M. Steininger, K. Kobs, P. Davidson, A. Krause, and A. Hotho, "Density-based weighting for imbalanced regression," *Mach. Learn.*, vol. 110, no. 8, pp. 2187–2211, Aug. 2021.

[27] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, Sep. 2022.

[28] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, and L. Dinges, "L2CS-Net: Fine-grained gaze estimation in unconstrained environments," in *Proc. 8th Int. Conf. Frontiers Signal Process. (ICFSP)*, Oct. 2023, pp. 98–102.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[30] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.*, 2005, pp. 878–887.

[31] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.

[32] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9268–9277.

[33] K. Cao, C. Wei, A. Gaidon, N. Aréchiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1567–1578.

[34] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2001, pp. 973–978.

[35] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.
[36] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, "SMOTE for regression," in *Proc. Portuguese Conf. Artif. Intell.*, 2013, pp. 378–389.
[37] P. Branco, L. Torgo, and R. P. Ribeiro, "SMOGN: A pre-processing approach for imbalanced regression," in *Proc. 1st Int. Workshop Learn. Imbalanced Domains, Theory Appl.*, 2017, pp. 36–50.
[38] J. Ren, M. Zhang, C. Yu, and Z. Liu, "Balanced MSE for imbalanced visual regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7916–7925.
[39] G. C. Gutiérrez-Tobal, D. Álvarez, F. Vaquerizo-Villar, A. Crespo, L. Kheirandish-Gozal, D. Gozal, F. del Campo, and R. Hornero, "Ensemble-learning regression to estimate sleep apnea severity using at-home oximetry in adults," *Appl. Soft Comput.*, vol. 111, Nov. 2021, Art. no. 107827.
[40] D. Kim, H. Yu, H. Lee, E. Beighley, M. Durand, D. E. Alsdorf, and E. Hwang, "Ensemble learning regression for estimating river discharges using satellite altimetry data: Central Congo river as a test-bed," *Remote Sens. Environ.*, vol. 221, pp. 741–755, Feb. 2019.
[41] R. Liu, Y. Liu, J. Duan, F. Hou, L. Wang, X. Zhang, and G. Li, "Ensemble learning directed classification and regression of hydrocarbon fuels," *Fuel*, vol. 324, Sep. 2022, Art. no. 124520.
[42] A. K. Chaudhary, R. Kothari, M. Acharya, S. Dangi, N. Nair, R. Bailey, C. Kanan, G. Diaz, and J. B. Pelz, "RITNet: Real-time semantic segmentation of the eye for gaze tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3698–3702.
[43] R. S. Kothari, A. K. Chaudhary, R. J. Bailey, J. B. Pelz, and G. J. Diaz, "EllSeg: An ellipse segmentation framework for robust gaze tracking," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 5, pp. 2757–2767, May 2021.
[44] T. Guo, Y. Liu, H. Zhang, X. Liu, Y. Kwak, B. I. Yoo, J.-J. Han, and C. Choi, "A generalized and robust method towards practical gaze estimation on smart phone," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1131–1139.
[45] H. Hu, C. Wu, K. Lin, and T. Liu, "HG-Net: Hybrid coarse-fine-grained gaze estimation in unconstrained environments," in *Proc. 9th Int. Conf. Virtual Reality (ICVR)*, May 2023, pp. 1–6.
[46] P. Blignaut and D. Wium, "Eye-tracking data quality as affected by ethnicity and experimental design," *Behav. Res. Methods*, vol. 46, no. 1, pp. 67–80, Mar. 2014.
[47] S. Serengil and A. Ozpinar, "A benchmark of facial recognition pipelines and co-usability performances of modules," *J. Inf. Technol.*, vol. 17, no. 2, pp. 95–107, 2024. [Online]. Available: https://dergipark.org.tr/en/pub/gazibtd/issue/84331/1399077
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[50] S. Park, S. De Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, "Few-shot adaptive gaze estimation," 2019, *arXiv:1905.01941*.
[51] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," 2017, *arXiv:1711.09017*.
[52] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," 2019, *arXiv:1908.03265*.
[53] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1135–1143.
[54] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision*. Boca Raton, FL, USA: CRC Press, 2022, pp. 291–326.
[55] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

JAN GLINKO received the B.S. and M.S. degrees in automatic control and robotics from Gdańsk University of Technology, in 2019 and 2020, respectively, where he is currently pursuing the Ph.D. degree in computer science with the Faculty of Electronics, Telecommunications and Informatics, Poland.

Since 2020, he has been a Research Assistant with the Department of Decision Systems and Robotics, Gdańsk University of Technology. His research interests include the personalization of appearance-based gaze estimation neural networks using meta-learning, deep neural networks, computer vision, and synthetic datasets.