GESCAM : A Dataset and Method on Gaze Estimation for Classroom Attention Measurement

Athul M. Mathew Arshad Ali Khan Thariq Khalid Riad Souissi
Elm Company, Saudi Arabia
{amathew, arkhan, tkadavil, rsouissi}@elm.sa
https://athulmmathew.github.io/GESCAM/

Abstract

Human gaze provides crucial insights into individual attention during social or educational interactions. Attention systems often rely on head and facial features to predict gaze direction, but reliable gaze target detection (GTD) requires rich contextual cues. These cues inform the system about an individual's position within a scene and the surrounding objects they might be interacting with. Our paper proposes attention measurement using GTD in educational classrooms, leveraging a synthetic dataset called GESCAM (Gaze Estimation based Synthetic Classroom Attention Measurement). This dataset was meticulously generated using 3D modelling, animation, simulation, and rendering techniques comprising 60,000 images with 650,000 instances of individuals (students, teachers) engaged in various activities, including looking at blackboard, notebooks, mobile phones etc. Our novel network trained on GESCAM proficiently identifies gaze fixations within complex classroom scenes, offering insights into human attention in classrooms across diverse contexts.

1. Introduction

Estimating the attention levels in a classroom using computer vision is important for numerous reasons including student engagement monitoring, identifying in-attentive marginalised students, and optimising classroom environments. The modern educational landscape is characterized by information overload. This necessitates the development of effective methods for educators to gauge student engagement, a crucial factor in learning outcomes. By understanding engagement levels, educators can adapt their pedagogical strategies accordingly. For example, if a significant portion of the class appears disengaged, educators can implement strategies to re-engage students and ultimately enhance learning outcomes. Promptly identifying signs of distraction or disengagement allows educators to offer timely support. This proactive approach can help prevent academic challenges from escalating and contribute to fostering optimal teacher-student engagement within an educational institution.

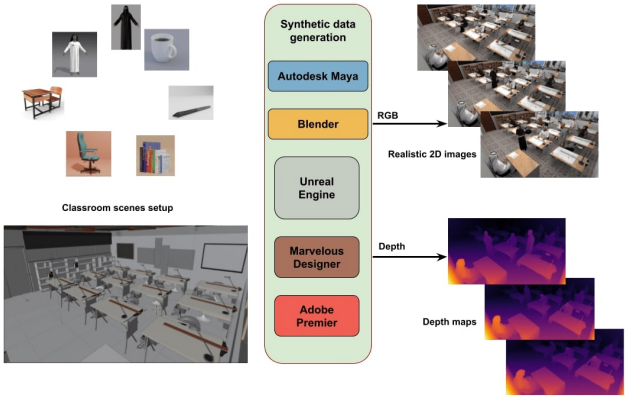

Figure 1. **Synthetic Classroom Generation via 3D Rendering.** This figure illustrates the process of synthesizing classroom videos using 3D assets. A suite of 3D rendering tools enables the creation of customized classroom scenarios and animations of individuals within those environments.

Existing attention measurement methods primarily rely on gaze estimation and tracking techniques. These techniques, such as gaze point or direction estimation [29], involve identifying where a person is looking based on facial features. However, they are prone to inaccuracies, particularly in challenging real-world conditions like low lighting or when subjects wear glasses. GTD [26] offers an alternative approach. It focuses on the relationship between a person's position within a scene and the surrounding objects within their field of view. Unlike gaze estimation, GTD aims to identify what a person is looking at, not just where their gaze is directed. However, capturing specific objects from a distance using real-world 2D gaze estimation methods often suffers from high error margins [29]. Addition-

ally, real-world data collection for training such systems can be expensive, time-consuming, and impractical, especially in classrooms where privacy concerns and logistics complicate data gathering. Synthetic datasets offer a solution to these limitations. They provide a virtually unlimited source of diverse data at a significantly lower cost and effort compared to real-world data collection.

This paper introduces the GESCAM dataset, a novel resource facilitating groundbreaking research in educational technology and personalized learning. GESCAM leverages GTD for classroom attention estimation. Our primary contribution lies in demonstrating the ability to synthesize a multitude of realistic classroom environments. This allows for a comprehensive understanding of student interactions with objects and the quantification of attention levels across diverse settings. Fig. 1 showcases the generation process of realistic 2D images and depth maps using 3D assets. We meticulously designed intricate classroom scenes with various objects, textures, and lighting conditions (detailed tool descriptions are provided in Section 3). This approach offers complete control over scene composition, object placement, camera viewpoints, and viewing angles, resulting in synthetic images that closely resemble real-world classrooms.

The synthetic data within GESCAM is meticulously annotated with bounding boxes, masks, class labels and gaze fixation points. This labeled ground truth facilitates the training and evaluation of GTD and attention measurement algorithms. The inherent flexibility and scalability of our data generation process enabled the creation of a robust dataset specifically tailored for classroom attention measurement tasks. Section 4 provides further details on training and evaluation of various algorithms with GESCAM dataset. The proposed GESCAM dataset empowers researchers to make significant advancements in the field of classroom attention measurement. It facilitates the development and validation of automated attention estimation systems, enabling deeper investigations into the relationship between gaze patterns and specific learning outcomes. This knowledge can be harnessed to design adaptive learning environments that react to individual student attention levels. Ultimately, GESCAM paves the way for personalized instruction, optimizing student engagement, comprehension, and ultimately revolutionizing our understanding of learning and enhancing the effectiveness of educational practices. Some example images from the dataset are shown in Fig. 2.

Our proposed neural network tackles the challenges of discerning head pose, gaze orientation, and pinpointing objects of interest within the scene. These objects can range from tablets and mobile phones to blackboards, teachers, and student interactions. Notably, our method exhibits reasonable accuracy in predicting gaze targets even when presented with the challenge of only seeing the back of a head from a distance. Our two-pathway architecture for GTD combines head information with the scene itself. This facilitates robust training and evaluation within an end-to-end inference pipeline. By enabling dense measurements of natural gaze behavior, our approach offers a promising avenue for understanding human attention measurement in educational settings.

The remainder of this paper is organized as follows. Sec. 2 reviews state-of-the-art attention estimation (AE) and gaze Estimation (GE) using GTD methods with focus on synthetic datasets. Sec. 3 walks through GESCAM dataset including our proposed architecture for synthetic data construction using various characters and artefacts orchestration and rendering technologies. Sec. 4 discusses our proposed neural network in greater detail. Sec. 5 discusses experimental requirements and evaluation criteria of a synthetic dataset based on established benchmarks using Area Under the Curve (AUC), Distance and Angular metrics etc. In Sec. 6, we summarize our research findings and, Sec. 7 sets out our goals for extending this work in the future.

2. Related Work

2.1. Gaze Estimation

Several research studies have explored the use of gaze estimation for measuring classroom attention. One of the challenging problems in attention estimation based on gaze target detection (GTD) and gaze estimation (GE) methods is that there is a reality gap between the physical and rendered data. To address the issue, various approaches have been proposed. Matching the simulated data with physical reality using high-quality rendering is one approach proposed by [32]. However, [14] suggests that using realistic RGB rendering alone has had limited success for transferring to real tasks. However [25] points out that incorporating realistic simulation of depth information can allow models, trained on rendered images to transfer reasonably well to real-world scenarios. Concomitantly, combining data from high-quality simulators with other approaches like fine-tuning can also reduce the number of labelled samples, required in the real-world data [28] specifically in a classroom environment.

Research on automatic gaze analysis can be categorized into two main areas: GE and GTD [4] [7] [27]. GE estimates the direction of a person's gaze, typically in 3D, and does not necessarily focus on precisely locating the object of their interest [37] [13]. Methods such as [24] estimate the gaze direction and do not identify the objects that are being attended to. On the other hand, [20] uses a head-mounted eye-tracker to estimate the user's point of gaze. Likewise, [31] detail a technique for determining the gaze direction of individuals in a scene by integrating video data

Figure 2. **Diverse Classroom Scenarios in the GESCAM Dataset.** This figure showcases a selection of samples from the GESCAM dataset, highlighting the variety of classroom environments captured. The images depict classrooms with different seating arrangements, lighting conditions, and student attention states. Notably, the diverse gaze directions demonstrate the dataset's ability to capture real-world classroom dynamics, where students may not always focus solely on the teacher.

with Inertial Measurement Unit (IMU) data. However, it's important to note that both of these methods primarily concentrate on GTD from a first-person perspective. In this paper, we focus on GTD with in-the-wild images, captured from a third-person viewpoint.

Similarly, [22] presented a new publicly available eye tracking video dataset to facilitate the research in point of gaze detection or any other related eye-tracking application. However, the dataset lacks the projection of 2D objects in terms of location and/or movement in 3D space which restricts the tracking to be carried out by a human eye alone. GTD has undergone significant advancements with the integration of computer vision technologies into human gaze research. It is increasingly recognized that in fields where precise iris or eye tracking is impractical, head pose emerges as a crucial feature for determining the human focus of attention, alongside other semantic cues. [7] have pinpointed three significant issues in previous research.

Initially, many research endeavors focus on analyzing gaze direction solely within a 2D framework, neglecting to incorporate depth information. A dataset lacking this dimensionality could severely hinder the effectiveness of algorithms trained on such data. Additionally, developing mapping functions solely from head position to gaze direction, without accounting for the interplay between eye and head movements, may compromise the overall generalization capability of attention measurement applications utilizing GTD. Therefore, a holistic comprehension of attention environments becomes imperative to recognize potential objects situated across multiple layers (using multiple modalities) along the subject's line of sight.

While many approaches learn the mapping function from head features to gaze direction using 2D visual cues [26] [3] [4] [27], estimating depth information from a RGB image is essential to accurately predicting the gaze target. We have covered this in-depth in our research work [21], and have provided this feature in the GESCAM dataset, being presented in this paper.

Moreover, prevailing datasets in gaze target detection and attention estimation, primarily emphasize individuals in the foreground for detecting gaze targets, resulting in gaze vectors of markedly larger scales. Conversely, classroom contexts typically entail gaze vectors of reduced magnitude owing to the close proximity of focal points. To reconcile this disparity, and fill the gap, we have addressed this issue in the GESCAM dataset. This pioneering resource is tailored to confront the intricacies of gaze target detection within controlled classroom settings. By encompassing an array of classroom layouts, character profiles, and activities, GESCAM provides a more authentic and domain-specific training environment for gaze estimation models. The dataset's configuration addresses the deficiencies of current resources by featuring densely populated classrooms, diminished gaze vector magnitudes that align with conventional classroom viewing behaviours, and the potential for interaction with multiple objects. We anticipate that GESCAM will serve as a catalyst for advancements in gaze target detection research within the realm of educational technology.

2.2. Gaze Estimation Datasets

Several datasets have been developed for gaze estimation tasks, each with its unique characteristics and limitations. EYEDIAP, created using Kinect sensors and HD cameras, features a few participants and synchronized RGB-D and HD streams [10]. OpenEDS is a large-scale collection of eye images obtained from VR head-mounted displays, comprising 12759 images with pixel-level annotations from 152 participants [11]. GazeCapture, focusing on first-person gaze estimation, includes data from 1450 participants cap-

tured using mobile phones, tablets, or other sensors [17]. MPIIGaze, another first-person gaze dataset, consists of 213,659 images collected from laptops used by 15 participants over three months [39]. Gaze360 offers a large-scale dataset for gaze tracking with 238 subjects and 3D gaze annotations across various head poses and distances [16].

ETH-XGaze is notable for extreme pose and gaze variation, featuring over one million images from 110 subjects captured using custom hardware and 18 SLR cameras [38]. RT-GENE addresses issues of subject-to-camera distance and head pose/gaze angle variations using eye-tracking glasses [9]. iSUN, deployed on Amazon Mechanical Turk, predicts saliency from 20608 webcam images [35]. CAT2000 focuses on eye movements across various image categories with 4000 images and 120 participants [1]. SALICON provides saliency annotations on the MSCOCO dataset [15]. GazeFollow and GOO datasets offer third-person perspectives, with GazeFollow curated from diverse sources like movies and social media [26], while GOO focuses on closed retail scenes with single-person gaze on objects [30]. VideoAttentionTarget includes gaze from multiple persons in video settings [2], but lacks CCTV angles. Our proposed GESCAM dataset aims to address these limitations, providing synthetic data covering varying subject-to-camera distances, head poses, illumination, and multiple gaze estimations within a scene.

2.3. Classroom Gaze Estimation

Various research endeavours have investigated gaze estimation's role in gauging classroom attention. For instance, a study utilized eye-tracking glasses to scrutinize student attention during presentations, shedding light on gaze data's utility in assessing teaching methodologies [12]

Gaze estimation using camera-based models in classrooms: [23] investigated the potential of using standard cameras instead of specialized eye trackers to estimate student gaze, offering a more scalable and cost-effective approach to assessing attention. Similarly, [34] combined gaze tracking with object detection algorithms to identify what are students looking at in the classroom, providing insights into the specific sources of their attention and potential distractions. Addressing the shortcomings of existing datasets for classroom attention analysis, our proposed dataset aims to provide a vast, diverse, and openly accessible resource, facilitating advancements in this crucial field.

3. GESCAM Dataset

This section details the comprehensive workflow employed for generating the **G**aze **E**stimation based **S**ynthetic **C**lassroom **A**ttention **M**easurement (GESCAM) dataset. An overview of the dataset generation process can be seen in Fig. 3

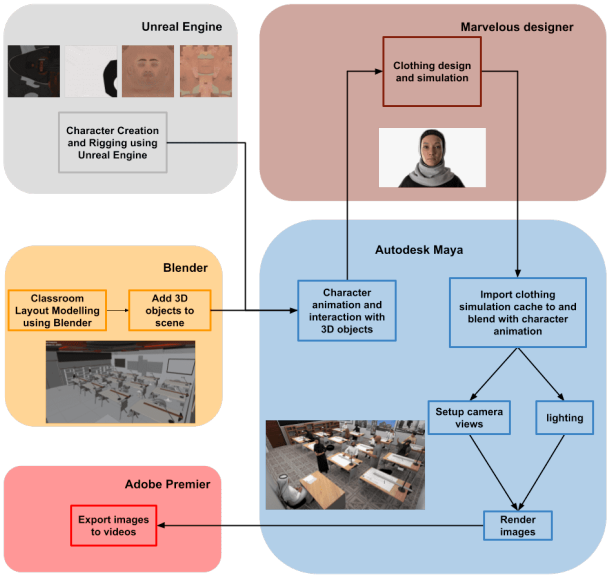

Figure 3. **GESCAM** dataset creation flowchart

3.1. Dataset Generation and Annotation

Constructing the Classroom Stage - 3D Modeling with Blender: Blender, a robust 3D modelling software, was used to meticulously design lifelike classroom settings. Its open-source nature, rich features, and strong community support make it a cost-effective and flexible option for crafting realistic environments. Leveraging Blender's extensive toolkit, a wide array of classroom elements was created, including furniture, walls, and windows. A core set of 3D models representing common classroom components served as the foundation, allowing for the generation of 20 distinct classroom layouts. These variations considered factors like furniture arrangement, visual stimuli, teaching methods, and lighting conditions to enhance dataset diversity and capture potential influences on student gaze patterns.

Populating the Classroom - Character Design with Unreal Engine: Diverse human characters were designed within Unreal Engine, a game development platform renowned for its industry-standard character creation tools. This software facilitated the development of characters encompassing a variety of ages, skin tones, genders, and clothing styles. This deliberate variation aimed to enhance the generalizability of the dataset's findings to real-world classrooms. Beyond the visual design, meticulous character rigging was performed within Unreal Engine. This process equipped the characters with the ability to perform natural and realistic movements, crucial for simulating authentic classroom behaviors within the animations.

Breathing Life into the Classroom - Animation and Rendering with Maya and Arnold: Autodesk Maya, a widely used 3D animation software, was employed to craft animations that mirrored common classroom scenarios. These animations depicted activities such as following a teacher's lecture, interacting with classmates, or glancing out the window. For each animation, the Arnold rendering engine, known for its high-fidelity output, was utilized to generate visually stunning and realistic classroom scenes.

Dressing the Characters - Attire Design and Simulation with Marvelous Designer: The attire for the characters was designed and simulated using Marvelous Designer, a software specifically designed for creating and simulating clothing and fabrics. This involved creating virtual garments and simulating their movement on the designed human characters. Costumes were designed and simulated to complement and blend in with the character animations seamlessly. This entailed careful coordination between the character rigging and the simulated clothing to ensure natural movement.

Capturing Different Perspectives - Multi Viewpoint Video Rendering: To provide researchers with the flexibility to analyze gaze patterns from various angles, each classroom scene was rendered from five distinct camera viewpoints. These viewpoints were strategically chosen to capture the environment from a range of perspectives, thus providing more versatility in the dataset.

Standardizing the Dataset - Video Specifications: For consistency, all videos adhered to identical specifications: 1-minute duration, 1920x1080 resolution (HD), and 10 frames per second. Adobe Premiere Pro then compiled rendered images from each viewpoint into final video files.

This meticulously designed workflow for generating the GESCAM dataset resulted in a rich resource for researchers investigating gaze patterns within classroom environment. The dataset generation process involved a combination of 3D modeling, animation, simulation, and rendering techniques. The dataset's diversity in classroom layouts, character profiles, animation scenarios, and camera viewpoints fosters its generalizability.

The GESCAM dataset provides comprehensive annotations for eye-gaze analysis. These annotations include hand-labeled bounding boxes for heads and gaze lines for all individuals within the dataset. Additionally, bounding boxes and classification labels are provided for all objects present in the scene. Depth maps are also available for each RGB image, enriching the data for potential 3D gaze estimation tasks. Some examples from the dataset alongside annotations are shown in Fig. 4. Our dataset consists of 60,000 images containing 650,000 samples of individuals with corresponding gaze targets. This data was collected across 20 distinct classroom layouts and under various camera viewing angles. Fig. 5 shows plots of metrics extracted from the GESCAM dataset.

3.2. Comparison to other datasets

Current gaze target detection datasets primarily focus on generic scenarios captured through in-the-wild images or videos. Examples include GazeFollow [26], which leverages images from diverse sources such as SUN [34], MS COCO [19], Actions 40 [36], and PASCAL [6] to train models for general gaze target detection. Similarly, VideoAttentionTarget [5] was created for video gaze target modeling using videos of interviews, sitcoms, reality shows and movies. While both GazeFollow (with 160,000 head bounding box annotations) and VideoAttentionTarget (with 164,541 annotations) offer considerable data size, they lack domain-specificity. Models trained on these datasets often perform well for general applications but require fine-tuning for adaptation to specific use cases. Addressing this limitation, datasets like Gaze On Objects (GOO) cater to specific domains like retail environments. GOO includes 192,000 synthetic images (GOO-Synth) and 9552 real images (GOO-Real) depicting humans gazing at various grocery items. Each image is annotated with head bounding boxes, gaze lines, object class labels, and segmentation masks for the observed items. However, the retail setting in GOO differs significantly from classrooms. Classrooms typically involve denser arrangements of people interacting with multiple objects simultaneously, unlike the sparser scenarios found in GazeFollow or VideoAttentionTarget. While the GOO dataset boasts a larger number of objects compared to others, it maintains a one-to-many person-to-object mapping. This may not accurately represent real-world classroom scenarios, where many individuals often interact with multiple objects simultaneously. Additionally, existing datasets often prioritize people in the foreground for gaze target detection tasks. This can lead to a bias towards larger gaze vector magnitudes. However, classroom settings involve a unique mix of near and far interactions, resulting in a wider range of gaze vector magnitudes.

GESCAM bridges the gap in gaze target detection for classrooms. It offers a realistic and domain-specific training ground by incorporating diverse layouts, characters, and activities. Unlike existing resources, GESCAM features densely populated classrooms, near and far gaze interactions, and multi-object interactions, reflecting real-world scenarios. This comprehensive dataset holds immense potential to propel gaze target detection research in educational technology.

4. Method

4.1. Baseline Methods

Gaze target detection has benefited significantly from pioneering works such as those by Recansens *et al*. [26], Lian

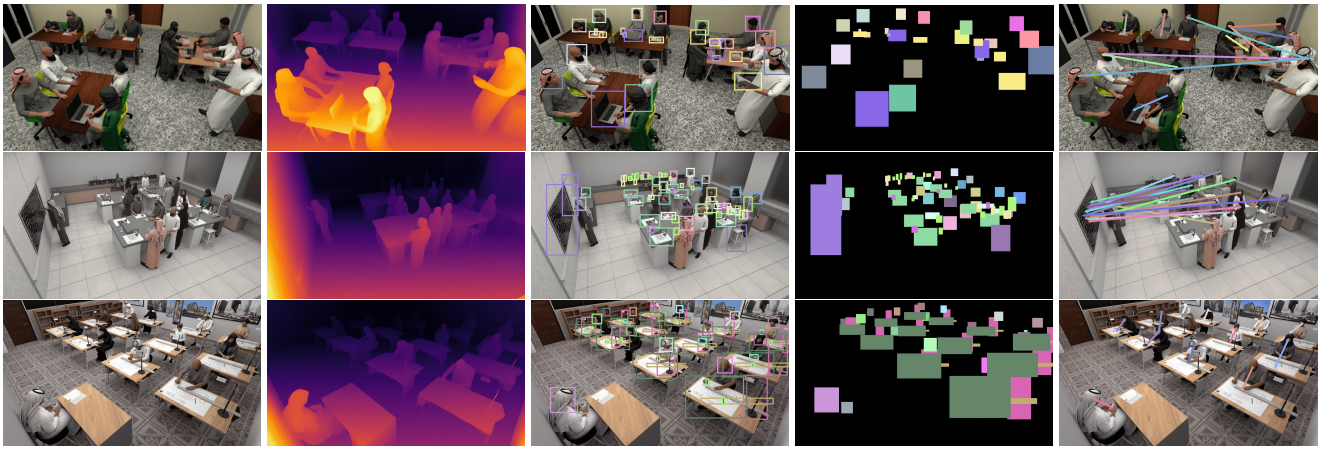

Figure 4. **Illustrative examples from the GESCAM dataset**. From left to right, each column displays rendered RGB images of classrooms captured from different viewpoints, corresponding depth maps for each RGB image, annotation of bounding boxes and labels for people and various objects within the scene, masks highlighting individual objects for precise identification and gaze lines indicating points of focus within the scene.

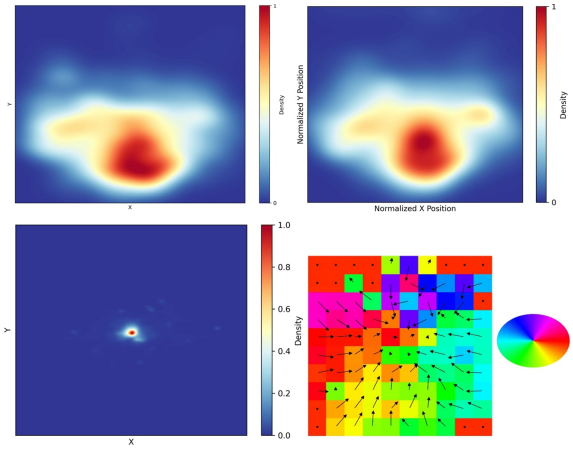

Figure 5. **Visualizations derived from GESCAM dataset**. The first row depicts the head position density and gaze fixation position density. The second row shows the normalized fixation relative to head position and average gaze direction (with directional color code) of the rendered people within the dataset.

et al. [18], and Chong *et al*. [5]. These early approaches often employed multi-stream architectures featuring separate pathways for processing the scene image, head image, and head location information. Building upon this foundation, Fang *et al*. [8], Tonini *et al*. [33] and recently Athul *et al*. [21] further enhanced performance by incorporating depth information as an additional modality within their network architectures.

These multi-stream architectures typically follow a common processing pipeline. The scene pathway extracts scene embeddings from the input scene image using a dedicated ”scene backbone.” Similarly, the head pathway utilizes the head image and head location data to extract head embeddings through a ”head backbone”. Fang *et al*., Tonini *et al*. and Athul *et al*. additionally extracted depth embeddings through a ”depth backbone”. Finally, the heatmap pathways leverage the combined scene, head (and optionally depth) embeddings to predict a gaze heatmap, where the peak point signifies the predicted gaze fixation.

In this work, we leverage the aforementioned methods by [26], [18], [5] as baseline models for training and evaluating GESCAM dataset. Our evaluation focused solely on models that operate exclusively on the RGB input modality and do not incorporate additional data streams like depth maps. Consequently, the evaluation was limited to such architectures, avoiding the introduction of supplementary processing pathways that could increase computational complexity. This selection allows us to establish a benchmark and compare the performance of our proposed network against these two-pathway techniques for gaze target detection.

4.2. GESCAM Network

Existing baseline approaches, while effective for general-purpose tasks, necessitate more sophisticated design elements for precise gaze object capture within classroom settings. Capturing the entire classroom often necessitates high field-of-view camera systems. Notably, prevalent neural network architectures commonly utilize a ResNet-50 backbone accepting a fixed-size input (256×256). This approach is suitable for datasets like GazeFollow and VideoAttentionTarget, where the subject of interest occupies a central position within a limited depth of field. However, this assumption is not valid in classroom scenarios due to the broader field of view and multiple students. To address this challenge, guiding the model towards identifying

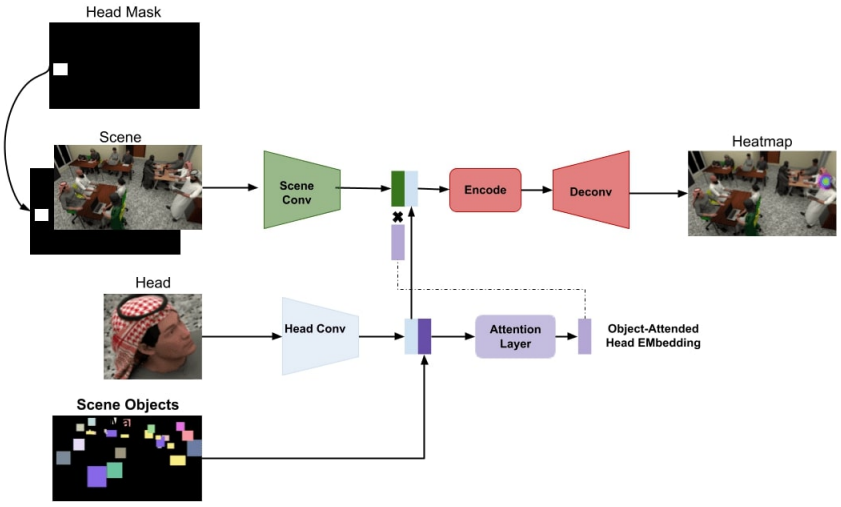

Figure 6. **Overview of GESCAM network.** Scene objects mask M conditions head pathway using a soft-attention layer. Object-attended head embedding H' modulates scene pathway. The output is a 2D heatmap superimposed on the scene image for visualization.

crucial regions becomes essential, particularly for accurate classroom attention estimation.

The model from [5] performed the best in our experiments as seen in Sec. 5.2. Building upon the success of this model, we adapted the network with minor modifications to cater to the specific requirements of classroom environments. This modified architecture, named the GESCAM network, demonstrates improved performance on the provided dataset. Our network incorporates scene objects mask M into the Head pathway, and guides the model to focus on salient objects of interest. The head embeddings H alongside M is passed through a learable soft-attention layer A.

$$H' = A(H \oplus M) \tag{1}$$

The object-attended head embedding H' finally modulates the scene embedding S. The modulated scene embedding S' is given as :

$$S' = S \otimes H' \tag{2}$$

where $\oplus$ denote concatenation operation and $\otimes$ represent elementwise multiplication operation.

The remainder of the network remains the same as in [5]. To improve the accuracy of the predicted gaze heatmaps, we incorporated an angular loss term $\mathcal{L}_{ang}$ in addition to the standard mean squared error loss $\mathcal{L}_{mse}$. The mathematical formulation of the angular loss is provided below:

$$\mathcal{L}_{ang} = 1 - \frac{(g_x, g_y) \cdot (p_x, p_y)}{\| (g_x, g_y) \|_2 \cdot \| (p_x, p_y) \|_2} \tag{3}$$

where (g_x, g_y) is the ground-truth gaze vector and (p_x, p_y) is the predicted gaze vector. The overall loss function is defined by:

$$\mathcal{L}_{tot} = \mathcal{L}_{mse} + \mathcal{L}_{ang} \tag{4}$$

5. Experiments

We quantitatively and qualitatively evaluated our network on the GESCAM datasets. We demonstrate that our method surpassed the performance of prior two-pathway methods across all metrics in Sec. 5.3.

5.1. Implementation Details

For controlled and consistent evaluation, all baseline methods were trained and evaluated on a single Nvidia RTX 4080 GPU. Facilitating transparency and reproducibility, we implemented them within a unified PyTorch codebase and employed pre-training methods from the corresponding publications. Standard data augmentation techniques (random crop, color manipulation, random flip, and head bounding box jittering) were applied during training.

5.2. Evaluation

To assess the effectiveness of each selected gaze target prediction model from the literature, we employed a comprehensive suite of evaluation metrics, each offering valuable insights into different aspects of performance.

Area Under the Curve (AUC) assesses the model's ability to differentiate true gaze locations from false positives. We compare the predicted gaze distribution (flattened output heatmap) with the ground truth heatmap (indicating gaze presence/absence). A Receiver Operating Characteristic (ROC) curve visualizes the model's performance in correctly identifying gaze and avoiding false positives. The AUC score (0-1) quantifies overall performance, with 1.0 indicating perfect agreement between predicted and actual gaze distributions. **L2 Distance (Dist.)** measures the Euclidean distance between the ground truth target location – the actual point of gaze – and the point of maximum intensity within the predicted gaze heatmap. To ensure a fair

Figure 7. **Qualitative results**. The red and green lines denote ground truth and predictions respectively. The head bounding box for each person is assigned a unique color for easier identification across the images.

comparison across images of varying sizes, we normalize the image height and width to 1. This normalization step accounts for potential discrepancies solely due to image dimensions. **Angular Error (Ang.)** quantifies the angular difference between the predicted gaze direction and the actual gaze vector. The gaze vector is calculated based on the face location and the gaze point. A lower angular error indicates a closer alignment between the predicted and actual gaze directions.

Method	AUC ↑	Dist. ↓	Angle ↓
Random	0.512	0.380	66.0
Center	0.553	0.235	51.1
Recansens *et al.*	0.906	0.157	40.9
Lian *et al.*	0.935	0.125	35.5
Chong *et al.*	0.938	0.112	36.6
GESCAM (M only)	0.941	0.110	33.5
GESCAM (M & L_{ang})	**0.943**	**0.109**	**32.9**

Table 1. Comparison of different methods on GESCAM dataset. The numbers in bold represent best results

5.3. Results

Experimental results are given in Tab. 1. **Random:** To establish a performance lower bound, we generated heatmaps with random values (standard normal distribution) and evaluated them against the ground truth. This provides a benchmark for comparing more complex networks. Networks that consistently outperform this random baseline demonstrate the ability to learn meaningful patterns from the training data. **Center:** The predicted gaze point is always fixed to be at center of the image.

Building upon the work of [5], which achieved the best performance among all the evaluated baselines, the GESCAM network demonstrates superior overall performance. However, a consistent observation across all methods is the relatively high angular error. This can be attributed to scenarios where individuals are positioned away from the camera, resulting in obscured facial features. Consequently, fixation point estimation becomes inaccurate for all models in such cases. Notably, the angular error metric amplifies this effect while distance and AUC metrics remain less affected. Figure 7 showcases example images with gaze point predictions from the GESCAM network.

6. Conclusion

This work presented GESCAM, a novel dataset specifically designed for classroom gaze estimation. Curated to encompass diverse classroom scenarios, GESCAM fills a critical gap in the absence of domain-specific datasets in the educational domain. After benchmarking existing 2-pathway models on GESCAM, it became evident that there is room for further performance improvement. Furthermore, we introduced the GESCAM network, a 2-pathway architecture demonstrating better performance on GESCAM dataset. We expect that this research will stimulate further investigation and progress in gaze estimation within the educational domain.

7. Future Work

In this paper, we focus on two-pathway networks as baseline methods. However, GESCAM dataset (including depth maps) allows us to explore additional modalities. Furthermore, the correlation between classroom attention estimation and GTD warrants further investigation. Future work could involve introducing new metrics and framing attention estimation as a function of GTD. Additionally, bridging the gap between synthetic and real-world settings remains a question for future research.

References

[1] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *Journal of Computer Vision*, 42(3):123–137, 2015. 4

[2] Alice Chong, Bob Lee, and Carol Wang. Detecting attended visual targets in video. *Journal of Computer Vision*, 42(3):123–137, 2020. 4

[3] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part V*, page 397–412, Berlin, Heidelberg, 2018. Springer-Verlag. 3

[4] E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg. Detecting attended visual targets in video. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5395–5405, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. 2, 3

[5] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5, 6, 7, 8

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 5

[7] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11385–11394, 2021. 2, 3

[8] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11390–11399, June 2021. 6

[9] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018. 4

[10] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. *Journal of Computer Vision*, 42(3):123–137, 2014. 3

[11] Stephan J. Garbin, Yiru Shen, Immo Schuetz, Robert Cavin, Gregory Hughes, and Sachin S. Talathi. Openeds: Open eye dataset. *arXiv preprint arXiv:1905.03702*, 2019. 3

[12] Enrique Garcia Moreno-Esteva and Markku S. Hannula. *Using gaze tracking technology to study student visual attention during teacher's presentation on board*, pages 1393–1399. Charles University in Prague, Faculty of Education and ERME, Czech Republic, 2015. ¡hal-01287672¿. 4

[13] Xinwei Guo, Yong Wu, Jingjing Miao, and Yang Chen. LiteGaze: Neural architecture search for efficient gaze estimation. *PLOS ONE*, 18(5):e0284814–, 5 2023. 2

[14] Stephen James and Edward Johns. 3d simulation for robot arm control with deep q-learning. 09 2016. 2

[15] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080. IEEE Computer Society, 2015. 4

[16] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[17] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[18] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 35–50, Cham, 2019. Springer International Publishing. 6

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5

[20] Meng Liu, Youfu Li, and Hai Liu. 3d gaze estimation for head-mounted eye tracking system with auto-calibration method. *IEEE Access*, 8:104207–104215, 2020. 2

[21] Athul Mathew, Arshad Khan, Thariq Khalid, Faroq AL-Tam, and Riad Souissi. Leveraging multi-modal saliency and fusion for gaze target detection. In *NeuRIPS 2023 Workshop on Gaze Meets ML*, 2023. 3, 6

[22] Christopher D McMurrough, Vangelis Metsis, Jonathan Rich, and Fillia Makedon. An eye tracking dataset for point of gaze detection. pages 305–308, 2012. 3

[23] Lars Ojinnaka and Derek Harter. *Gaze estimation using a camera-based model in a classroom*. PhD thesis, 06 2020. 4

[24] Daniel Parks, Ali Borji, and Laurent Itti. Augmented saliency model using automatic 3d head pose detection and learned gaze following in natural scenes. *Vision research*, 116:113–126, 2015. 2

[25] Benjamin Planche, Ziyan Wu, Kai Ma, Shanhui Sun, Stefan Kluckner, Terrence Chen, Andreas Hutter, Sergey Zakharov, Harald Kosch, and Jan Ernst. Depthsynth: Real-time realistic synthetic data generation from cad models for 2.5d recognition, 2017. 2

[26] Adria Recasens*, Aditya Khosla*, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. * indicates equal contribution. 1, 3, 4, 5, 6

[27] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443, 2017. 2, 3

[28] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games, 2016. 2

[29] Shashimal Senarath, Primesh Pathirana, Dulani Meedeniya, and Sampath Jayarathna. Retail gaze: A dataset for gaze estimation in retail environments. In *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, pages 1040–1044. IEEE, 2022. 1

[30] John Smith and Emily Lee. Goo: A dataset for gaze object prediction in retail environments. *Journal of Computer Vision*, 42(3):123–137, 2023. 4

[31] Sanket Kumar Thakur, Cigdem Beyan, Pietro Morerio, and Alessio Del Bue. Predicting gaze from egocentric social interaction videos and imu data. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 717–722, 2021. 2

[32] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017. 2

[33] Francesco Tonini, Cigdem Beyan, and Elisa Ricci. Multimodal across domains gaze target detection. New York, NY, USA, 2022. Association for Computing Machinery. 6

[34] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 4, 5

[35] Pingmei Xu, Krista A. Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam-based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015. 4

[36] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International Conference on Computer Vision*, pages 1331–1338, 2011. 5

[37] Xucong Zhang, Michael Xuelin Huang, Yusuke Sugano, and Andreas Bulling. Training person-specific gaze estimators from user interactions with multiple devices. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018. 2

[38] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. *arXiv preprint arXiv:2007.15837*, 2020. 4

[39] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. *arXiv preprint arXiv:1504.02863*, 2015. 4