Old Dominion University

ODU Digital Commons

2022

Customer Gaze Estimation in Retail Using Deep Learning

Shashimal Senarath

Primesh Pathirana

Dulani Meedeniya

Sampath Jayarathna
Old Dominion University

Original Publication Citation

Customer Gaze Estimation in Retail Using Deep Learning

SHASHIMAL SENARATH[1], (Student Member, IEEE),
PRIMESH PATHIRANA[1], (Student Member, IEEE),
DULANI MEEDENIYA [1], (Senior Member, IEEE), AND
SAMPATH JAYARATHNA[2], (Member, IEEE)
[1]Department of Computer Science and Engineering, University of Moratuwa, Moratuwa 10400, Sri Lanka
[2]Department of Computer Science, College of Science, Old Dominion University, Norfolk, VA 23529, USA

Corresponding author: Dulani Meedeniya (dulanim@cse.mrt.ac.lk)

ABSTRACT At present, intelligent computing applications are widely used in different domains, including retail stores. The analysis of customer behaviour has become crucial for the benefit of both customers and retailers. In this regard, the concept of remote gaze estimation using deep learning has shown promising results in analyzing customer behaviour in retail due to its scalability, robustness, low cost, and uninterrupted nature. This study presents a three-stage, three-attention-based deep convolutional neural network for remote gaze estimation in retail using image data. In the first stage, we design a mechanism to estimate the 3D gaze of the subject using image data and monocular depth estimation. The second stage presents a novel three-attention mechanism to estimate the gaze in the wild from field-of-view, depth range, and object channel attentions. The third stage generates the gaze saliency heatmap from the output attention map of the second stage. We train and evaluate the proposed model using benchmark GOO-Real dataset and compare results with baseline models. Further, we adapt our model to real-retail environments by introducing a novel Retail Gaze dataset. Extensive experiments demonstrate that our approach significantly improves remote gaze target estimation performance on GOO-Real and Retail Gaze datasets.

INDEX TERMS Computer vision, deep learning, gaze estimation, retail customer behaviour.

I. INTRODUCTION

In today's world, retail stores are becoming smarter with the availability of numerous data and the power to analyze them autonomously. Even with the rise of online shopping, most of the physical retail stores use smart applications for the purchasing process [1]. Several techniques and devices have been introduced to automate the shopping process and analyze shoppers' behaviour inside stores. At the same time, the shopping experience is a key consideration towards the success of a retail business, which affects the performance of customer satisfaction, customer purchase probability, and customer loyalty [2]–[4].

In order to improve the shopping experience and maximize business profits, it is essential to capture and and analyze the customer's behaviours without interfering their natural

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

shopping journey [5], [6]. Various solutions have introduced for customer behaviour analyzis in retail using developments in computer vision technology. For instance, counting the number of people and detecting the hot spots in retail [6] and public [7], and tracking shoppers' emotion [5] are such applications. However, the existing solutions only capture coarse touch-points of a shopper's journey and vulnerable to unconstrained environment settings. With the adaptation of computer vision technologies in gaze estimation, there has been eye tracking-based solutions for customer behaviour analysis in retail as well [1], [8]. Moreover, there are solutions based on virtual reality devices and head-mounted displays, wearable eye tracker based solutions [9], and non-intrusive 3D eye tracking solutions [10]. However, these solutions do not completely satisfy the retailers due to high cost of 3D eye tracking solutions, unscalability of wearable, and head-mounted display-based solutions, and manual calibration of eye tracking systems.

The concept of Gaze Following, which was introduced by Recasense *et al.* [11], refers to the identification of the object being looked at by a person, given the scene image. This concept has been extended by Tomas *et al.* [8] and presented the idea of Gaze Object Prediction in retail, refers to the task of predicting the bounding box for a human's gazed-at object. Both these concepts only require gaze estimation from the scene image, and it avoids the need to wear special types of devices to capture the eye gaze and remove the restrictions of manual calibration. Thus, they are economically feasible solutions for retailers. This concept is known as the remote human gaze estimation in retail.

Remote gaze estimation in retail is a problem in the wild that requires further gaze estimation from back-head images with partial and total eye occlusion. Estimating customer gaze in retail is a novel concept that has not been given enough focus in the research literature. However, the concept holds much promise in retail to effortlessly and securely analyze the shopper behaviour in stores.

This paper presents deep learning-based remote gaze estimation models for the retail environment. We propose a three-stage deep CNN based on attention mechanisms and hand-designed features for retail, remote back-head gaze saliency estimation. We present a comprehensive methodology of our model design process starting from an end-to-end solution to the proposed three-stage architecture. We develop four design solutions and introduce the novel object channel and depth channel features to improve the accuracy of gaze saliency estimation in retail. The object channel, which is a hypothetical gaze distribution of gaze generated from retail product item boundaries, helps the models to narrow down their gaze estimation point search space. Subsequently, the depth channel encodes the scene depth and helps the models to overcome the issue of estimating correct retail shelf depth.

Based on the hypothesized design solutions, we develop two model architectures to encode the object and depth channels separately. First, the hypothetical gaze distribution model, as defined in our previous work [12], represents the concept of the object channel. The model has surpassed the existing benchmark Area Under the Curve (AUC) and Angular error baselines on the GOO dataset [8], showing the importance of the object channel. Second, the Face3D model represents the depth channel and the novel concept of remote gaze estimation in 3D vector space. In this model, we present the idea of monocular depth estimation for encoding depth information in remote gaze estimation in retail. The model significantly improves AUC, L2-distance, and Angular error criteria on the GOO dataset.

With the success of the two models, we combine the two concepts and design the three-stage, depth-based dual attention model to estimate the gaze saliency accurately and robustly. The three stages of the proposed model are named as, gaze and depth estimation, dual attention module, and heatmap generator. In the first stage, the network generates a 3D gaze vector, given the head image of the subject. Then it decomposes the 3D information into a 2D gaze target and

a depth channel. Motivated by the stufy of Fang *et al.* [13], we design and modify the dual attention module to capture the specific parameters in a retail environment. Specifically, we incorporate third attention, the object channel presented in the hypothetical gaze distribution model, for the dual attention module. The results of this model on the GOO-Real dataset shows superiority over the other models. Hence, we transfer learn this model's knowledge to the Retail Gaze dataset to test its applicability in a real retail environment. The proposed depth-based dual attention model can estimate the gaze saliency in retail environments with a high accuracy, using only, in-the-wild image data.

Our contributions are summarized as follows:

- We implement a novel mechanism for remote human gaze estimation in retail, that does not require manual user calibration and does not interfere the natural shopper journey.
- We design and develop a novel deep convolutional neural network (CNN) for accurate, remote gaze saliency estimation in retail using back head images.
- We present the novel depth channel and object channel to improve the accuracy of gaze saliency estimation in retail.

The paper is structured as follows. Section II explores the related work on gaze target estimation, gaze following, gaze object prediction, and gaze estimation in retail. Section III presents a comprehensive elaboration of the proposed approach for remote gaze saliency estimation in retail. Following this, Section IV discusses the experiments carried out and the datasets and evaluation criteria used. Section V presents the obtained results and Section VI compares the existing studies and possible future research directions. Finally, Section VII concludes the paper.

II. RELATED WORK

We explore the appearance-based gaze estimation approaches with deep learning (DL), focusing on gaze estimation in retail. Appearance-based gaze estimation solutions with DL have been studied extensively in the past decade [10], [14]–[16]. These solutions have been proven to perform well under extreme unconstrained environmental conditions in remote gaze estimation like partial and total eye occlusion, lighting condition variations, camera to subject distance variations, subject diversity, and camera capture angle variations [10], [13], [17]. With the adaptation of deep-learning-based solutions, there is a high promise to robustly estimate gaze in unconstrained environments with the above-mentioned considerations. This section presents the related work on gaze target estimation, gaze following, 3D gaze estimation, and gaze object prediction as shown in Fig. 1. A summary of the related studies is presented in Table 1.

A. GAZE TARGET ESTIMATION

Gaze target estimation is a well-established area in the literature. It is defined as locating the point being gazed at by the subject in either 2D image coordinates or 3D real-world

FIGURE 1. Gaze estimation tasks: (a) Gaze target estimation, (b) Gaze following, (c) Gaze object prediction, (d) Gaze estimation in retail.

coordinates [11], [17]–[21]. Moreover, this can be categorized as gaze target estimation in controlled environments and in-the-wild, based on the nature of the application environment. Researchers have explored gaze target estimation in the wild in multiple application domains such as sports training, gaze target identification of the public for targeted advertising in digital signage [1], gaze target localization of a crowd [22], shoppers gaze estimation [1].

Among several related studies, Kellnhofer *et al.* [17] have presented Gaze360, which is a large-scale gaze-tracking dataset and a deep CNN based method for 3D gaze estimation in the wild. This dataset includes a range of gaze and head poses, diverse capture environments that enhance the wild nature of the dataset environments, and a large subject diversity. Gaze360 dataset consists of 172,000 images, which includes 238 subjects captured in 7 different locations. Moreover, the dataset is annotated with 3D gaze vectors in real-world coordinates. The gaze estimation problem in retail requires back-head image data annotated with gaze targets to train and evaluate gaze models. The authors of this dataset have considered the partial eye occlusion scenario when collecting the dataset, which makes Gaze360 a suitable dataset for gaze estimation model training in retail.

Recent work in gaze target estimation has been presented by Fang *et al.* [13], to simulate the gaze estimation behaviour of humans in 3D space. The authors have introduced a three-stage solution, a first stage to estimate the 3D gaze orientation from the head of the subject, a second stage introducing a novel dual attention module to select the correct field of view mask based on depth information, and a third stage to estimate the gaze target in 2D image coordinates. Their work has shown better results for estimating gaze targets in general scenes compared to related literature. It closely resembled the nature of how humans follow gaze and have achieved an angular error of 11.1% for front-head remote gaze estimation. Our introduced model architecture is highly inspired by their work to generate dual attention-based features in retail gaze estimation.

B. GAZE FOLLOWING
The concept of gaze following has been first introduced by Recasense *et al.* [11]. They have defined gaze following as

the task of identifying the object being looked at by the subject, given the scene image data. This study generalized the gaze target estimation in the wild by introducing Gaze-Follow [11], which is a large-scale, benchmark dataset. It is annotated with the 2D image coordinates of the looking point of humans' in images. The dataset includes a large subject and scene diversity of people performing daily activities. In addition, they have introduced the human-inspired gaze saliency estimation model architecture, which has substantially improved the accuracy of saliency estimation even with eye occluded images. The proposed deep CNN architecture consists of two pathways: the gaze pathway and the saliency pathway to mimic the human nature of gaze following. Moreover, they have introduced a new method called the shifted grids to robustly estimate the gaze target by carrying out multiple classification problems instead of directly regressing it in 2D coordinates.

Following the work of Recasense *et al.* [11], several studies have extended the concept of gaze following to predict the gaze targets in videos and handle the gaze target predictions when the subjects are looking at objects outside of the captured scene [23], [24], [26]. Lian *et al.* [24], have encoded the gaze direction into multi-scale gaze direction regions to mimic the behaviour of a human in gaze following. Moreover, they have used a feature pyramid network [27], to regress the gaze saliency heatmap due to their success in object detection. In a similar work, Chong *et al.* [23], have presented a novel attention-based deep CNN approach to detect gaze targets in videos. Their study has extended the Gaze Follow dataset with out-of-frame gaze target annotations and further introduced VideoAttentionTarget dataset to estimate gaze saliency in videos. The introduced spatio-temporal CNN uses an attention layer to bridge the gap between the scene and head pathways and allows to control the scene pathway via the head pathway. Our attention mechanisms are mostly inspired by this study. Most of the gaze following domain work is complementary to ours; however, they are not specific to the retail environment. They have not considered gaze saliency prediction from back-head images. Directly applying these models in retail leads to several new issues, which we have solved by introducing hand-designed features specifically designed towards retail environments.

TABLE 1. Summary of Related Work: GTE - Gaze target estimation, GF - Gaze following, GOP - Gaze object prediction, GE - Gaze estimation.

| Study	Dataset	GTE	GF	GOP	GE in Retail	Approach\Technique
Bermejo et al. [1]	UcoHead, Own dataset	✓	-	-	✓	CNN (Coarse-to-Fine)
Recasense et al. [11]	Gaze Follow	✓	✓	-	-	CNN with shifted grids
Tomas et al. [8]	GOO	✓	✓	✓	✓	Existing CNN models
Kellnhofer et al. [17]	Gaze360	✓	-	-	✓	CNN-LSTM
Fang et al. [13]	Gaze360, Gaze Follow, VideoAttentionTarget	✓	-	-	-	Attention-based CNN
Chong et al. [23]	Gaze Follow, VideoCoAtt, VideoAttentionTarget	✓	✓	-	-	CNN-LSTM
Lian et al. [24]	Gaze Follow	✓	✓	-	-	Static-CNN
Kodama et al. [22]	Own dataset	✓	✓	-	-	Static-CNN
Khamis et al. [25]	Own dataset	✓	-	-	-	Tobii Rex Eye Tracker

C. GAZE OBJECT PREDICTION

Gaze object prediction is a novel concept in gaze estimation, and a more applicable approach in retail. The concept was recently introduced by Tomas *et al.* [8], and defined as the task of predicting the bounding box of the gazed at the object by the subject. Mainly they have introduced a dataset named Gaze On Objects. It is a large dataset for gaze object prediction in retail, consisting of 2D gaze point annotations, gazed at object bounding boxes and gazed at object segmentation masks. This includes a large synthetic dataset consists of 192,000 images of 20 different synthetic human subjects collected from 50 virtual camera angles and a small real dataset with 9,552 images of 100 human subjects collected from two different camera angles. Their work has further presented benchmark results on state-of-the-art gaze estimation models provided by Recasense *et al.* [11], Lian *et al.* [24], and Chong *et al.* [23] in their gaze following work. This study has served as a starting point for our study in gaze saliency estimation in retail. Further, we added complex hand-designed features specifically designed for retail and incorporated attention-based mechanisms for accuracy enhancements of gaze saliency estimation from back-head images in retail.

D. GAZE ESTIMATION IN RETAIL

The concept of gaze estimation in retail has not been explored in-depth in the literature. Bermejo *et al.* [1] have introduced EyeShopper, which is a system to track the gaze of shoppers in retail using 3D gaze estimation. The proposed deep CNN estimates the shopper gaze when facing away from the camera in real-world 3D coordinates. Authors have used a simple model architecture with ResNet-18 [28] backbone and by incorporating a coarse-to-fine approach in machine learning to estimate gaze pitch and yaw directly from head images. Even though the presented study has achieved better results for back-head gaze estimations, their work requires personal calibration of shoppers to map the 3D gaze targets into 2D image coordinates. In retail, it is not practical to carry out personal calibration of shoppers, which will interrupt the natural viewing experiences.

In another study, Kellnhoffer *et al.* [17], have applied the Gaze360 model to estimate the attention in a supermarket. They have presented the Gaze360 model as a Spatio-temporal deep neural network based on bidirectional long short-term
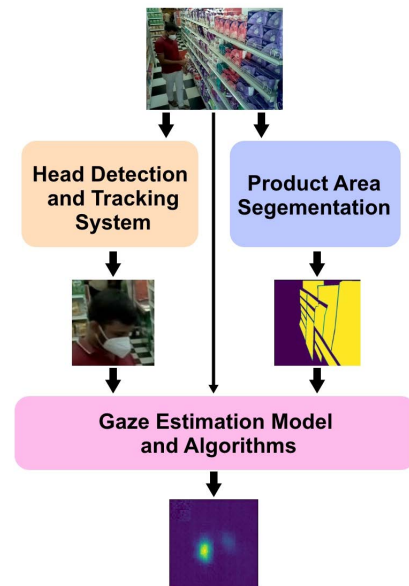
FIGURE 2. Overview of proposed framework.

memory (LSTM) architecture to predict the gaze target in real-world 3D coordinates. However, they have only carried out their experiment on a mock retail shelf with objects placed wide apart, which cannot be considered a real retail environment. Furthermore, as their work depends on 3D gaze estimation, it would require personal calibration of subjects to map the gaze targets into 2D image coordinates.

III. METHODOLOGY
A. GAZE ESTIMATION FRAMEWORK

The proposed gaze estimation framework consists of three stages as shown in Fig. 2. In the first stage, we use a head detection and tracking module to extract head image of the person in a given scene image. Here, the trajectories and segmentation masks are provided as input to the gaze estimation model. The output head bounding box from this module is cropped from the raw image and fed into the gaze model as the head image. Since traditional face detection algorithms fail to detect total or half-face occlusions, we build a custom head detection and tracking system for detecting and tracking customer heads in a sequence of frames. In the second stage

we perform the pixel-wise OR operation between either the product item bounding box annotations or product area segmentation mask annotations to manually generate the object channel. The object channel is a mask, which represents the object locations using 1s and the background using 0s. In the third stage, the head tracker module output and the object channel are provided as input to the gaze estimation model.

Our approach to the gaze estimation problem in retail is an appearance-based solution. Moreover, we use an appearance-based approach with DL due to their ability to overcome challenges in an unconstrained environment such as eye and head occlusions, illumination condition variations, subject differences, and significant head motion. The CNN is the main DL architecture to extract low-level automatically and high-level features from retail gaze images. We name our attention-based deep CNN architecture as the depth-based dual attention (DDA) model.

Our DDA model combines two sub-model architectures: the hypothetical gaze distribution model and the depth-based Face3D model. The hypothetical gaze distribution model aims to predict the gaze heatmap of the scene given the scene image, the hypothetical gaze distribution, the head channel, and the head crop image. The depth-based Face3D model is designed to output a 3D gaze direction vector in image coordinates given the scene and head crop images. The hypothetical gaze distribution feature allowed the model to narrow down its gaze estimation point search space, allowing it to get more accurate gaze fixations from the scene. The depth-based face3D model introduces the monocular depth estimation feature to enhance gaze point estimation accuracy. The proposed depth-based dual attention model is a complex hybrid approach of the two models.

B. HEAD DETECTOR AND TRACKER

1) HEAD DETECTION

Most DL-based gaze following networks use head or face image as an input. Therefore, detecting the head or face is essential in gaze following tasks. Most of the existing gaze following studies have used traditional face detection algorithms to detect face bounding boxes, which is a well-established area of research. However, some gaze estimation domains such as in retail, total or half head, or eye occlusions, may happen, which makes it challenging to detect the face or head. In order to address this face or head detection issue, we used DL-based object detection algorithms. We explored several object detection models such as single-shot detector (SSD), recent major You only look once (YOLO) versions with one stage object detection algorithms and Fast R-CNN, Mask R-CNN two-stage object detection algorithms. Considering the accuracy and Speed (frames per second (FPS)), we have selected the YOLOv5 small model (YOLOv5s6), as an object detection model. We have trained a head detector using the YOLOv5 detector from the scratch with the real datasets Retail gaze [29] and Gaze on Object [8]. These datasets contain head bounding

boxes of the customers, and have challenging total and half occlusion head bounding boxes to detect. These datasets consist of head box annotations for 13,474 frames. We train the model with Stochastic Gradient Descent (SGD) optimizer, initial learning rate $1*10^{-2}$ and SGD momentum 0.937. First, we trained the detector using the Gaze on Object dataset with the training set of 2,450 images validation set of 1633 images for five epochs. Then, we pre-trained the model using the Retail Gaze dataset with the training set of 2,745 images validation set of 589 images for five epochs.

2) HEAD TRACKING

Object detection models identifies objects bounding box over the frame. However, they do not address the concept of object permanence between frames. In gaze estimation, there is a need to track the subject to get a sequence of gaze estimation in the entire video, and head detection is not sufficient. We track the person's head in the entire video or sequence of frames. There are traditional object tracking methods such as meanshift, optical flow and DL-based methods such as Recurrent YOLO (ROLO), DeepSORT for object tracking. Considering the popularity and wide usage in multi-object tracking, we have selected the object tracking framework DeepSORT, which is an extension of the SORT algorithm. It integrates the appearance data of objects to improve associations. Data association combines an extra appearance measure based on the pre-trained CNNs, and re-identifies the tracks after a duration of a long occlusion. Since we do not have the frame-by-frame head bounding boxes in both Retail Gaze and Gaze on Object datasets, we have used the pre-trained DeepSORT. In this Deep-SORT algorithm, the CNN is trained on a large dataset with human re-identification, implemented by deep cosine metric learning.

C. GAZE ESTIMATION

The development process of the attention-based deep CNN is discussed in this section qualitatively. In the gaze estimation literature, there exist several approaches to estimate the gaze in the wild using DL like multi-task CNN, LSTM-based CNNs, static CNNs, capsule networks, CNNs with shifted grids [11], [13], [16], [23], [24], [30]. As our problem is less researched in the previous literature, we designed and developed the gaze estimation model specifically for the retail environment from the scratch by incorporating concepts from the literature. For this purpose, we designed four abstract design solutions, improving one by one to accurately estimate gaze saliency heatmaps. We initially developed an end-to-end learning design solution as shown in Fig. 3 (a), to the problem by giving the input as the scene image and directly outputting a saliency heatmap of the gaze. Due to the complex nature of the gaze estimation problem in retail and with dataset limitations, an end-to-end model was insufficient to capture the required accuracy for gaze estimation.

Our second design solution was inspired by how humans used to follow the gaze of other people and the work of Chong *et al.* [23]. When a person predicts the looking
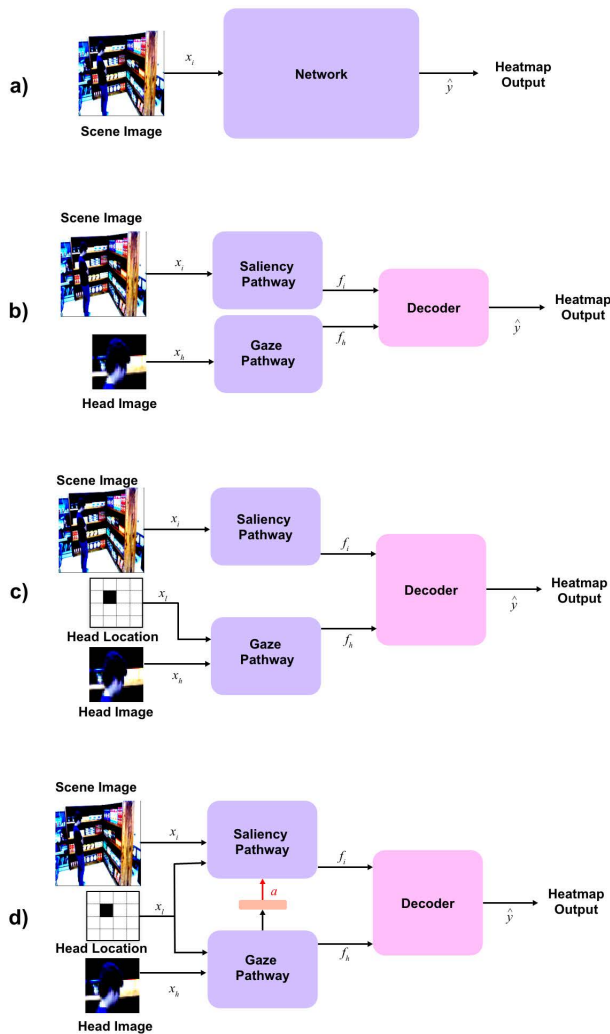
FIGURE 3. Gaze Estimation Solutions for retail: (a) End-to-end solution, (b) Human inspired two-pathway solution, (c) Hypothetical Gaze Distribution solution, (d) Depth-based solution.

direction of another human, they first look at the head or eyes to predict the region of view. Then, reason about saliency objects in their perspective to estimate the looking point. This model contains two different CNN pathways, the saliency pathway for the scene image and the gaze pathway for the head image cropped from the scene image, as depicted in Fig. 3 (b). We incorporated the head location channel into the gaze pathway to improve the learning of the gaze direction in the gaze pathway. Moreover, the attention mechanism introduced by Chong *et al.* [23], was added to regulate the scene pathway by the gaze pathway. This allowed the model to focus on the scene frame, where the head is oriented and connected the two pathways through the attention layer. Finally, the features from the two pathways were combined and fed into the decoder to generate the gaze saliency heatmap. The human-inspired two pathway solution improved the accuracy of saliency estimation. However, as the model architecture was specifically designed for gaze estimation with

front-head images in daily activity scenes, the model was under-performing in estimating out-of-scene gaze targets and the correct retail shelf depth.

We developed a third design solution to improve the accuracy of out-of-scene gaze target estimations. The model architecture is depicted in Fig. 3 (c), which is an extension to the second design solution by adding a new hand-designed feature called the object channel. The object channel is a hypothetical distribution of the gaze generated from retail product item boundaries. This feature helped the model to narrow down its gaze estimation point search space and increased the accuracy of out-of-scene gaze targets. Hypothetical gaze distribution model, the complete model architecture developed based on this solution, is discussed in Section III-D.

A fourth design solution was developed to overcome the issue of estimating the correct retail shelf depth. All the previous solutions explored the gaze direction in 2D representation and hardly encoded the gaze depth-channel. Hence, it is difficult to capture the exact spatial data due to the limited awareness of the scene depth. The introduced depth-based solution uses a depth pathway and a 3D gaze pathway as shown in Fig. 3 (d). Due to the unavailability of depth image data in gaze estimation datasets, we used monocular depth estimation to incur the scene depth. This model uses multi-task learning, learns 3D gaze estimation with different datasets, and uses a pre-trained monocular depth estimation network. This depth feature and 3D gaze were combined to predict the gaze heatmap. This design solution estimated the scene depth and improved the accuracy of gaze saliency estimation. The implemented model architecture for this solution is discussed in Section III-E.

Combining all the hand-designed features introduced in the previous solution, we developed a final design solution which is the depth-based dual attention (DDA) model, shown in Fig. 6. In this approach, we used more hand design components such as field of view (FOV) generator, monocular depth estimation, 3D gaze directions, depth selector, and object channel to help the model improve gaze saliency estimation. We aggregated two parallel attention components, namely a FOV generator for FOV attention and a depth rebasing component for depth attention. This solution also used multi-task learning on different datasets to train different model components. The combined final solution further improved the accuracy of gaze saliency estimation.

D. HYPOTHETICAL GAZE DISTRIBUTION MODEL

We have developed the object channel in the hypothetical gaze distribution model prior to the final Depth-based dual attention model. Hypothetical Gaze Distribution architecture consists of three main components: scene pathway, head pathway, and a shifted grids classification. The model design is given in Fig. 4. The head pathway, which is a CNN feature extractor, computes the head feature map from the head image. Scene pathway computes the scene feature map by taking input, as the concatenation of the scene image, head position channel, and hypothetical gaze

distribution (object channel). Based on the features of the head, we applied an attention mechanism similar to the study by Chong *et al.* [23], to pay greater attention to scene characteristics that are more likely to be looked at. A fully connected layer, which models the attention mechanism, is used to compute the attention map using concatenated head feature map and head position channel. Thus, we performed the concatenation of three 2D feature maps to produce a stacked feature map. Few existing models [11], [23], have provided head position as a spatial reference, allowing the model to learn quicker, and we followed that method in this model.

In addition, we found that the hypothetical gaze distribution assists the model to learn faster and get accurate gaze fixation from the scene. The hypothetical gaze distribution is a binary image of product items boundaries, with white pixels representing object boundary boxes and black pixels representing the other area of the image. We used multi-model predictions to predict the fixation point that is introduced in [11]. Their proposed shifted grids, which predict overlapping outputs from the model improved the confidence of the classification. Finally, we calculated the average of the shifted outputs to get the final prediction. Inputs of this model are (224×224) size scene image, head image, head position channel and hypothetical gaze distribution. Input to the scene pathway is a ($3 \times 224 \times 224$) size feature map, a concatenation of the scene image, head position channel, and hypothetical gaze distribution. We used pre-trained VGG-16 [31] as the feature extractor for both head and scene pathways.
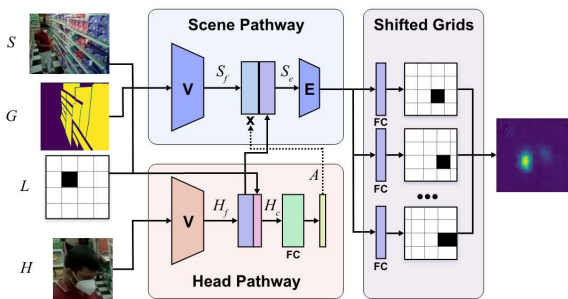

FIGURE 4. Hypothetical gaze distribution model architecture.

E. DEPTH BASED Face3D MODEL

As shown in Fig. 5, the depth based Face3D Model is a depth-based approach that uses monocular depth estimation and 3D gaze estimation for gaze estimation. It consists of 3D gaze estimation pathway and a monocular depth estimation network. The 3D gaze estimation pathway consists of a CNN feature extractor that computes the head feature map and three fully connected layers to compute the 3D gaze vector $g(g_x, g_y, g_z)$. Computed 3D gaze vector then mapped into 2D gaze point on the image plane using the depth information from the monocular depth estimation network and the head location. Inputs of this model were (224×224) scene image and head image. We used pre-trained ResNet-50 [32], as the
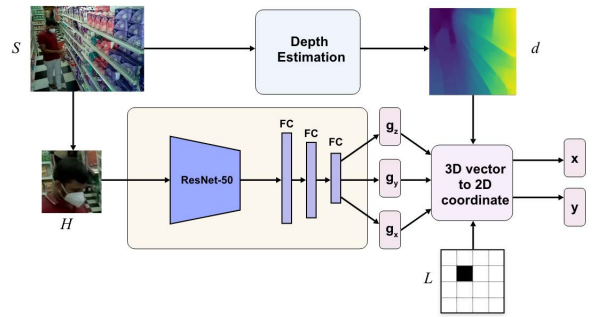

FIGURE 5. Depth based Face3D model architecture.

feature extractor and MiDaS [33], as the monocular depth estimation network.

F. DEPTH BASED DUAL ATTENTION MODEL

We developed the DDA model by combining the concepts of hypothetical gaze distribution model and depth-based face3D model. The architecture of the model is depicted in Fig. 6. There are several main components in this model, such as 3D gaze estimator from the Face3D model, object channel from the hypothetical gaze distribution model, monocular depth estimator, scene image feature extractor and dual attention module. This model consists of four stages. In the first stage, the 3D gaze estimator estimates the 3D gaze direction using the head of the image, and the monocular depth estimator estimates the depth of the scene. The second stage is the field of view mask generator that generates a field of view mask of the person. In the third stage, the depth range selector divides the depth into ranges. In the final stage, we get the scene features using CNN feature extractor and feeds these dual attentions features, hypothetical gaze distribution and the scene feature into CNN heatmap generator to generate the heatmap. We generate the pixel-level gaze location and target gaze object using this heatmap.

1) 3D GAZE ESTIMATION

The 3D gaze estimation component is similar to the depth-based Face3D model. This 3D gaze estimator outputs a normalized 3D gaze vector $g(g_x, g_y, g_z)$, where g_x and g_y are in the image plane and g_z is for depth direction. As shown in Fig 6, We feed this gaze vector to the rectification layer denoted by L, a fully connected layer which learns the right adjustment to the FoV mask generator. In addition, the rectification layer overcome catastrophic forgetting in multi-task learning by mapping different domains of the 3D gaze into the same domain. The depth range selector and FoV mask generator, and depth range selector use these 3D gaze vector values as their inputs.

2) DEPTH ESTIMATION

Face3D model gets the advantages of depth information to improve the gaze estimation. In this architecture, the depth range selection component use scene depth as an input.
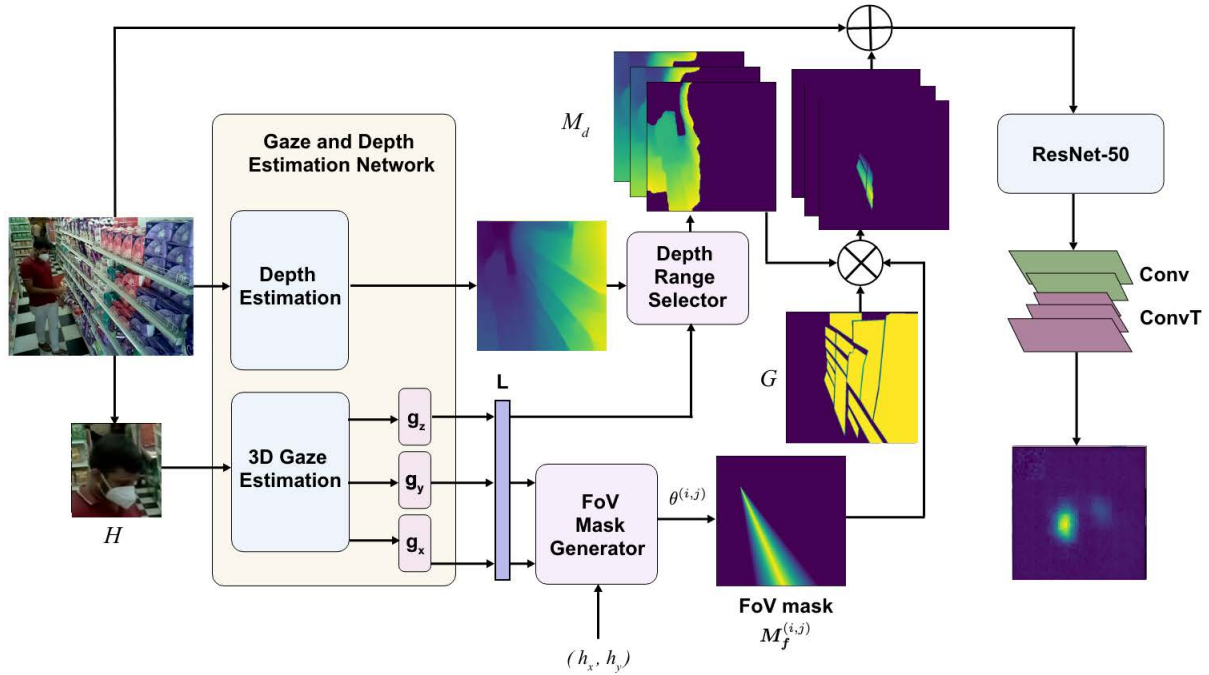
FIGURE 6. Depth-based dual attention model architecture.

We used a state-of-the-art monocular depth estimation network to extract the depth map of the scene. This component takes the normalized RGB scene image as input and output the normalised depth map (d).

3) DEPTH RANGE SELECTOR

We introduced the depth range selector component to improve the scene depth understanding of our model. First, we infer the depth map (d) from the depth estimation component and compute the depth of person head location (d_h) as in (1), where (i, j) is a pixel index of the head bounding box and N is the number of indexes.

$$d_h = \frac{1}{N} \sum_{(i,j)} d^{(i,j)} \quad (1)$$

Then, we calculated the gaze depth level (μ) as given in (2), where (g_z) is the gaze vector in for depth direction.

$$\mu = d_h + g_z \quad (2)$$

Next, We divided the depth to similar size of depth section to get a depth range δ as in (3), where α decide the depth sections.

$$\delta = (max(d) - min(d))/\alpha \quad (3)$$

Then, we selected three different depth sections from the gaze depth level as in (4). Therefore, the considered three depth sections cover multiples of two(M_d^1), four(M_d^2), and
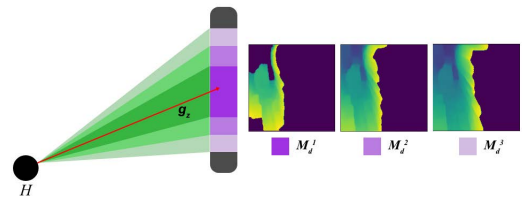

FIGURE 7. Three depth sections selected by depth range selector.

six(M_d^3) of the depth range(δ), respectively as shown in 7.

$$M_d = \begin{cases} (\mu - \delta, \mu + \delta) \\ (\mu - 2\delta, \mu + 2\delta) \\ (\mu - 3\delta, \mu + 3\delta) \end{cases} \quad (4)$$

4) FIELD-OF-VIEW MASK GENERATOR

Field-of-view (FoV) mask generator component generates the person's image level field-of-view region. This generated plane-polarized region is an infinitely extended solid cone shape starting from the head location of the person. Image plane level gaze direction from the 3D gaze estimator (g_x, g_y) and the head location (h_x, h_y) are the input of the FoV mask generator and output is the FoV attention map (M_f). First, we computed the angular difference θ between gaze direction and the vector from one point to head location, as given in (5), where (i, j) is the coordinate of each point in the (M_f).

$$\theta^{(i,j)} = \arccos\left(\frac{(i - h_x, j - h_y) \cdot (g_x, g_y)}{||(i - h_x, j - h_y)||_2 \cdot ||(g_x, g_y)||_2}\right) \quad (5)$$

Since, fixation points are more likely to have smaller θ values, we assigned more weights to the data points closer to the estimated sight line and less to further. Therefore, the FoV mask can be generated as in (6), where α is the field of view angle. We set α to 12 and achieve a viewing angle of 60.

$$M_f^{(i,j)} = max\left(1 - \frac{\alpha\theta^{(i,j)}}{\pi}, 0\right) \tag{6}$$

Then, in order to create the dual attention map, we aggregated the FoV attention map and depth attention map as given in (7). After that, to enhanced the saliency estimation we aggregated this dual attention map with hypothetical gaze distribution G (object channel) to create hypo dual attention map as in (8), where $\otimes$ denotes the element wise product.

$$M_{dual} = M_f \otimes M_d \tag{7}$$
$$M_{hypo} = M_{dual} \otimes G \tag{8}$$

5) SALIENCY ESTIMATION
In this stage, we concatenated the scene image and hypo dual attention map and fed into a feature extractor, which is a pre-trained CNN feature extractor. The output feature map from the feature extractor is then fed into the heatmap generator. Heatmap generator consists of two convolution layers followed by three de-convolutional layers to generate the gaze heatmap. The maximum value point in this heatmap is the estimated gaze point. We use mean square error (MSE) for heatmap regression loss.

Inputs for this model are (224×224) size scene image, head image, head location point in the image plane. We used the pre-trained Face3D model as our 3D gaze estimation network, MiDaS [33] as the monocular depth estimation network and pre-trained ResNet-50 [32] as the feature extractor.

The presented models are implemented on PyTorch framework using python language. We used two pre-trained feature extractors, ResNet-50 [32] and VGG-16 [31] in model implementations. All these feature extractors are pre-trained on ImageNet [34].

All the models are trained with Adam optimizer, a learning rate of $1 * 10^{-5}$ and batch size of 32 on Colab pro and used data augmentation such as random crops and colour profiles. In order to prevent overfitting, we used specific patient values for early stopping.

IV. EXPERIMENTS
A. DATASETS
This study used multiple gaze estimation datasets to train and evaluate different components in the DL models. For the task of gaze estimation in 2D image coordinates, we used the GazeFollow dataset [11] and the GOO dataset [8]. Subsequently, for gaze estimation in 3D real-world coordinates, the Gaze360 dataset was used. Moreover, due to the limitations in retail gaze estimation datasets, we introduced a new real-world retail gaze estimation dataset with annotation novelties that are specifically designed to retail. Each dataset

is discussed in depth under this section. Further, Fig. 8 shows sample images from each dataset.

Gaze360: Gaze360 [17], is a large-scale gaze tracking dataset, annotated with 3D gaze in real-world coordinates. The authors have collected the dataset primarily for gaze target estimation in the wild with a wide range of head poses, significant variation of indoor and outdoor settings, lighting variations, and large subject diversity. The dataset contains 172,000 images collected with 238 different subjects in five indoor and two outdoor locations. An essential feature of the dataset is the significant amount of back-head images. The authors have captured a gaze yaw variation of approximately $+/-140°$, including partially occluded eye images.

Gaze Follow: Gaze Follow is another important dataset used in our study. Recasense *et al.* [11], have introduced the Gaze Follow dataset along with the human gaze following concept. Gaze Follow is another large-scale dataset consisting of images with people performing daily activities, annotated with their gaze targets in 2D image coordinates within the image itself. The dataset was collected using several primary datasets that contain people as a source of images; hence it is an image set of humans doing different daily activities. This multi-user gaze dataset consists of 122,143 images of 130,339 people.

Gaze On Objects (GOO): The GOO dataset [8] is the most suitable dataset in the literature for 2D remote gaze estimation in retail environments, currently. This has two parts, a large synthetic dataset (GOO-Synth) consisting of 192,00 images and a small real dataset (GOO-Real) consisting of 9,552 images. The GOO-Real dataset has been collected in a mock-up retail environment by placing grocery items belonging to 24 different classes on shelves. They have used 100 different subjects to capture a large subject diversity. However, authors have only used two camera capture angles which is a significant limitation considering the real-world applicability of the dataset in retail. Furthermore, this dataset is a single-user gaze dataset annotated with gaze targets in 2D image coordinates, gazed at object bounding box, and gazed at object segmentation mask. The GOO-Synth dataset has been created with a natural looking duplicated scenes used in GOO-Real in a synthetic environment. It consists of images captured from 50 different virtual camera angles of 20 synthetic person models interacting with the scene. Authors have included significant subject diversity in the dataset by varying the human model's skin tone, gender, body form, and outfit parameters. Furthermore, the dataset contains a vast scene diversity of 38,400 different synthetic scenes.

Retail Gaze: Retail Gaze is a dataset for remote gaze estimation in real-world retail environments that includes retail product category area segmentation annotations. This dataset is collected and processed by the authors of this study. We introduced the novel area segmentations, as they make more practical sense for gaze object prediction in retail. This dataset consists of images in a local supermarket environment, where each image contains a human gazing upon an object or an area on a shelf. The images capture the
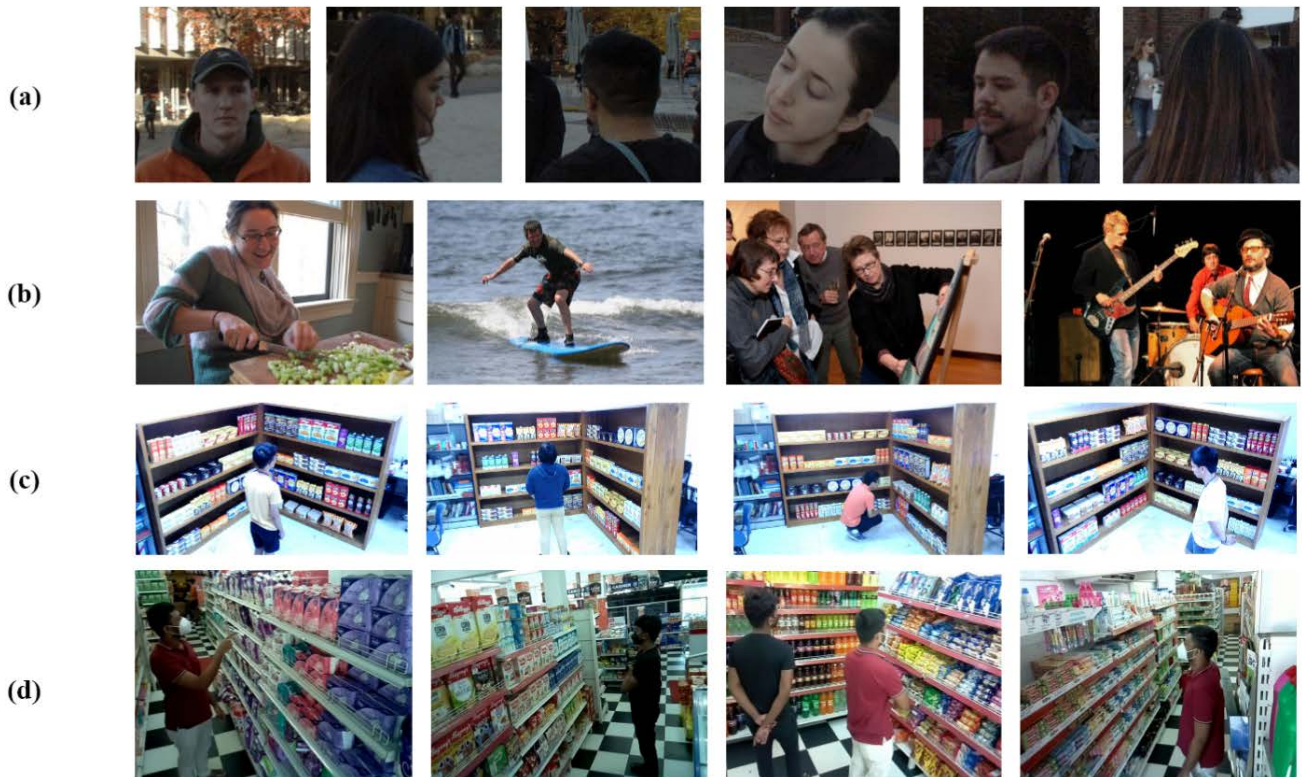
FIGURE 8. Datasets: (a) Gaze360, (b) Gaze Follow, (c) GOO-Real, (d) Retail Gaze.

third-person view of the scene. Each image is annotated with the gaze target in 2D image coordinates, the bounding box of the person's head, and segmentation masks of product areas being gazed at. It consists of 3,922 images of two subjects, captured from twelve different camera capture angles, with each image consisting of a shelf completely packed with different retail product items.

Moreover, we have used shelves that consisted of products belonging to the same categories in the same area to replicate the real nature of retail shelves. The dataset consists of 2,745 images in the train set, 589 in the test set, and 588 in the validation set with 70%, 15%, and 15% as split ratios. In the collection process, we have focused on capturing the real-world retail environment conditions by collecting the images under controlled light conditions in the supermarket without any external light sources and collecting images from diverse camera capture angels. Furthermore, each participant was instructed on which area to look at on the shelf using a predetermined pattern. The annotation process used these predetermined patterns when annotating the ground-truth gaze target. The head box and the product category areas were manually annotated. A comprehensive overview of the dataset is presented in our previous paper [29]. We have made the dataset available to the public for research use [35].

B. EVALUATION METRICS
We use three main performance metrics namely Area Under ROC curve (AUC), L2-distance, and Angular error to assess the presented gaze estimation models. The AUC criteria presented by Judd *et al.* [36] to predict the performance of saliency maps using gaze fixations is used as the first evaluation metric. L2 distance and Angular error are two primary metrics that are widely used to evaluate the performance of gaze estimation in 2D image coordinates. L2 distance is defined as the mean Euclidean distance between the gaze predictions and their respective ground-truth gaze target in 2D image coordinates. As given in (9), where gt_x_i, gt_y_i denote the ground truth gaze annotations and x_i, y_i refer to gaze predictions in 2D image coordinates. Unlike the L2 distance, the angular error is used to evaluate 2D and 3D gaze estimation performance. The angular difference between the predicted gaze direction vector and the ground-truth gaze direction vector is defined as the angular error in the literature.

$$L_2 distance = \frac{1}{n} \sum_{i=1}^{n} \sqrt{(gt\_x_i - x_i)^2 + (gt\_y_i - y_i)^2} \quad (9)$$

C. TRANSFER LEARNING APPROACHES
Transfer learning approaches are used to train models with limited resources and high efficiency. We have incorporated transfer learning approaches for the model training processes to overcome the gaze estimation dataset limitations in retail. Gaze Follow dataset provided a better starting point to train some of the models due to its generic nature of diverse scene images and wide range of head pose variations, including back-head images. Furthermore, an early stopping technique

TABLE 2. Dataset summary.

Dataset	Year	Subjects	Total	Annotations	Type	Environment
Gaze360 [17]	2019	238	172,000	3D gaze direction	Single user	Miscellaneous
Gaze Follow [11]	2015	130,339	122,143	2D gaze gaze target	Multi user	Miscellaneous
GOO-Real [8]	2021	100	9,552	2D gaze target, gazed at object bounding box, gazed at object segmentation mask	Single user	Mock retail environment
Retail Gaze [34]	2022	2	3,922	2D gaze target, gazed at object segmentation, head bounding box	Single user	Real retail environment

TABLE 3. Transfer learning approach.

Model	Trained Dataset	Epochs	AUC	L2-distance	Angular Error
Hypothetical Gaze Distribution	GOO-Real	5	0.897	0.175	33.512
Hypothetical Gaze Distribution	Gaze Follow + GOO-Real	13, 25	0.94	0.14	25.22
Hypothetical Gaze Distribution	Gaze Follow + GOO-Syth + GOO-Real	6, 5, 10	0.90	0.18	27.18
Face3D	GOO-Real	75	0.94	0.15	20.6
Depth-based Dual Attention	GOO-Real	18	0.95	0.14	19.3

was used to avoid overfitting during model training using Gaze Follow and GOO validation sets.

Our three models are separately trained on the GOO dataset with or without pre-training on the Gaze Follow dataset. The obtained results are shown in Table 3, with the parameters and associated values used for the model experiments. Also, it shows the performance test results evaluated on the combined datasets. The models which received transfer learning are shown in the trained dataset column with the used datasets in the corresponding trained order. The number of epochs trained for each dataset is shown in the epochs columns in the same order as the dataset. Thereafter, the models were tested on the GOO dataset test set for model validity. We observed that the transfer learning approach improved the results of the hypothetical gaze distribution model. However, the model performance degraded when transfer learned with the GOO-synthetic dataset due to overfitting to the synthetic environment. The Face3D and the proposed DDA models were not transferred learned, because they achieved better results on the GOO-Real dataset. Transfer learning on these models remains as potential future work.

Since the main focus of this study is to apply remote gaze estimation in real-retail environments, we apply transfer learning on the Retail Gaze dataset. Depth-based dual attention model, the best performing model on the GOO-Real dataset, was selected as the model architecture. We plot the learning curves of this model for train loss and validation loss as shown in Fig. 9 and Fig. 10, respectively.

We plot the train and validation loss of the model on the GOO-Real dataset in green colour and transfer learning train and validation loss on the Retail Gaze dataset in red colour. The model loss was calculated using the standard MSE loss function. We observed that the training loss and validation loss on the GOO-Real dataset significantly reduce with the number of epochs. However, the transfer learning losses only reduce in a small quantity. A potential reason is a model learning most features from learning on the GOO-Real dataset itself, which eventually reduces the number of epochs to be transferred learned on, causing minimal losses. Furthermore,
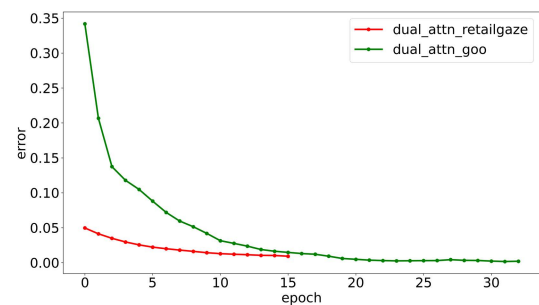

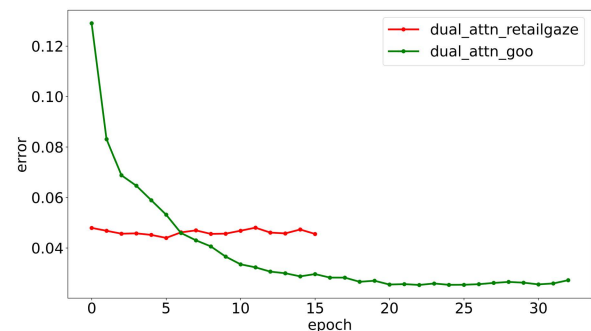**FIGURE 9.** Depth-based dual attention (dda) model train loss.

FIGURE 10. Depth-based dual attention (DDA) model validation loss.

the transfer learning was stopped at 15 epochs with a patience value of 10, because it's validation loss started to increase after 5 epochs.

D. GAZE OBJECT PREDICTION

The task of gaze object prediction is highly applicable in the retail industry. This task was first introduced by Tomas *et al.* [8] in his GOO paper, and it is complex due to the unconstrained environmental parameters in retail. With accurate gaze heatmap data generated from our DDA model, we implemented a rule-based algorithm for gaze object prediction. The inputs to our algorithm are the predicted gaze

heatmap and the corresponding segmentations masks. The element wise matrix multiplication between the heatmap and each segmentation mask is then computed to calculate the product sum. The mask with the highest sum was taken as the gazed objects bounding box. The pseudocode of the algorithm is shown in the Algorithm 1.

Algorithm 1 Gaze Object Prediction Algorithm

$h \leftarrow$ *predicted heatmap*
layers $\leftarrow$ *list of segmentation masks*
seg $\leftarrow$
max_product $\leftarrow -\infty$
for *layer* in *layers* **do**
    product $\leftarrow$ *sum(h * layer)*   ▷ element wise product sum
    if *max_product* < *product* **then**
        max_product $\leftarrow$ *product*
        seg $\leftarrow$ *layer*
    end if
end for

Even though this approach was straightforward, the resulting gaze object prediction accuracy was low, around 30%. The noisy pixels included in the heatmap could produce a higher product sum when multiplied with a larger mask, leading to false predictions. To overcome this issue, first we clipped the heatmap using a threshold value and set the low intensity pixels to zero. The threshold was selected as 150 by conducting extensive experiments. The introduced clipping mechanism improved the accuracy of the algorithm. We obtained a better accuracy of 75% on the Retail Gaze test set. The accuracy was calculated using the Equation 10.

$$Accuracy = \frac{PositiveObjectPredictions}{TotalPredictions} \quad (10)$$

V. RESULTS ANALYSIS

A. HEAD DETECTOR EVALUATION

The presented head detector model is trained on both GOO and Retail gaze datasets using transfer learning approach. Fig. 11 shows the obtained evaluation matrices recall, precision and mean average precision (mAP). Here, mAP@p, where $p \in (0, 1)$ denotes the increase of Intersection over Union (IoU) for object detection from 0.5 to a 1. Considering the model trained on GOO dataset, the metrics recall, precision and mAP@.5 quickly get closer to the value 1 with a small number of epochs. Here, mAP@[.5,0.95] get near to 0.8, indicating that the mAP over different IoU thresholds increase from 0.5 to 0.95, in steps of 0.05.

B. RESULTS ANALYSIS USING GOO DATASET

GOO is the current benchmark dataset available for gaze saliency estimation in retail. We evaluated the performance of the proposed DL models on the GOO dataset. We have done an in-depth analysis of the model results by using the standard evaluation metrics in 2D gaze estimation, AUC,

TABLE 4. Model results comparison on GOO-real dataset.

Model	AUC	L2-distance	AE
Recasense et al. [11]	0.9	0.185	37.2
Lian et al. [24]	0.89	0.16	32.9
Chong et al. [23]	0.88	0.16	30.3
Hypothetical Gaze Distribution	0.94	0.14	25.2
Face3D	0.94	0.15	20.6
Depth-based Dual Attention (DDA)	0.95	0.14	19.3

L2-distance, and Angular error (AE) to quantitatively evaluate the models.

Table 4 summarizes the experimental results of the presented three models on the GOO dataset. We compared the results of our models with the state-of-the-art gaze saliency estimation model results on the GOO dataset [11], [23], [24].

Accordingly, the proposed models in this study surpass the state-of-the-art gaze saliency estimation model results on all evaluation metrics. Moreover, we observed the best performance of the proposed DDA model, which combines the two concepts implemented in the hypothetical gaze distribution and Face3D model, in all three evaluation criteria. This model achieved an angular error of 19.3°, which is an improvement of 33% compared to the work of Chong *et al.* [23]. Furthermore, the model achieves a relative improvement of 7% for AUC and 12.5% for L2-distance, showing the superiority and applicability of the model for remote gaze estimation from back-head images. A potential reason is the integration of hand-designed features specifically designed to capture the features in a retail environment like the object channel and scene depth understanding.

Moreover, several examples of failure cases from the baseline Chong *et al.* [23] model and the corresponding success cases from our proposed Depth-based Dual Attention model are shown in Fig. 12. These examples depict how the proposed model learned to predict out-of-shelf gaze targets correctly and, accurately estimate the target depth. The first-row in the left side images illustrate a test case where the baseline model incorrectly predicted the gaze target outside the retail shelf. The first-row in the right-side images show the result for the test case from our proposed model. The first image represents the designed object channel, which helped the model to narrow down its gaze estimation point search space to avoid predicting targets outside of the shelf. The second-row in the left side images illustrate a test case where the baseline model fails to predict the correct depth of the gaze target. The proposed model accurately estimates the target depth using monocular depth estimation, shown in the second-row in the right-hand side. Despite the extreme eye occlusion conditions in back-head scenarios, our proposed model correctly estimates the gaze target within the retail shelf with the guidance of object channel and depth features.

C. EVALUATION OF THE PROPOSED DDA MODEL

The proposed depth-based dual attention model is evaluated using the novel Retail Gaze dataset. In order to assess the task
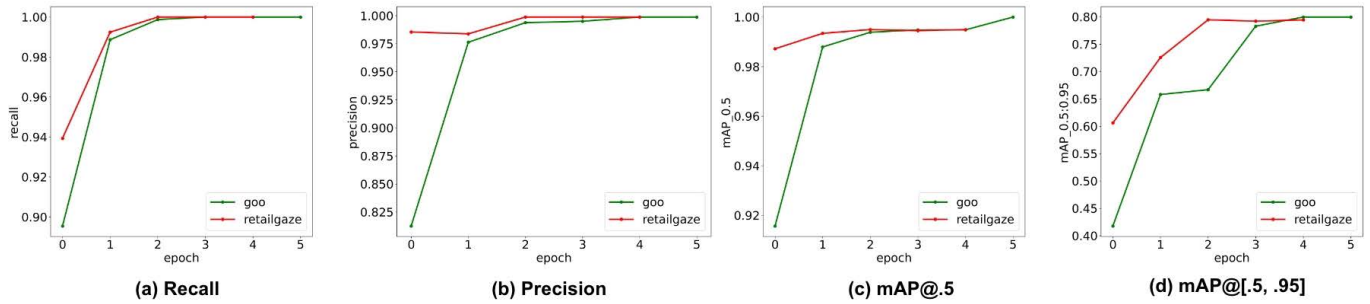
(a) Recall **(b) Precision** **(c) mAP@.5** **(d) mAP@[.5, .95]**

FIGURE 11. Evaluation of head detector model with GOO and Retail gaze datasets using transfer learning.
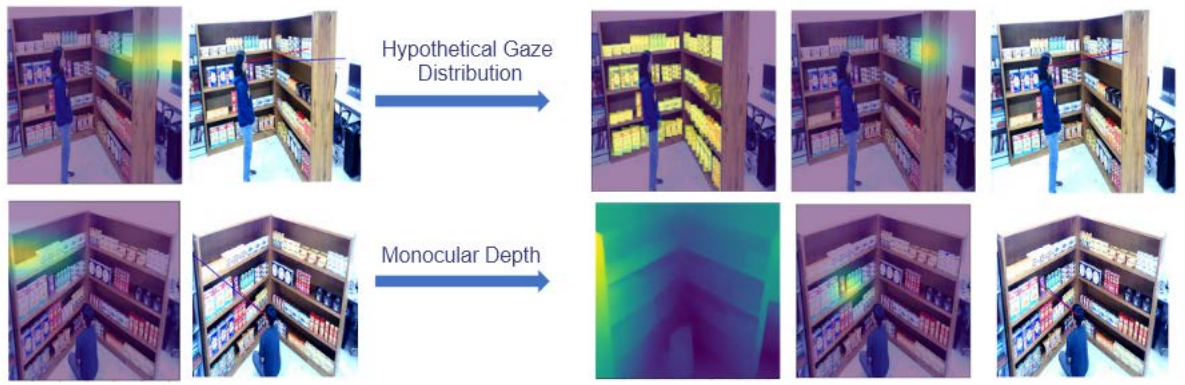

FIGURE 12. Sample results: Left: failure cases from the baseline Chong *et al.* [23] model, Right: corresponding success cases from the proposed DDA model. Red Arrow: Ground truth gaze direction, Blue Arrow: Predicted gaze direction.

of remote gaze saliency estimation on this dataset, we compared the results against the state-of-the-art baselines [11], [23], [24]. We trained all the models using the same transfer learning approach and same early stopping approach to ensure the fairness of the results. The models received GOO-Real pre-training and then transfer learning with the Retail Gaze dataset until convergence using early stopping to avoid overfitting. The obtained results are summarized in Table 5. We observed that our proposed model surpasses state-of-the-art baseline models on all three performance metrics. Our model achieves an angular error of 15.3°, which is an improvement of 57.9% compared to the model proposed by Chong *et al.* [23]. The obtained results clearly show the superiority of the proposed solution and applicability in real-retail environments. This significant improvement is the introduced object channel, which reduced the gaze estimation point search space and incorporated gaze target depth information into the model.

D. GAZE OBJECT PREDICTION RESULTS ANALYSIS

We introduced gaze object prediction, the main experimental task performed on the output heatmap data from the DDA model. Fig. 13 shows few success and failure cases of gaze object prediction. Since the gaze object prediction algorithm is an optimized, simple rule based approach, we can

TABLE 5. Model results comparison on retail Gaze dataset.

Model	AUC	L2-distance	Angular Error
Recasense et al. [11]	0.79	0.25	36.4
Lian et al. [24]	0.52	0.42	62.4
Chong et al. [23]	0.73	0.24	35.2
Proposed DDA model	0.95	0.15	15.3

assume that the accuracy of gaze heatmap estimation directly resemble with the accuracy of gaze object prediction. Hence, the two success cases show how the proposed DDA model robustly estimates gaze saliency heatmap and correctly predicts the gazed-at segmentation mask. The two failure cases show incorrect gazed-object segmentation mask prediction due to the inaccuracy of the generated saliency heatmap. Here, the reason for the left failure case can be hypothesized as, lack of temporal gaze point data in the model. Rather, the right failure case shows a general failure scenario due to a large angular error in the predicted gaze point.

VI. DISCUSSION

A. STUDY CONTRIBUTIONS

In this study, we explored the problem of remote gaze estimation in retail environments. With the rapid adaptation of computer vision and DL technique in gaze estimation literature, we identified the potential of applying this to

FIGURE 13. Gaze Object Prediction Results: (a) Success, (b) Failure.

accurately estimate the customer gaze saliency in unconstrained retail environment settings like eye and head occlusion, large subject variations, diverse lighting conditions. Our main contribution of the study is the introduced Depth-based dual attention (DDA) model, a novel deep CNN for accurately estimating gaze saliency maps in real-world retail environments. In the methodology section, we showed a comprehensive overview of the model design using four design solutions inspired by the related studies. In order to adapt the proposed gaze estimation model to a retail environment, we proposed two novel features, object channel and the monocular depth channel, which were implemented in the presented Hypothetical gaze distribution and Face3D models, respectively. These two concepts were combined in the final model. Importantly, the introduced Retail Gaze dataset [35], for benchmarking gaze saliency estimation in real-retail environment is another main contribution of our study. The dataset contains 3922 labelled images, captured from 12 different camera angles and 2 different subjects. Also, the images contain images with varied retail-product diversity.

Moreover, the camera angle is a contributing factor to the performance of the model. For instance, a camera angle that captures totally occluded eyes of the user, can hinder the performance of the model. In order to address this issue, we captured majority of the images from camera angles that captured a side view of the user. The scope of this study is mainly based on eye gaze estimation. Since the camera angle tracking is not the main scope of this research, the experiments related to the impact of camera angle on the performance of the eye gaze estimation are not considered in this study. In addition, our research scope has not focused on the time complexity of the model and up to the authors' knowledge, the current state-of-the-art models for gaze estimation have not stated time consumption metrics.

B. COMPARISON WITH EXISTING STUDIES

To the best of our knowledge, there are no existing state-of-the-art models that are directly related to eye gaze estimation

in retail environment. Therefore, in this section, we conducted multiple experiments to improve the performance of our models, and presented thorough quantitative and qualitative analysis on the obtained results. First, we evaluated the obtained results on the benchmark GOO-Real dataset to validate the model implementations. The obtained results in Table 4 shows the superiority of our models compared to the state-of-the-art models, by surpassing them in all three performance metrics. Next the results of the proposed final model was compared against the sate-of-the-art models on the Retail Gaze dataset. In additon, the results shown in Table 5, confirmed the superiority of the model compared to the related work. The model achieved an improvement of 57.9% for angular error criteria, compared to the baseline Chong *et al.* [23] model.

C. FUTURE POSSIBLE EXTENSIONS

With the promising results obtained in the experiments, we observe the significant potential that yields in our approach for retail. However, in real-retail environments, often multiple subjects are present in the same scene. This could lead to complex scenarios like subject occlusion, which was not considered in our research. The concept of multi-user gaze estimation holds a high promise in gaze saliency estimation in retail for accurate and robust predictions. The introduced retail gaze datasets could be extended as multi-user gaze estimation datasets, and DL architectures can be further improved to handle multi-user gaze estimation. Furthermore, incorporating the temporal aspect of gaze estimation is another possible future extension to our work.

VII. CONCLUSION

Remote gaze saliency estimation in retail is a novel concept that has a significant potential towards building innovative retail stores. In this study, we researched the application of remote gaze saliency estimation for non-interruptive, low-cost, and scalable customer behaviour analysis in retail. We proposed a Depth-based Dual Attention model, a three-stage, three-attention-based deep CNN for gaze saliency estimation from back-head images in the wild. We developed four design solutions to comprehensively represent the parameters of gaze saliency estimation problem in retail and introduced the novel object channel and depth-rebasing components as hand-designed features, designed in our two preceding model architectures and then combined in the final model.

Extensive quantitative and qualitative analysis on the benchmark GOO-Real dataset demonstrates the superiority of the proposed models and the importance of our introduced hand-designed components. Our proposed solution improved 33% for angular error compared to the current best work in the literature. Furthermore, we introduced Retail Gaze, a real-world retail gaze saliency estimation dataset, to ensure the validity and applicability of our proposed solution in real retail environments. The proposed solution achieved an angular error of 15.3° on the Retail Gaze dataset, which

demonstrates that it performs favourably in real retail environments.

As future work the depth-based dual attention model can be extended to handle temporal information which could improve the performance of remote gaze estimation. Furthermore, we suggest to tackle the complex task of multi-user gaze estimation in a retail environment. Unavailability of multi-user retail gaze datasets and handling subject occlusions are current barriers to this task.

REFERENCES

[1] C. Bermejo, D. Chatzopoulos, and P. Hui, "EyeShopper: Estimating shoppers' gaze using CCTV cameras," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2765–2774, doi: 10.1145/3394171.3413683.

[2] D. Grewal, M. Levy, and V. Kumar, "Customer experience management in retailing: An organizing framework," *J. Retailing*, vol. 85, no. 1, pp. 1–14, Mar. 2009, doi: 10.1016/j.jretai.2009.01.001.

[3] C. Ofir and I. Simonson, "The effect of stating expectations on customer satisfaction and shopping experience," *J. Marketing Res.*, vol. 44, no. 1, pp. 164–174, Feb. 2007. [Online]. Available: https://www.jstor.org/stable/30162462

[4] D. W. Wallace, J. L. Giese, and J. L. Johnson, "Customer retailer loyalty in the context of multiple channel strategies," *J. Retailing*, vol. 80, no. 4, pp. 249–263, Jan. 2004, doi: 10.1016/j.jretai.2004.10.002.

[5] A. Generosi, S. Ceccacci, and M. Mengoni, "A deep learning-based system to track and analyze customer behavior in retail store," in *Proc. IEEE 8th Int. Conf. Consum. Electron. Berlin (ICCE-Berlin)*, Sep. 2018, pp. 1–6, doi: 10.1109/ICCE-Berlin.2018.8576169.

[6] V. Nogueira, H. Oliveira, J. A. Silva, T. Vieira, and K. Oliveira, "RetailNet: A deep learning approach for people counting and hot spots detection in retail stores," in *Proc. 32nd Conf. Graph., Patterns Images (SIBGRAPI)*, Rio de Janeiro, Brazil, 2019, pp. 155–162, doi: 10.1109/SIBGRAPI.2019.00029.

[7] D. Lian, X. Chen, J. Li, W. Luo, and S. Gao, "Locating and counting heads in crowds with a depth prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 4, 2021, doi: 10.1109/TPAMI.2021.3124956s.

[8] H. Tomas, M. Reyes, R. Dionido, M. Ty, J. Mirando, J. Casimiro, R. Atienza, and R. Guinto, "GOO: A dataset for gaze object prediction in retail environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3119–3127, doi: 10.1109/CVPRW53098.2021.00349.

[9] Y. Li, M. Liu, and J. Rehg, "In the eye of the beholder: Gaze and actions in first person video," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 15, 2021, doi: 10.1109/TPAMI.2021.3051319.

[10] 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, doi: 10.1016/j.eswa.2022.116894.

[11] T. Recasens, A. Aditya, K. Carl, and V. Antonio, "Where are they looking?" in *Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2015, pp. 199–207, doi: 10.5555/2969239.2969262.

[12] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Single-user 2D gaze estimation in retail environment using deep learning," in *Proc. 2nd Int. Conf. Adv. Res. Comput. (ICARC)*, Feb. 2022, pp. 206–211, doi: 10.1109/ICARC54489.2022.9754167.

[13] Y. Fang, J. Tang, W. Shen, W. Shen, X. Gu, L. Song, and G. Zhai, "Dual attention guided gaze target detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11385–11394, doi: 10.1109/CVPR46437.2021.01123.

[14] A. A. Akinyelu and P. Blignaut, "Convolutional neural network-based methods for eye gaze estimation: A survey," *IEEE Access*, vol. 8, pp. 142581–142605, 2020, doi: 10.1109/ACCESS.2020.3013540.

[15] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16495–16519, 2017, doi: 10.1109/ACCESS.2017.2735633.

[16] P. L. Mazzeo, D. D'Amico, P. Spagnolo, and C. Distante, "Deep learning based eye gaze estimation and prediction," in *Proc. 6th Int. Conf. Smart Sustain. Technol. (SpliTech)*, Sep. 2021, pp. 1–6, doi: 10.23919/SpliTech52315.2021.9566413.

[17] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6911–6920, doi: 10.1109/ICCV.2019.00701.

[18] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2019, doi: 10.1109/TPAMI.2017.2778103.

[19] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: Passive eye contact detection for human-object interaction," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2013, pp. 271–280, doi: 10.1145/2501988.2501994.

[20] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 334–352, doi: 10.1007/978-3-030-01249-6_21.

[21] 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.*, Glasgow, U.K., 2020, pp. 365–381, doi: 10.1007/978-3-030-58558-7_22.

[22] Y. Kodama, Y. Kawanishi, T. Hirayama, D. Deguchi, I. Ide, H. Murase, H. Nagano, and K. Kashino, "Localizing the gaze target of a crowd of people," in *Proc. 14th Asian Conf. Comput. Vis. (ACCV)*, Perth, WA, Australia, 2018, pp. 15–30, doi: 10.1007/978-3-030-21074-8_2.

[23] E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg, "Detecting attended visual targets in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5395–5405, doi: 10.1109/CVPR42600.2020.00544.

[24] D. Lian, Z. Yu, and S. Gao, "Believe it or not, we know what you are looking at," in *Proc. Asian Conf. Comput. Vis.*, Perth, WA, Australia, 2019, pp. 35–50, doi: 10.1007/978-3-030-20893-6_3.

[25] M. Khamis, A. Hoesl, A. Klimczak, M. Reiss, F. Alt, and A. Bulling, "EyeScout: Active eye tracking for position and movement independent gaze interaction with large public displays," in *Proc. 30th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2017, pp. 155–166, doi: 10.1145/3126594.3126630.

[26] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg, "Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 397–412, doi: 10.1007/978-3-030-01228-1_24.

[27] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125, doi: 10.1109/CVPR.2017.106.

[28] 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, doi: 10.1109/CVPR.2016.90.

[29] S. Senarath, P. Pathirana, D. Meedeniya, and S. Jayarathna, "Retail gaze: A dataset for gaze estimation in retail environments," in *Proc. Int. Conf. Decis. Aid Sci. Appl. (DASA)*, Mar. 2022, pp. 1040–1044, doi: 10.1109/DASA54658.2022.9765224.

[30] B. Mahanama, Y. Jayawardana, and S. Jayarathna, "Gaze-Net: Appearance-based gaze estimation using capsule networks," in *Proc. 11th Augmented Hum. Int. Conf.*, May 2020, pp. 18–21, doi: 10.1145/3396339.3396393.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[32] 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, doi: 10.1109/CVPR.2016.90.

[33] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022, doi: 10.1109/TPAMI.2020.3019967.

[34] 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, doi: 10.1109/CVPR.2009.5206848.

[35] P. Pathirana and S. Senarath. *Retail Gaze: Gaze Estimation in Retail Environment*. Accessed Feb. 16, 2022. [Online]. Available: https://www.kaggle.com/dulanim/retailgaze

[36] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113, doi: 10.1109/ICCV.2009.5459462.

SHASHIMAL SENARATH (Student Member, IEEE) received the B.Sc.Eng. degree (Hons.) from the Department of Computer Science and Engineering, Faculty of Engineering, University of Moratuwa, Sri Lanka. His main research interests include computer vision and deep learning.

PRIMESH PATHIRANA (Student Member, IEEE) received the B.Sc.Eng. degree (Hons.) from the Department of Computer Science and Engineering, Faculty of Engineering, University of Moratuwa, Sri Lanka. His main research interests include computer vision and deep learning.

DULANI MEEDENIYA (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of St. Andrews, U.K. She is a Professor of computer science and engineering with the University of Moratuwa, Sri Lanka. She is the Director of the Bio-Health Informatics Group, where she engages in many collaborative research. She is a coauthor of 100+ publications in indexed journals, peer-reviewed conferences and international book chapters. She serves as a reviewer, a program committee, and an editorial team member in many international conferences and journals. Her main research interests include software modeling and design, bio-health informatics, deep learning and technology-enhanced learning. She is a fellow of HEA (U.K.), MIET, a member of ACM, and a Chartered Engineer registered at EC, U.K.

SAMPATH JAYARATHNA (Member, IEEE) received the Ph.D. degree in computer science from Texas A&M University College Station, in 2016. He is an Assistant Professor of computer science with Old Dominion University, Norfolk, Virginia, USA, where he is associated with the Web Science and Digital Libraries (WS-DL) Research Group. His research interests include machine learning, information retrieval, data science, eye tracking, and brain–computer interfacing. He was a recipient of the 2021 National Science Foundation CAREER Award. He is a member of ACM and Sigma XI.

• • •