Gaze Estimation via the Joint Modeling of Multiple Cues

Wenhe Chen, Hui Xu, Chao Zhu, Xiaoli Liu, Yinghua Lu, Caixia Zheng, and Jun Kong

***Abstract*—How to automatically predict people's gaze has attracted attention in the field of computer vision and machine learning. Previous studies on this topic set many constraints, such as restricted scenarios and strict and complex inputs. To mitigate these constraints to predict the gaze of people in more general scenarios, we propose a three-pathway network (TPNet) to estimate gaze via the joint modeling of multiple cues. Specifically, we first design a human-centric relationship inference (HCRI) module to learn the object-level relationship between the target person and the surrounding persons/objects in a scene. To the best of our knowledge, this is the first time that the object-level relationship is introduced into the gaze estimation task. Then, we construct a novel deep network with three pathways to fuse multiple cues, including scene saliency, object-level relationships and head information, to predict the gaze target. In addition, to extract the multilevel features during network training, we build and embed a micropyramid module in TPNet. The performance of TPNet is evaluated on two gaze estimation datasets: GazeFollow and DLGaze. A large number of quantitative and qualitative experimental results verify that TPNet can obtain robust results and significantly outperform the existing state-of-the-art gaze estimation methods. The code of TPNet will be released later.**

***Index Terms*—Gaze Estimation, Three-pathway Network, Multiple Cues Fusion, Human-centric Relationship Inference, Image Understanding, Computer Vision.**

I. Introduction

In the process of infant growth, the visual perceptual system is developed very early on. A baby who is a few months old can lock its gaze onto others in a very sensitive way [1]. With increasing age, human beings can not only exactly follow the gaze of other people but also quickly identify the target at which they are looking [2]. This extraordinary ability of human beings is very important for understanding human intentions and analyzing human social activities in human-to-human interactions or human-to-object interactions [3]. For example, when a person is crossing the road while staring at a mobile phone (as shown in Fig. 1(a)), we infer that there is a safety hazard for him. When people are shopping (as shown in Fig. 1(b)), by following their gaze, we can identify the products that they are interested in. Estimating other people's gaze is an innate ability of human beings, but is it possible for a computer to achieve gaze estimation with machine learning techniques? This question has attracted the attention of researchers and has become a significant research issue in the community of computer vision and machine learning.

In the field of computer vision, gaze estimation is used to automatically estimate people's gaze from an image or a video and predict the target at which they are looking [3], which is beneficial for many practical applications, e.g., virtual reality [4], human-robot interactions [5],[6], behavior monitoring [7] and gaming [8]. However, there is only a limited amount of research work on this topic, and those studies set many constraints to make gaze estimation tasks simple, e.g., restricting the scenarios of gaze estimation to people looking at each other only [9], requiring that a person's face is available for detection [10], or utilizing the ground-truth data of eye-tracking or multiple images as additional input [11],[12]. These constraints greatly limit the application of gaze estimation, as the frontal face of people does not always appear in an image or a video (as shown in Fig. 1(a), (c), and (d)), and the eye-tracking data or multiple images are not easy to acquire[3]. Hence, there is an urgent need to develop gaze

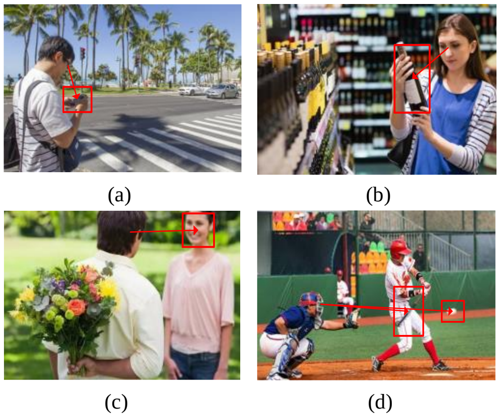


Fig. 1. Several examples of estimating gaze in a natural scene. (a) A person is crossing the street with a mobile phone in his hand. (b) A person is shopping in a supermarket. (c) Two persons are talking. (d) The people are playing baseball; the two arrows and the two boxes bounding the baseball and baseball bat indicate that the left person's gaze point is ambiguous.

This work was supported in part by the National Natural Science Foundation of China under Grants 61672150, 61702092 and 61907007, in part by the Fund of the Jilin Provincial Science and Technology Department under Grants 20200201199JC, 20200401081GX, 20200401086GX, 20190201305JC and 20180201089GX, and in part by the Fundamental Research Funds for the Central Universities under Grants 2412020FZ031, 2412020FZ029 and 2412019FZ049. (Corresponding Authors: Caixia Zheng, Jun Kong.)

W. Chen, H. Xu, C. Zhu, and C. Zheng are with the College of Information Sciences and Technology, Northeast Normal University, Changchun 130117, China (e-mail: chenwh256@nenu.edu.cn; xuh504@nenu.edu.cn; zhuc377@nenu.edu.cn; zhengcx789@nenu.edu.cn).

X. Liu is with the Department of Chemical & Biomolecular Engineering, National University of Singapore, Singapore 117585, Singapore (e-mail: chelxi@nus.edu.sg).

Y. Lu and J. Kong are with the Institute for Intelligent Elderly Care, College of Humanities & Sciences of Northeast Normal University, Changchun 130117, China (e-mail: kongjun@nenu.edu.cn; luyh@nenu.edu.cn).

C. Zheng and J. Kong are with the Key Laboratory of Applied Statistics of MOE, Northeast Normal University, Changchun 130024, China.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3071621, IEEE Transactions on Circuits and Systems for Video Technology

estimation methods that can address more general scenarios.

Recently, to execute gaze estimation without the above restrictive assumptions, Recasens *et al.* [13] proposed a method that jointly employs gaze field estimation and salient object detection to predict the gaze direction and gaze point in a single-view image. Specifically, this method uses head images to estimate the gaze direction to avoid the problem of face detection and considers the saliency of the objects in a scene to estimate the candidate targets of a gaze. However, although this method does not require the appearance of the frontal face, it cannot accurately predict the gaze point of a person when the person's face is totally occluded (as shown in Fig. 1(c)). In addition, it will generate ambiguous gaze point estimation results, such as the case shown in Fig. 1(d). In fact, the problems of gaze point estimation shown in Fig. 1(c) and (d) are very difficult to solve; even humans struggle to precisely judge the gaze point shown in Fig. 1(c) and (d) without using the scene content and the relationship between the persons/objects.

In a natural scene, there is a strong spatial geometric relationship between the target person and the persons/objects being interacted with, which can provide a strong additional basis for inferring people's gaze target. Hence, investigating high-level semantic relationships are very significant, as the result of doing so is the further improvement of the accuracy of gaze estimation, but this is ignored by the method proposed in [3],[13],[14]. Furthermore, to the best of our knowledge, no research has introduced this object-level relationship into the task of gaze estimation. Therefore, the motivation of our work is to design a gaze estimation model that simultaneously fuses a person's information (e.g., the head position), scene saliency and the high-level relationships between persons/objects to predict the gaze point of the person that appears in the image. To achieve this goal, we construct a gaze estimation model with three pathways. Specifically, the first pathway predicts the coarse gaze area based on the head close-up image and head position; the second pathway adopts the full image as the input to obtain the saliency map of the objects in the scene; and the third pathway is used to learn the high-level semantic relationships between the persons/objects in the original image. These three pathways are then fused to obtain the final gaze field, in which the maximum value of the location is defined as the predicted gaze point.

The main contributions of this paper are as follows:

1) Develop a novel three-pathway network (TPNet) to effectively combine head information, scene saliency and relationship cues for gaze estimation.

2) Design a human-centric relationship inference (HCRI) module to capture the high-level semantic relationships between the person, whose we want to predict his gaze, and the surrounding persons/objects. To the best of our knowledge, this is the first study to introduce semantic relation information into the task of gaze estimation.

3) Construct a micropyramid module inspired by the feature pyramid network (FPN). The micropyramid module can cause the network to capture the different level features simultaneously.

4) Test TPNet, a deep network with only a few parameters that must be tuned, which makes it robust against different gaze estimation tasks, on the GazeFollow and DLGaze datasets and compare it with some baselines and state-of-the-art methods. The comparison results effectively validate the effectiveness of TPNet.

II. Related Work

A. Gaze Estimation

Gaze estimation predicts the gaze target of a person in an image/video, which is beneficial for understanding human actions and human intentions in an image/video. In recent years, many gaze estimation methods [3],[13]-[24] have been proposed. Arar *et al.* [16] employed the face and eye detection/tracking method to obtain eye features, e.g., the glint and pupil center, to estimate the gaze of a person. Lai *et al.* [17] utilized two cameras to capture eye images and then adopted the glint and contour features of the eye to track the person's gaze. Lian *et al.* [18] proposed a multiview multitask gaze estimation method based on the multiview captured eye images. Zhu and Ramanan [10] proposed a face detector to model each facial landmark to detect the direction of a person's gaze. Corcoran *et al.* [8] combined face detection and eye-tracking to predict gaze by detecting key points in the eye. Sugano *et al.* [19] used facial feature points to reconstruct 3D eye images to learn a gaze estimator. Duffner and Garcia [23] proposed an unsupervised incremental learning method based on low-level features to predict the visual focus of the person in the video. These methods have achieved good performances for gaze estimation, but they still have several limitations. For example, persons must be facing the camera, and facial contours and eyes must be visible. To address this problem, Recasens *et al.* [13] first combined head pose and scene saliency information to predict the gaze of a person in a scene and then built a public gaze estimation dataset called GazeFollow to test the effectiveness of the proposed method. This method has none of the above constraints; it only needs to input a single image to predict the gaze direction and the gaze object of the person in an image, which makes it a great breakthrough in the field of gaze estimation. Hence, in subsequent periods of time, many algorithms [3],[14],[22] have been proposed based on the idea of Recasens *et al.* [13]. Recently, Zhuang *et al*. [24] first constructed a group gaze dataset, and then proposed a two-stream framework to predict the common gaze point of a group of persons. Although these existing studies have made some progress in gaze estimation, there is no work that considers learning the relationship among persons/objects in a scene to predict gaze. Hence, how to introduce high-level semantic relationships to the gaze estimation task to effectively improve the estimation accuracy is the main motivation of this paper, and that also is the main difference between our work and the abovementioned works.

B. Visual Saliency Detection

Visual saliency detection is highly correlated with gaze estimation, but they are still two different tasks. The goal of visual saliency detection is to determine the regions in an image that attract the attention of observers outside this image, while the goal of gaze estimation is to predict the location in an image being looked at by the person in the image [22]. Visual saliency detection has become a research hotspot in recent years. In the early stage, saliency detection methods were developed based on a low-level hand-crafted feature, e.g., Itti *et al.* [25] first employed multiple scales of low-level features to generate a saliency map of an image. Judd *et al.* [26] proposed a saliency detection model that combines low-, middle-, and high-level image features. Borji [27] combined bottom-up low-level saliency features with top-down advanced visual features to learn the gaze area of the eye. Driven by the superior performance of deep learning models, many visual saliency detection methods have been presented based on deep convolutional neural networks. For example, Pan *et al.* [28] proposed a new saliency detection model based on data-driven metrics, which has two parts inspired by a generative adversarial network (GAN): a generator used to generate the saliency map and a discriminator used in the true/false judgment of the detected saliency map and the ground truth. There are also many other methods [29]-[32] developed based on deep learning that have achieved significant success in the field of visual saliency detection. The abovementioned saliency detection methods are based only on the features of an image, irrespective of the gaze prediction task. In this paper, we want to utilize the saliency of an image to predict the gaze object of a person; hence, the saliency map will be influenced by the direction of the gaze.

C. *Relation Inference*

Relation inference, which models object/person relationships at the instance level in an image or a video, can provide useful information for diverse tasks, e.g., object detection, video activity recognition and scene understanding [33]-[37]. Most of the early work utilized statistical methods to model the object or region co-occurrence [38],[39] as the object relationship. In recent years, based on the breakthrough results of object detection [40]-[43], instance-level graph relation inference [44]-[46] has gradually attracted attention. Hu *et al.* [47] proposed a lightweight object-relational module that simultaneously processes a set of objects and models their relationships based on the appearance features and geometry information. Liu *et al.* [48] proposed a structure inference network for object detection, which learns the potential relationship between objects by a graphical model to improve the object detection performance. Xu *et al.* [49] proposed a generic framework based on iterative message passing to generate a scene graph, which models the objects in an image and their relationships. In addition to the abovementioned supervised relation inference methods, some unsupervised and weakly supervised methods [50],[51] have been proposed to infer the relationships between objects/persons. Relation inference is often used for object detection or scene understanding, and many studies have proved that considering relation inference can significantly improve the performance of a model; however, there is almost no research introducing the relationship between objects into the task of gaze estimation. Gaze estimation predicts the gaze direction and gaze target object of a person in a scene, which is generally related to object detection and scene understanding. Hence, we believe that the relationship between objects/persons is also beneficial for improving the accuracy of gaze estimation.

III. The Architecture of TPNet

We propose a new three-pathway network (TPNet), which is a deep neural network consisting of three pathways. Each pathway in TPNet can solve a different subproblem, and they are fused to solve the final gaze estimation problem. Specifically, the gaze pathway takes the head close-up picture and the position coordinates of the corresponding head as inputs to roughly predict where a person is looking. The saliency pathway takes the original image as input to estimate the salient objects in a scene. The relation pathway takes the original image as input to obtain the high-level semantic relationships between the person and his surrounding objects/persons in a scene. To learn different scale features of the original image, the saliency and the relation pathways separately add a micropyramid module. In addition, it is worth noting that all the components of TPNet are divisible, hence, our proposed components can be easily transplanted to the design of new networks for other tasks. The framework of TPNet is shown in Fig. 2.

A. *Gaze and Saliency Pathways*

In the gaze pathway, we input a close-up image of the head into the Alexnet network [52] to learn its feature representation and then concatenate the head position information with the head image representation to regress a gaze prediction field. The position and close-up image of the head of the person can be detected in the entire image by the single shot multibox detector (SSD) method [53].

When we estimate the gaze target of persons in a scene, we first infer their gaze direction and then consider whether there are salient objects in the estimated gaze direction. Hence, we believe that the most salient object in the direction of the gaze is the most likely gaze target object. Based on this idea, in the saliency pathway, the entire image is input into the Alexnet network, in which the convolutional layers are retained but the fully connected layers are removed. This process is carried out because we use a micropyramid module to replace the fully connected layers to fuse the features learned by the convolutional layers. The micropyramid module can integrate the multiscale information of the learned feature and transform Alexnet into a fully convolutional network to detect the salient regions in a scene much more effectively. The details of the micropyramid module are given in Section III-C.

B. *Relation Pathway*

The image of a scene does not simply contain a set of objects/persons; it also implies scene contextual information and the relationships between objects/persons [36]. The object-level relationship is useful to identify and locate the gaze target objects of a person. To be specific, when we estimate the gaze of a person, especially when that person’s face cannot be seen clearly, we first unconsciously form the high-level semantic relationships among the objects/persons in the scene through the position of the person, hand-held items, head posture, and locations of surrounding objects and then make a reasonable judgment based on these high-level semantic relationships. For example, when we see a person crossing the road with a mobile phone in hand and his head looking down (as shown in Fig. 1(a)), we naturally think that he is looking at

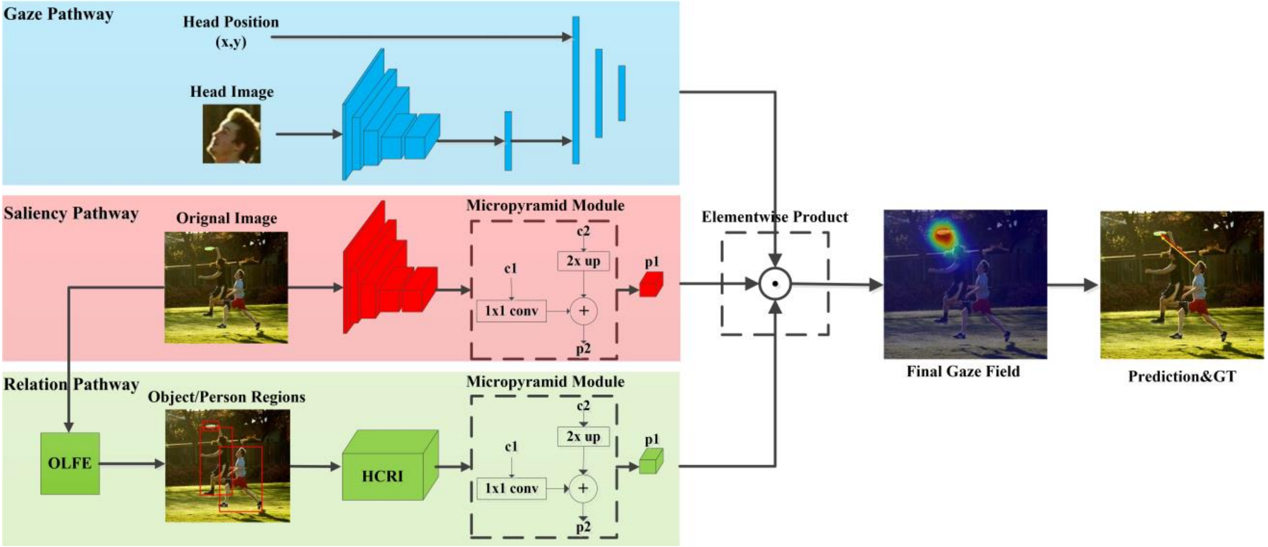


Fig. 2. The framework of the TPNet architecture. TPNet has three main components: the gaze pathway for roughly predicting the gaze area, the saliency pathway for obtaining the saliency map of a scene, and the relation pathway for capturing the high-level relationship between objects/persons. The outputs of the three pathways are fused to estimate the gaze point of the person in an image. In the picture "Prediction&GT", the red line indicates the ground truth of the gaze, and the yellow line indicates our predicted gaze.

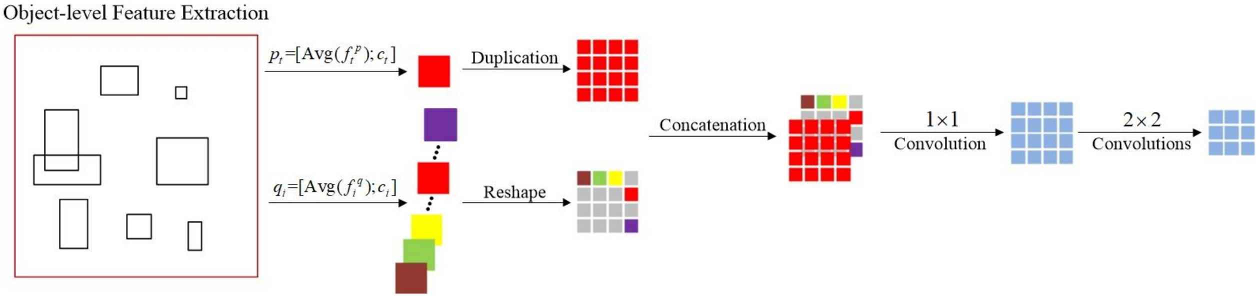

Object-level Feature Extraction
p_t=[Avg(f_t^p); c_t]
Duplication
q_i=[Avg(f_i^q); c_i]
Reshape
Concatenation
1×1 Convolution
2×2 Convolutions

Fig. 3. The architecture of HCRI. First, RPN proposals and ROI-Pooling features extracted by SIN are used to generate the feature representations p_t and q_t for the target person and the surrounding objects/persons, respectively. Then, p_t is duplicated to make the amount of p_t equal to the amount of q_i. Finally, all duplicated p_t and all q_i (i=1,2,…,M+N) are reshaped and concatenated as the inputs of the relationship feature extraction operator $E_\theta(\cdot)$, which includes several convolutional layers, to infer the human-centric relationship.

the mobile phone even if we cannot see the person's face and eyes. Hence, we hope the proposed TPNet can also capture the object-level relationships to improve the accuracy of the gaze estimation. To achieve this, we propose a relation pathway, which is composed of two key modules: object-level feature extraction and human-centric relationship inference.

1) Object-level feature extraction (OLFE). Structure inference network (SIN) [48] is an object detection method that can effectively infer the object locations in the image. Here, we employ SIN to obtain the potential person/object regions and the corresponding features from a given image. Specifically, we first use SIN to extract the region proposals by the region proposal network (RPN) and the corresponding region appearance features by the ROI-Pooling layer, respectively. Then, we fuse the position coordinate of each RPN region proposal and its corresponding ROI-Pooling feature as the object-level feature representation.

2) Human-centric relationship inference (HCRI). Inspired by [54] and [55], we design a module called HCRI to learn the high-level semantic relationship between the target person (the person whose gaze we want to predict) and the surrounding persons/objects. To be specific, HCRI can learn the pair-wise relationship between the target person and all surrounding persons/objects in the scene based on the object-level feature representations. The structure of HCRI is shown in Fig. 3.

Given an image I, we first employ OLFE to extract some local regions $Q = \{Q_1, Q_2, \dots, Q_{M+N}\}$ representing all persons/objects in I. Suppose that Q contains M persons $P = \{P_1, P_2, \dots, P_M\}$ and N objects $O = \{O_1, O_2, \dots, O_N\}$, we utilize HCRI to infer the pair-wise high-level semantic relationship between the selected target person P_t and the surrounding persons/objects by the following formula:

$$HCRI(P_t) = \{E_\theta(p_t, q_i) : Q_i \in Q\} \tag{1}$$

where p_t and q_i denote the object-level feature representations of P_t and Q_i respectively, $E_\theta(\cdot)$ is a relationship feature extractor with the parameter θ. Next, we will introduce the details of p_t, q_i and $E_\theta(\cdot)$.

To learn the human-centric relationship based on both appearance and geometric location information of persons/objects, as we mentioned earlier in OLFE, we combine the regional appearance features and the region proposals extracted by SIN to produce p_t and q_i. The specific forms of p_t and q_i are described as follows:

$$p_t = [\text{Avg}(f_t^p); c_t] \text{ and } q_i = [\text{Avg}(f_i^q); c_i] \tag{2}$$

where f_t^p and f_i^q represent the ROI-Pooling appearance features of P_t and Q_i respectively, c_t and c_i denote the position coordinates of the corresponding RPN proposals for P_t and Q_i respectively, $\mathrm{Avg}(\cdot)$ is an average pooling operator which scales and normalizes f_t^p and f_i^q.

In actual implementation, $E_\theta(\cdot)$ is set as several convolution operations. To be specific, $E_\theta(\cdot)$ first adopts a 1×1 convolutional layer to efficiently calculate the pair-wise relationship between the target person P_t and each Q_i in Q. Then, to capture the high-level semantic relationship, $E_\theta(\cdot)$ applies a set of 2×2 convolutional layers to leverage the information from the neighboring relationships.

The proposed HCRI in the relation pathway is inspired by [54] and [55], but HCRI has several main differences compared to [54] and [55]. First, the relation inference models proposed by [54] and [55] are for the video action classification and visual question answering tasks, but our HCRI is designed for the gaze estimation task. Second, the relation inference models proposed by [54] and [55] are based on the 'abstract' object-level relationship inference, that is, they don't detect the candidate persons/objects in the image but directly divide the entire convolutional feature map into regularly individual feature cells and treat each cell as an 'abstract object'. While the object-level relationship inferred by our HCRI is the 'instance' relationship since we use an object detection network to extract the instance regions. Third, [54] and [55] only utilize the appearance feature as the feature representation to learn the relation, while our HCRI simultaneously employs the appearance feature and position coordinate as the object representation to learn the relationship information. Finally, different from [54] adopting 3×3 convolution operator, HCRI employs 2×2 convolution operator because that it can be executed quickly and obtain a better performance [56], which has been proved in our experiments.

At the end of the relation pathway, we embed a micropyramid module to further capture the multiscale information of the learned relationship feature to improve the performance of TPNet.

C. Micropyramid Module

The feature pyramid has been widely used in hand-crafted feature extraction because it can extract features and identify objects at different scales. However, because the learning feature pyramid requires many computing resources, it has not been commonly applied to deep learning. Recently, Lin *et al.* [57] proposed a feature pyramid network (FPN) with a laterally connected top-down structure to construct high-level semantic maps at multiple scales. The FPN takes advantage of the inherent multiscale structure of deep convolutional neural networks; thus, it needs to construct only a small number of additional boundaries, which greatly reduces the computational cost. In the proposed TPNet, we build a general-purpose feature extractor named the micropyramid module inspired by the FPN to capture the multilevel (from high-level to low-level) feature representations. Specifically, we use a micropyramid module to replace the fully connected layers of Alexnet in the saliency pathway, which can not only improve the accuracy of the gaze estimation but also increase the computing speed of TPNet. Besides, we also embed a micropyramid module into the relation pathway to improve the performance of TPNet. The specific structure of the micropyramid module is shown in Fig. 4.

In Fig. 4, "c1" represents the output of the last convolutional layer, and "c2" is obtained from "c1" through a convolutional layer. "c2" and "c1" are first processed by "upconv" (the upper convolutional layer) and "lateral" (the lateral connection layer), respectively, and then fused by an addition operation to obtain "p2". The final output "p1", which contains rich information, is gained by executing "smooth" (the smoothing layer) on "p2". It should be mentioned that we inserted only the micropyramid module into the saliency pathway and relation pathway and experimentally demonstrated the effectiveness of this setting.

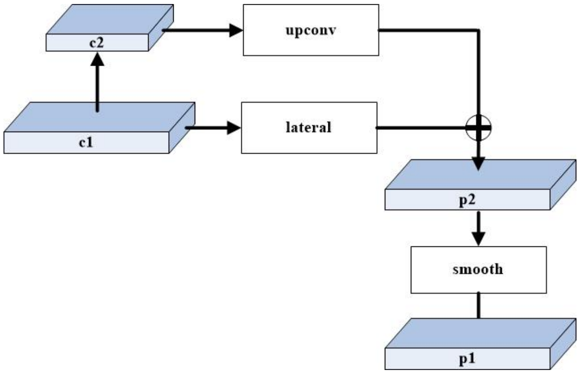


Fig. 4. The architecture of the micropyramid module. The micropyramid module uses the inherent multiscale structure of neural networks to fuse the high-level (c2) and low-level (c1) features and finally outputs a semantic feature map (p1).

D. Fusion of the three pathways

Since the proposed gaze, saliency, and relation pathways can separately obtain the different information, we integrate them into one unified framework to solve the problem of gaze estimation. When fusing the three pathways to learn the final gaze field, we hope that the large activations appear in the locations where all output of the three pathways have large activations. To achieve this, we tested several different fusion strategies in the experiments and selected a relatively superior fusion strategy, i.e., elementwise multiplication, to combine the three pathways to form TPNet. Denoting the original image as x_i, the head close-up image as x_h, and the head position coordinate as x_p, the fusion of the three pathways is described as follows:

$$H = F(G(x_h, x_p) \cdot S(x_i) \cdot R(x_i)) \tag{3}$$

where F represents a fully connected layer and G, S and R denote the outputs of the gaze, saliency and relation pathways, respectively. The final output gaze field is H, which is the possibility distribution of the predicted gaze. In H, the possibility of those regions that people tend to look at is close to 1, and conversely, the possibility is close to 0.

We visualize the outputs of each pathway and the results of the whole model to show the information learned by them. Fig. 5 gives the visualization results. The visualization heatmaps are obtained by resizing the outputs of the last layers of each pathway to the size of the original input image. It should be noted that our TPNet can predict the gaze of each person in the image, but we only randomly select one person and show the predicted gaze of him for clarity. The visualization results in Fig. 5 demonstrate that 1) the gaze pathway learns a rough gaze area, the saliency pathway learns the locations of saliency

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3071621, IEEE Transactions on Circuits and Systems for Video Technology

objects existing in the scene and the relation pathway infers the important object regions which have strong semantic relationship with the people whose we want to predict his gaze; 2) the final output gaze field involves the precise gaze location, which proves that combining the gaze pathway, saliency pathway and relation pathway by elementwise multiplication to estimate gaze is valid; 3) the proposed TPNet has no constraints on the input images, that is, it can address gaze prediction in different scenes even when the resolution of the face is low or the person is not facing the camera.

E. Implementation Details

In the gaze pathway, we adopt Alexnet pretrained on the ImageNet dataset [58] and replace the two original fully connected layers in Alexnet with one fully connected layer of size 500 to learn the features of the head close-up image. These learned feature maps concatenated with the coordinates of the head positions are subsequently processed by the three remaining fully connected layers, which are 400, 200, and 169 in size.

In the saliency pathway, we modify Alexnet pretrained on the Place365 dataset [59] to learn the features of the whole image. To be specific, only the first five convolutional layers of Alexnet are retained, and a 3×3 convolutional layer is added; then, they are connected with a micropyramid module. In the micropyramid module, “upconv” is composed of an upsampling and a 2×2 convolutional layer, and the “lateral” and “smooth” convolutional layers have sizes of 1×1 and 3×3, respectively.

In the relation pathway, we input the original image into the SIN model pretrained on the PASCAL VOC [60] and MS COCO [61] datasets to extract 256 RPN proposals and 256 corresponding ROI-Pooling features. To reduce the computational burden, we apply a 1×1 convolutional layer to transform $7 \times 7 \times 512$ ROI-Pooling features into $7 \times 7 \times 1$ features which are denoted as f_t^p and f_i^q in Eq. (2). Then, we concatenate f_t^p and f_i^q with the 4-dimensional position coordinates c_t and c_i by Eq. (2) to produce p_t and q_i, respectively. The sizes of p_t and q_i are $1 \times 1 \times 5$. When we want to predict a person’s gaze, the relation pathway will employ HCRI to infer the relationship between this target person and the surrounding persons/objects to improve the accuracy of gaze prediction. In HCRI, the head position information of the target person is used to select a RPN proposal covering or nearest to this head position as the target person P_t in Eq. (1). To parallelly learn all pair-wise relationships in an image, p_t is first duplicated 256 times so that each duplicated p_t can be paired with each q_i to learn the human-centric relationship. Then, all duplicated p_t and all q_i (*i*=1,2,…,256) are combined and reshaped to 16×16, respectively (This operation is shown as the red blocks and colorful blocks in Fig. 3). $E_\theta(\cdot)$ in Eq. (1) consists of one 1×1 convolutional layer and three 2×2 convolutional layers. The detailed structure of the micropyramid module is the same as the one adopted in the saliency pathway.

In all pathways, the activation function of the last layer of the network is the sigmoid function, and all the other activation functions are ReLU. The sizes of outputs of the three pathways are 13×13, which are fused by the elementwise product to obtain the final predicted gaze location.

We use PyTorch to implement TPNet. To train the model well, we augment the training image data by flipping and random cropping and resize the images to 227×227. All input data are normalized to the range [0, 1]. In the process of model training, the stochastic gradient descent (SGD), in which the momentum is set to 0.9, is adopted to optimize the network parameters. The number of iterations is set to 200 epochs, and the batch size is set to 50. Due to the large difference between the two experimental datasets, we use different initial learning rates and weight decays for them. Since the GazeFollow dataset is relatively large, we set the learning rate of the model to 5×10^{-3} and set the weight decay as 20% for every 10 epochs. For the DLGaze dataset, to achieve better convergence of the network, we set the learning rate as 3×10^{-3} without weight decay.

IV. Experiments

A. Experimental Setup

1) Datasets: We utilize two large-scale gaze estimation datasets, i.e., GazeFollow [13] and DLGaze [3], to evaluate the performance of the proposed TPNet. The GazeFollow dataset includes a total of 130,339 people in 122,143 images that are collected from several large datasets, e.g., MS COCO [61], ImageNet [58], and Places[59]. This dataset is very challenging since the scenarios in the images are various and the people in the images are performing very different activities. GazeFollow has the annotations about the head position of the person in the entire image, and it provides the standard training/testing split (approximately 4,782 people in the dataset for testing and the rest for training); we follow this split in our experiments. The DLGaze dataset consists of 95,000 images in total, which are frames collected from 86 videos. This dataset includes the different activities of 16 persons in 4 real scenes, and it is also a very challenging dataset because there are many occlusions and severe lighting changes. We randomly divide the DLGaze dataset into a training/testing set according to the ratio of 7/3 in the experiments. DLGaze has no annotation about the head position of the person in the entire image, but it has the annotation of the eye position. Hence, we use the eye position as the center to crop a sub-image as the close-up image of the head of the person. This can simplify the calculation for obtaining the close-up image of the head compared to adopting the SSD method to detect the head of the person.

2) Evaluation Metrics: We use four criteria, i.e., **AUC** (area under the curve), **AvgDist**, **MinDist**, and **AvgAng**, to evaluate the gaze estimation models. The AUC is the area under the receiver operating characteristic curve, which is calculated according to [26]. AvgDist is the average Euclidean distance between our predictions and the corresponding ground-truth annotations when assuming that each image is 1×1 in size. MinDist is the minimum distance between our predictions and the corresponding ground-truth annotations. AvgAng is the average angular difference between the gaze vectors of our predictions and the gaze vectors of the corresponding ground-truth fixations. The gaze vectors are computed by using the eye positions from the ground-truth annotations. The criteria AvgDist, AvgAng, and MinDist are generated according to [13].

3) Baseline Methods: Similar to [3],[13], we compare TPNet against the following baselines to verify its validity. **Random:** The predicted gaze heatmap is a Gaussian map that is generated by the random mean and covariance, and the predicted gaze point is the position where the value of the Gaussian map is maximum. **Center:** The predicted gaze point is defined as the center of the image. **Fixed Bias:** The predicted gaze of a test image is the average of gaze points from the training images in which the head locations are similar to those of this test image. **Judd *et al.*** [26]: This method predicts a person's gaze inside the image by a saliency model, which is proposed based on the low-, middle-, and high-level features. **SalGAN** [28]: This estimates the gaze by using a saliency detection method consisting of a generator and a discriminator. **Recasens *et al.*** [13]: This is a gaze prediction model including a gaze pathway and a saliency pathway. **Lian *et al.*** [3]: This is a two-stage framework that learns the gaze direction field and uses heatmap regression to track the gaze point of persons in an image. **Chong *et al.*** [14]: This is a multitask learning architecture that leverages three different datasets to detect people's general visual attention in images.

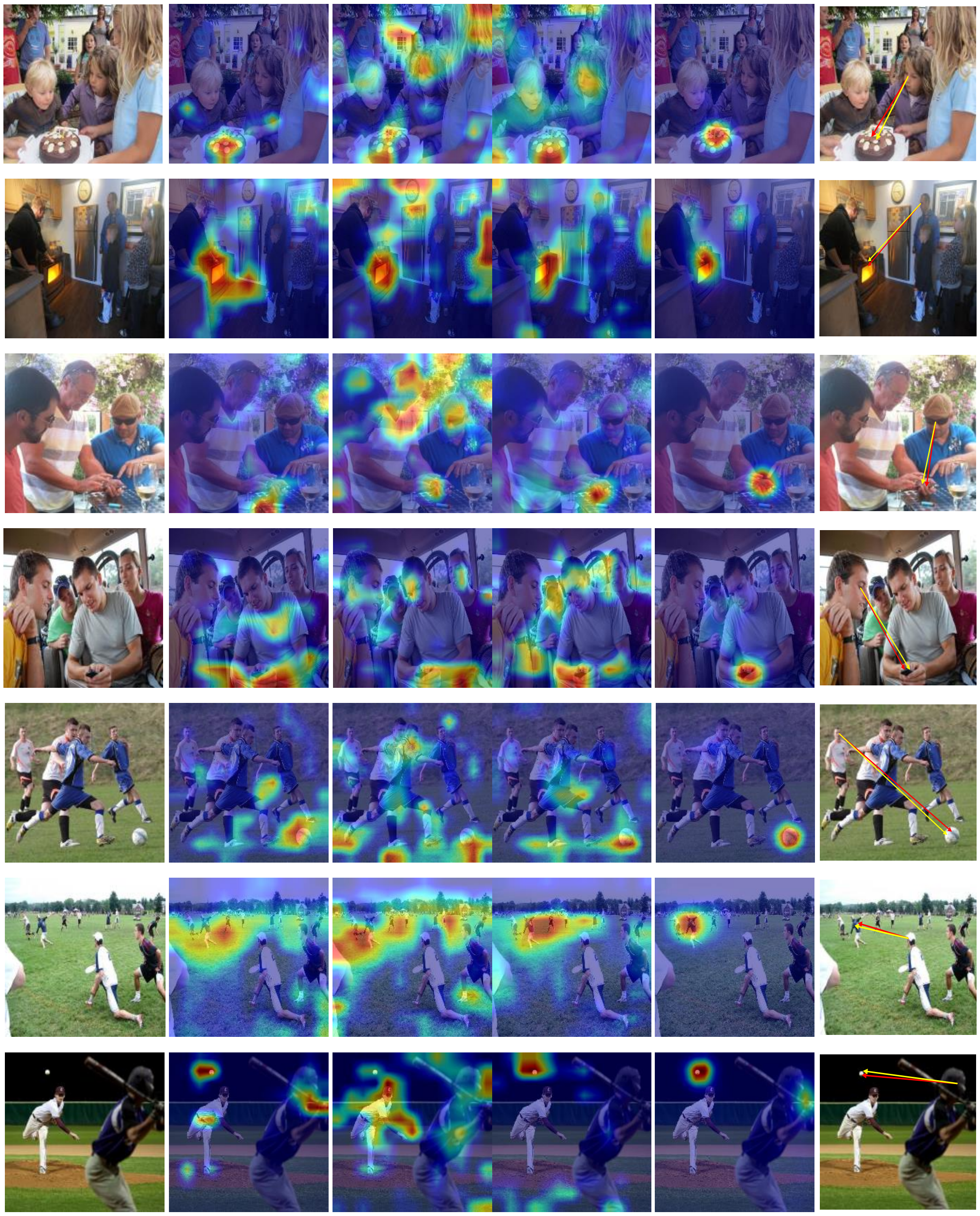


Fig. 5. Visualization of the output of different components of TPNet. The images in the first column are the original images. The second, third, and fourth columns show the visualization of TPNet's gaze, saliency, and relation pathway, respectively. The fifth column shows the final gaze field of TPNet. The last column is the gaze prediction result, in which the red line indicates the ground truth of the gaze and the yellow line indicates our predicted gaze.

B. *Results and Analysis*

1) Quantitative Analysis: In this section, we compare TPNet with the baseline methods, and the comparison results are shown in Table I and Table II. From these two tables, we can observe the following points: 1) The performances of the Random, Center and Fixed Bias methods are generally lower than those of the other methods because these three baselines are relatively simple. 2) Judd *et al.* [26] and SalGAN [28] gain relatively good results compared to the Random, Center and Fixed Bias methods, because they utilize scene saliency information as the cues to predict the gaze, which illustrates that scene saliency is helpful in the gaze estimation task. 3) The performances of Recasens *et al.* [13], Chong *et al.* [14] and Lian *et al.* [3] are further improved compared to those of Judd et al. [26] and SalGAN [28], because they consider not only scene saliency but also the head pose to estimate gaze. 4) Our proposed TPNet outperforms all baselines significantly under all the evaluation metrics, because in addition to the scene saliency and head information, TPNet introduces human-centric relationships and micropyramid modules in the gaze estimation, which fuse more useful information for the gaze estimation. 5) Comparing the results obtained on the DLGaze dataset and GazeFollow dataset, we can see that although the results of the Random, Center and Fixed Bias methods are basically the same, the results of most of other baselines on the DLGaze dataset are generally better than the results on the GazeFollow dataset, which may be because the DLGaze dataset being smaller and more strictly labeled than the GazeFollow dataset. 6) Compared to the results of **One Human** [13] (where humans predicted the person's gaze point in the images from the testing dataset, which can be regarded as the upper bound of the machine learning method) on the GazeFollow dataset, there is still a gap between the performance of the machine learning methods and human prediction. In the future, we will further improve the gaze estimation method.

2) Qualitative Analysis: To further demonstrate the performance of TPNet, we visualize the learned gaze field and compare it to the method proposed by Recasens *et al.* [13], since it is a state-of-the-art method in the community of gaze prediction and our TPNet is inspired by it. Fig. 6 shows the visualization results. Though our TPNet can estimate the gaze of each person in the image, we only randomly select one person and visualize the predicted gaze of him in Fig. 6 for the sake of clarity. The images in the first column of Fig. 6 display two challenges in gaze estimation: occlusion (self-occlusion) and ambiguous attention. In the first three images of the first column, there are different degrees of occlusion of the face, and some images do not even have any face information; hence, the judgment of the gaze direction is difficult. In the last three images of the first column, there may be more than one salient object in the scope of the gaze; therefore, determining the target point of the gaze is challenging.

From Fig. 6, we find that compared with Recasens *et al.* [13], TPNet achieves better performance and generalization, which illustrates that TPNet can better capture people's intentions in a scene because the high-level relationship between the target person and the surrounding persons/objects is modeled and combined with the gaze and saliency cues. Specifically, when dealing with the occlusion (self-occlusion) problem, even if no face information can be used, TPNet can better predict the gaze target through the combination of attention, saliency, and relationship. When addressing the problem of ambiguous attention, TPNet can also obtain a more reasonable selection of the gaze target.

TABLE I
COMPARISON OF TPNET WITH OTHER METHODS ON THE GAZEFOLLOW DATASET

Methods	AUC	AvgDist.	MinDist.	AvgAng.
Random	0.504	0.484	0.391	69.0°
Center	0.633	0.313	0.230	49.0°
Fixed bias	0.674	0.306	0.219	48.0°
Judd *et al.* [26]	0.711	0.337	0.250	54.0°
SalGAN [28]	0.848	0.238	0.192	36.7°
Recasens *et al.* [13]	0.878	0.190	0.113	24.0°
Chong *et al.* [14]	0.896	0.187	0.112	-
Lian *et al.* [3]	0.906	0.145	0.081	17.6°
TPNet	**0.908**	**0.136**	**0.074**	**16.5°**
One Human [13]	0.924	0.096	0.040	11.0°

TABLE II
COMPARISON OF TPNET WITH OTHER METHODS ON THE DLGAZE DATASET

Methods	AUC	AvgDist.	MinDist.	AvgAng.
Random	0.505	0.480	0.390	68.7°
Center	0.634	0.315	0.229	48.8°
Fixed bias	0.674	0.305	0.219	48.0°
Judd *et al.* [26]	0.789	0.287	0.214	46.0°
SalGAN [28]	0.861	0.225	0.181	35.5°
Recasens *et al.* [13]	0.906	0.170	0.101	21.4°
Lian *et al.* [3]	0.910	0.158	0.088	18.6°
TPNet	**0.912**	**0.152**	**0.081**	**18.0°**

3) Ablation Study Analysis: To prove and better understand the importance of each module in the proposed TPNet, we test the performances of the different modules of TPNet. In particular, since the basic component of TPNet is the framework proposed by Recasens *et al.* [13], we use Recasens *et al.* [13] as the baseline architecture and give the results of different networks constructed by gradually adding different modules to the baseline. These constructed networks are as follows: 1) **Baseline+micropyramid:** introduce micropyramid modules to the different pathways of the baseline architecture; 2) **Baseline+relation:** add the relation pathway to the baseline architecture; and 3) **Baseline+relation+micropyramid:** first bring the relation pathway into the baseline architecture, and then combine the micropyramid modules and all pathways. The results of the ablation study are listed in Table III-VI, in which "Gazepath", "Salpath", and "Rnpath" represent the gaze pathway, saliency pathway, and relation pathway, respectively. From Tables III and IV, it can be seen that the gaze estimation accuracy will decline only when the micropyramid module is added to the gaze pathway, while the accuracy is improved when the micropyramid module is added to the saliency pathway. Among them, the best setting is introducing only the micropyramid module into the saliency pathway (we adopt this method in our proposed TPNet). The reason for this features of phenomenon is that the micropyramid module can integrate the target objects of different sizes at multiple levels to obtain high-level semantic information to enhance the discriminability of the network, which is beneficial for target detection and instance segmentation. This characteristic is more consistent with the requirement of the saliency pathway. Therefore, adding the micropyramid module to the saliency pathway can largely improve the performance.

Tables V and VI show that adding the relation pathway to the network can effectively improve the performance, which proves that the proposed HCRI is very significant. On this basis, we further introduce the micropyramid module into all the pathways, and the results reveal that the best performance is gained when adding micropyramid modules to the saliency pathway and relation pathway while keeping the gaze pathway unchanged, which is the final architecture of our TPNet.

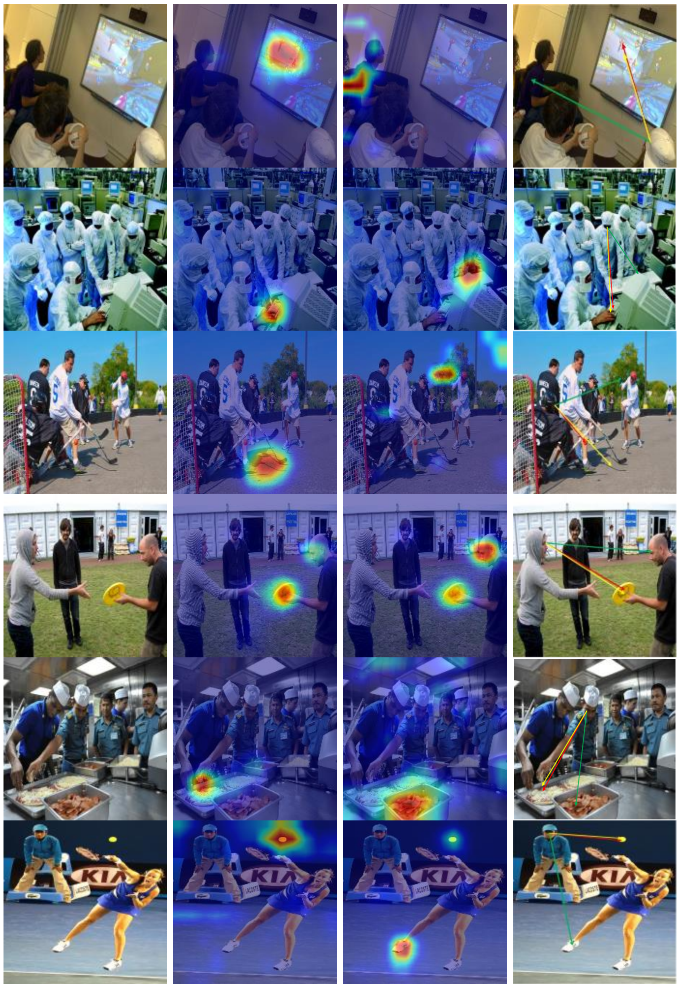


Fig. 6. Qualitative results. The images in the first column are the original images. The second column and the third column show the visualization of the gaze estimated by TPNet and Recasens *et al.* [13], respectively. The last column compares the final predicted gaze of TPNet to that of Recasens *et al.* [13], in which the red line indicates the ground truth, the yellow line indicates the predicted gaze of TPNet and the green line indicates the gaze predicted by Recasens *et al.* [13].

TABLE III
ABLATION STUDY ON THE GAZEFOLLOW DATASET

Methods	AUC	AvgDist.	MinDist.	AvgAng.
Baseline	0.878	0.190	0.113	24.0°
Baseline+micropyramid (Gazepath)	0.850	0.215	0.135	28.6°
Baseline+micropyramid (Salpath)	0.885	0.186	0.112	23.3°
Baseline+micropyramid (Salpath, Gazepath)	0.860	0.220	0.136	27.1°
TPNet	0.908	0.136	0.074	16.5°

TABLE IV
ABLATION STUDY ON THE DLGAZE DATASET

Methods	AUC	AvgDist.	MinDist.	AvgAng.
Baseline	0.906	0.170	0.101	21.4°
Baseline+micropyramid (Gazepath)	0.865	0.194	0.122	25.0°
Baseline+micropyramid (Salpath)	0.907	0.169	0.100	21.3°
Baseline+micropyramid (Salpath, Gazepath)	0.867	0.190	0.116	24.4°
TPNet	0.912	0.152	0.081	18.0°

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3071621, IEEE Transactions on Circuits and Systems for Video Technology

TABLE V
ABLATION STUDY ON THE GAZEFOLLOW DATASET

Methods	AUC	AvgDist.	MinDist.	AvgAng.
Baseline	0.878	0.190	0.113	24.0°
Baseline+relation	0.906	0.140	0.080	17.4°
Baseline+relation+micropyramid (Gazepath)	0.886	0.174	0.113	22.8°
Baseline+relation+micropyramid (Salpath)	0.907	0.138	0.075	16.9°
Baseline+relation+micropyramid (Rnpath)	0.907	0.139	0.075	16.8°
Baseline+relation+micropyramid (Salpath, Rnpath)	0.908	0.136	0.074	16.5°
Baseline+relation+micropyramid (All Pathways)	0.890	0.170	0.109	22.1°
TPNet	0.908	0.136	0.074	16.5°

TABLE VI
ABLATION STUDY ON THE DLGAZE DATASET

Methods	AUC	AvgDist.	MinDist.	AvgAng.
Baseline	0.906	0.170	0.101	21.4°
Baseline+relation	0.910	0.156	0.086	18.7°
Baseline+relation+micropyramid (Gazepath)	0.907	0.165	0.097	20.3°
Baseline+relation+micropyramid (Salpath)	0.911	0.154	0.085	18.5°
Baseline+relation+micropyramid (Rnpath)	0.911	0.153	0.083	18.2°
Baseline+relation+micropyramid (Salpath, Rnpath)	0.912	0.152	0.081	18.0°
Baseline+relation+micropyramid (All Pathways)	0.909	0.159	0.090	19.1°
TPNet	0.912	0.152	0.081	18.0°

4) Fusion Strategy Analysis: Here, we test several final fusion strategies for integrating the three pathways, e.g., addition, concatenation and multiplication. Tables VII and VIII give the performances of the different fusion strategies. From Tables VII and VIII, we can observe that the addition operation is the worst, while the results of concatenation are equivalent to the results of the multiplication effect. To select a better strategy between concatenation and multiplication, we further test the speed of convergence of these two fusion strategies on two datasets. Fig. 7 gives the convergence curves, which show the value of loss versus the number of epochs. From the convergence curves, we can find that the multiplication fusion strategy converges faster; thus, we choose it as the final fusion strategy in our TPNet.

TABLE VII
RESULTS OF DIFFERENT FUSION STRATEGIES ON THE GAZEFOLLOW DATASET

Fusion Strategy	AUC	AvgDist.	MinDist.	AvgAng.
Addition	0.904	0.144	0.080	18.6°
Concatenation	0.908	0.136	0.074	16.5°
Multiplication	0.908	0.136	0.074	16.5°

TABLE VIII
RESULTS OF DIFFERENT FUSION STRATEGIES ON THE DLGAZE DATASET

Fusion Strategy	AUC	AvgDist.	MinDist.	AvgAng.
Addition	0.909	0.158	0.090	18.7°
Concatenation	0.912	0.152	0.081	18.0°
Multiplication	0.912	0.152	0.081	18.0°

5) Comparison of the Micropyramid Module and FPN: Our proposed micropyramid module is inspired by FPN [57]. Here, we compare micropyramid and FPN to illustrate why we design micropyramid rather than directly adopt FPN in TPNet.

The main difference between the micropyramid module and FPN is that the former has fewer layers than the latter. Hence, the micropyramid module can be viewed as a lightweight version of FPN. This enabled the micropyramid module to be more flexible, plug-and-play and more suitable for the architecture of TPNet. Specifically, the micropyramid module can be inserted into the saliency pathway and relation pathway, while FPN only can be embedded into the saliency pathway to improve the performance of TPNet. FPN cannot be integrated into the relation pathway because HCRI in the relation pathway is with small changes in the size of convolutional layers, which does not meet the structure requirements of FPN.

Tables IX and X show the performances of the micropyramid module and FPN when inserting them into the saliency pathway of TPNet. From these two tables, it can be seen that the micropyramid module can achieve the same performance as FPN, but the structure of micropyramid is simpler, which is beneficial for computational efficiency. Besides, FPN cannot be added into the relation pathway due to its structure, while our micropyramid module can be integrated into the relation pathway to further improve the network performance. Adding the micropyramid module into the saliency and relation pathways simultaneously is the final configuration of our TPNet, which can obtain the best performances, as shown in the last rows of Tables IX and X.

TABLE IX
COMPARISON OF MICROPYRAMID AND FPN ON THE GAZEFOLLOW DATASET

Methods	AUC	AvgDist.	MinDist.	AvgAng.
Micropyramid (Salpath)	0.907	0.138	0.075	16.9°
FPN (Salpath)	0.907	0.138	0.075	16.9°
TPNet	0.908	0.136	0.074	16.5°

TABLE X
COMPARISON OF MICROPYRAMID AND FPN ON THE DLGAZE DATASET

Methods	AUC	AvgDist.	MinDist.	AvgAng.
Micropyramid (Salpath)	0.911	0.154	0.085	18.5°
FPN (Salpath)	0.911	0.154	0.085	18.5°
TPNet	0.912	0.152	0.081	18.0°

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3071621, IEEE Transactions on Circuits and Systems for Video Technology

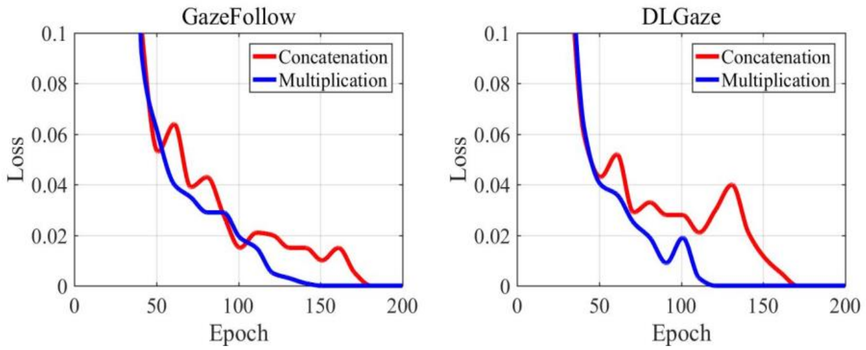


Fig. 7. Comparison of the convergences of two fusion strategies on two datasets

V. Conclusion

Following the gaze of other people is an innate ability of human beings. It is very important in understanding the behavior of other people and inferring their intentions. In this paper, we propose TPNet to estimate what a person is looking at. We first consider extracting the high-level semantic relationship between peoples/objects in a scene and then combining it with the scene saliency context and head information to predict the gaze target of a person in an image. To capture the multilevel features learned during the networking training, a micropyramid module is developed and embedded in the structure of TPNet. We test TPNet on two gaze estimation datasets. The experimental results show that TPNet can address the problem of ambiguity and occlusion in gaze estimation and achieves a state-of-the-art performance compared with that of the existing method. In addition, the success of TPNet demonstrates the importance of relation information in the gaze estimation task, which was ignored by previous research.

References

[1] J. S. Werner and M. Perlmutter, “Development of visual memory in infants,” *Advances in Child Development and Behavior*, pp. 1-56, 1979.

[2] M. L and and B. Tatler, “Looking and acting: vision and eye movements in natural behaviour,” New York, NY, USA: Oxford University Press, 2009.

[3] D. Lian, Z. Yu, and S. Gao, “Believe It or Not, We Know What You Are Looking At!” in *Asian Conference on Computer Vision*, Perth, Australia, 2018, pp. 35-50.

[4] A. Patney *et al.*, "Perceptually-based foveated virtual reality," in *ACM SIGGRAPH*, Jul. 2016, pp. 1-2.

[5] L. Fridman, B. Reimer, B. Mehler, and W. T.Freeman, “Cognitive load estimation in the wild,” in *Proc. Conf. Human Factors in Computing Systems*, Montreal, Canada, 2018, pp. 1-9.

[6] T. Guo *et al.*, “A Generalized and Robust Method Towards Practical Gaze Estimation on Smart Phone,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Seoul, South Korea, 2019, pp. 1131-1139.

[7] A. Bulling, J. A. Ward, H. Gellersen, and G. Troster, “Eye movement analysis for activity recognition using electrooculography,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 741-753, Apr. 2010.

[8] P. M. Corcoran, F. Nanu, S. Petrescu, and P. Bigioi, “Real-time eye gaze tracking for gaming design and consumer electronics systems,” *IEEE Trans Consumer Electronics*, vol. 58, no. 2, pp. 347-355, Jul. 2012.

[9] M. J. Marín-Jiménez, A. Zisserman, M. Eichner, and V.Ferrari, “Detecting people looking at each other in videos,” *Int. J. Comput. Vis.*, vol. 106, no. 3, pp. 282-296, Feb. 2014.

[10] X. Zhu, and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 2879-2886.

[11] A. Fathi, Y. Li, and J. M. Rehg, “Learning to recognize daily actions using gaze,” in *European Conference on Computer Vision*, Florence, Italy, 2012, pp. 314-327.

[12] H. S. Park, E. Jain, and Y. Sheikh, “Predicting primary gaze behavior using social saliency fields,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 3503-3510.

[13] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, “Where are they looking?” in *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 2015, pp. 199-207.

[14] E. Chong *et al.*, “Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency,” in *European Conference on Computer Vision*, Munich, Germany, 2018, pp. 383-398.

[15] D. Lian *et al.*, “RGBD based gaze estimation via multi-task CNN,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, USA, 2019, pp. 2488-2495.

[16] N. M. Arar *et al.*, “A Regression-Based User Calibration Framework for Real-Time Gaze Estimation,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2623-2638, Dec. 2017.

[17] C. Lai, S. Shih, and Y. Hung, “Hybrid Method for 3-D Gaze Tracking Using Glint and Contour Features,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 25, no. 1, pp. 24-37, Jan. 2015.

[18] D. Lian *et al.*, “Multiview multitask gaze estimation with deep convolutional neural networks,” *IEEE Trans. on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3010-3023, Oct. 2019.

[19] Y. Sugano, Y. Matsushita, and Y. Sato, “Learning-by-synthesis for appearance-based 3d gaze estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 1821-1828.

[20] E. Wood *et al.*, “Rendering of eyes for eye-shape registration and gaze estimation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 3756-3764.

[21] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “Appearance-based gaze estimation in the wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 4511-4520.

[22] P. Wei *et al.*, “Where and why are they looking? jointly inferring human attention and intentions in complex tasks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6801-6809.

[23] S. Duffner and C. Garcia, “Visual Focus of Attention Estimation With Unsupervised Incremental Learning,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 26, no. 12, pp. 2264 - 2272, Dec. 2016.

[24] N. Zhuang *et al.*, “MUGGLE: MUlti-Stream Group Gaze Learning and Estimation,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3637 - 3650, Oct. 2020.

[25] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[26] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, 2009, pp. 2106-2113.

[27] A. Borji, “Boosting bottom-up and top-down visual features for saliency estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 438-445.

[28] J. Pan *et al.*, “Salgan: Visual saliency prediction with generative adversarial networks,” arXiv preprint arXiv:1701.01081, 2017.

[29] K. Zhang and Z. Chen, “Video Saliency Prediction Based on Spatial-Temporal Two-Stream Network,” IEEE Trans. on Circuits and

Systems for Video Technology, vol. 29, no. 12, pp. 3544-3557, Dec. 2019.

[30] T. Zhao and X. Wu, “Pyramid feature attention network for saliency detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 3085-3094.

[31] Z. Bylinskii *et al.*, “Where should saliency models look next?” in *European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 809-824.

[32] T. Wang *et al.*, “Detect globally, refine locally: A novel approach to saliency detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 3127-3135.

[33] X. Chen, L. J. Li, L. Fei-Fei, and A. Gupta, “Iterative visual reasoning beyond convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7239-7248.

[34] J. Gu *et al.*, “Learning region features for object detection,” in *European Conference on Computer Vision*, Munich, Germany, 2018, pp. 381-395.

[35] Y. H. H. Tsai *et al.*, “Video relationship reasoning using gated spatio-temporal energy graph,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 10424-10433.

[36] R. Krishna, I. Chami, M. Bernstein, and L. Fei-Fei, “Referring relationships,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6867-6876.

[37] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, “Neural motifs: Scene graph parsing with global context,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 5831-5840.

[38] C. Galleguillos, A. Rabinovich, and S. Belongie, “Object categorization using co-occurrence, location and appearance,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA,2008, pp. 1-8.

[39] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr, “Graph cut based inference with co-occurrence statistics,” in *European Conference on Computer Vision*, Heraklion, Crete, Greece, 2010, pp. 239-253.

[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., Columbus, OH, USA, 2014, pp. 580-587.

[41] R. Girshick, “Fast r-cnn,” in *Proc. IEEE Int. Conf.Comput. Vis*., Santiago, Chile, 2015, pp. 1440-1448.

[42] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 2015, pp. 91-99.

[43] J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 7263-7271.

[44] H. Hu *et al.*, “Learning structured inference neural networks with label relations,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 2960-2968.

[45] F. Sung *et al.*, “Learning to compare: Relation network for few-shot learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., Salt Lake City, UT, USA, 2018, pp. 1199-1208.

[46] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, “Temporal relational reasoning in videos,” in *European Conference on Computer Vision*, Munich, Germany, 2018, pp. 803-818.

[47] H. Hu *et al.*, “Relation networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 3588-3597.

[48] Y. Liu, R. Wang, S. Shan, and X. Chen, “Structure inference net: Object detection using scene-level context and instance-level relationships,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6985-6994.

[49] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2018, pp. 5410-5419.

[50] S. Van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber, “Relational neural expectation maximization: Unsupervised discovery of objects and their interactions,” arXiv preprint arXiv:1802.10353, 2018.

[51] X. Yang, H. Zhang, and J. Cai, “Shuffle-then-assemble: Learning object-agnostic visual relationship features,” in *European Conference on Computer Vision*, Munich, Germany, 2018, pp. 36-52.

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, 2012, pp. 1097-1105.

[53] W. Liu *et al.,* “SSD: Single Shot MultiBox Detector,” in *European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 21-37.

[54] C. Sun *et al.*, “Actor-centric relation network,” in *European Conference on Computer Vision*, Munich, Germany, 2018, pp. 318-334.

[55] A. Santoro *et al*., “A simple neural network module for relational reasoning,” in *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 4967-4976.

[56] K. He and J. Sun, “Convolutional neural networks at constrained time cost,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 5353-5360.

[57] T. Y. Lin *et al.*, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2117-2125.

[58] O. Russakovsky *et al*., “Imagenet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211-252. Dec. 2015.

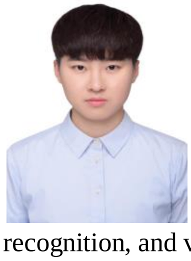
[59] B. Zhou *et al.*, “Learning deep features for scene recognition using places database,” in *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 2014, pp. 487-495.

[60] M. Everingham *et al.*, “The pascal visual object classes (voc) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303-338, Jun. 2010.

[61] T. Y. Lin *et al.*, “Microsoft coco: Common objects in context.” in *European Conference on Computer Vision*, Zurich, Switzerland, 2014, pp. 740-755.

Wenhe Chen received the B.S. degrees from the College of Computer Science and Technology, Beihua University, China, in 2015 and the M.S. degree from School of Computer Science, Northeast Electric Power University, China, in 2018. He is currently pursuing doctor's degrees in the College of Information Science and Technology, Northeast Normal University, China. His research interests include pattern recognition, machine learning, and image/video understanding.

Hui Xu received the B.S. degrees from the College of Software, Harbin University of Science and Technology, China, in 2015. She is currently pursuing doctor's degrees in the College of Information Science and Technology, Northeast Normal University, China. Her research interests include pattern recognition, machine learning, and image/video understanding.



ChaoZhu received the B.S. degrees from the College of Software, Changchun University of Science and Technology, China, in 2017. She is currently pursuing master's degrees in the College of Information Science and Technology, Northeast Normal University, China. Her research interests include pattern recognition, and video understanding.

Xiaoli Liu received the B.S. and M.S. degrees from the College of Computer Science and Information Technology, Northeast Normal University, China, in 2010 and 2013, respectively, and the Ph.D. degree from the College of Computer Application Technology, Northeastern University, China, in 2018. She is currently a post-doctor with the College of Chemical & Biomolecular Engineering, National University of Singapore. Her research interests are in machine learning, data mining, algorithm optimization, deep learning, and their applications in complex real-world learning problems, including problems in computer vision, natural language processing, medical imaging, bioinformatics, and text analysis. From 2015 to 2017, she studied as a Visiting Ph.D. Student with the University of Minnesota, Twin Cities, Minneapolis, USA.

Yinghua Lu received the B.S. degree in computer science from Jilin University, the M.S. degree from Utsunomiya University, Japan, and the Ph.D. degree in computer science from Jilin University. He is currently the President with the College of Humanities & Sciences of Northeast Normal University, Jilin, China. He is currently a Professor with the School of Information Science and Technology, Northeast Normal University. His research interests include artificial intelligence, pattern recognition, and machine learning.

Caixia Zheng received the B.S. and M.S. degrees from the College of Computer Science and Information Technology, Northeast Normal University, China, in 2009 and 2012, respectively, and the Ph.D. degree from the School of Mathematic and Statistic, Northeast Normal University in 2016. She is currently a Lecturer with the College of Information Science and Technology, Northeast Normal University. Her research interests include pattern recognition, machine learning, and image/video understanding. From 2013 to 2014, she studied as a Visiting Ph.D. Student with the University of California, Davis, CA, USA.

Jun Kong received the B.S. and M.S. degrees from the Department of Mathematics, Northeast Normal University, China, in 1992 and 1997, respectively, and the Ph.D. degree from the College of Mathematics, Jilin University, in 2001. From 2003 to 2004, he was a Visiting Scholar with Edith Cowan University, Perth, WA, Australia. He iscurrently a Professor with the College of Information Science and Technology, Northeast Normal University. His research interests include artificial intelligence, digital image processing, pattern recognition, machine learning, biometrics, and information security.