TOWARDS ADAPTING CLIP FOR GAZE OBJECT PREDICTION

Dazhi Chen, Gang Gou

State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University

ABSTRACT

This paper aims to investigate the problem of gaze object prediction in single images.We propose an application-friendly network based on CLIP for gaze location detection and gaze object detection.To avoid domain bias, we utilize a shallow feature adapter that transfers pre-trained features to target-oriented ones.Secondly, we introduce a pooling attention block to exploit the joint representation of multimodal elements, reducing gaze point deviation.Additionally, we introduce a loss that measures the prediction quality by comparing the distribution difference between the model's predictions heatmaps and the ground truth.Extensive experiments demonstrate the superior performance of our model compared to previous models.We will provide the method code at: https://github.com/fadaishaitaiyang/CCLIP.git.

Index Terms— Deep learning approaches, Human-centered computing, Applied computing

1. INTRODUCTION

Gaze following aims to accurately monitor the angle at which individuals gaze upon objects.Consequently, precise gaze following prediction algorithms have substantial potential in diverse domains, including human-object interaction [1] and human action detection [2]. In the initial stages, [3] prognosticated the gaze area by extracting head posture and gaze direction through depth models. [4] on the other hand, amalgamated saliency maps with human head posture and gaze direction, effectively predicting the observer's gaze area.

Currently, CLIP[5] leverages pre-trained knowledge of visual and textual language to exhibit impressive capabilities in diverse downstream tasks, including image classification, object detection, and semantic segmentation.

In contrast to the previously mentioned work, this study aims to investigate how the CLIP model can be used for gaze object prediction tasks without pre-training. However, this does not imply a simple integration of CLIP into the existing gaze object prediction model. The critical objective is to explore a method to leverage both visual and semantic prior information for individual images. In this paper, we propose an application-friendly network for gaze location detection and gaze object detection. Additionally, we introduce a cross-modal interaction mechanism that exploits visual cues to capture fine-grained information. To enhance the transfer of pre-trained features to target-oriented features, we address domain bias issues by introducing a feature adapter. Moreover, we propose a pooling attention block to investigate joint representations of multimodal elements, effectively reducing gaze deviation. Lastly, we propose a loss function that measures prediction quality by comparing the distribution differences between the model-generated predictions and the ground truth.

2. METHODOLOGY

Given a scene image I_s and a head position mask H, the head image I_h is usually obtained by cropping the scene image I_s . The purpose of gaze object prediction is to forecast the locations and objects that humans fixate on. Initially, we elaborate on the adaptation of CLIP for gaze object prediction. Then, we present the technical intricacies of the complete model. Figure 1 portrays an overview of our approach.

2.1. Adapting CLIP for Gaze Object Prediction

CLIP [5], which gathers 400 million image-text pairs for pre-training without human annotations, has exhibited considerable potential in acquiring transferable knowledge and comprehending open visual concepts. To effectively employ CLIP for the gaze object prediction task, we propose a methodology depicted in Figure 1. This methodology extracts image and text embeddings from the image encoder and text encoder of the CLIP model. Subsequently, we introduce a cross-modal interaction mechanism, devised using visual cues, for capturing fine-grained information. To facilitate the transfer of pre-trained features into target-oriented features, we introduce a feature adapter specifically designed for the gaze object prediction task.

Image Encoder. The pre-trained ResNet50 from CLIP is employed as the image encoder, generating an embedding vector for each input pixel. When given the scene image $I_s = R^{H\times W\times 3}$ and the head image $I_h = R^{H\times W\times 3}$, this encoder produces globality image embeddings $I_s^\star = R^{\hat{H}\times\hat{W}\times C}$ and $I_h^\star = R^{\hat{H}\times\hat{W}\times C}$, where C is set to 2048. Moreover, $\hat{H} = \frac{H}{l}$, $\hat{W} = \frac{W}{l}$ (where l represents the downsampling

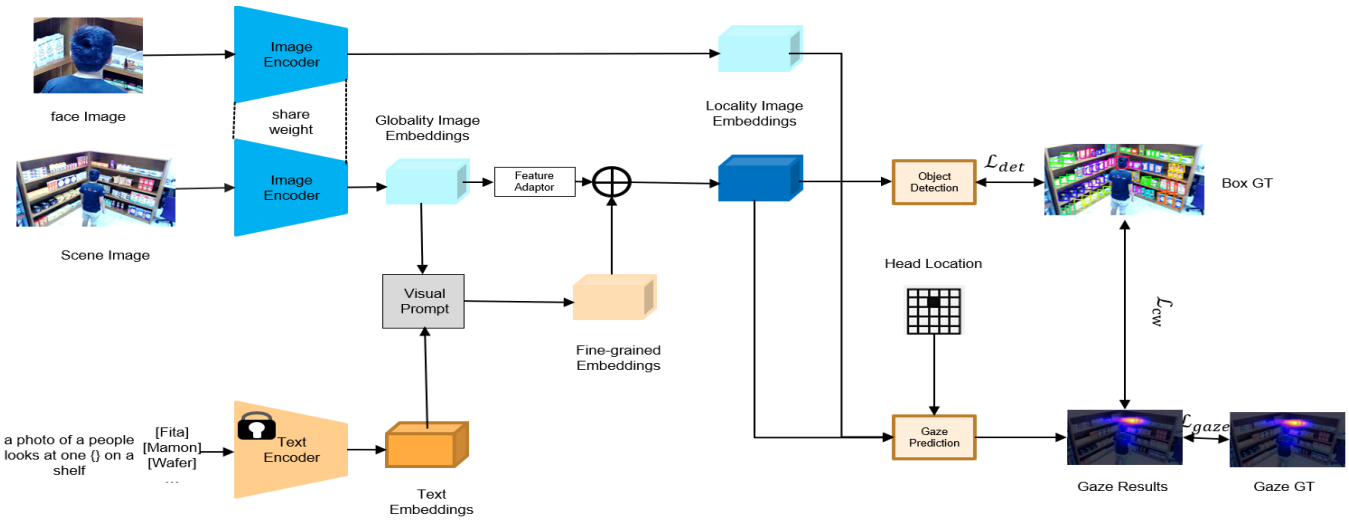

Fig. 1. Overview of the proposed method

ratio, typically set to 32). The expression is provided below:

$$\begin{aligned} I_s^* &= ImageEncoder\left(I_s\right) \\ I_h^* &= ImageEncoder\left(I_h\right) \end{aligned} \tag{1}$$

Text Encoder. We utilize the frozen pre-trained text encoder from CLIP to provide prior language knowledge for the gaze object prediction task. Specifically, this text encoder takes K-class prompts as input, embedding them into the continuous vector space R^C. Consequently, we obtain text embeddings $T = \{t_1, \cdots, t_k\} \subset R^{k \times c}$, where each $t_i \subset R^c$. Setting K to 24 aligns with the 24 categories in the gaze object prediction task. In contrast to the original model's usage of templates like "a photo of a [CLS]," we predefine the language prompt as "a photo of a person looking at one [CLS] on a shelf". The output for the text encoder can be defined as follows:

$$T = TextEncoder\left(Text\right) \subset R^D \tag{2}$$

Where D represents the dimension of text embeddings, which is set to 1024 in this article.

Visual Prompt. We have developed a visual prompt to capture fine-grained information from textual and visual features.To model the cross-modal interactions between visual embeddings (Q) and text embeddings (K, V), we utilized a transformer-based attention mechanism.As a result, the visual cue $\hat{I}$ is capable of learning a dual representation of textual and visual information.The formalization is as follows:

$$\hat{I} = TD\left(Q = I_s^*, K = T, V = T\right) \subset R^{\hat{H} \times \hat{W} \times C} \tag{3}$$

Where TD represents the TransformDecoder.

Feature Adaptor. Because the datasets used for gaze object prediction and pretraining the image encoder have different distributions, we utilize a feature adapter F_θ to transfer pre-trained features to the target domain.Global image embeddings I_s^* are mapped to adapted features $I_s^{:}$ by the feature adapter F_θ.

$$I_s^{:} = F_\theta\left(I_s^*\right) \subset R^{\hat{H} \times \hat{W} \times C} \tag{4}$$

We experimentally find that a multi-layer perceptron (MLP) yields good performance.Based on visual cues, the adaptive feature $I_s^{:}$ is equipped with $\hat{I}$ to generate text-aware locality image embeddings I for object detection and gaze prediction:

$$I = I_s^{:} + \hat{I} \tag{5}$$

Object Detection. Due to the presence of numerous small objects in the retail scene, employing a comprehensive feature pyramid structure may not guarantee favorable outcomes.Consequently, we opted to utilize the features extracted from the initial three stages of the ResNet50 model and subsequently merged them.Specifically, we applied upsampling to the features obtained from the subsequent stage and subsequently concatenated them with the existing features.The fusion of these features from the three stages yielded a feature map equivalent in size to the one produced by the initial stage.This methodology facilitates the detection of small objects.

2.2. Gaze prediction

Figure 2 illustrates that the gaze prediction network utilizes two gaze-specific features, denoted as I and I_h^*, along with the head position map H, as input for predicting gaze outcomes.

In traditional gaze following tasks, the head position is conventionally integrated with the scene image prior to extracting comprehensive scene features. In contrast, [6] enriches this approach by incorporating depth images to acquire more robust semantic scene features. However, the application of this method to gaze object prediction tasks may result

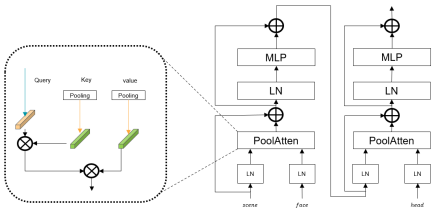

Fig. 2. Detailed network architecture of our pooled attention block

in deceptive object detection . Consequently, we employ the 'head-delay' approach, which furnishes the gaze prediction network with head position cues.

Building upon the aforementioned enhancements, we performed gaze object prediction following the approach proposed by [7].We integrated two gaze-specific features along with the raw head position map, collectively inputting them into a pooling attention module(PAM).The pooling attention module in our model comprises two primary components.The initial component emphasizes the generation of robust features by leveraging interactions with the two gaze-specific features.The second component establishes connections between the features derived from the initial part and the head position, allowing for gaze position perception within the feature map.

2.3. Loss function

Gaze object prediction tasks in research generally necessitate high-quality gaze heatmaps to produce precise outcomes.For instance, the [8] model achieves an AUC score of 84% while maintaining an Average Distance (Avg.Dist) of 0.321.While the magnitude of this error value may appear negligible, such inaccuracies can manifest as conspicuous deviations from the gaze point, consequently yielding incorrect gaze object predictions.A gaze heatmap of superior quality can accurately capture the point of gaze fixation.As a solution to the aforementioned problem, we introduce a loss function using 1-order Wasserstein distance.The Wasserstein distance quantifies the quality of predictions by comparing the distribution differences between the predicted heatmap and the ground truth.

In particular, a matrix G of the same dimensions as the predicted heatmap , with all elements set to zero.Subsequently, a mask is created by assigning the value of 1 to the corresponding region in matrix G, which corresponds to the coordinates of the ground truth bounding box $b = (x_1, y_1, x_2, y_2)$ within the predicted heatmap, indicating the presence of the target.The remaining regions are retained as 0, denoting the background.Here, (x_1, y_1) and (x_2, y_2) represent the coordinates of the top-left and bottom-right corners, respectively.Afterwards, we aim to find the optimal matching between matrix G and the predicted heatmap.The specific formula is expressed as follows:

$$P = argminW_1\left(dgm_{pre}, G\right) = argmin \sum_{(u,v)\subset P} \|u - v\| \tag{6}$$

where $u \subset dgm_{pre}$, $v \subset G$ and the Wasserstein distance here uses the L-norm distance between each pair of points $(u, v) \subset P$, for each possible match m between diagrams dgm_{pre} and dgm_{gt}.Once P is found, calculate the squared distance of each matched point:

$$\mathcal{L}_w = \sum_{(u,v)\subset P} \|u - v\|^2 \tag{7}$$

then $\mathcal{L}_w$ is integrated with cross entropy loss $\mathcal{L}_{ce}$ to obtain the final objective function $\mathcal{L}_{cw}$ for optimization:

$$\mathcal{L}_{cw} = \mathcal{L}_w + \mathcal{L}_{ce} \tag{8}$$

Since our detection branch is built upon the YOLOv4 detection head, we utilize the identical methodology to calculate the detection loss, $\mathcal{L}_{det}$. This approach simultaneously takes into account the detection confidence scores, class classification scores, and bounding box regression.Within the gaze prediction branch, we initially generate a ground truth heat map based on the ground truth gaze point coordinates. Subsequently, we determine the heat map loss, $\mathcal{L}_{gaze}$, through the calculation of the mean squared error (MSE) loss between the predicted and ground truth heat maps.The overall training loss comprises these three components:

$$\mathcal{L}_{total} = \mathcal{L}_{cw} + \mathcal{L}_{gaze} + \mathcal{L}_{det} \tag{9}$$

3. EXPERIMENTS

3.1. Setups

Datasets. In this paper, we employ the GOO dataset[9] to assess the efficacy of our proposed approach. The GOO dataset comprises comprehensive annotations encompassing foveated points, gaze objects, and bounding boxes for a diverse range of 24 categories. The GOO-Synth subset showcases a collection of 192,000 composite images, while the GOO-Real subset comprises 9,552 real-world images.

Implementation details. Our proposed methodology incorporates the pretrained image encoder ResNet50 of the CLIP. For the object detection branch, we employ Non-Maximum Suppression (NMS) with a threshold of 0.3 to eliminate redundant bounding boxes while retaining the first 100 boxes per image. In the gaze estimation branch, we apply a Gaussian blur with a kernel size of 3 to seamlessly fuse it with the actual box's gaze point. To optimize the network, we employ the Adam optimization algorithm, conducting a total of 100 epochs. The batch size is set to 64, and the initial learning rate is set to 10^{-4}.

Table 1. Gaze estimation performance on the GOO-Synth

Method	Venue	gaze estimation	
		AUC	Avg.Dist
Random	-	49.7	0.454
Recasens[3]	AINIPS'15	92.9	0.162
Lian [8]	ACCV'18	95.4	0.107
Chong[7]	CVPR'20	95.2	0.075
Wang[10]	CVPR'22	95.7	0.073
Tu [11]	arxiv'23	96	0.071
ours	-	**96.2**	**0.071**

Evaluation Metrics. For object detection, we use Average Prediction (AP) as our metric following the previous method. For the gaze estimation, we used two metrics as per previous work, Heatmap Area Under Curve (AUC %), and Average distance (Avg.Dist.). AUC is the confidence level of the heat map to evaluate the predicted heat map versus the real map. Avg.Dist predicts the Euclidean distance between gaze positioning and the true gaze point.

3.2. Comparison with State-of-the-Arts

We compare ours method with the current method, as shown in Table 1, Table 2, where Table 1 uses the GOO-Synth subset and Table 2 uses the GOO-Real subset. These comparisons also included the standard gaze analysis baseline, random. Random represents generating a heat map of each pixel by sampling values from a Gaussian distribution.

Figure 3 presents the exquisite visual outcomes achieved by ours approach. Consequently, this culminates in an accurate forecast of the gaze objects.

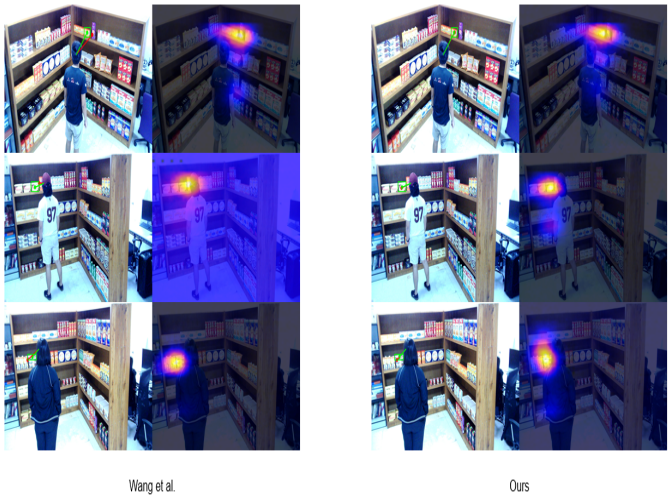

Fig. 3. Sample predicted points and heatmaps between our method and the Wang's method.

Table 2. Gaze estimation performance and object detection on the GOO-Real

Method	gaze estimation		object detection		
	AUC	Avg.Dist	AP	AP50	AP75
Recasens et al. [3]	85	0.22	-	-	-
Lian et al. [8]	84	0.321	-	-	-
Chong et al. [7]	79.6	0.252	-	-	-
Tonini et al. [6]	91.8	0.164	-	-	-
YOLOv4 [12]	-	-	43.69	84.02	43.59
Wang [10]	90.28	0.133	52.25	91.92	55.34
Tu [11]	93.14	0.1138	58.26	**97.31**	58.42
ours	**94.07**	**0.1023**	**58.38**	96.86	**64.27**

Table 3. Ablation studies on GOO-Real, we report the performance of gaze estimation.

Setups	gaze estimation	
	AUC	Avg.Dist
#a w/o Feature adaptor	91.35	0.1452
#b w/o PAM	83.24	0.1324
#c w/o visual prompt	92.17	0.1375
#d w/o loss	91.34	0.2218
ours	94.07	0.1023

3.3. Ablation Studies

The comprehensive results of our ablation experiments are presented in Table 3. Primarily, we observed a significant performance decrease when PAM is removed, confirming the essential role of PAM in feature fusion.Furthermore, the performance declines when the Feature adapter module is removed, providing additional validation of its effectiveness.Removal of the loss leads to a significant increase in Avg.Dist, providing further evidence that our proposed loss function accurately localizes the gaze point.

4. CONCLUSION

This paper explores the effective application of CLIP for gaze object prediction tasks.Firstly, a cross-modal interaction mechanism was designed using visual cues to capture fine-grained information.Simultaneously, a feature adapter was introduced to transfer pre-trained features into target-oriented features within the target domain.Secondly, a pooling attention block was proposed to explore the combined representation of multimodal elements, thus minimizing gaze point deviation.Moreover, a loss function was introduced to quantify prediction quality by comparing the distribution differences between the model-generated predictions and the ground truth labels.The experimental results demonstrate that our method outperforms the state-of-the-art approaches.

5. REFERENCES

[1] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli, "Interact as you intend: Intention-driven human-object interaction detection," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1423–1432, 2019.

[2] Nataliya Shapovalova, Michalis Raptis, Leonid Sigal, and Greg Mori, "Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[3] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba, "Where are they looking?," *Advances in neural information processing systems*, vol. 28, 2015.

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

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[6] Francesco Tonini, Cigdem Beyan, and Elisa Ricci, "Multimodal across domains gaze target detection," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 420–431.

[7] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg, "Detecting attended visual targets in video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5396–5406.

[8] Dongze Lian, Zehao Yu, and Shenghua Gao, "Believe it or not, we know what you are looking at!," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 2019, pp. 35–50.

[9] Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto, "Goo: A dataset for gaze object prediction in retail environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3125–3133.

[10] Binglu Wang, Tao Hu, Baoshan Li, Xiaojuan Chen, and Zhijie Zhang, "Gatector: A unified framework for gaze object prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19588–19597.

[11] Danyang Tu, Wei Shen, Wei Sun, Xiongkuo Min, and Guangtao Zhai, "Joint gaze-location and gaze-object detection," *arXiv preprint arXiv:2308.13857*, 2023.

[12] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.