TransGOP: Transformer-Based Gaze Object Prediction

Binglu Wang[1, 2], **Chenxi Guo**[1], **Yang Jin**[1], **Haisheng Xia**[3], **Nian Liu**[4*]

[1]Xi'an University of Architecture and Technology
[2]Beijing Institute of Technology
[3]University of Science and Technology of China
[4]Mohamed bin Zayed University of Artificial Intelligence
{wbl921129, guochenxix, jin91999}@gmail.com, hsxia@ustc.edu.cn, liunian228@gmail.com

Abstract

Gaze object prediction aims to predict the location and category of the object that is watched by a human. Previous gaze object prediction works use CNN-based object detectors to predict the object's location. However, we find that Transformer-based object detectors can predict more accurate object location for dense objects in retail scenarios. Moreover, the long-distance modeling capability of the Transformer can help to build relationships between the human head and the gaze object, which is important for the GOP task. To this end, this paper introduces Transformer into the fields of gaze object prediction and proposes an end-to-end Transformer-based gaze object prediction method named TransGOP. Specifically, TransGOP uses an off-the-shelf Transformer-based object detector to detect the location of objects and designs a Transformer-based gaze autoencoder in the gaze regressor to establish long-distance gaze relationships. Moreover, to improve gaze heatmap regression, we propose an object-to-gaze cross-attention mechanism to let the queries of the gaze autoencoder learn the global-memory position knowledge from the object detector. Finally, to make the whole framework end-to-end trained, we propose a Gaze Box loss to jointly optimize the object detector and gaze regressor by enhancing the gaze heatmap energy in the box of the gaze object. Extensive experiments on the GOO-Synth and GOO-Real datasets demonstrate that our TransGOP achieves state-of-the-art performance on all tracks, *i.e.*, object detection, gaze estimation, and gaze object prediction. Our code will be available at https://github.com/chenxiGuo/TransGOP.git.

Introduction

Predicting where a person is looking at a screen or an object has important applications in the real world, such as detecting goods of interest to people in retail scenarios, and fatigue detection is possible in autonomous driving, and in the medical field, it can help some patients with mobility or speech impairment to express their intentions (Kleinke 1986; Land and Tatler 2009; Yu et al. 2022; de Belen et al. 2023).

Previous research on gaze-related topics has primarily focused on gaze estimation (GE) task, which predicts heatmaps or points showing the location of human gaze objects. However, GE models fall short of identifying the exact location and category of the gaze object. Since the gaze object is closely linked to human behavior, it is important to identify the object due to its significant practical implications. Tomas *et al.* (Tomas et al. 2021) proposed the gaze object prediction (GOP) task, which aims to predict both the location and category of the human gaze object. To facilitate research in the GOP task, they introduced the first dataset, the GOO dataset, which consists of images of people looking at different objects in retail scenarios. Compared to GE, the GOP task is more challenging as it requires richer information for model predictions. Wang *et al.* (Wang et al. 2022a) proposed GaTector, the first model for the GOP task, which combines a CNN-based object detector (YOLOv4 (Bochkovskiy, Wang, and Liao 2020)) with a CNN-based gaze prediction branch (Chong et al. 2020). GaTector was trained and evaluated on the GOO dataset and achieved state-of-the-art performance on the GOP task.

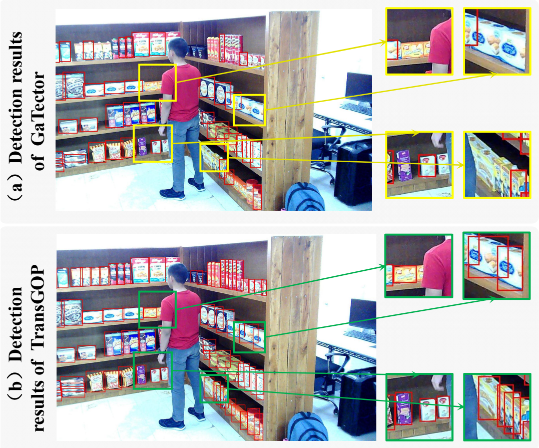
(a) Detection results of GaTector
(b) Detection results of TransGOP

Figure 1: Object Detection results of GaTector (Wang et al. 2022a) (a) and our TransGOP (b) when IoU threshold is 0.75. TransGOP predicts the object location more accurately than the GaTector, especially for the objects that are close to human or goods shelves.

The GOP task is highly related to the object detection task, as both tasks aim to accurately localize objects in

[*]Nian Liu is the corresponding author.

images. The accuracy of GOP is highly dependent on the accuracy of the object detector. CNN-based object detection has been extensively studied and achieved good performance scores (Bochkovskiy, Wang, and Liao 2020; Wang, Bochkovskiy, and Liao 2023). However, in recent years, Transformer-based object detection methods have received increasing research attention (Carion et al. 2020; Li et al. 2022b). In this paper, we found that Transformer-based object detectors perform better than CNN-based object detectors in object-dense retail scenes (see Fig. 1). Transformer-based object detectors are more effective at handling dense object scenes due to the attention mechanism, which provides them with long-range modeling capability. There is no research on introducing Transformer-based methods into the GOP task. Compared to traditional methods, we believe that the capacity to capture long-range feature dependencies of Transformer methods could establish a better attention relationship between the human and the gaze object, further improving prediction accuracy.

In this paper, we introduce TransGOP, an end-to-end method for GOP tasks based on the Transformer architecture. TransGOP comprises two branches, an object detector and a gaze regressor, as shown in Fig. 2 (a). The object detector takes the entire image as input and detects the location and category of objects using an off-the-shelf Transformer-based object detector. The gaze regressor takes the head image and scene image as input and predicts the gaze heatmap. Specifically, in the gaze regressor, we design a Transformer-based gaze autoencoder to establish long-range dependencies of gaze-related features. To improve gaze heatmap regression, in the gaze autoencoder, we propose an object-to-gaze attention mechanism that enables the gaze autoencoder queries to learn global-memory position knowledge from the object detector (see in Fig. 2 (b)).

To facilitate end-to-end training of the model, we propose a Gaze Box loss that jointly optimizes the object detector and gaze regressor. As illustrated in Fig. 2 (c), our gaze box loss can further optimize the generation of gaze heatmaps through GT gaze boxes so that they can reflect object information. Moreover, the gradient backpropagation of gaze box loss is propagated to the object detector backbone through scene features in the gaze fusion module to achieve joint optimization. Since the Transformer architecture can make the object detector end-to-end trained without any post-processing operation, our TransGOP would be the first end-to-end approach for the GOP task. Results on the GOO-Synth and GOO-Real datasets demonstrate that TransGOP outperforms existing state-of-the-art GOP methods by significant margins. To summarize, the contribution of this paper is fourfold:

- We introduce the Transformer mechanism into the gaze object prediction task and propose an end-to-end model TransGOP.
- We propose an object-to-gaze cross-attention mechanism to establish the relationship between the object detector and gaze regressor.
- We propose a gaze box loss to jointly optimize the object detector and gaze regressor.
- Extensive experiments on the GOO-Synth and GOO-Real datasets show that TransGOP outperforms the state-of-the-art GOP methods.

Related Works

Gaze Estimation

Gaze estimation has important applications in many fields (Kleinke 1986; Land and Tatler 2009), which aims to estimate where are people looking by taking eye or face images as input (Yin et al. 2022; Balim et al. 2023). Gaze point estimation (Krafka et al. 2016; He et al. 2019), gaze following (Judd et al. 2009; Leifman et al. 2017; Zhao et al. 2020; Zhu and Ji 2005), and 3D gaze estimation (Zhang et al. 2015, 2017; Cheng, Lu, and Zhang 2018; Park, Spurr, and Hilliges 2018) are sub-tasks of gaze estimation. The gaze-following task was first proposed by Recasens *et al.* (Recasens et al. 2015) who also released the dataset publicly. Chong *et al.* (Chong et al. 2020) proposed a cross-frame gaze-following model for video, which achieved a significant score metric. Recently, some GE works (Cheng and Lu 2021; Guo, Hu, and Liu 2022; Tu et al. 2022; Yu et al. 2021) introduced the transformer model. Tu *et al.* (Tu et al. 2022) and Tonini *et al.* (Tonini et al. 2023) proposed transformer-based end-to-end GE model, which aims to estimate the human gaze heatmap but can not detect the bounding boxes and the categories of the gaze objects.

Gaze Object Prediction

Different from the GE task, the GOP task predicts not only the human gaze heatmap but also the location and category of the gaze object. The GOP task is first proposed by Tomas *et al.* (Tomas et al. 2021), who also contribute a novel dataset, *i.e.*, the GOO dataset which consists of a large number of synthetic images (GOO-Synth dataset) and a smaller number of real images (GOO-Real dataset) of people gazing object in a retail environment. However, Tomas *et al.* do not propose a model to resolve the GOP problem. Afterward, Wang *et al.* (Wang et al. 2022a) propose the first unified framework GaTector for the GOP task which utilizes a CNN-based object detector (Bochkovskiy, Wang, and Liao 2020) to detect the objects and design another CNN-based gaze prediction branch (Chong et al. 2020) to predict the gaze heatmap. To further improve the performance of gaze estimation, GaTector proposed an energy aggregation loss to supervise the range of gaze heatmaps.

In this paper, we want to propose a Transformer-based GOP model due to the long-distance modeling capability of the Transformer can help to build human-object relationships, which can improve the performance of the GOP task.

Object Detection

Object detection (OD) is a fundamental task in computer vision. CNN-based object detectors can be classified into anchor-based (He et al. 2015; Girshick 2015; Law and Deng 2018; Duan et al. 2019; Bochkovskiy, Wang, and Liao 2020) and anchor-free (Huang et al. 2015; Yang et al. 2020; Zhou, Zhuo, and Krahenbuhl 2019; Tian et al. 2022; Kong et al. 2020) methods.

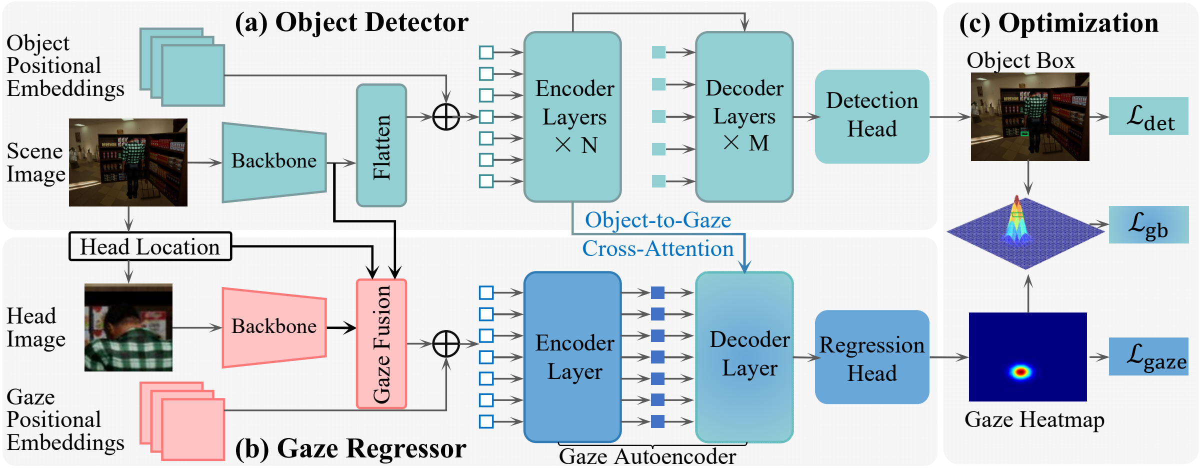

Figure 2: Overview framework of our TransGOP method. (a) The object detector in TransGOP is the of-the-shelf Transformer-based object detection method that detects object location and category. (b) The gaze regressor feeds the fused feature into the Transformer-based gaze autoencoder to predict the gaze heatmap. (c) The optimization of TransGOP consists of three parts: the object detection loss $\mathcal{L}_{\text{det}}$ for optimizing the object detector, the gaze regression loss $\mathcal{L}_{\text{gaze}}$ for optimizing the gaze regressor, and the gaze box loss $\mathcal{L}_{\text{gb}}$ to jointly optimize the object detector and gaze regressor.

Recently, Transformer has also been widely applied to resolve the object detection task (Vaswani et al. 2017; Dai et al. 2021; Fang et al. 2021; Liu et al. 2022b; Li et al. 2022b). DETR (Carion et al. 2020) is the first Transformer-based method that treats object detection as a set sequence prediction problem. Many subsequent DETR series methods (Zhu et al. 2020; Meng et al. 2021; Wang et al. 2022b; Liu et al. 2022a; Li et al. 2022a; Zheng et al. 2023) are trying to resolve the problem of DETR about slow convergence and low precision. DINO (Zhang et al. 2022) achieves better performance and faster speed of convergence than previous DETR-like models by using contrastive denoising training methods and excellent query design strategy.

In this paper, we find that Transformer-based object detectors can predict more accurately than CNN-based object detectors, especially in dense object retail scenarios, so we want to propose a Transformer-based method for the GOP task to better utilize this advantage.

Method

Given a scene image and a head image, our goal is to predict the category, bounding box, and gaze heatmap for the human gaze object. In this section, we first present the overall framework of TransGOP, then we introduce the object detector and the gaze regressor in detail. Finally, we give a detailed introduction to the proposed gaze box loss function.

Overview

As illustrated in Fig. 2, the proposed TransGOP consists of an object detector and a gaze regressor. The object detector is a Transformer-based object detection method, which takes the whole scene image as input and predicts categories and locations for all objects. The gaze regressor has a Transformer-based gaze autoencoder, which takes the head image, scene image, and head location map as input and generates queries with gaze information, which will be regressed to the gaze heatmap by the regression head. In inference, the gaze object is determined by the value of gaze heatmap energy in the predicted object boxes. The overall loss function during the training process is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \alpha\mathcal{L}_{\text{gaze}} + \beta\mathcal{L}_{\text{gb}}, \tag{1}$$

where α and β are the weights of the gaze heatmap loss and the gaze box loss, respectively.

For the Transformer-based object detector, we directly use an existing method DINO (Zhang et al. 2022). The object detector takes the whole image as input and aims to predict the object category and bounding box. The loss of the object detector is denoted by $\mathcal{L}_{\text{det}}$.

This paper proposes a novel Transformer-based gaze regressor to predict the gaze heatmap. The gaze regressor first extracts the head feature of the person in the image, then the fused feature the head feature, scene feature, and head location map as the input of the Transformer-based gaze autoencoder. The gaze heatmap is optimized by $\mathcal{L}_{\text{gaze}}$.

Finally, we propose a new gaze box loss function $\mathcal{L}_{\text{gb}}$ to jointly optimize the object detector and the gaze regressor, and make the whole framework be trained end-to-end.

Transformer-based Object Detector

As we illustrate in Fig. 2 (a), a DETR-like object detector usually consists of a backbone to extract semantic features, multiple layers of Transformer encoders, multiple layers of Transformer decoders, and a prediction head. In this paper, we use DINO (Zhang et al. 2022) as our object detector, which achieves remarkable results with a convergence speed similar to previous CNN-based methods, surpassing the accuracy of other DETR series models.

The loss of our object detector $\mathcal{L}_{\text{det}}$ consists of a classification loss and a box regression loss. Following the setting

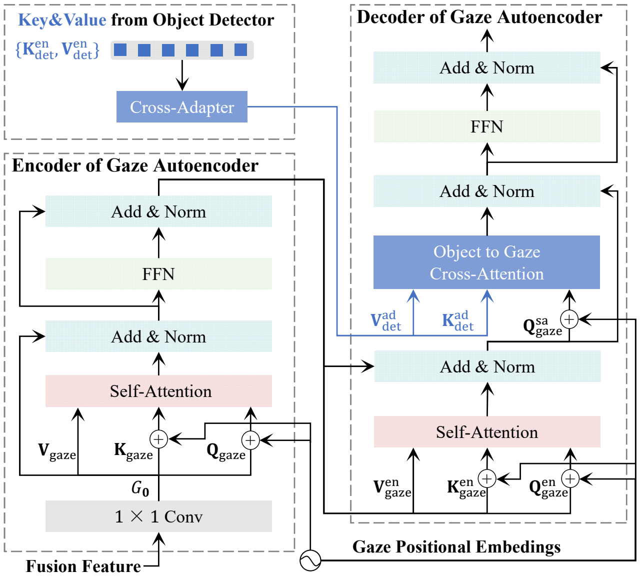

Figure 3: Details of the Transformer-based gaze autoencoder and the object-to-gaze cross-attention in the gaze regressor.

of DINO, we employ focal loss (Lin et al. 2017) as the classification loss to address the imbalanced positive and negative samples and use L1 and GIoU loss (Rezatofighi et al. 2019) to supervise the regression of the predicted box.

It is notable that the key-value pairs $\{\mathbf{K}^{\text{en}}_{\text{det}}, \mathbf{V}^{\text{en}}_{\text{det}} \in \mathbb{R}^{D\times M}\}$ from the object encoder are fed to the decoder layer of the gaze autoencoder to make the query in the gaze regressor decoder can perceive objects in the scene. Where $D = 256$ is the hidden size of tokens, and $M = 1045$ is the sum of spatial scale for multi-scale features in DINO. Moreover, our proposed *TransGOP can use any DETR-like method as the object detector, as long as it can provide keys and values from the encoder.*

Transformer-based Gaze Regressor

The gaze regressor aims to predict a gaze heatmap for the human gaze object. As illustrated in Fig. 2 (b), we use the head image, scene image, and head location map as the input of the gaze regressor. The gaze regressor consists of a gaze backbone, a gaze fusion module, a single-layer Transformer-based gaze autoencoder, and a gaze predictor.

Input. As illustrated in the left part of Fig. 2 (b), the scene image is resized into the size of $3 \times H_0 \times W_0$, where $H_0, W_0 = 224$. The head location map is a binary map with the same size as the scene image, where the value in the gaze box is 1 and 0 otherwise. The head image is obtained by cropping the scene image according to the head location.

Gaze feature fusion. In this paper, we use the gaze fusion module proposed in the GaTector (Wang et al. 2022a) to fuse features from the head image and scene image. The scene image and head image are fed into two independent Resnet50 backbones to extract the salient feature of the scene and head direction feature, respectively. The head location map is fed into five convolutional layers to generate head location features that have the same spatial resolution as image features (Wang et al. 2022a). Then, the head features and scene features are fused with the help of the head location features. Specifically, the gaze feature module first stacks the head feature with the head location map and feeds it into a linear layer to generate an attention map that encodes directional cues. By computing the inner product between the attention map and the scene feature map, a fused gaze feature $\mathbf{F}_{\text{fuse}} \in \mathbb{R}^{C\times H\times W}$ is generated, which can enhance the region that is highly related to the gaze behavior. Typical values we set are $C = 256, H, W = 15$.

Encoder. Different from GaTector which directly predicts the gaze heatmap from the fused feature, we propose a Transformer-based gaze autoencoder to build the long-range relationship for better gaze heatmap prediction results. As illustrated in Fig. 3, a 1×1 convolution operation is used to reduce the channel dimension of the fused gaze feature $\mathbf{F}_{\text{fuse}}$ from C to the hidden size D, and get a new feature map $\mathbf{G}_0 \in \mathbb{R}^{D\times H\times W}$. The spatial dimensions of $\mathbf{G}_0$ are collapsed into one dimension and result in a $D \times HW$ feature map. Then, we generate queries, keys, and values $\{\mathbf{Q}_{\text{gaze}}, \mathbf{K}_{\text{gaze}}, \mathbf{V}_{\text{gaze}} \in \mathbb{R}^{D\times HW}\}$ for the encoder, which consists of a multi-head self-attention module and a feed-forward network (FFN). Gaze positional embeddings are also added to the input of each attention layer.

Decoder and Object-to-gaze cross-attention. The decoder layer of the gaze autoencoder mainly consists of a self-attention block and an object-to-gaze cross-attention block. As shown in Fig. 3, we use the encoded queries, keys, and values $\{\mathbf{Q}^{\text{en}}_{\text{gaze}}, \mathbf{K}^{\text{en}}_{\text{gaze}}, \mathbf{V}^{\text{en}}_{\text{gaze}} \in \mathbb{R}^{D\times HW}\}$ as the input of the self-attention block, gaze positional embeddings are also added to queries and keys of each self-attention layer.

To improve gaze heatmap regression, this paper proposes an object-to-gaze cross-attention mechanism to make gaze autoencoder learn more accurate position information about the gaze object. As illustrated in Fig. 3, we first fed key-value pairs from the encoder of object detector, *i.e.*, $\{\mathbf{K}^{\text{en}}_{\text{det}}, \mathbf{V}^{\text{en}}_{\text{det}}\}$, into a cross-adapter module to optimizes the keys and values $\{\mathbf{K}^{\text{ad}}_{\text{det}}, \mathbf{V}^{\text{ad}}_{\text{det}}\}$ to provide more specific information about the gaze object. $\{\mathbf{K}^{\text{ad}}_{\text{det}}, \mathbf{V}^{\text{ad}}_{\text{det}}\}$ are used as the input of the object-to-gaze cross-attention module in the gaze decoder layer. Moreover, the output queries from the self-attention module in the decoder—$\mathbf{Q}^{\text{sa}}_{\text{gaze}}$—are used as the input queries of the cross-attention module, and gaze positional embeddings are also added to $\mathbf{Q}^{\text{sa}}_{\text{gaze}}$. We use the object-to-gaze cross-attention mechanism to make $\mathbf{Q}^{\text{sa}}_{\text{gaze}}$ learn global-memory position knowledge from the object detector, so can help to predict a more accurate gaze heatmap for the gaze object.

Gaze prediction. After the gaze autoencoder, we fed the decoded features into a regression head to predict the gaze heatmap. In this paper, we use the same regression head as the GaTector (Wang et al. 2022a) and generate ground truth gaze heatmaps $\mathbf{T} \in \mathbb{R}^{H_{\text{T}}\times W_{\text{T}}}$ by Gaussian blurred gaze points (p_x, p_y):

$$\widetilde{\mathbf{T}} = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{(x-q_x)^2}{\sigma_x^2} + \frac{(y-q_y)^2}{\sigma_y^2}\right)\right], \tag{2}$$

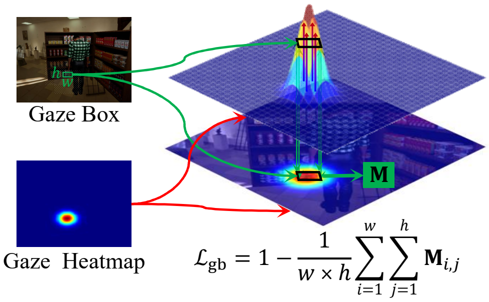
Gaze Box
h
w
M
Gaze Heatmap
$\mathcal{L}_{gb} = 1 - \frac{1}{w \times h}\sum_{i=1}^{w}\sum_{j=1}^{h}\mathbf{M}_{i,j}$

Figure 4: Illustration of the gaze box loss.

$$\mathbf{T}_{i,j} = \frac{\widetilde{\mathbf{T}}_{i,j}}{\max(\widetilde{\mathbf{T}})}, \tag{3}$$

where $\widetilde{\mathbf{T}}$ is the Gaussian-blurred heatmap and $\max(\widetilde{\mathbf{T}})$ is its maximum value. σ_x and σ_y indicate the standard deviation, which we follow Chong *et al.* (Chong et al. 2020) to set $\sigma_x = 3, \sigma_y = 3$. H_{T} and W_{T} represent the height and width of the heatmap.

Suppose the predicted gaze heatmaps is $\mathbf{M} \in \mathbb{R}^{H_{\mathrm{T}} \times W_{\mathrm{T}}}$, we use the mean square between $\mathbf{M}$ and $\mathbf{T}$ calculate the gaze heatmap loss:

$$\mathcal{L}_{\mathrm{gaze}} = \frac{1}{H_{\mathrm{T}} \times W_{\mathrm{T}}} \sum_{i=1}^{H_{\mathrm{T}}} \sum_{j=1}^{W_{\mathrm{T}}} (\mathbf{M}_{i,j} - \mathbf{T}_{i,j})^2. \tag{4}$$

Gaze Box loss

Since DETR-like object detectors can achieve end-to-end training without any extra post-processing operations, we want to make our proposed Transformer-based gaze regressor can be trained together with the DETR-like object detector to obtain an end-to-end framework for the GOP task. In this paper, we propose a gaze box loss to jointly optimize the object detector and gaze regressor.

In previous work (Wang et al. 2022a), an energy aggregation loss was proposed to optimize the gaze heatmap regression process by supervising the overall distribution of the heatmap to concentrate in the gaze box. However, we find that the actual predicted gaze heatmap area is usually much larger than the area of the gaze box, indicating that the energy aggregation loss has a very limited role in optimizing the gaze regressor. Moreover, because the gradient backpropagation of the gaze heatmap will affect the object detector backbone through the scene features in the gaze fusion, Table 6 illustrates that the energy aggregation loss could cause a decrease in object detection performance.

To this end, we want to propose a novel loss to provide a positive effect for both the object detector and gaze regressor. As shown in Fig. 4, the proposed gaze box loss aims to focus on improving the heatmap energy in the gaze box:

$$\mathcal{L}_{\mathrm{gb}} = 1 - \frac{1}{h \times w} \sum_{i=x_1}^{x_2} \sum_{j=y_1}^{y_2} \mathbf{M}_{i,j}. \tag{5}$$

where h, w are the height and width of the ground truth gaze box. The gaze box loss aims to maximize the high-energy value part of the predicted gaze heatmap distributed within the gaze box, to accurately reflect object position information. Experimental results in Table 6 demonstrate that our gaze box loss has a positive effect on both object detection and gaze heatmap regression process.

Methods	mSoC	mSoC$_{50}$	mSoC$_{75}$	mSoC$_{95}$
GaTector	67.9	98.1	86.2	0.1
TransGOP	**92.8**	**99.0**	**98.5**	**51.9**

Table 1: Gaze object prediction results on GOO-Synth.

Methods	mSoC	mSoC$_{50}$	mSoC$_{75}$	mSOC$_{95}$
No Pre-train				
GaTector	62.4	95.1	73.5	0.2
TransGOP	82.6	98.3	93.5	15.3
Pre-trained on GOO-Synth				
GaTector	71.2	97.5	88.7	0.4
TransGOP	**89.0**	**98.9**	**97.5**	**33.2**

Table 2: Gaze object prediction results on GOO-Real.

Experiments

Datasets & Settings

All experiments were conducted on GOO-Synth and GOO-Real datasets (Tomas et al. 2021). TransGOP is trained for 50 epochs, with an initial learning rate of $1e-4$, and the learning rate is reduced by 0.94 times every 5 epochs. We use AdamW as our optimizer. For the gaze autoencoder, we set the hidden size to 256 and employ 200 decoder queries. In Eq.1, the loss weight α is 1000 and β is 10. The input size of the image is set to 224×224 and the predicted heatmap size is 64×64. All experiments are implemented based on the PyTorch and one GeForce RTX 3090Ti GPU.

Metrics We use the Average Precision (AP) to evaluate the object detection performance. For gaze estimation, we adopt commonly used metrics including AUC, L2 distance error (Dist.), and angle error (Ang.).

For the gaze object prediction, we utilize mSoC[1] metric that can measure differentiation even when the predicted box and ground truth box do not overlap with each other.

Comparison to State-of-the-Art

Gaze object prediction. Table 1 and Table 2 show the performance comparison with the GaTector (Wang et al. 2022a) on the GOO-Synth and GOO-Real datasets respectively.

As shown in Table 1, TransGOP achieves better results, which is 24.9% mSoC higher than Gatector (92.8% *vs.*67.9%). The performance of GaTector drops sharply with increasingly strict mSoC constraints, while TransGOP can achieve comparable results. On the GOO-Real dataset in

[1] The mSoC metric is proposed by Wang *et al.* in the latest arXiv version.URL: https://arxiv.org/abs/2112.03549.
The code can be available at https://github.com/CodeMonsterPHD/GaTector-A-Unified-Framework-for-Gaze-Object-Prediction.git.

	Methods	AP	AP$_{50}$	AP$_{75}$	AP$_{95}$
GOO-Synth	YOLOv4	46.6	88.1	46.8	0.1
	YOLOv7	54.4	98.3	54.4	0.2
	GaTector	56.8	95.3	62.5	0.1
	Deformable DETR	81.0	98.3	92.8	14.9
	DINO	83.7	98.5	93.9	21.7
	TransGOP	**87.6**	**99.0**	**97.3**	**25.5**
GOO-Real	YOLOv4	43.7	84.0	43.6	0.1
	YOLOv7	57.3	96.4	63.2	0.3
	GaTector	52.2	91.9	55.3	0.1
	Deformable DETR	81.3	96.0	93.4	11.3
	DINO	82.8	98.7	94.6	13.5
	TransGOP	**84.1**	**98.8**	**94.7**	**20.1**

Table 3: Object detection results on GOO-Synth and GOO-Real datasets.

Methods	AUC↑	Dist.↓	Ang.↓
Random	0.497	0.454	77.0
Recasens	0.929	0.162	33.0
Lian	0.954	0.107	19.7
Chong	0.952	0.075	15.1
GaTector	0.957	**0.073**	14.9
TransGOP	**0.963**	0.079	**13.3**

Table 4: Gaze estimation results on GOO-Synth.

Methods	AUC↑	Dist.↓	Ang.↓
No Pre-train			
GaTector	0.927	0.196	39.50
TransGOP	0.947	0.097	16.73
Pre-trained on GOO-Synth			
GaTector	0.940	0.087	14.79
TransGOP	**0.957**	**0.081**	**14.71**

Table 5: Gaze estimation results on GOO-Real.

Table 2, TransGOP can achieve a comparable performance of 82.6% mSoC without pre-train. GaTector requires pre-training to achieve better performance (71.2% mSoC) on the GOO-Real dataset. This demonstrates the powerful modeling and feature extraction capabilities rooted in the Transformer, which enable our model to exhibit better GOP performance and generalization capability.

Object detection. Comparison with CNN-based object detectors YOLOv4 (Bochkovskiy, Wang, and Liao 2020), YOLOv7 (Wang, Bochkovskiy, and Liao 2023), GaTector (Wang et al. 2022a) and Transformer-based object detectors Deformable DETR (Zhu et al. 2020), DINO (Zhang et al. 2022) on the GOO-Synth dataset and GOO-Real datasets in Table 3. The global modeling capability of the Transformer makes Transformer-based methods have better performance in dense object scenes, such as the GOO-Synth dataset and GOO-Real datasets, than CNN-based methods. TransGOP also achieves good object detection performance, and although our object detector is based on DINO, through the joint optimization capability of gaze box loss, TransGOP object detection performance exceeds DINO.

Gaze estimation. Comparison with SOTA GE methods (Recasens et al. 2017; Lian, Yu, and Gao 2018; Chong et al. 2020; Wang et al. 2022a) on the GOO-Synth dataset is summarized in Table 4. Our TransGOP achieves current SOTA performance with AUC (0.963) and angular error (13.30°). This is attributed to the capability of our proposed gaze autoencoder to establish long-distance gaze relationships, while the object-to-gaze cross-attention mechanism enables it to learn global-memory position knowledge, thereby enhancing gaze estimation performance. The L2 distance error of TransGOP slightly lags behind GaTector by 0.006 (0.079 *vs.* 0.073). Because our gaze box loss makes the predicted gaze heatmap prefer to reflect the object position information, this results in a slightly worse L2 distance error calculated using gaze points. Meanwhile, Our TransGOP also outperforms GaTector on the GOO-Real dataset in Table 5.

Ablation Studies and Model Analysis

Ablation study about each component. We conducted ablation studies in Table 6 to analyze the effectiveness of our proposed **gaze autoencoder** (GA) and **gaze box loss** ($\mathcal{L}_{\text{gb}}$) on the GOO-Synth dataset. In the first row of Table 6, we build a strong baseline by combining DINO with a CNN-based gaze regressor (Chong et al. 2020) for comparison.

We first compare our $\mathcal{L}_{\text{gb}}$ to the energy aggregation loss ($\mathcal{L}_{\text{eng}}$) proposed in Gatector (Wang et al. 2022a). Based on the results, $\mathcal{L}_{\text{eng}}$ can enhance GE and GOP performance of the baseline but leads to a 2.5% decrease in object detection AP (81.2% *vs.*83.7%). However, our $\mathcal{L}_{\text{gb}}$ can jointly optimize performance gains for both GE and OD. The baseline with the gaze autoencoder (in Table 6 fourth line) achieves a 4.9% mSoC improvement in GOP performance (89.8% *vs.*84.9%), which demonstrates that the gaze autoencoder can establish long-distance gaze relationships, thus predict the more accurate gaze object. Our complete model can achieve 92.8% mSoC. From the experimental results, our proposed methods in TransGOP further improve the performance significantly.

Ablation study about cross-adapter. Table 7 reports the effectiveness of the cross-adapter in the gaze autoencoder. The results show that the cross-adapter can improve the performance of GOP (92.9% vs 90.8%), which is mainly because the cross-adapter can optimize the task conversion from the key-value pair of the object detector to the gaze autoencoder.

Comparative between different cross-attention mechanisms. In Table 8, we compare the effectiveness of different cross-attention mechanisms, *i.e.*, the gaze-to-object and object-to-gaze mechanism refers to obtaining key-value pairs from the different components. The results show that the key-value pairs from the gaze autoencoder have a limited effect on the object detector, while the global-memory position knowledge of the object detector can optimize the performance of the gaze autoencoder.

Baseline	Loss		GA	Gaze Object Prediction				Object Detection				Gaze Estimation		
	$\mathcal{L}_{\text{eng}}$	$\mathcal{L}_{\text{gb}}$		mSoC	mSoC$_{50}$	mSoC$_{75}$	mSOC$_{95}$	AP	AP$_{50}$	AP$_{75}$	AP$_{95}$	AUC↑	Dist.↓	Ang.↓
√				84.9	98.5	93.2	27.3	83.7	98.5	93.9	21.7	0.949	0.095	17.90
√	√			85.4	98.5	95.1	28.7	81.2	98.3	92.7	20.2	0.952	0.091	16.60
√		√		90.4	99.0	97.5	41.2	84.1	98.7	95.1	22.7	0.958	0.089	15.93
√			√	89.8	98.9	97.2	38.6	85.0	98.9	96.2	23.2	0.955	0.092	14.31
√		√	√	**92.8**	**99.0**	**98.5**	**51.9**	**87.6**	**99.0**	**97.3**	**25.5**	**0.963**	**0.079**	**13.30**

Table 6: Ablation comparison about each component on the GOO-Synth dataset.

Setups	Gaze Object Prediction				Object Detection				Gaze Estimation		
	mSoC	mSoC$_{50}$	mSoC$_{75}$	mSoC$_{95}$	AP	AP$_{50}$	AP$_{75}$	AP$_{95}$	AUC↑	Dist.↓	Ang.↓
w/o cross-adapter	90.8	99.0	97.8	41.8	85.8	98.0	96.5	23.8	0.960	0.084	14.20
w/ cross-adapter	**92.8**	**99.0**	**98.5**	**51.9**	**87.6**	**99.0**	**97.3**	**25.5**	**0.963**	**0.079**	**13.30**

Table 7: Ablation study about cross-adapter on the GOO-Synth dataset.

Setups	Gaze Object Prediction				Object Detection				Gaze Estimation		
	mSoC	mSoC$_{50}$	mSoC$_{75}$	mSoC$_{95}$	AP	AP$_{50}$	AP$_{75}$	AP$_{95}$	AUC↑	Dist.↓	Ang.↓
gaze-to-object	79.4	93.1	89.3	13.2	70.0	94.4	79.4	4.2	0.860	0.173	23.43
object-to-gaze	**92.8**	**99.0**	**98.5**	**51.9**	**87.6**	**99.0**	**97.3**	**25.5**	**0.963**	**0.079**	**13.30**

Table 8: Comparison of the different cross-attention mechanisms on the GOO-Synth dataset.

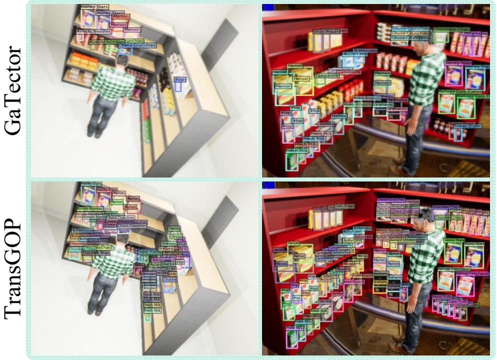

Figure 5: Object detection visualization of GaTector and TransGOP when IoU is 0.75.

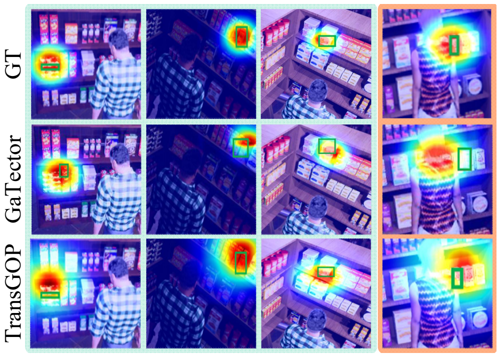

Figure 6: Gaze object prediction visualization results of GaTector and TransGOP.

Visualization

Visualization about object detection. Fig. 5 shows the qualitative results of object detection for GaTector and our proposed TransGOP when IoU is 0.75. Our TransGOP outperforms GaTector in detecting objects located at the intersection of people and goods or the edge of the shelf.

Visualization about gaze object prediction. As shown in Fig. 6, the GaTector predicts a relatively accurate heatmap, but due to its poor object detection performance, it cannot accurately locate the gaze box. In contrast, scene our TransGOP can learn precise location information from the object detector, improving the accuracy of gaze object prediction. The last column is failure cases.

Conclusion

This paper proposes an end-to-end Transformer-based gaze object prediction method named TransGOP, which consists of an object detector and a gaze regressor. The object detector is a DETR-like method to predict object location and category. Transformer-based gaze autoencoder is designed to build the long-range human gaze relationship in the gaze regressor. Meanwhile, to improve the regression of gaze heatmap, we propose an object-to-gaze cross-attention mechanism that utilizes the key-value pairs from the object detector as the input of the cross-attention block in the decoder layer of gaze autoencoder, so can help the model to predict more accurate gaze heatmap. Moreover, we also propose a novel gaze box loss that only focuses on the gaze energy in the gaze box to jointly optimize the performance of the object detector and gaze regressor. Finally, comprehensive experiments on the GOO-Synth and GOO-Real dataset demonstrates the effectiveness of our TransGOP.

References

Balim, H.; Park, S.; Wang, X.; Zhang, X.; and Hilliges, O. 2023. EFE: End-to-end Frame-to-Gaze Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2687–2696.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, 213–229. Springer.

Cheng, Y.; and Lu, F. 2021. Gaze Estimation Using Transformer. arxiv:2105.14424.

Cheng, Y.; Lu, F.; and Zhang, X. 2018. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Eur. Conf. Comput. Vis.*, 100–115.

Chong, E.; Wang, Y.; Ruiz, N.; and Rehg, J. M. 2020. Detecting attended visual targets in video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 5396–5406.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021. Up-detr: Unsupervised pre-training for object detection with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1601–1610.

de Belen, R. A.; Eapen, V.; Bednarz, T.; and Sowmya, A. 2023. Using visual attention estimation on videos for automated prediction of Autism Spectrum Disorder and symptom severity in preschool children. *medRxiv*, 2023–06.

Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; and Tian, Q. 2019. Centernet: Keypoint triplets for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 6569–6578.

Fang, Y.; Liao, B.; Wang, X.; Fang, J.; Qi, J.; Wu, R.; Niu, J.; and Liu, W. 2021. You only look at one sequence: Rethinking transformer in vision through object detection. *Adv. Neural Inform. Process. Syst.*, 34: 26183–26197.

Girshick, R. 2015. Fast r-cnn. In *Int. Conf. Comput. Vis.*, 1440–1448.

Guo, H.; Hu, Z.; and Liu, J. 2022. MGTR: End-to-End Mutual Gaze Detection with Transformer. *ACCV*.

He, J.; Pham, K.; Valliappan, N.; Xu, P.; Roberts, C.; Lagun, D.; and Navalpakkam, V. 2019. On-device few-shot personalization for real-time gaze estimation. In *Int. Conf. Comput. Vis. Worksh.*, 0–0.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9): 1904–1916.

Huang, L.; Yang, Y.; Deng, Y.; and Yu, Y. 2015. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*.

Judd, T.; Ehinger, K.; Durand, F.; and Torralba, A. 2009. Learning to predict where humans look. In *Int. Conf. Comput. Vis.*, 2106–2113. IEEE.

Kleinke, C. L. 1986. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1): 78.

Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; and Shi, J. 2020. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.*, 29: 7389–7398.

Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; and Torralba, A. 2016. Eye tracking for everyone. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2176–2184.

Land, M.; and Tatler, B. 2009. *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press.

Law, H.; and Deng, J. 2018. Cornernet: Detecting objects as paired keypoints. In *Eur. Conf. Comput. Vis.*, 734–750.

Leifman, G.; Rudoy, D.; Swedish, T.; Bayro-Corrochano, E.; and Raskar, R. 2017. Learning gaze transitions from depth to improve video saliency estimation. In *Int. Conf. Comput. Vis.*, 1698–1707.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022a. Dn-detr: Accelerate detr training by introducing query denoising. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 13619–13627.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022b. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, 280–296. Springer.

Lian, D.; Yu, Z.; and Gao, S. 2018. Believe it or not, we know what you are looking at! In *ACCV*, 35–50. Springer.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2980–2988.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022a. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022b. Swin transformer v2: Scaling up capacity and resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 12009–12019.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional detr for fast training convergence. In *Int. Conf. Comput. Vis.*, 3651–3660.

Park, S.; Spurr, A.; and Hilliges, O. 2018. Deep pictorial gaze estimation. In *Eur. Conf. Comput. Vis.*, 721–738.

Recasens, A.; Khosla, A.; Vondrick, C.; and Torralba, A. 2015. Where are they looking? *Adv. Neural Inform. Process. Syst.*, 28.

Recasens, A.; Vondrick, C.; Khosla, A.; and Torralba, A. 2017. Following gaze in video. In *Int. Conf. Comput. Vis.*, 1435–1443.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 658–666.

Tian, Z.; Chu, X.; Wang, X.; Wei, X.; and Shen, C. 2022. Fully Convolutional One-Stage 3D Object Detection on LiDAR Range Images. *arXiv preprint arXiv:2205.13764*.

Tomas, H.; Reyes, M.; Dionido, R.; Ty, M.; Mirando, J.; Casimiro, J.; Atienza, R.; and Guinto, R. 2021. Goo: A dataset for gaze object prediction in retail environments. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 3125–3133.

Tonini, F.; Dall'Asen, N.; Beyan, C.; and Ricci, E. 2023. Object-aware Gaze Target Detection. *arXiv preprint arXiv:2307.09662*.

Tu, D.; Min, X.; Duan, H.; Guo, G.; Zhai, G.; and Shen, W. 2022. End-to-End Human-Gaze-Target Detection with Transformers. arxiv:2203.10433.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30.

Wang, B.; Hu, T.; Li, B.; Chen, X.; and Zhang, Z. 2022a. GaTector: A Unified Framework for Gaze Object Prediction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 19588–19597.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7464–7475.

Wang, Y.; Zhang, X.; Yang, T.; and Sun, J. 2022b. Anchor detr: Query design for transformer-based detector. In *AAAI*, 2567–2575.

Yang, H.; Deng, R.; Lu, Y.; Zhu, Z.; Chen, Y.; Roland, J. T.; Lu, L.; Landman, B. A.; Fogo, A. B.; and Huo, Y. 2020. CircleNet: Anchor-free glomerulus detection with circle representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 35–44. Springer.

Yin, P.; Dai, J.; Wang, J.; Xie, D.; and Pu, S. 2022. NeRF-Gaze: A Head-Eye Redirection Parametric Model for Gaze Estimation. *arXiv preprint arXiv:2212.14710*.

Yu, L.; Zhou, X.; Wang, L.; and Zhang, J. 2022. Boundary-Aware Salient Object Detection in Optical Remote-Sensing Images. *Electronics*, 11(24): 4200.

Yu, Q.; Xia, Y.; Bai, Y.; Lu, Y.; Yuille, A. L.; and Shen, W. 2021. Glance-and-Gaze Vision Transformer. In *Advances in Neural Information Processing Systems*, volume 34, 12992–13003. Curran Associates, Inc.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *Int. Conf. Learn. Represent.*

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-based gaze estimation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 4511–4520.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(1): 162–175.

Zhao, H.; Lu, M.; Yao, A.; Chen, Y.; and Zhang, L. 2020. Learning to draw sight lines. *Int. J. Comput. Vis.*, 128(5): 1076–1100.

Zheng, D.; Dong, W.; Hu, H.; Chen, X.; and Wang, Y. 2023. Less is More: Focus Attention for Efficient DETR. *arXiv preprint arXiv:2307.12612*.

Zhou, X.; Zhuo, J.; and Krahenbuhl, P. 2019. Bottom-up object detection by grouping extreme and center points. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 850–859.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zhu, Z.; and Ji, Q. 2005. Eye gaze tracking under natural head movements. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 1, 918–923. IEEE.