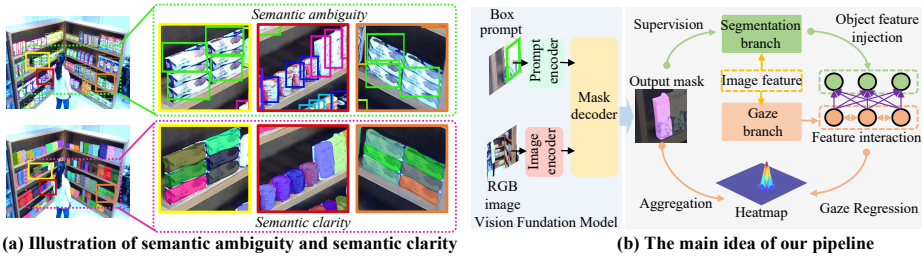
# Boosting Gaze Object Prediction via Pixel-level Supervision from Vision Foundation Model

Yang Jin[1], Lei Zhang[2], Shi Yan[2], Bin Fan[3], and Binglu Wang[1]†

[1] Xi'an University of Architecture and Technology, Xi'an, China
{jin91999, wbl921129}@gmail.com † Corresponding author
[2] School of Automation, Northwestern Polytechnical University, Xi'an, China
zl_hnly@163.com, yanshi@mail.nwpu.edu.cn
[3] National Key Lab of General AI, School of Intelligence Science and Technology, Peking University, Beijing, China
binfan@pku.edu.cn

**Abstract.** Gaze object prediction (GOP) aims to predict the category and location of the object that a human is looking at. Previous methods utilized box-level supervision to identify the object that a person is looking at, but struggled with semantic ambiguity, *i.e.*, a single box may contain several items since objects are close together. The Vision foundation model (VFM) has improved in object segmentation using box prompts, which can reduce confusion by more precisely locating objects, offering advantages for fine-grained prediction of gaze objects. This paper presents a more challenging gaze object segmentation (GOS) task, which involves inferring the pixel-level mask corresponding to the object captured by human gaze behavior. In particular, we propose that the pixel-level supervision provided by VFM can be integrated into gaze object prediction to mitigate semantic ambiguity. This leads to our gaze object detection and segmentation framework that enables accurate pixel-level predictions. Different from previous methods that require additional head input or ignore head features, we propose to automatically obtain head features from scene features to ensure the model's inference efficiency and flexibility in the real world. Moreover, rather than directly fuse features to predict gaze heatmap as in existing methods, which may overlook spatial location and subtle details of the object, we develop a space-to-object gaze regression method to facilitate human-object gaze interaction. Specifically, it first constructs an initial human-object spatial connection, then refines this connection by interacting with semantically clear features in the segmentation branch, ultimately predicting a gaze heatmap for precise localization. Extensive experiments on GOO-Synth and GOO-Real datasets demonstrate the effectiveness of our method. The code will be available at https://github.com/jinyang06/SamGOP.

**Keywords:** Gaze object prediction · Vision foundation model · Object segmentation · Space-to-object gaze regression




**(a) Illustration of semantic ambiguity and semantic clarity** **(b) The main idea of our pipeline**

**Fig. 1:** **(a)** Box-level supervision often fails to localize objects in dense settings precisely and leads to semantic ambiguity problems, whereas pixel-level supervision excels by providing clear semantic distinction through pixel-by-pixel predictions. **(b)** Vision foundation models can produce instance masks, thereby segmentation features can be used to improve the gaze regression branch's spatial perception, and the gaze object mask can help the heatmap focus on the gaze object.

## 1 Introduction

The objects stared at by humans contain important semantic information, which can reveal a person's behavior and state of mind. Therefore, identifying the objects that people are looking at has a wide range of applications. For example, in medical diagnosis, whether a child focuses on an object can potentially reveal whether he/she has autism or visual impairment [4,26,32,34]. In assisted driving scenarios, determining what object the driver is looking at can assess their attention and provide corresponding driving guidance [5,15,23]. In general, detecting the object gazed at by humans realizes the interaction between human sight and specific objects, fostering a richer semantic understanding of the scene.

Tomas *et al.* [35] first proposed the gaze object detection task and the benchmark datasets GOO-Synth and GOO-Real, aiming to predict the object boxes and categories stared at by humans. Recently, Wang *et al.* [40, 41] conducted an in-depth exploration of this task. Despite achieving promising results, they still suffer from the following limitations: **1)** They typically rely on box-level supervision, which is prone to suffer semantic ambiguity in dense object scenes. Especially when multiple objects are densely stacked or placed adjacent to each other, the box-level supervision algorithm may struggle to accurately separate them, as illustrated in Fig. 1(a). **2)** The GOO [35] dataset only contains pixel-level annotations of synthetic images, leading to difficulties in generalizing to fine-grained predictions in real scenes. **3)** They usually require additional head information as input, affecting the efficiency and flexibility of the model. While global modeling using the Transformer [38,39] can alleviate this problem, it lacks special attention to head features, which are crucial for gaze direction perception. **4)** They directly fuse features to regress the gaze heatmap, lacking a full understanding of the spatial location and subtle detailed information of objects.

Recently, vision foundational models (VFM) have gained significant attention due to their powerful generalization capabilities, which can achieve accurate pixel-level segmentation using a variety of prompts, *i.e.*, boxes, points, or masks.

In real scenes with small and densely packed objects, VFM can generate accurate boundaries and pixel-level masks (See Fig 3) by utilizing boxes as prompt, which reduces semantic ambiguity and thus opens opportunities for gaze object segmentation. Motivated by this, we propose a more challenging task, gaze object segmentation (GOS). This task involves inferring the pixel-level mask of objects captured by human gaze behavior, helping to further enhance the understanding of human gaze behavior. Furthermore, we establish a GOS framework specifically for this task, which integrates the pixel-level supervision provided by VFM into gaze object prediction to alleviate the aforementioned limitations of previous methods [40, 41].

In the proposed model, we introduce the unified detection and segmentation Transformer, *i.e.* MaskDINO [20], to simultaneously obtain candidate gaze object boxes, segmentation masks, and head boxes of the subject, which are indispensable for enabling the model to infer in the real world and achieve accurate detection. Instead of requiring additional head input [40, 41] or ignoring head features by global modeling [38, 39], we propose a RoI reconstruction module to automatically obtain head region features based on holistic features and head box provided by the unified Transformer. This helps capture gaze direction information during gaze regression to improve model efficiency and flexibility. Subsequently, a space-to-object gaze regression strategy is developed to enhance the perception of the spatial location of objects and the understanding of detailed information. Specifically, inspired by the ability of humans to infer others' gaze objects, a dual attention fusion module is presented to establish an initial human-object gaze spatial connection. Then, we propose an object feature interaction module to refine the gaze spatial connection through semantically clear feature interaction between gaze regression features and object mask features, thus strengthening the modeling of the human-object gaze relationship. In the end, an energy aggregation loss is used to further guide the regression of the gaze heatmap. By focusing on pixel-level details and progressively improving the gaze representation, it further reduces semantic ambiguity, leading to the generation of more precise heatmaps for gazed objects.

Our main contributions could be summarized as follows:

- We introduce the vision foundation model into gaze object prediction to provide pixel-level supervision and propose a more challenging gaze object segmentation task, which focuses on identifying the pixel-level masks of objects captured by human gaze behavior.
- We present an all-in-one end-to-end framework for gaze object detection and segmentation without relying on additional head-related inputs, capable of simultaneously addressing gaze estimation, object detection, and segmentation. To the best of our knowledge, our work is the first to tackle gaze object prediction from the image segmentation perspective.
- We propose a space-to-object gaze regression approach that enhances the interaction between object and gaze branches, incrementally improving the modeling of the human-object gaze relationship and alleviating semantic ambiguity.

## 2 Related Works

**Vision Foundation Model.** The vision foundation models (VFMs) [19, 24, 48] are typically built on large-scale datasets using self-supervised or semi-supervised methods, which have strong generalizability and zero-shot transfer capabilities. Segment Anything Model (SAM) [19] received attention as a foundation model for image segmentation. Its portability enables it to be flexibly applied to various downstream tasks. Then, the medical foundation model MedSAM [24] extends the foundation model to medical diagnosis. In this article, we choose SAM [19] to obtain high-quality masks to boost gaze estimation and gaze object prediction.
**Gaze Object Detection.** The gaze object detection is first proposed in [35], which aims to predict the box and categories of the object stared at by humans in an RGB image. Tomas *et al*. [35] released a benchmark dataset for GOP. Wang *et al*. [41] first proposed a GOP method named GaTector. However, GaTector requires additional head-related prior information, which hinders its application in the real world. Recently, the Transformer-based method [40] achieves comparable performance. Although previous methods are effective to some extent, they can only provide instance-level box outputs, leading to potential ambiguities or misjudgments in scenarios with small and densely packed objects.
**Gaze Target Detection.** Early research on gaze focused on gaze direction estimation [2, 5–7, 16, 18, 27, 28, 31, 42, 44, 46, 47, 49], which only predict the gaze vector and ignores the impact of the surrounding environment on the sight, resulting in limited application scenarios. Therefore, to capture higher-level gaze-related semantic information, researchers propose gaze target detection [1, 8–10, 12–14, 17, 21, 22, 25, 30, 34, 36, 37, 43], which involves to infer the specific gaze point. Recasens *et al*. [29] introduced the GazeFollow dataset and first proposed gaze target detection. Recently, Tu *et al*. [38, 39] achieved end-to-end gaze target detection using Transformer, with a relatively low detection efficiency. Despite advancements in gaze target estimation, previous methods struggle with gaze regression due to spatial perception ambiguity. In contrast, our space-to-object regression strategy helps the model gradually converge to a certain gaze point, simplifying learning and enhancing detailed interaction with the object.

## 3 Method

Given an RGB image, our model aims to precisely predict the pixel-level mask of the object humans are looking at. We first present the overall structure of our method, then introduce the VFM-based mask supervision generation method, and finally describe each module of the proposed model in detail.

### 3.1 Overview

As shown in Fig. 2, our method consists of four components: a feature extractor for specific feature extraction, a detection and segmentation module predicting candidate gaze objects boxes, masks, and human head location, a gaze regressor

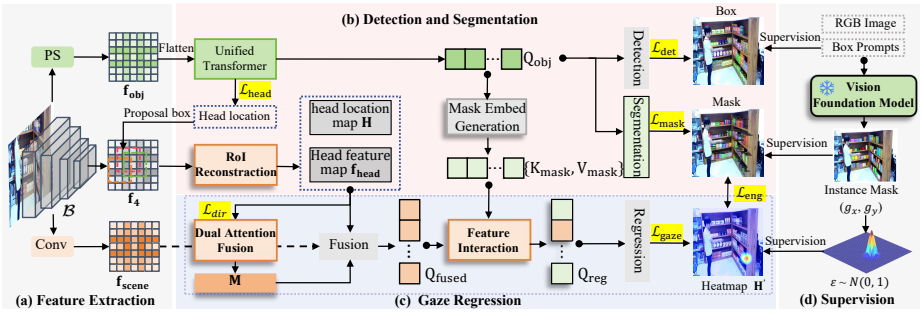



**Fig. 2: Overview of the proposed model. (a)** The feature extraction module extracts features for detection and regression branches. **(b)** The detection and segmentation branch identifies object and human head positions. **(c)** After obtaining head features, the gaze regression branch progressively refines its output: 1) employing a dual attention fusion module for initial human-object correlations; 2) leveraging a feature interaction module to incorporate semantically clear object-aware insights from the segmentation branch; 3) ultimately predicting the gaze heatmap. **(d)** Supervision signals are applied to both branches only during training.

producing gaze object heatmap with a space-to-object strategy, and a supervision information generation module creating pixel-level masks and ground truth heatmap during training. At training, the overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \alpha\mathcal{L}_{\text{dir}} + \beta\mathcal{L}_{\text{gaze}} + \gamma\mathcal{L}_{\text{eng}} \tag{1}$$

where $\alpha$, $\beta$, and $\gamma$ are the weights of the gaze direction loss, gaze heatmap loss, and energy aggregation loss respectively. Before model training, we choose SAM to generate pixel-level masks for the objects. See Sec. 3.2 for details.

**Feature extraction.** As shown in 2(a), given an RGB image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$, we first use the backbone $\mathcal{B}$ to extract multi-scale features $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4\}$. In previous work [41], pixel shuffle [33] significantly improved small object detection by enlarging feature maps for detailed feature capture. Therefore, we integrate it into dense prediction tasks, *i.e.*, image segmentation. In our model, the multi-scale features $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4\}$ are fed into the pixel shuffle module to generate object-specific features $\mathbf{f}_i^{obj} = \phi^{obj}(\mathbf{f}_i, \eta)$, where $\mathbf{f}_i$ denotes the $i$-th feature layer, $\eta$ denotes the scale factor, $\phi^{obj}(\cdot)$ denotes the pixel shuffle module, which consists of a pixel shuffle operation and a convolution layer that transforms the feature channels to be consistent with the original features of the unified Transformer [20]. Finally, the scene saliency features $\mathbf{f}_{\text{scene}} \in \mathbb{R}^{1024\times 7\times 7}$ are extracted by a scene residual block using high dimensional features $\mathbf{f}_4 \in \mathbb{R}^{2048\times 7\times 7}$.

**Unified detection and segmentation Transformer.** In this paper, we employ a unified Transformer for candidate gaze object detection, segmentation, and human head positioning, integrating MaskDINO [20] for candidate gaze object detection, segmentation, and the generation of pixel-level mask feature embeddings and a head decoder for head box regression, which facilitate accurate head-related feature extraction in head feature reconstruction module and

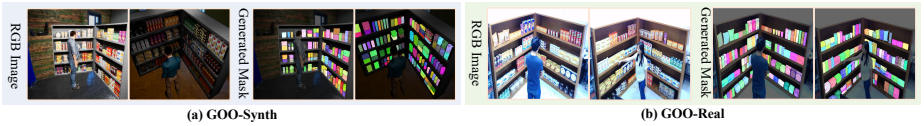



**Fig. 3:** Illustration of instance mask generated by VFM.

accurate prediction in real-world settings. The training loss is defined as:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{det}}^{\text{obj}} + \mathcal{L}_{\text{mask}}^{\text{obj}} + \mathcal{L}_{\text{det}}^{\text{head}} \tag{2}$$

$\mathcal{L}_{\text{det}}^{\text{obj}}$ and $\mathcal{L}_{\text{mask}}^{\text{obj}}$ is the object detection and segmentation loss respectively. $\mathcal{L}_{\text{det}}^{\text{head}}$ is the head decoder loss, which is consistent with the $\mathcal{L}_{\text{det}}^{\text{obj}}$. See Sec. 3.3 for details.

**Space-to-object gaze regression.** We employ a space-to-object regression strategy for gaze heatmap prediction. The module begins by representing human gaze behavior with a gaze vector. Subsequently, a dual attention fusion module establishes an initial human-object spatial connection. Finally, pixel-level object location knowledge from the mask branch is introduced to achieve semantically clear feature interaction, refining the gaze heatmap regression while alleviating semantic ambiguity. Optimization involves $\mathcal{L}_{\text{dir}}$ and $\mathcal{L}_{\text{gaze}}$ for gaze direction and gaze heatmap, respectively. The energy aggregation loss $\mathcal{L}_{\text{eng}}$ guides the gaze heatmap to focus on the gaze object mask. See Sec. 3.4 for details.

### 3.2 Instance mask generation via VFM.

Given an RGB image $\mathcal{I}$, we choose the vision foundation model SAM [19] to obtain instance masks for model training. Specifically, image $\mathcal{I}$ and prompt $\mathcal{P}_{box}$ are first fed into the image encoder $\Psi_{Image}$ and prompt encoder $\Psi_{prompt}$, respectively. Then, the mask $\mathcal{P}_{mask}$ and confidence $\mathcal{S}$ are output by the mask decoder $\Phi_{mask}$, which can be formulated as:

$$\mathcal{P}_{\text{mask}}, \mathcal{S} = \Phi_{mask}(\Psi_{Image}(\mathcal{I}), \Psi_{prompt}(\mathcal{P}_{box})) \tag{3}$$

Instance segmentation by using boxes produces accurate masks for most objects, but accurate segmentation of objects with unclear texture or edge features remains challenging. Therefore, we segment again using $\mathcal{P}_{mask}$ as a dense prompt to produce final mask $\mathcal{M}_{\text{mask}}$, which can be formulated as:

$$\mathcal{M}_{\text{mask}}, \mathcal{S} = \Phi_{mask}(\Psi_{Image}(\mathcal{I}), \Psi_{prompt}(\mathcal{P}_{mask})) \tag{4}$$

Finally, $\mathcal{M}_{mask}$ is used to train the gaze object segmentation model. The generated masks are shown in Fig. 3.

### 3.3 Unified detection and segmentation Transformer

The unified detection and segmentation Transformer aims to detect candidate gaze objects and locate human heads while extracting high-level pixel mask

features, which comprises a detection and segmentation Transformer, a head decoder, and a RoI feature reconstruction module. As shown in Fig. 2(b), the object-specific features $\{\mathbf{f}_1^{obj}, \mathbf{f}_2^{obj}, \mathbf{f}_3^{obj}, \mathbf{f}_4^{obj}\}$ are fed it to detect object boxes, categories, masks, and human head boxes.

**Object detection and segmentation** The global attention of Transformer ensures powerful long-distance modeling and global perception, achieving excellent detection performance through layer-by-layer self-attention refinement. We use MaskDINO [20] as a foundational detection and segmentation structure due to its robust representation capabilities and fast convergence speed.

**Head decoder** Our model is designed for real-world gaze object detection and segmentation without relying on head priors. We introduce a head decoder based on the unified Transformer, using a 3-layer self-attention decoder to generate proposals for head-related feature reconstruction. During training, we filter head queries with IoU $> 0.5$ as positives. During inference, boxes are selected based on confidence. Similar to other DETR-like detectors [45, 50], the head decoder uses cross-entropy, $L_1$, and GIoU loss for classification and box regression.

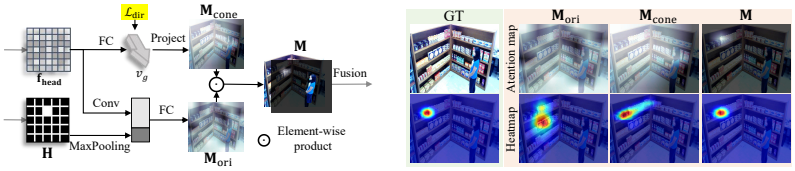
**Head feature reconstruction** Previous methods typically extract head features from manually cropped head images, limiting model inference efficiency and flexibility in the real world. To this end, we design a RoI feature reconstruction module to reconstruct head-related features based on holistic scene features and the head box from the unified Transformer. In this module, we first use the head bounding box $R = (x_1, y_1, x_2, y_2)$, which represents the upper left and lower right corner coordinates, as the region proposal to crop the regions of interest (ROIs). Then, those candidate ROIs are refined by ROIAlign [11], which maps the head region to the corresponding pixels in the holistic image features $\mathbf{f}_4 \in \mathbb{R}^{2048\times7\times7}$ and produces a fixed-size feature map $\mathbf{f}_{\text{head}} \in \mathbb{R}^{2048\times7\times7}$. For head location map generation, following previous works [9, 10, 29, 41] to generate a binary image, setting the head area to 1 and the rest to 0. Additionally, a specific residual block extracts gaze saliency features $\mathbf{f}_{\text{gaze}} \in \mathbb{R}^{1024\times7\times7}$, further enhancing the model's attention to the gaze direction.

### 3.4 Space-to-object gaze regression

We propose a space-to-object gaze regression approach to gradually refine the human-object gaze relationship. Fig. 2(c) shows our gaze regression branch with a dual attention fusion module establishing an initial human-object spatial connection and a feature interaction module strengthening this connection using semantically clear feature interaction.

**Spatial perception for gaze objects.** When humans infer what another person is staring at, they usually first observe the gaze direction of the others and search for the gazed object along this direction. Inspired by this, we propose the dual attention fusion module for the spatial perception of gaze objects, establishing an initial human-object spatial connection.

As shown in Fig. 4(a), we first use global average pooling to obtain the global features of $\mathbf{f}_{\text{head}}$, followed by fully connected layers to transform the features into




(a) The detailed architecture of the dual attention fusion module

(b) Comparison of different attention map generation methods in dual attention fusion module

**Fig. 4:** Illustration of spatial perception for gaze objects.

a 2D gaze vector $\mathbf{v}_{\mathrm{g}}$. To achieve spatial perception, we project it into a gaze cone map $\mathbf{M}_{\mathrm{cone}} \in \mathbb{R}^{1\times 7\times 7}$, activating the area around the gaze direction:

$$\mathbf{M}_{\mathrm{i,j}}^{\mathrm{cone}} = Max(0, \cos(\mathbf{v}_{\mathrm{g}}, \mathbf{v}_{\mathrm{i,j}})) \tag{5}$$

where $\mathbf{v}_{\mathrm{g}}$ and $\mathbf{v}_{\mathrm{i,j}}$ respectively denotes the predicted gaze vector and the vector composed of the eye point and each point on the gaze field. During training, $\mathbf{v}_{\mathrm{g}}$ is optimized by $\mathcal{L}_{\mathrm{dir}}$ to obtain a more accurate gaze direction and activation map. $\mathcal{L}_{\mathrm{dir}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{dir}} = 1 - \sum_{i=1}^{n} (\mathbf{v}_{\mathrm{g}}^{i} \cdot \mathbf{v}_{\mathrm{t}}^{i}) \tag{6}$$

where vector $\mathbf{v}_{\mathrm{g}}$ and $\mathbf{v}_{\mathrm{t}}$ denote the predicted gaze direction and ground truth, respectively. During backpropagation, $\mathcal{L}_{\mathrm{dir}}$ adjusts parameters of head feature $\mathbf{f}_{\mathrm{head}}$, providing more direction information for gaze regression. To boost scene attention, we adopt the method from GaTector [41] to generate an original attention map $\mathbf{M}_{\mathrm{ori}} \in \mathbb{R}^{1\times 7\times 7}$, activating scene salience area based on head properties. By multiplying $\mathbf{M}_{\mathrm{cone}}$ with $\mathbf{M}_{\mathrm{ori}}$ via element-wise dot product, we generate the final spatial perception map $\mathbf{M}$:

$$\mathbf{M} = \mathbf{M}_{\mathrm{ori}} \odot \mathbf{M}_{\mathrm{cone}} \tag{7}$$

This compels the model to search spaces more likely gazed at by humans along gaze direction, establishing an initial human-object gaze spatial relationship.

**Semantically clear feature interaction.** Previous methods [40, 41] usually directly fuse features to regress the gaze heatmap based on spatial saliency, lacking interaction with object details and facing challenges in accurately predicting heatmap for the gazed object. In our method, we introduce a feature interaction module, which leverages semantically clear mask knowledge from VFM to achieve feature interaction, refining human-object gaze relationship modeling.

As shown in Fig. 2(c), we initially extract detailed mask features from the mask branch. Subsequently, the fused feature $\mathbf{f}_{\mathrm{fuse}}$ is utilized for pixel-level attention perception. Given that candidate gaze objects can be located anywhere in the image, the Transformer's global modeling excels in accomplishing this task. Therefore, utilize a transformer layer, akin to DETR [3], to capture the global relationship between human gaze behavior and object masks. For fine-grained

object mask features, a three-layer MLP is applied to process the decoder query $\mathcal{Q}_{\text{obj}}$ of the unified Transformer, generating mask embeddings:

$$\mathcal{Q}_{\text{mask}} = \mathbf{MLP}(\mathcal{Q}_{\text{obj}}) \tag{8}$$

where the $\mathcal{Q}_{\text{mask}}$ denote the generated mask embedding. We first apply a layer self-attention for $\mathbf{f}_{\text{fuse}}$:

$$\mathbf{f}_{\text{e}} = \mathbf{FFN}(\mathbf{SelfAttn}(\mathbf{f}_{\text{fuse}})) \tag{9}$$

where $\mathbf{f}_{\text{fuse}}$, SelfAttn$(\cdot)$, and $\mathbf{f}_{\text{e}}$ denote the fused feature $\mathbf{f}_{\text{fuse}}$, self-attention operation, and encoded $\mathbf{f}_{\text{fuse}}$, respectively. Then, the $\mathcal{Q}_{\text{mask}}$ is perceived by the $\mathbf{f}_{\text{e}}$ as key-value pairs, *i.e.*, $\{\mathbf{K}_{\text{mask}}, \mathbf{V}_{\text{mask}}\}$:

$$\mathbf{f}_{\text{reg}} = \mathbf{FFN}(\mathbf{CrossAttn}(\mathbf{f}_{\text{e}}, \{\mathbf{K}_{\text{mask}}, \mathbf{V}_{\text{mask}}\})) \tag{10}$$

where $\mathbf{f}_{\text{reg}}$ and CrossAttn$(q, kv)$ denote the regression feature and cross-attention operation, respectively. This operation refines the perception of the object by providing a pixel-level understanding of the gaze object in each token of the regression feature. It effectively alleviates semantic ambiguity and strengthens the ability of the model to represent the human-object gaze relationship.

**Ultimate gaze regression.** Finally, following previous work [41] to generate the ground truth gaze heatmap and use $L_2$ loss to supervise gaze heatmap regression. $\mathcal{L}_{\text{eng}}$ proposed by [40] guides gaze heatmap energy aggregation towards the gaze object. We use the gaze object mask for more accurate error measurements. Space-to-object gaze regression approach gradually refines the human-object gaze behavior representation, regressing a more accurate gaze heatmap.

## 4 Experiments

### 4.1 Experiments Settings

Datasets and implementation details are given in **supplementary material**.

**Metrics:** Following previous work [40,41], we use AUC (Area Under the Curve), L2 distance, and Angle error for gaze estimation evaluation, and mSoC (min Single over Closure)[4] under different thresholds for gaze object prediction evaluation. Following MaskDINO [20], AP (Average Precision) is used for instance segmentation and object detection evaluation. To evaluate the gaze object segmentation, we use masks to replace boxes when calculating mSoC (details refer to the **supplementary material**). For a fair comparison, we reported results for both real-world and non-real-world settings. The non-real-world settings involve inputting additional head images and head location maps while real-world settings only use scene images. Furthermore, to compare with GTR [39], we also adopt the evaluation method proposed by [39] [5]. This method computes gaze estimation metrics only for results with L2 distance $< 0.15$ and head box IoU $> 0.5$, which is an incomplete evaluation for AUC, L2 distance, and Angle error.

[4] mSoC metric proposed by Wang *et al*. URL: `https://arxiv.org/abs/2112.03549`.

[5] Evaluation method proposed by Tu *et al*., URL: `https://ieeexplore.ieee.org/document/10262016`.

**Table 1:** Gaze object detection results on GOO-Synth and GOO-Real datasets. $*$ indicates using additional head detector to achieve real-world settings.

| | Method | GOO-Synth | | | | GOO-Real No Pretrain | | | | GOO-Real Pretrain | | | | Params | FPS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mSoC | 50 | 75 | 95 | mSoC | 50 | 75 | 95 | mSoC | 50 | 75 | 95 | | |
| Non-Real | GaTector [41] | 67.94 | 98.14 | 86.25 | 0.12 | 62.40 | 95.10 | 73.50 | 0.20 | 71.20 | 97.50 | 88.80 | 2.70 | 60.78M | 14.1 |
| | TransGOP [40] | 92.80 | 99.00 | 98.50 | 51.90 | 82.60 | 97.80 | 89.40 | 6.50 | 89.00 | 98.90 | 97.50 | 33.20 | 94.03M | 13.7 |
| | Ours | 91.19 | 93.97 | 93.46 | 73.27 | 81.75 | 98.78 | 96.43 | 7.13 | 81.89 | 98.99 | 96.63 | 7.28 | 80.99M | 14.5 |
| Real | GaTector [41]$*$ | 56.19 | 88.08 | 66.18 | - | 40.87 | 82.05 | 33.54 | 0.30 | 66.40 | 96.22 | 80.49 | 0.40 | 75.73M | 9.3 |
| | TransGOP [40]$*$ | 91.60 | 99.00 | 98.10 | 48.70 | 79.70 | 97.80 | 91.40 | 11.20 | 83.20 | 98.80 | 95.30 | 14.40 | 106.37M | 11.6 |
| | Ours | 90.77 | 93.92 | 93.45 | 69.58 | 81.35 | 98.95 | 96.02 | 2.61 | 81.60 | 99.03 | 96.54 | 6.03 | 84.26M | 14.1 |

**Table 2:** Ours gaze object segmentation results on GOO-Real and GOO-Synth.

| Settings | GOO-Synth | | | | GOO-Real No Pretrain | | | | GOO-Real Pretrain | | | | Params | FPS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mSoC | 50 | 75 | 95 | mSoC | 50 | 75 | 95 | mSoC | 50 | 75 | 95 | | |
| Non-Real | 82.68 | 93.70 | 91.68 | 26.83 | 86.39 | 98.95 | 97.44 | 20.07 | 89.39 | 98.80 | 98.02 | 33.03 | 80.99M | 14.5 |
| Real | 82.73 | 93.67 | 91.79 | 26.85 | 85.94 | 98.99 | 97.29 | 16.70 | 86.96 | 98.94 | 97.75 | 22.21 | 84.26M | 14.1 |


**Fig. 5:** Comparison of TransGOP [40] and our method inference in the real world.

### 4.2 Comparison with SOTA

**Gaze object detection and segmentation.** Table 1 shows the performance of gaze object detection on GOO-Synth and GOO-Real. In real-world settings, previous methods [40, 41] significantly increase parameters and slow down the inference speed due to the additional head detector (As shown in Fig. 5 (a)) and they can only achieve box-level prediction. In contrast, although our method is slightly inferior to TransGOP [40] in detection performance, our model runs significantly more efficiently because it integrates the acquisition of head features into the overall framework (As shown in Fig. 5 (b)) and can output bounding boxes and segmentation masks simultaneously. The above results also show that our space-to-object gaze regression method can effectively model the human-object gaze relationship. Table 2 shows our gaze object segmentation performance in real-world and non-real-world scenarios. On GOO-Synth, our model achieves 82.68% and 82.73% mSoC in real-world and non-real-world settings. On GOO-Real, without pre-training, our model performs well in both scenarios (86.39% mSoC and 85.94% mSoC). After pre-training, it reaches 89.39% mSoC and 86.96% mSoC. The above experimental results show that the pixel-level mask knowledge provided by VFMs can effectively alleviate semantic ambiguity while achieving pixel-level prediction. Even under the multi-task learning framework, we still achieve an FPS of 14.1. This is mainly attributed to the automatic

**Table 3:** Instance segmentation results on GOO-Synth and GOO-Real datasets

| | Method | GOO-Synth Dataset | | | | GOO-Real Dataset | | | | Params |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AP | 50 | 75 | 95 | AP | 50 | 75 | 95 | |
| Non-Real | Ours | 79.33 | 93.51 | 91.09 | 11.72 | 82.91 | 98.85 | 95.09 | 11.22 | 80.99M |
| Real | MaskDINO [20] | 70.85 | 95.54 | 83.31 | 1.81 | 77.05 | 98.38 | 90.45 | 4.26 | 52.03M |
| | Ours | 78.56 | 96.00 | 90.56 | 8.43 | 82.86 | 99.02 | 95.57 | 10.09 | 84.26M |

**Table 4:** Object detection results on GOO-Synth and GOO-Real datasets. $*$ indicates using additional head detector to achieve real-world settings.

| | Method | GOO-Synth Dataset | | | | GOO-Real Dataset | | | | Params |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AP | 50 | 75 | 95 | AP | 50 | 75 | 95 | |
| Non-Real | GaTector [41] | 56.80 | 95.30 | 62.50 | 0.10 | 52.25 | 91.92 | 55.34 | 0.10 | 60.78M |
| | TransGOP [40] | 87.60 | 99.00 | 97.30 | 25.50 | 77.20 | 97.80 | 89.40 | 6.50 | 94.03M |
| | Ours | 88.33 | 93.88 | 93.08 | 50.70 | 77.32 | 98.65 | 93.91 | 1.93 | 80.99M |
| Real | GaTector [41]$*$ | 43.50 | 86.23 | 37.52 | - | 29.21 | 74.02 | 16.13 | - | 75.73M |
| | TransGOP [40]$*$ | 86.00 | 98.80 | 96.20 | 23.80 | 73.30 | 96.70 | 86.00 | 4.10 | 106.37M |
| | MaskDINO [20] | 72.93 | 96.26 | 88.13 | 2.04 | 74.44 | 98.72 | 91.19 | 0.19 | 52.03M |
| | Ours | 87.60 | 93.72 | 73.07 | 45.53 | 75.20 | 98.94 | 93.20 | 0.76 | 84.26M |

incorporation of head features, avoiding the need for an additional head detector and thus enhancing the efficiency of the model.

**Object detection and segmentation.** Tables 3 and 4 show our model's instance segmentation and object detection performance. For a fair comparison, we adjusted the input size of MaskDINO [20] to match ours ($224 \times 224$). In Table 3, our model outperforms MaskDINO by 8.66 AP on GOO-Real (70.85 vs. 79.51) and by 5.81 AP on GOO-Synth (77.05 vs. 82.86) in real-world scenarios, with similar advantages in non-real-world settings of our model. These results show the challenge faced by the original MaskDINO [20] in capturing detailed features in scenes with dense objects and the effectiveness of the pixel shuffle of our model in capturing detailed features. Table 4 further demonstrates the superiority of our model in object detection over previous methods and MaskDINO [20].

**Gaze estimation.** Tables 5 and 6 show the gaze estimation performance on GOO-Synth and GOO-Real, respectively. In Table 5, our method outperforms GTR [39] and TransGOP [40] by 23.9 mAP and 10.6 mAP, respectively. Even directly evaluating all prediction results in the real world, our model reduces L2 distance and angle errors compared to TransGOP [40] by 0.015 and 1.8, respectively. Applying non-real-world methods to the real world significantly degrades performance, highlighting our real-world model is more efficient. Table 6 verifies our method on GOO-Real, achieving state-of-the-art results without pre-training. Even with direct inference in a real-world setting, our method outperforms previous non-real-world methods. After pre-training, our model exhibits more pronounced performance advantages. These results show the effectiveness

**Table 5:** Gaze estimation results on GOO-Synth dataset. † indicates the evaluation method proposed by [39].

| | Method | AUC | Dist.↓ | Ang.↓ | mAP |
|---|---|---|---|---|---|
| Non-Real | GazeFollow [29] | 0.929 | 0.162 | 33.0 | - |
| | Lian [22] | 0.954 | 0.107 | 19.7 | - |
| | VideoAttention [9] | 0.952 | 0.075 | 15.1 | - |
| | GaTector [41] | 0.957 | 0.073 | 14.9 | - |
| | GTR [39] | 0.960 | 0.071 | 14.5 | - |
| | TransGOP [40] | 0.963 | 0.079 | 13.3 | - |
| | Ours | 0.947 | 0.072 | 14.8 | - |
| Real | GazeFollow [29]† | 0.832 | 0.317 | 42.6 | 0.468 |
| | Lian [22]† | 0.903 | 0.153 | 28.1 | 0.454 |
| | VideoAttention [9]† | 0.912 | 0.143 | 24.5 | 0.489 |
| | GaTector [41]† | 0.918 | 0.139 | 24.5 | 0.510 |
| | GTR [39]† | 0.962 | 0.068 | 14.2 | 0.597 |
| | TransGOP [40]† | 0.977 | 0.070 | 11.5 | 0.730 |
| | TransGOP [40] | 0.945 | 0.106 | 20.7 | - |
| | Ours† | 0.978 | 0.057 | 10.2 | 0.836 |
| | Ours | 0.938 | 0.091 | 18.9 | - |

**Table 6:** Gaze estimation results on GOO-Real dataset. † indicates the evaluation method proposed by [39].

| | | Method | AUC | Dist.↓ | Ang.↓ | mAP |
|---|---|---|---|---|---|---|
| No Pretrain | Non-Real | GazeFollow [29] | 0.850 | 0.220 | 44.4 | - |
| | | Lian [22] | 0.840 | 0.321 | 43.5 | - |
| | | VideoAttention [9] | 0.796 | 0.252 | 51.4 | - |
| | | GaTector [41] | 0.927 | 0.196 | 39.5 | - |
| | | Tonini [36] | 0.918 | 0.164 | - | - |
| | | TransGOP [40] | 0.947 | 0.097 | 16.7 | - |
| | | Ours | 0.943 | 0.078 | 12.9 | - |
| | Real | Ours† | 0.971 | 0.059 | 9.8 | 0.810 |
| | | Ours | 0.944 | 0.088 | 14.7 | - |
| Pretrain | Non-Real | GazeFollow [29] | 0.903 | 0.195 | 39.8 | - |
| | | Lian [22] | 0.890 | 0.168 | 32.6 | - |
| | | VideoAttention [9] | 0.889 | 0.150 | 29.1 | - |
| | | GaTector [41] | 0.940 | 0.087 | 14.8 | - |
| | | TransGOP [40] | 0.957 | 0.081 | 14.7 | - |
| | | Ours | 0.963 | 0.073 | 12.4 | - |
| | Real | Ours† | 0.978 | 0.051 | 8.8 | 0.824 |
| | | Ours | 0.949 | 0.082 | 13.8 | - |

**Table 7:** Ablation study about each component on GOO-Real

| PS | RoI-Rec. | DAF | FIM | EAL | Gaze Object Segmentation | | | Instance Segmentation | | | Gaze Estimation | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | mSoC | $mSoC_{50}$ | $mSoC_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AUC | Dist.↓ | Ang.↓ |
| | | | | | 79.00 | 98.48 | 92.50 | 74.24 | 98.00 | 87.62 | 0.854 | 0.304 | 54.2 |
| ✓ | | | | | 84.50 | 98.75 | 96.56 | 80.53 | 98.68 | 93.67 | 0.839 | 0.334 | 59.8 |
| ✓ | ✓ | | | | 84.63 | 98.67 | 96.72 | 81.70 | 98.83 | 94.66 | 0.910 | 0.114 | 20.5 |
| ✓ | ✓ | ✓ | | | 85.44 | 98.92 | 96.96 | 81.58 | 98.76 | 94.29 | 0.919 | 0.095 | 16.8 |
| ✓ | ✓ | ✓ | ✓ | | 85.89 | 98.80 | 97.14 | 82.45 | 98.92 | 94.82 | 0.932 | 0.089 | 14.9 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 85.94 | 98.99 | 97.20 | 82.86 | 99.02 | 95.57 | 0.944 | 0.088 | 14.7 |

of our space-to-object gaze regression method in modeling human-object gaze relationships while achieving more accurate gaze heatmap regression.

### 4.3 Ablation study and model analysis

**Ablation study about the effect of each component.** In Table 7, we conduct various ablation studies on the GOO-Real dataset. We first establish a baseline using MaskDINO [20] and a gaze regression branch similar to TransGOP [40], and then gradually add the proposed modules and demonstrate their contribution to overall model performance. **(i) Pixel shuffle (PS).** After adding the pixel shuffle module, instance segmentation performance improved by 6.29% AP (80.53% vs 74.24%) compared to baseline, indicating that large feature maps significantly enhance detailed features in dense prediction tasks. **(ii) RoI Reconstruction (RoI-Rec.).** The RoI reconstruction module reduces angle error by 39.3 and L2 distance by 0.22, demonstrating the importance of head-related features for gaze direction perception and the ability of the module to learn accurate gaze features in real-world settings. **(iii) Dual attention fusion (DAF).**

**Table 8:** Analysis about dual attention fusion module.

| Original Attention | Gaze Cone | Gaze Object Segmentation | | | Instance Segmentation | | | Gaze Estimation | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | mSoC | $mSoC_{50}$ | $mSoC_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AUC | Dist.↓ | Ang.↓ |
| ✓ | | 85.51 | 98.86 | 97.17 | 82.08 | 98.87 | 94.78 | 0.938 | 0.089 | 15.6 |
| | ✓ | 84.13 | 98.78 | 96.35 | 81.30 | 98.78 | 93.69 | 0.926 | 0.097 | 16.3 |
| ✓ | ✓ | 85.94 | 98.99 | 97.20 | 82.75 | 99.01 | 95.17 | 0.944 | 0.088 | 14.7 |

**Table 9:** Analysis about feature interaction module

| Encoder Query | Decoder Query | Mask Embeding | Gaze Object Segmentation | | | Instance Segmentation | | | Gaze Estimation | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | mSoC | $mSoC_{50}$ | $mSoC_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AUC | Dist.↓ | Ang.↓ |
| ✓ | | | 85.79 | 98.38 | 96.44 | 81.28 | 98.87 | 94.87 | 0.928 | 0.095 | 16.0 |
| | ✓ | | 85.20 | 98.21 | 97.01 | 81.65 | 98.86 | 94.65 | 0.949 | 0.093 | 15.1 |
| | | ✓ | 85.94 | 98.99 | 97.20 | 82.86 | 99.02 | 95.57 | 0.944 | 0.088 | 14.7 |

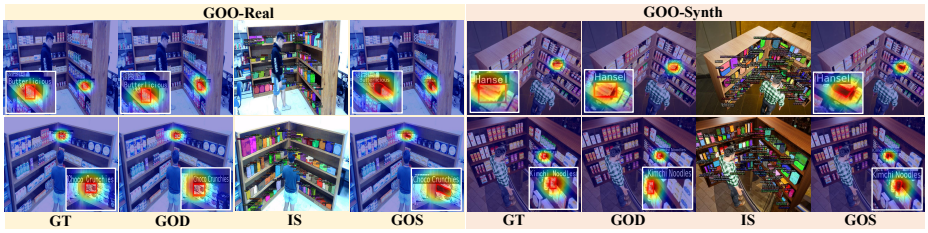
Introducing the dual attention fusion improves gaze heatmap quality (AUC: 0.919 vs 0.910), demonstrating that perceiving gaze object spatial location beforehand produces more accurate heatmaps and eases regression. **(iv) Feature interaction module (FIM).** The feature interaction module improves gaze estimation, reducing angle error by 1.9 and L2 distance by 0.006. Gaze object segmentation achieves an 85.89 mSoC, showing that pixel-level object information enhances feature interaction and human-object gaze relationship modeling. **(v) Energy aggregation loss (EAL).** Adding energy aggregation loss improves the performance of all three sub-tasks, showing that the gaze object mask effectively guides the regression of the gaze heatmap towards the gaze object.

**Analysis about dual attention fusion module.** Table 8 analyzed the effect of different spatial perception methods. Compared to using original attention $\mathbf{M}_{\text{ori}}$ or gaze cone $\mathbf{M}_{\text{cone}}$ alone, our dual attention fusion performs better. This indicates that combining spatial search along the gaze direction with human head-based scene saliency is more effective. Fig. 4(b) shows that element-wise fusion of $\mathbf{M}_{\text{ori}}$ and $\mathbf{M}_{\text{cone}}$ creates an initial human-object spatial connection.

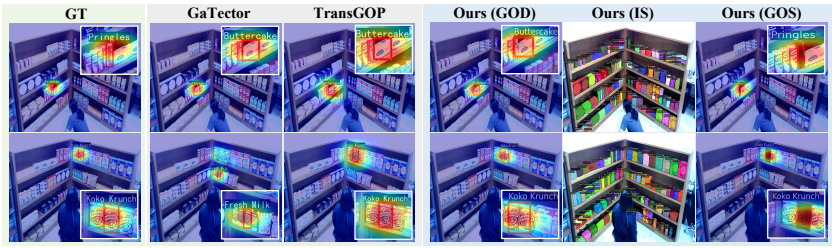
**Analysis about feature interaction module.** In Table 9, we evaluate the efficacy of different object information for gaze regression. Injecting the global encoder query of MaskDINO [20] results in 85.79% mSoC for gaze object segmentation but poor gaze estimation performance. Refining the decoder query with mask supervision reduces the angle error by 0.9, indicating better object representation. Using mask embedding with mask features during gaze regression achieves optimal performance, demonstrating that semantically clear mask features enhance gaze object positioning and the human-object gaze relationship.

### 4.4 Qualitative Results.

Fig. 6(a) presents qualitative results in the real-world setting. In object-dense scenes, our method accurately predicts the gaze object mask without relying on head-related cues. In Fig. 6(b), compared to previous methods, our method ac-




**(a)** Visualization of **our method** in **Real-World** setting.




**(b)** Visualization of **our method and previous box-level supervision methods** in **Non-Real-World** setting.

**Fig. 6:** Qualitative comparison of ground truth gaze object (GT), gaze object detection (GOD), instance segmentation (IS), gaze object segmentation (GOS)

curately locates adjacent objects using a pixel-level mask, effectively alleviating semantic ambiguity. See **supplementary material** for more visualization.

## 5 Conclusion

In this paper, we introduce the gaze object segmentation task, which aims to infer pixel-level masks for objects captured by human eyesight, and establish a framework based on the pixel-level supervision from VFM to solve this task while alleviating the semantic ambiguity existing in previous box-level supervision methods. In our method, we choose SAM to generate pixel-level supervision information for the model training. The RoI reconstruction module is proposed to reconstruct head features directly from holistic features using the head box provided by the unified Transformer, which improves the inference efficiency and flexibility of our model in the real world. To boost the representation of the human-object gaze relationship, we propose a space-to-object gaze regression strategy, which first uses the dual attention fusion module to establish a human-object spatial connection, and then a semantically clear mask feature is injected into the gaze regression to strengthen the human-object gaze relationship modeling. Extensive experiments on GOO-Real and GOO-Synth demonstrate the effectiveness of our method in various settings.

**Limitation:** The proposed method requires box supervision as prompts for the SAM when generating masks, which limits its application in a wider range of scenarios without strong location priors, more flexible approaches for generating masks via the latest VFM models need to be explored in future work.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China. (No. 82272020).
Bin Fan is sponsored by China National Postdoctoral Program for Innovative Talents (No. BX20230013).

## References

1. Bao, J., Liu, B., Yu, J.: ESCNet: Gaze target detection with the understanding of 3D scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14126–14135 (2022)
2. Cai, X., Zeng, J., Shan, S., Chen, X.: Source-free adaptive gaze estimation by uncertainty reduction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22035–22045 (2023)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. pp. 213–229 (2020)
4. Chen, W., Li, R., Yu, Q., Xu, A., Feng, Y., Wang, R., Zhao, L., Lin, Z., Yang, Y., Lin, D., et al.: Early detection of visual impairment in young children using a smartphone-based deep learning system. Nature Medicine **29**(2), 493–503 (2023)
5. Chen, Y., Nan, Z., Xiang, T.: FBLNet: Feedback loop network for driver attention prediction. In: Int. Conf. Comput. Vis. pp. 13371–13380 (2023)
6. Cheng, Y., Lu, F.: Gaze estimation using transformer. In: Int. Conf. Pattern Recog. pp. 3341–3347 (2022)
7. Cheng, Y., Lu, F., Zhang, X.: Appearance-based gaze estimation via evaluation-guided asymmetric regression. In: Eur. Conf. Comput. Vis. pp. 100–115 (2018)
8. Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., Rehg, J.M.: Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In: Eur. Conf. Comput. Vis. pp. 383–398 (2018)
9. Chong, E., Wang, Y., Ruiz, N., Rehg, J.M.: Detecting attended visual targets in video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5396–5406 (2020)
10. Gupta, A., Tafasca, S., Odobez, J.M.: A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5041–5050 (2022)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2961–2969 (2017)
12. Hu, Z., Yang, D., Cheng, S., Zhou, L., Wu, S., Liu, J.: We know where they are looking at from the RGB-D camera: Gaze following in 3D. IEEE Trans. Instrum. Meas. **71**, 1–14 (2022)
13. Hu, Z., Yang, Y., Zhai, X., Yang, D., Zhou, B., Liu, J.: GFIE: A dataset and baseline for gaze-following from 2D to 3D in indoor environments. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8907–8916 (2023)
14. Hu, Z., Zhao, K., Zhou, B., Guo, H., Wu, S., Yang, Y., Liu, J.: Gaze target estimation inspired by interactive attention. IEEE Trans. Circuit Syst. Video Technol. **32**(12), 8524–8536 (2022)
15. Huang, T., Fu, R.: Driver distraction detection based on the true driver's focus of attention. IEEE Trans. Intell. Transp. Syst. **23**(10), 19374–19386 (2022)
16. Jin, S., Wang, Z., Wang, L., Bi, N., Nguyen, T.: ReDirTrans: Latent-to-latent translation for gaze and head redirection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5547–5556 (2023)

17. Jin, T., Yu, Q., Zhu, S., Lin, Z., Ren, J., Zhou, Y., Song, W.: Depth-aware gaze-following via auxiliary networks for robotics. Eng. Appl. Artif. Intell. **113**, 104924 (2022)
18. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: Int. Conf. Comput. Vis. pp. 6912–6921 (2019)
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Int. Conf. Comput. Vis. pp. 4015–4026 (2023)
20. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3041–3050 (2023)
21. Li, Y., Shen, W., Gao, Z., Zhu, Y., Zhai, G., Guo, G.: Looking here or there? gaze following in 360-degree images. In: Int. Conf. Comput. Vis. pp. 3742–3751 (2021)
22. Lian, D., Yu, Z., Gao, S.: Believe it or not, we know what you are looking at! In: Asian Conf. Comput. Vis. pp. 35–50 (2018)
23. Lv, K., Sheng, H., Xiong, Z., Li, W., Zheng, L.: Improving driver gaze prediction with reinforced attention. IEEE Trans. Multimedia **23**, 4198–4207 (2020)
24. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
25. Miao, Q., Hoai, M., Samaras, D.: Patch-level gaze distribution prediction for gaze following. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 880–889 (2023)
26. Mundy, P., Sigman, M., Kasari, C.: A longitudinal study of joint attention and language development in autistic children. J. Autism Dev. Disord. **20**(1), 115–128 (1990)
27. Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation. In: Eur. Conf. Comput. Vis. pp. 721–738 (2018)
28. Park, S., Zhang, X., Bulling, A., Hilliges, O.: Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In: Proc. ACM Symp. Eye Track. Res. Appl. pp. 1–10 (2018)
29. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? Adv. Neural Inform. Process. Syst. **28** (2015)
30. Recasens, A., Vondrick, C., Khosla, A., Torralba, A.: Following gaze in video. In: Int. Conf. Comput. Vis. pp. 1435–1443 (2017)
31. Ruzzi, A., Shi, X., Wang, X., Li, G., De Mello, S., Chang, H.J., Zhang, X., Hilliges, O.: GazeNeRF: 3d-aware gaze redirection with neural radiance fields. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9676–9685 (2023)
32. Senju, A., Johnson, M.H.: Atypical eye contact in autism: Models, mechanisms and development. Neurosci. Biobehav. Rev. **33**(8), 1204–1214 (2009)
33. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1874–1883 (2016)
34. Tafasca, S., Gupta, A., Odobez, J.M.: ChildPlay: A new benchmark for understanding children's gaze behaviour. In: Int. Conf. Comput. Vis. pp. 20935–20946 (2023)
35. Tomas, H., Reyes, M., Dionido, R., Ty, M., Mirando, J., Casimiro, J., Atienza, R., Guinto, R.: Goo: A dataset for gaze object prediction in retail environments. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 3125–3133 (2021)
36. Tonini, F., Beyan, C., Ricci, E.: Multimodal across domains gaze target detection. In: Int. Conf. Multimodal Interact. pp. 420–431 (2022)

37. Tonini, F., Dall'Asen, N., Beyan, C., Ricci, E.: Object-aware gaze target detection. In: Int. Conf. Comput. Vis. pp. 21860–21869 (2023)
38. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: End-to-end human-gaze-target detection with transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2192–2200 (2022)
39. Tu, D., Shen, W., Sun, W., Min, X., Zhai, G., Chen, C.: Un-Gaze: A unified transformer for joint gaze-location and gaze-object detection. IEEE Trans. Circuit Syst. Video Technol. (2023)
40. Wang, B., Guo, C., Jin, Y., Xia, H., Liu, N.: Transgop: Transformer-based gaze object prediction. AAAI Conf. Artif. Intell. (2024)
41. Wang, B., Hu, T., Li, B., Chen, X., Zhang, Z.: GaTector: A unified framework for gaze object prediction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19588–19597 (2022)
42. Wang, K., Ji, Q.: Real time eye gaze tracking with 3D deformable eye-face model. In: Int. Conf. Comput. Vis. pp. 1003–1011 (2017)
43. Wang, X., Zhang, H., Wang, Z., Nie, W., Yang, Z., Ren, W., Xu, Q., Xu, X., Liu, H.: Dual regression-enhanced gaze target detection in the wild. IEEE Trans. Cybern. **54**(1), 219–229 (2024)
44. Wang, Z., Zhao, J., Lu, C., Yang, F., Huang, H., Guo, Y., et al.: Learning to detect head movement in unconstrained remote gaze estimation in the wild. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 3443–3452 (2020)
45. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. Int. Conf. Learn. Represent. (2022)
46. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4511–4520 (2015)
47. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 162–175 (2017)
48. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J.: Fast segment anything. arXiv preprint arXiv:2306.12156 (2023)
49. Zhu, W., Deng, H.: Monocular free-head 3D gaze tracking with deep learning and geometry constraints. In: Int. Conf. Comput. Vis. pp. 3143–3152 (2017)
50. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: Int. Conf. Learn. Represent. (2021)