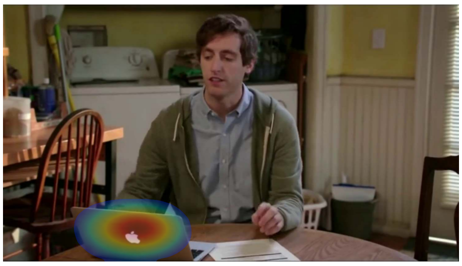
Object-aware Gaze Target Detection

Francesco Tonini[1,2] Nicola Dall'Asen[1,3] Cigdem Beyan[1] Elisa Ricci[1,2]

[1]University of Trento, Trento, Italy [2]Fondazione Bruno Kessler, Trento, Italy [3]University of Pisa, Pisa, Italy

Can we predict ***where*** and ***what*** a person is looking at?

Where: predict the image region on which the person is looking.



What: if a person is looking at an object, predict box and class of it.

✅ **Single end-to-end** method for person and gazed-object detection.

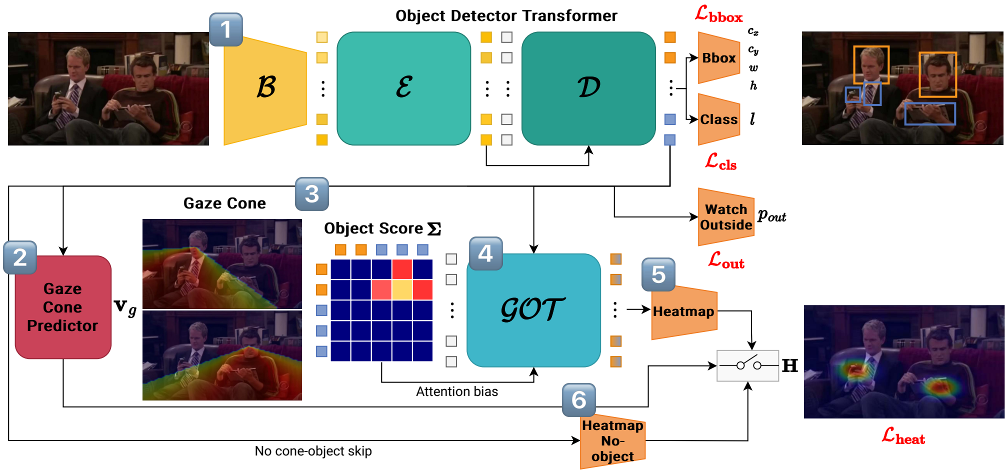
✅ Detect the gaze of **all people** in a **single forward pass.**

✅ Detect **heads** and **objects** with a single object backbone

❓ Predict **object gaze scores** for **each person's** gaze.

❓ Estimate a person's gaze in **absence** of objects.

OUR PROPOSAL - GAZE OBJECT TARGET DETECTOR



1 **Detect** and classify **objects/heads** in the image.

2 Predict the **2D/3D gaze cone** (field-of-view) for each head.

3 Calculate the **probability** that an **object is gazed by a person** based on the gaze cone scores.

4 Model the relationships for **each head-object pair.**

Learnable queries

Object features

Object score

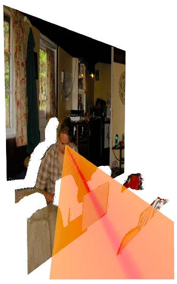
$$\text{softmax}\left(\frac{Q\;K^T+\Sigma}{\sqrt{d_k}}\right)V$$

Object features

5 **Predict** gaze **heatmap**, **object box** and **class**.

6 If **no object** is gazed, we predict a gaze heatmap from **head features only**.

2D cone

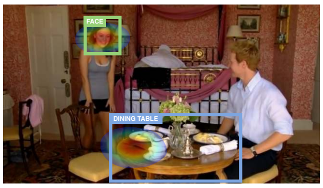


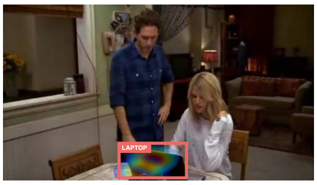
3D cone

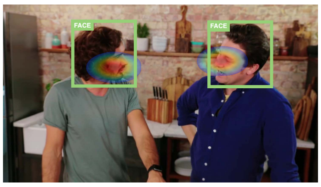
PAPER

CODE

QUALITATIVE RESULTS

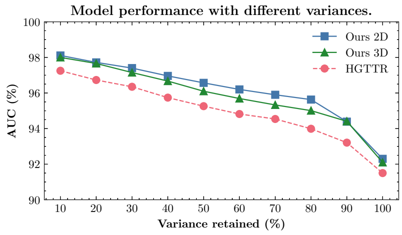


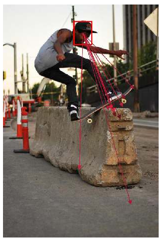




QUANTITATIVE RESULTS & THE EFFECTS OF VARIANCE IN ANNOTATIONS

Method	Modalities	Multiperson Gaze	GazeFollow			VideoAttentionTarget		
				Distance ↓		*In frame*		*Out of frame*
			AUC ↑	Avg.	Min.	AUC ↑	Dist. ↓	AP ↑
Recasens et al.	R	✗	0.804	0.233	0.124	-	-	-
Chong et al.	R + T	✗	0.902	0.142	0.082	0.812	0.146	0.849
Tonini et al.	R + D	✗	0.894	0.165	-	0.894	0.182	-
Tu et al.	R	✓	0.917	0.133	0.069	0.904	0.126	0.854
Ours	R	✓	**0.922**	0.072	0.033	0.923	**0.102**	**0.944**
Ours	R + D	✓	**0.922**	**0.069**	**0.029**	**0.933**	0.104	0.934





❗ Due to the **low consensus** across annotators, we evaluate our method under different levels of **variance** across individual gaze annotation.