

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053

## A.2 DATASET STATISTICS

The proposed PixelGaze dataset contains a total of 77.5k images with pixel-level gaze target annotations. Fig. 1 provides an overview of the dataset, including example annotations and the category distribution of gaze targets. The top part of Fig. 1 shows several examples of the pixel-level annotations in the PixelGaze dataset, demonstrating the diversity of scenes and objects. The bottom part of Fig. 1 illustrates the category distribution of gaze targets, highlighting the wide range of object categories present in the dataset.

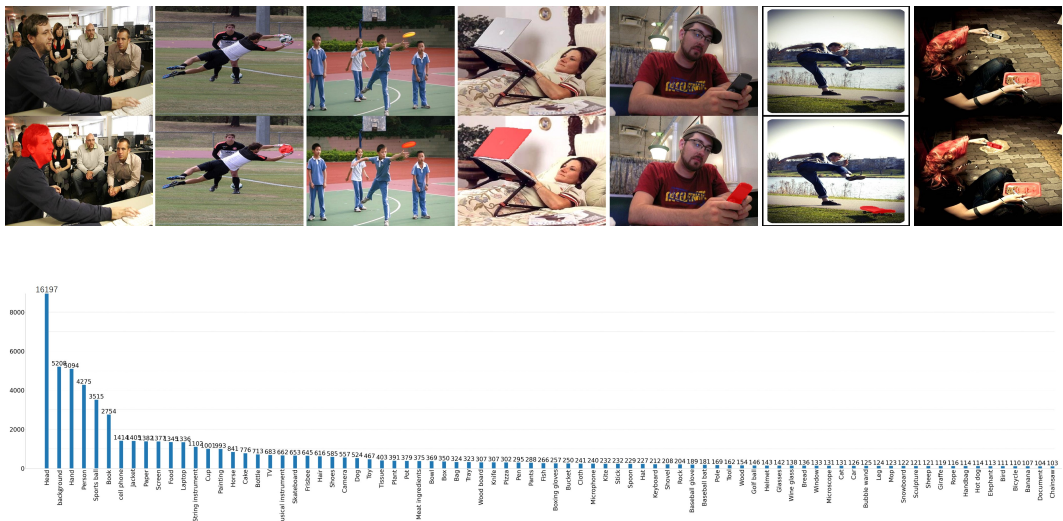


Figure 1: Overview of the PixelGaze dataset. *Top*: annotation examples, and *Bottom*: category distribution.

Table 1 presents a summary of critical features compared to the conventional gaze following dataset. The table shows that the existing dataset lacks pixel-level annotations with varied scenes and diverse objects. In contrast, we propose the GazeSeg dataset, which extends its scope to clearer semantics and accurate localization. Specifically, GazeSeg includes 77.5k images of varied natural scenes such as kitchens, sports, meetings, exhibitions, etc. Gaze targets are classified into 270 diverse common categories, including book, cellphone, person, head, and ball.

Table 1: The features and statistics of existing benchmarks and the proposed GazeSeg benchmark.

Benchmark	Type	Frames	Scenes	Class	Annotation
GazeFollow (Recasens et al., 2015)	Image	122.1k	Varied	-	Center point
VideoAttentionTaget (Chong et al., 2020)	Video	71.7k	Varied	-	Center point
ChildPlay (Tafasca et al., 2023)	Video	12.0k	Varied	-	Center point
GOO-real (Tomas et al., 2021)	Image	9.5k	Retail	24	Pixel-level
<b>GazeSeg (Ours)</b>	Image	77.5k	Varied	270	Pixel-level

### A.2.2 DATASET PROPERTIES

GazeSeg is built with a variety of objects in diverse scenes. The images in the dataset come from the publicly available Gazefollow (Recasens et al., 2015) dataset, which is collected from commonly used datasets in the field of computer vision, e.g. MS-COCO (Lin et al., 2014), SUN (Xiao et al., 2010), PASCAL(Everingham et al., 2010), ImageNet(Deng et al., 2009). The proposed benchmark

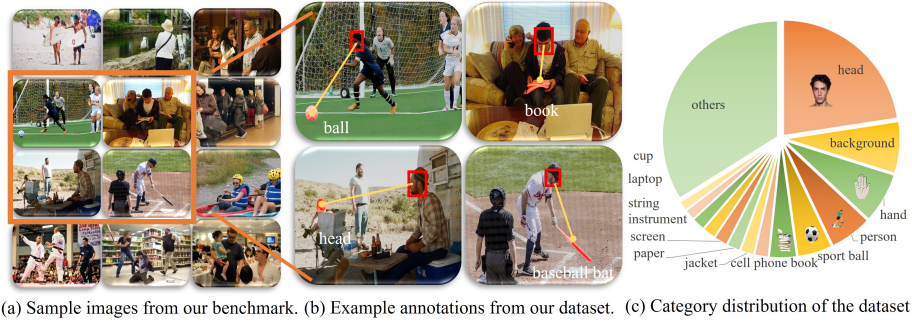


Figure 2: The properties of the proposed GazeSeg benchmark. (a) The sample images in GazeSeg cover diverse scenes. (b) The gaze targets in GazeSeg are annotated with pixel-level masks. (c) The distribution of gaze target categories in GazeSeg follows a long-tailed pattern.

inherits the diverse gaze targets characteristic of the GazeFollow dataset, with a total of 270 annotated visual target categories, including body parts, household items, sports equipment, and more. The distribution of these categories follows a long-tailed pattern, as shown in Fig. 2 (c), with a large number of categories accounting for less than 1%, which adds to the difficulty of the dataset. Compared to the existing datasets, the proposed benchmark is the first work that conducts pixel-level annotations and experiments in diverse scenarios and offers a large data volume.

#### A.2.3 CRITERIA FOR COLLECTION AND ANNOTATION

The **GazeSeg** dataset is constructed based on the **GazeFollow** dataset, with additional pixel-level masks and semantic category annotations. As illustrated in Fig. 3, the annotation details are as follows:

**(1) Annotation Process.** We start from the gaze points provided in GazeFollow. Potential target objects are first detected using YOLO, and initial masks are generated using SAM (Segment Anything Model). Then, human annotators use AnyLabeling<sup>1</sup> to refine and verify these masks.

**(2) Annotator Participation.** A total of 15 annotators participated in the annotation process. For the training set, each image was annotated by one annotator. For the test set, each image was independently reviewed by three annotators to ensure reliability.

**(3) Category Selection Strategy.** When multiple semantic interpretations exist for the gaze target (e.g., "head" vs. "whole person"), we prioritize finer-grained labels. For example, if the gaze point falls on the head, we annotate the "head" region; if it falls on other body parts, we annotate the entire "person".

**(4) Handling Ambiguous Cases.** In the training set, images with highly ambiguous or indeterminate gaze targets were excluded from GazeSeg to reduce noise. In the test set, all images were retained (including ambiguous cases) to support comprehensive evaluation. For cases with multiple gaze targets, up to three target objects were retained in the test set. During evaluation, if the model correctly segments any one of them, it is considered a success.

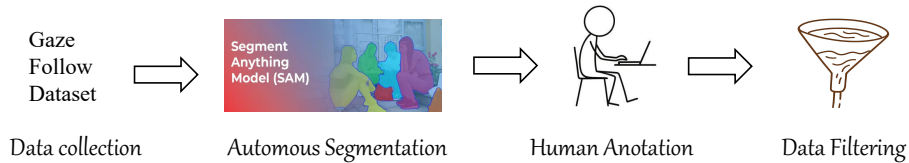


Figure 3: The data annotation pipeline of the proposed GazeSeg benchmark.

#### A.3 PERFORMANCE ON CHILDPLAY

Gaze as a nonverbal cue, can provide rich information about individuals, helping to infer human intentions and emotions. Especially for children, behaviors such as eye contact or joint attention are important indicators for diagnosing developmental disorders. We evaluated our approach in the

<sup>1</sup><https://github.com/vietanhdev/anylabeling>

Table 2: Results on the ChildPlay dataset.

Method	Venue	AUC $\uparrow$	L2 $\downarrow$	P.Head $\uparrow$
(Gupta et al., 2022)	CVPRW'22	0.919	0.113	0.694
(Tafasca et al., 2023)	ICCV'23	0.935	0.107	0.663
(Tafasca et al., 2024)	CVPR'24	-	0.106	0.600
(Ryan et al., 2024)(ViT-B)	CVPR'25	0.949	0.106	<b>0.715</b>
(Ryan et al., 2024)(ViT-L)	CVPR'25	<u>0.951</u>	<u>0.101</u>	0.662
Ours	-	<b>0.956</b>	<b>0.099</b>	<u>0.698</u>

Table 3: Results on the GOO-Synth dataset.

Method	Venue	GOO-Synth			Params $\downarrow$
		AUC $\uparrow$	Dist. $\downarrow$	mAP $\uparrow$	
SamGOP (Jin et al., 2024)	ECCV'24	94.7	0.072	2.70	80.99M
Ours	-	<b>96.3</b>	<b>0.069</b>	<b>2.85</b>	<b>30.57M</b>

ChildPlay dataset (Tafasca et al., 2023) to explore its potential in autism screening applications. This is an autism screening collection of carefully curated video clips of children playing and interacting with adults in uncontrolled environments (e.g. nursery schools, treatment centers, pre-schools, etc.). We describe performance using the Looking At Head Precision metric (P.Head) (Tafasca et al., 2023). From Table 2, in both scenarios our method outperforms existing methods, indicating its potential value in more advanced applications such as sentiment analysis.

#### A.4 PERFORMANCE ON GOO-SYNTH

We further validate our method on the GOO-Synth dataset (Tomas et al., 2021), which is a synthetic gaze following dataset collected in retail environments. As shown in Table 3, our method outperforms the existing SOTA method SamGOP (Jin et al., 2024) by a large margin across all metrics while using significantly fewer parameters, demonstrating the effectiveness and efficiency of our method.

#### A.5 ABLATION STUDY ON $\mathcal{L}_{mask}$

To validate the effectiveness of the proposed  $\mathcal{L}_{mask}$  loss function, we conduct ablation studies by integrating it into three representative gaze following methods: Chong (Chong et al., 2020), Song (Song et al., 2024), and Ryan (Ryan et al., 2024). As shown in Table 4, incorporating  $\mathcal{L}_{mask}$  consistently improves performance across all metrics for each method, demonstrating its effectiveness in enhancing gaze following models.

#### A.6 ABLATION STUDY ON TRAINING STRATEGIES

We conduct an ablation study to compare the two-stage training strategy with an end-to-end training approach. The results are presented in the Table 5. Specifically, the end-to-end approach is implemented by jointly optimizing both the gaze estimation and segmentation tasks simultaneously. The two-stage training strategy involves first training the gaze estimation model, followed by training the segmentation decoder using the features extracted from the pre-trained gaze model. We can observe from Table 5 that the two-stage training strategy outperforms the end-to-end approach across all evaluation metrics, i.e. AUC, Average Distance, mIoU, and Accuracy. It indicates that the two-stage training strategy is more effective in optimizing the model for the gaze target segmentation task.

#### A.7 IMPACT OF DECODER MODULE DEPTH

Empirically, increasing the depth of the decoder, i.e., stacking more layers, is likely to improve performance at the cost of more computational costs. We kept the depth of SAM’s mask architecture unchanged, and validated the effect of model depth on the multitask model architecture with the

Table 4: Ablation studies on the proposed  $\mathcal{L}_{mask}$  loss function.

Method	Venue	$\mathcal{L}_{mask}$	GazeSeg		
			AUC $\uparrow$	Avg Dist. $\downarrow$	Min Dist. $\downarrow$
Chong (Chong et al., 2020)	CVPR'20	-	0.9042	0.1510	0.0872
Chong (Chong et al., 2020) + $\mathcal{L}_{mask}$		✓	<b>0.9136</b>	<b>0.1494</b>	<b>0.0863</b>
Song (Song et al., 2024)	VI'24	-	0.9268	0.1161	0.0541
Song (Song et al., 2024) + $\mathcal{L}_{mask}$		✓	<b>0.9283</b>	<b>0.1140</b>	<b>0.0528</b>
Song (Song et al., 2024) (ft)	VI'24	-	0.9419	<b>0.1062</b>	<b>0.0463</b>
Song (Song et al., 2024) (ft) + $\mathcal{L}_{mask}$		✓	<b>0.9453</b>	0.1069	0.0465
Ryan (Ryan et al., 2024)	CVPR'25	-	0.9530	0.1064	0.0471
Ryan (Ryan et al., 2024) + $\mathcal{L}_{mask}$		✓	<b>0.9537</b>	<b>0.1070</b>	<b>0.0473</b>

Table 5: Ablation study on training strategies.

Method	AUC $\uparrow$	Avg Dist. $\downarrow$	mIoU $\uparrow$	Acc. $\uparrow$
End-to-end	0.949	0.099	34.41	43.86
<b>Two-stage (Ours)</b>	<b>0.953</b>	<b>0.092</b>	<b>34.86</b>	<b>45.10</b>

gaze and heatmap modules. As shown in Table 6, we try several different combinations of the depths of the two branching decoders and find that the 6-layer gaze decoder and the 6-layer heatmap decoder achieved the best balance of performance and computational cost. As the depth of the model continues to increase, the performance does not rise significantly and there is a decrease in segmentation and recognition. Based on our observation, the model first executes the gaze module, causing the encoder to be overly biased towards information about people to be detected.

Table 6: Ablation study on the depth of the decoder modules.  $N_g$  and  $N_h$  represent the number of layers in the gaze and heatmap decoders, respectively.

$N_g$	$N_h$	mIoU $\uparrow$	Acc $\uparrow$	AuC $\uparrow$	Min Dist $\downarrow$
2	2	26.15	39.5	0.915	0.058
4	2	24.52	37.5	0.907	0.062
2	4	24.61	37.8	0.910	0.060
4	4	26.34	39.7	0.919	0.056
6	4	28.75	41.9	0.928	0.051
4	6	32.11	44.3	0.940	0.046
<b>6</b>	<b>6</b>	<b>34.86</b>	<b>45.10</b>	<b>0.953</b>	<b>0.043</b>
8	8	32.21	44.7	0.941	0.045

## A.8 VISUALIZATION OF SUCCESSFUL AND FAILURE CASES

To provide a more intuitive understanding of the PixelGaze method, we present visualizations of both successful and failure cases in Fig. 4. In the successful cases, the PixelGaze method accurately localizes and segments the gaze targets, demonstrating its robust adaptability across various scenes and target categories. In the failure cases, we observe common challenges, e.g. complex environmental angles that interfere with the model’s judgment capabilities. These challenges can even pose difficulties for human observers in certain situations.

## A.9 VISUAL COMPARISON WITH EXISTING METHODS

Existing methods often rely on coarse 2D attention modeling or simplified 3D geometric assumptions. e.g., the attention maps in the VAT method (Chong et al., 2020) and the native gaze cone (Lian et al., 2018) struggle to effectively capture target locations and are limited in the granularity of image parsing. In contrast, our proposed method benefits from the joint modeling of 3D scene geometry and human gaze direction, along with the explicit incorporation of a pixel-level semantic segmentation mechanism. This enables effective prediction of the 3D Field of View (FoV) of individuals in the scene, which is then used to generate precise heatmaps. Finally, leveraging the our designed decoder architecture, PixelGaze effectively accomplishes target segmentation and recognition. Fig. 5 presents a visual comparison between our method and existing approaches, highlighting the superior performance of PixelGaze in accurately localizing and segmenting gaze targets.

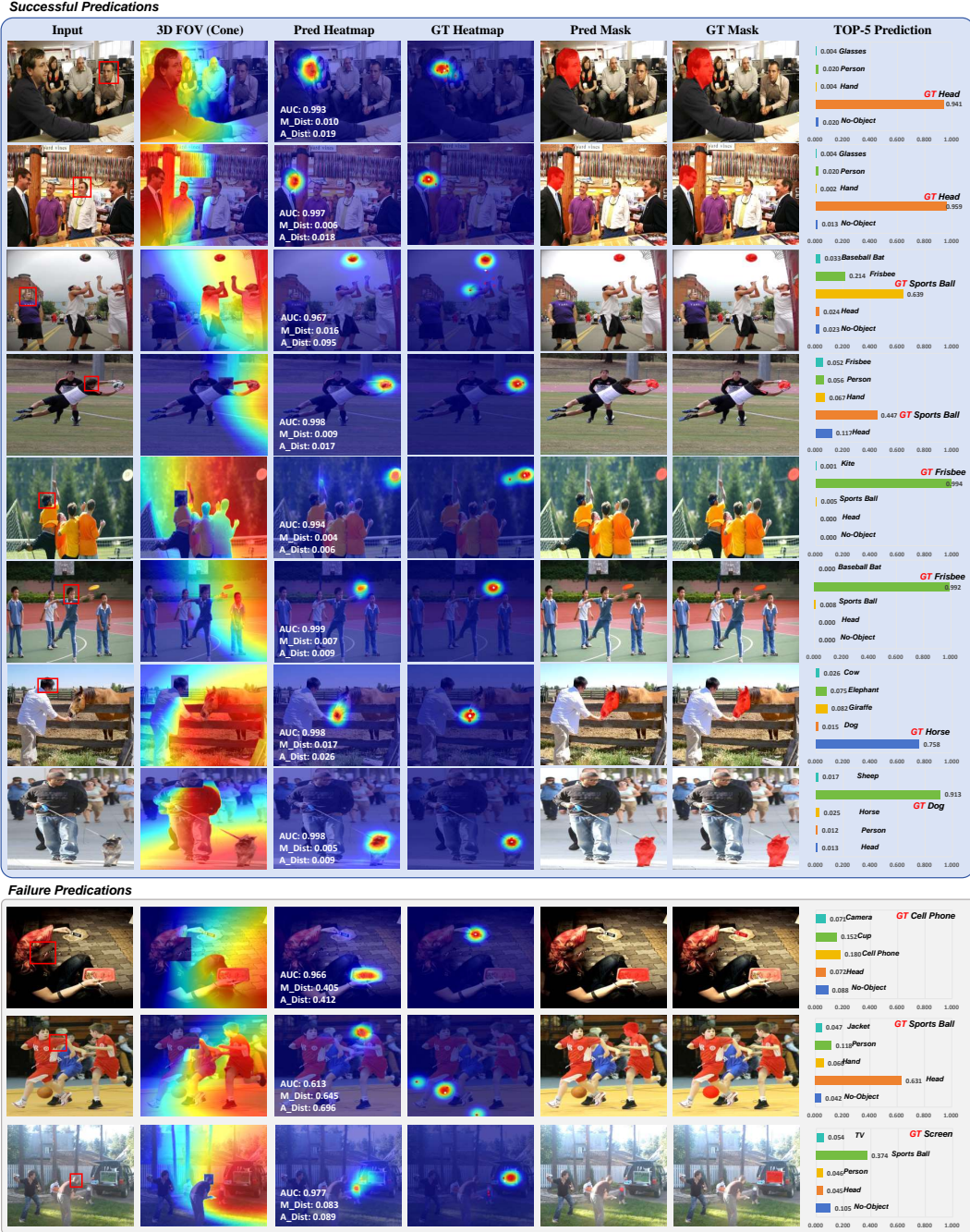


Figure 4: Visual demonstration of the PixelGaze method, including both successful samples and failure samples.

#### A.10 IMPACT OF SEGMENTATION AND RECOGNITION ON PIXEL-LEVEL PREDICTION

We conducted a parameter sensitivity analysis on the loss weights  $\lambda_1$  and  $\lambda_2$  in pixel-level prediction  $\mathcal{L}_{pred}$ . The  $\lambda_1$  denotes the weight of segmentation loss  $\mathcal{L}_{seg}$ , and  $\lambda_2$  denotes the weight of recognition loss  $\mathcal{L}_{cls}$ . The results are shown in Tab. 7. We can observe that the  $\mathcal{L}_{seg}$  and  $\mathcal{L}_{cls}$  have a significant impact on the final performance. When  $\lambda_1$  is too small, the segmentation performance drops significantly, which in turn affects the recognition performance. Similarly, when  $\lambda_2$  is too small, the recognition performance drops significantly. Therefore, we set  $\lambda_1$  and  $\lambda_2$  to 100 and 10, respectively, to achieve a balance between segmentation and recognition performance.

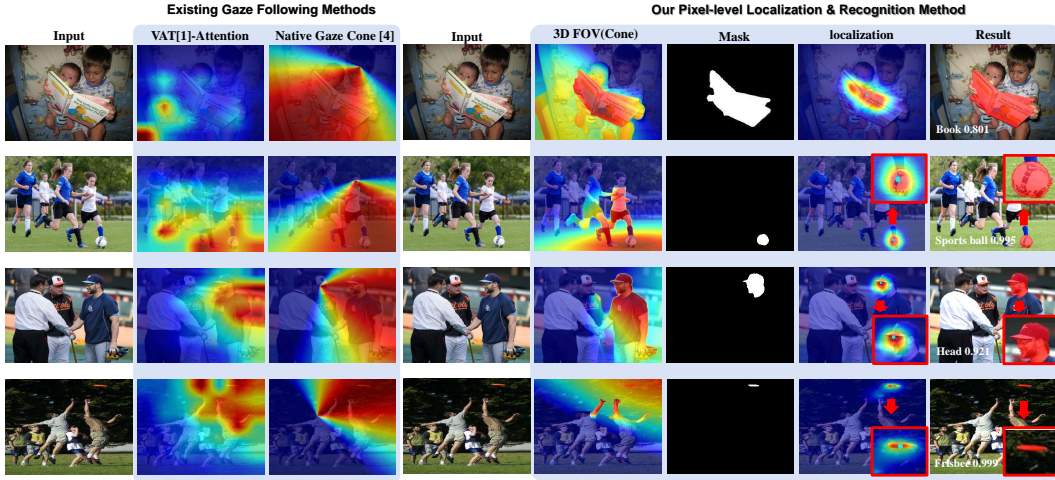


Figure 5: Visual comparison of gaze cone and localization results for (Chong et al., 2020) (VAT), (Lian et al., 2018) (NC) and our method.

Table 7: Ablation study on the loss weights  $\lambda_1$  and  $\lambda_2$ .

$\lambda_1$	$\lambda_2$	mIOU $\uparrow$	Acc $\uparrow$
50	10	33.87	44.35
85	5	33.90	44.85
100	10	<b>34.86</b>	<b>45.10</b>
100	5	33.76	41.56
100	20	33.50	44.09
20	33.93	44.18	150

## REFERENCES

- Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5396–5406, 2020.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.
- Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5041–5050, 2022.
- Yang Jin, Lei Zhang, Shi Yan, Bin Fan, and Binglu Wang. Boosting gaze object prediction via pixel-level supervision from vision foundation model. In *Proceedings of the European Conference on Computer Vision*, 2024.
- Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Proceedings of the Asian Conference on Computer Vision*, pp. 35–50, 2018.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014.
- Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Advances in Neural Information Processing Systems*, pp. 1–9, 2015.

- 
- Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. Gaze-llc: Gaze target estimation via large-scale learned encoders. *arXiv preprint arXiv:2412.09586*, 2024.
- Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: gaze following with interaction features in vision transformers. *Visual Intelligence*, 2(1):1–15, 2024.
- Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children’s gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20935–20946, 2023.
- Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Sharingan: A transformer architecture for multi-person gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2008–2017, June 2024.
- Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3125–3133, 2021.
- Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pp. 3485–3492. IEEE Computer Society, 2010.