Cutting-Edge Techniques in Gaze Target Detection Through Computer Vision

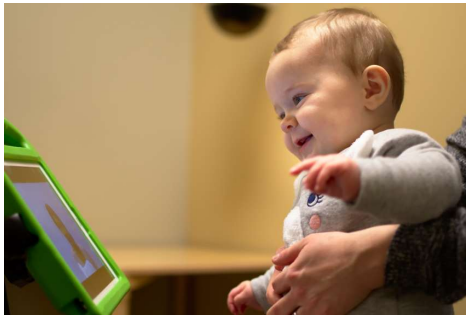
Cigdem Beyan
Dept. of Computer Science
University of Verona, Italy

Why Gaze?

- A fundamental form of **non-verbal** communication
  - Reflects a person's **attention**, **intention**, and **emotional state**.
    - A "shy gaze" involves averted eye contact, often signaling **discomfort** or **self-consciousness**.
  - Crucial for understanding how **gestures** are interpreted and understood in **human interactions***.

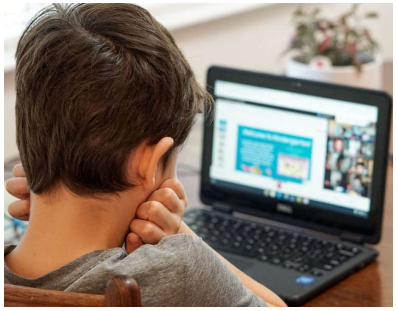
*Beyan et al., Co-Located Human-Human Interaction Analysis using Nonverbal Cues: A Survey, ACM Computing, 2023

(Some) Applications of Automated Gaze Analysis



Mental Health Monitoring: Autism

Driver Monitoring Systems: Distraction Detection



Education and Learning: Attention Analysis

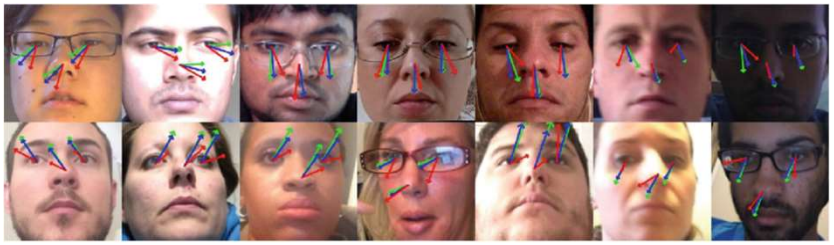
Human-Computer Interaction

Img. credit: geoGAZElab@ETH

Marketing and Consumer Research

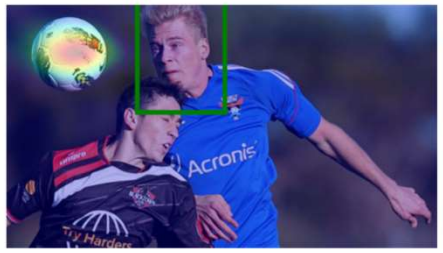
Gaze Analysis w/ Computer Vision

- Two tasks:
  1) **Gaze estimation**: Determines the direction of a person's gaze in 3D space, estimating horizontal and vertical angles, and the depth of focus.



Che et al., CVPR 2022

2) **Gaze target detection (gaze following)**: Identifies the specific point or area a person is looking at in 2D/3D.



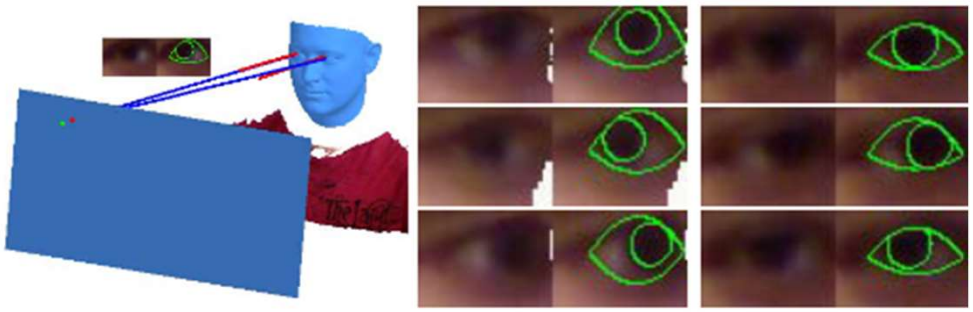
Tonini et al., ICCV 2023

Benefits & Differences

Gaze estimation

- Can be very accurate
- Applies to dedicated [illegible] set-ups
  - May use simple geometr[illegible] reasoning

With respect to



Mora et al., CVPR 2014

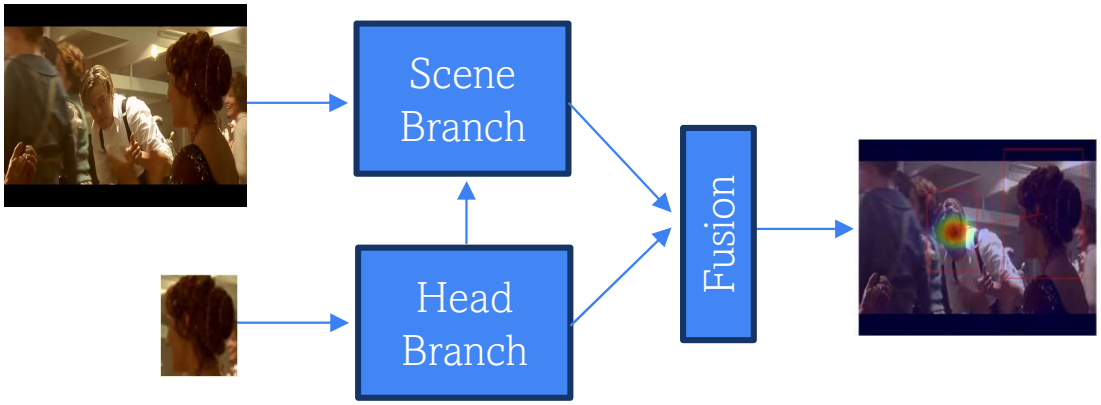
Gaze target detection

Today!

- Works on arbitrary sc[illegible]er than frontal images: generaliz[illegible]on
- Results in **spatial location**, not only a direction
- Can determine the category of the target gaze: **semantic info**

Tonini et al., ICCV 2023

GTD Standard Approach



Recasens et al., NeurIPS 2015
Chong et al., ECCV 2018, CVPR 2020
Lian et al., ACCV 2018
Fang et al., CVPR 2021
Bao et al., CVPR 2022
Tonini et al., ICMI 2022
Qiaomu et al., WACV 2023…...

Inputs: Scene image + head crop + head location

Head Branch: processes the person's head to infer gaze direction

Scene Branch: processes the scene to capture saliency and the context. Uses the head location information. The attention mechanism by considering the head features.

Fusion: Combines information to predict the gaze heatmap.

GTD Standard Approach + What Else?

- Adjusting for **video** processing

(Chong et al., CVPR 2020)

- Adjusting to predict **out-of-frame** gaze detection

(Chong et al., CVPR2020)

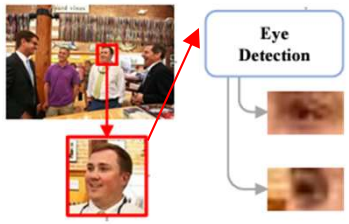
- Injection **depth** information; improves the performance

(Fang et al., CVPR 2021; Jin et al., 2022; Miao et al., WACV 2023; Tonini et al., ICMI 2022)

- Injecting **eye location;** improves

the performance by improving direction estimation

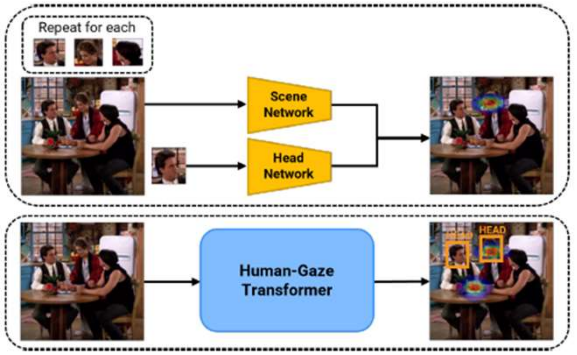
(Fang et al., CVPR 2021)

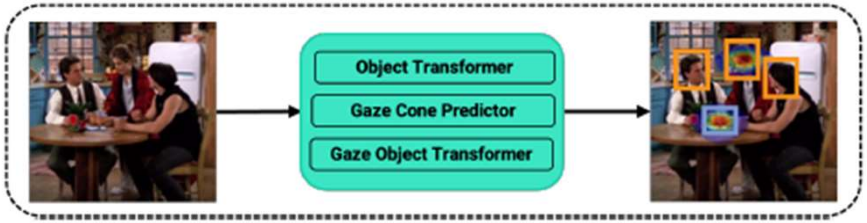


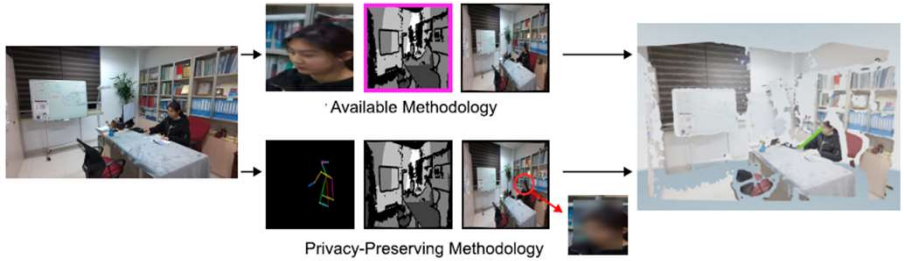
- Injecting **body pose;** is a good alternative (Fang et al., CVPR 2021; Gupta et al., CVPRw 2022)

GTD Standard Approach: Disadvantages

- Single gaze target detection at a time, **inefficient** for **multiple** people
  - Tu et al., CVPR 2022; a **transformer-based** approach
- Not **directly applicable** to human-human or human-robot interaction analysis
  - Tonini et al., ICCV 2023; **transformer-based** approach, which also performs **gazed-object classification** and **localization.**
- Performs only in 2D & **not privacy-preserving**
  - Toaiari et al., ECCVw 2024; a pipeline preserving privacy (no head crops and blurred scene, pose-based) to predict the gaze location in both 2D and 3D.





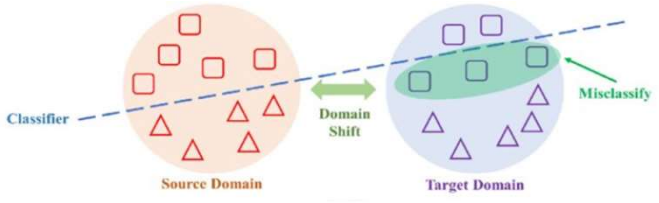


UNIVERSITÀ di VERONA

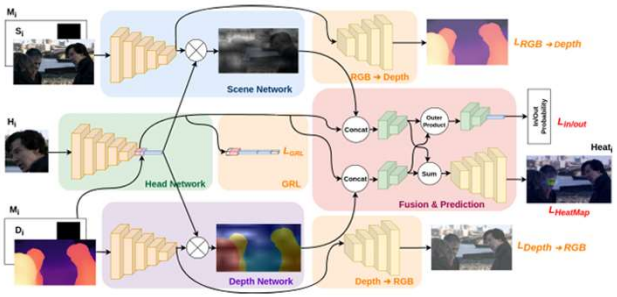
GTD Standard Approach & Extensions: Key Takeaways

- **Multimodality** matters!
  - Especially depth information, but also body pose and/or eye location
- **Transformer-based** methods are SOTA
- Most recent trend: exploring the zero-shot capabilities of **Vision-Language Models** (Gupta et al., CVPRw 2024)

GTD Challenges



- GTD models trained on one dataset often struggle to **generalize** to new datasets in **zero-shot** settings.
  - Performance drop up to 30% in AUC (Tonini et al., ICMI 2022).



- **Unsupervised domain adaptation** can be a solution to some extent, but the gap with fine-tuning is still noticeable (Tonini et al., ICMI 2022). ➔ In practice, we perhaps need labelled data to be used for **training**

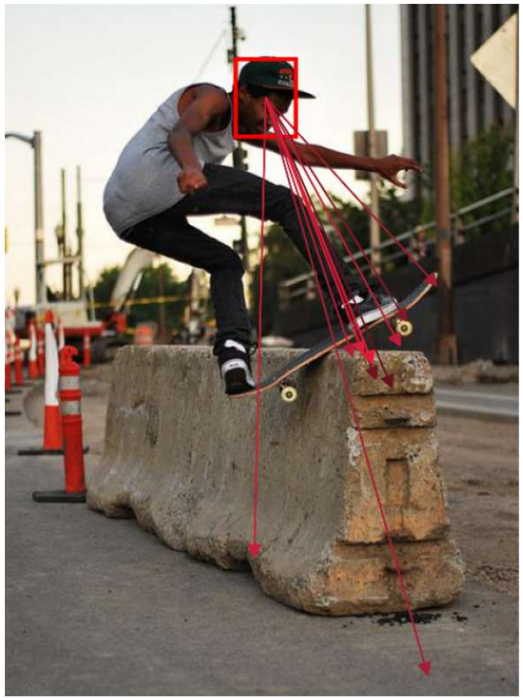
- Labelling gaze data: **highly tedious and time-consuming**

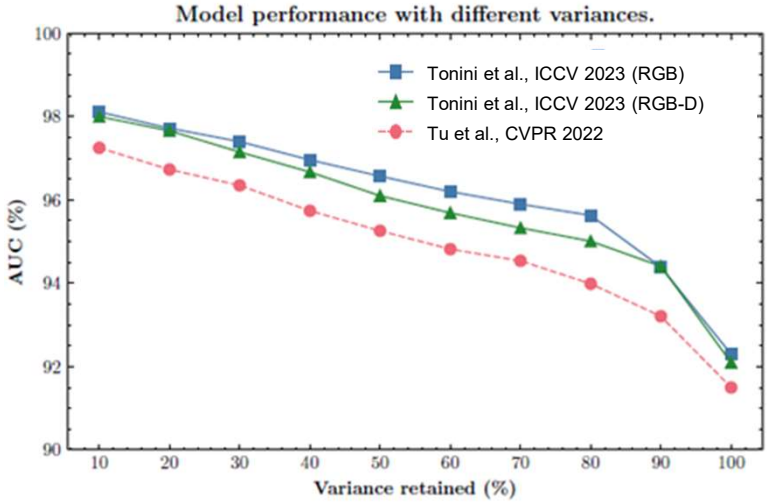
"*Labor times ranging from **ten seconds to one minute** for **every second** of gaze data*"
-- Agtzidis, et al., 2020; Gutiérrez et al., 2018

*"An experienced human annotator may require **two to more than ten times** the duration of a video to accurately label the gaze"* -- Erel et al., 2023

GTD Challenges

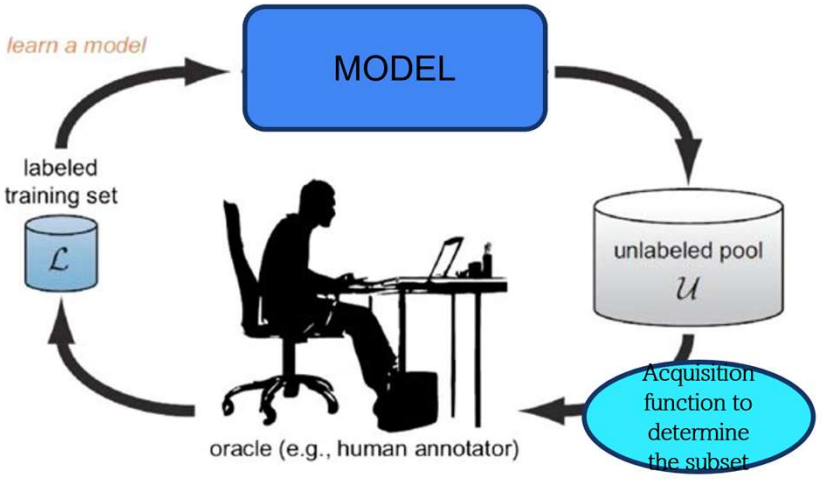
- **Low annotation reliability** due to subjectivity, causing inconsistencies in labelling the same image.





Research Question & Approach

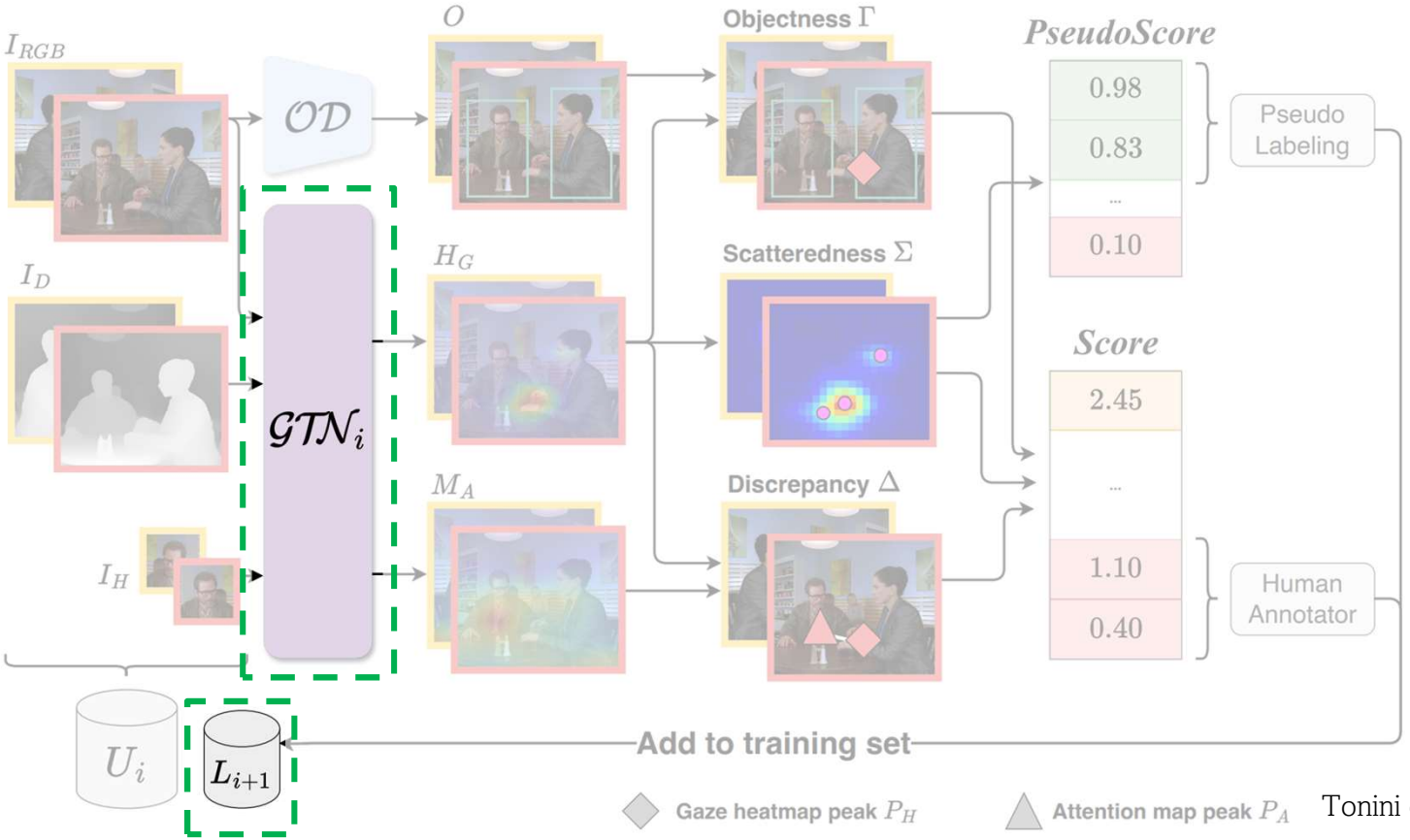
- Can we reduce reliance on labelled training data for gaze detection without sacrificing performance?
- Given a gaze target detection network,
  - Aim: to **effectively select a subset** of the data that **optimizes model training**.
- Our Solution: based on Active Learning
  - Among a large pool of unlabelled data, choose those that are the most informative to be labelled.
  - Iterative approach.
  - Most informative is defined by an acquisition function.



Tonini et al., AL-GTD: Deep Active Learning for Gaze Target Detection, In Proc. of ACM Multimedia, 2024

Our pipeline – Pre-training the gaze model on random data

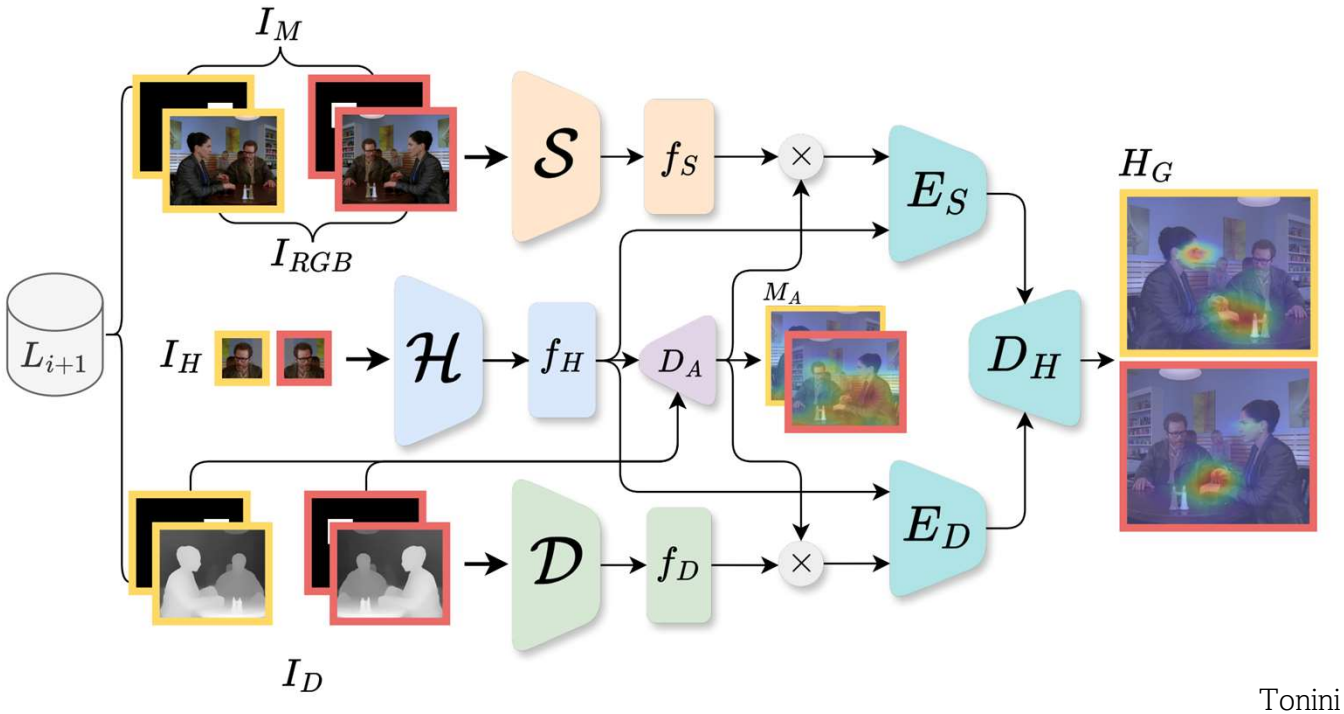
Train a gaze target detection model on a small random subset of labelled samples L and obtain predictions for unlabelled samples U.



Tonini et al., ACM Multimedia, 2024

Training the Gaze Target Detection model

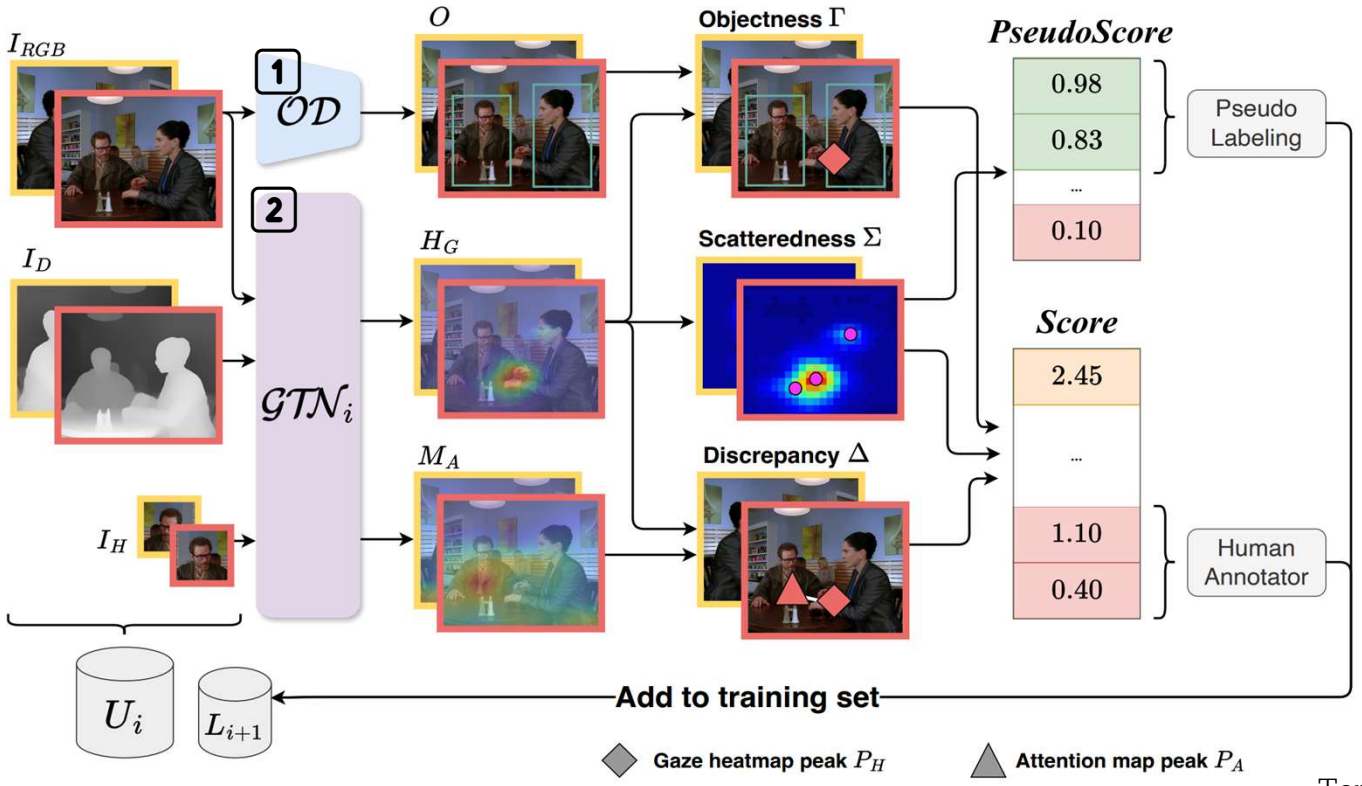
Follows standard recipe of GTD training. To alleviate prediction inconsistency, we perform strong augmentations and perform self-supervised learning among augmentations and the original image.



Tonini et al., ACM Multimedia, 2024

Our pipeline - Active Learning for Gaze Target Detection

1. Detect **objects** in the image.
2. Extract a **gaze heatmap** and **attention map** of the scene.



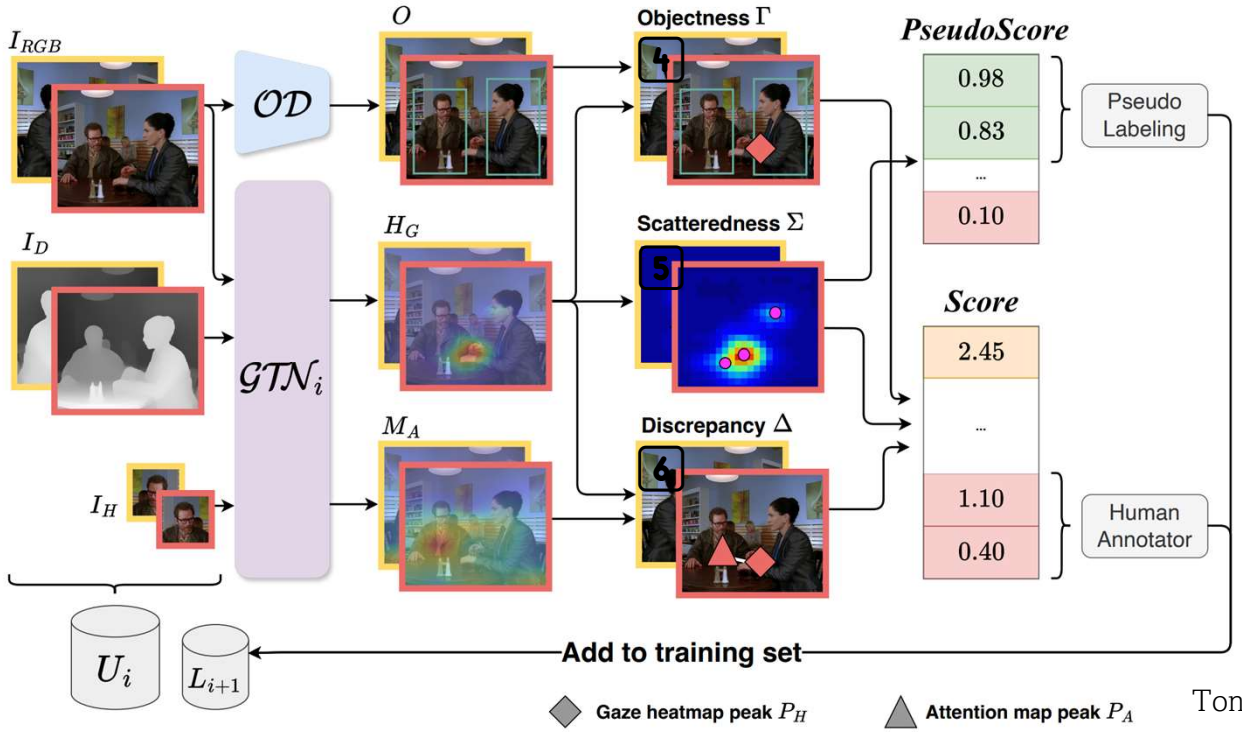
Tonini et al., ACM Multimedia, 2024

Our pipeline – AL Acquisition Function

4 **Objectness score:** penalize predictions centred on foreground objects.

5 **Scatteredness score:** penalize sparse activation of the network output corresponding to prediction uncertainty

6 **Discrepancy score:** penalize differences in network predictions among augmented versions of the same image.



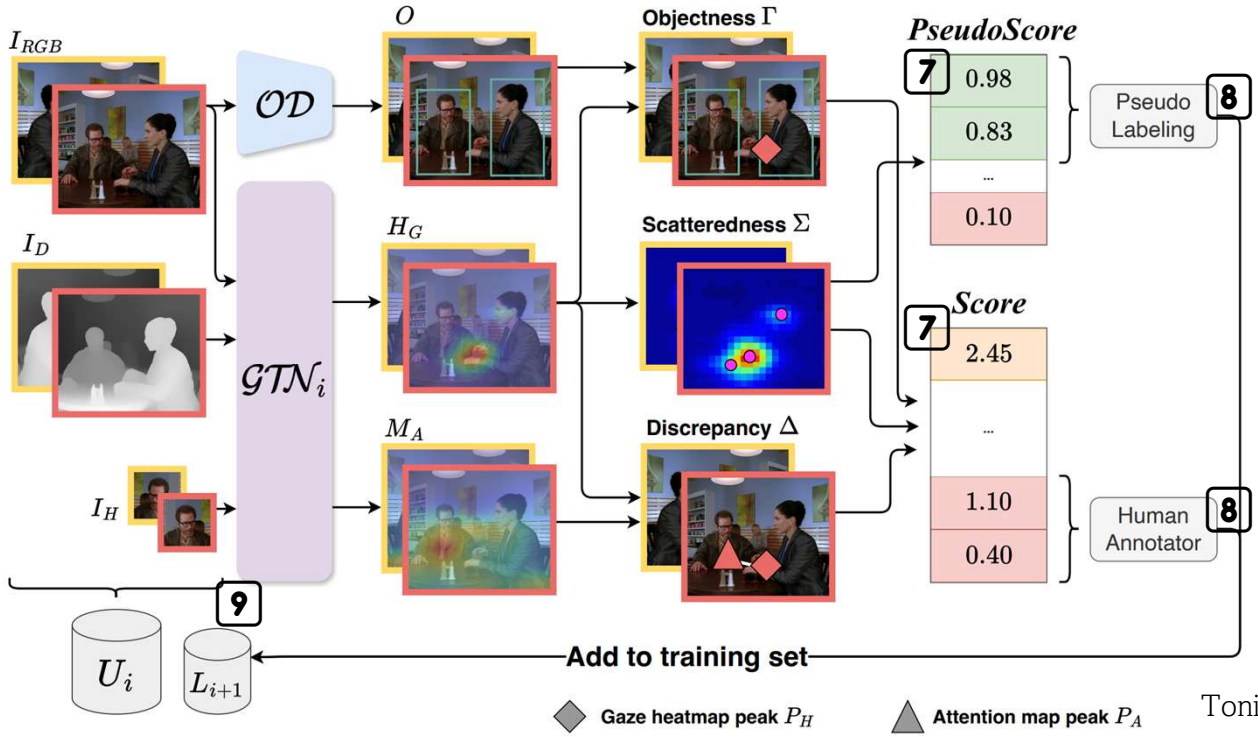
Tonini et al., ACM Multimedia, 2024

Our pipeline - Active Learning for Gaze Target Detection

7 The **score** and **pseudo score** are calculated.

8 **Oracle annotates the most informative samples**, while those with high pseudo scores are pseudo labelled.

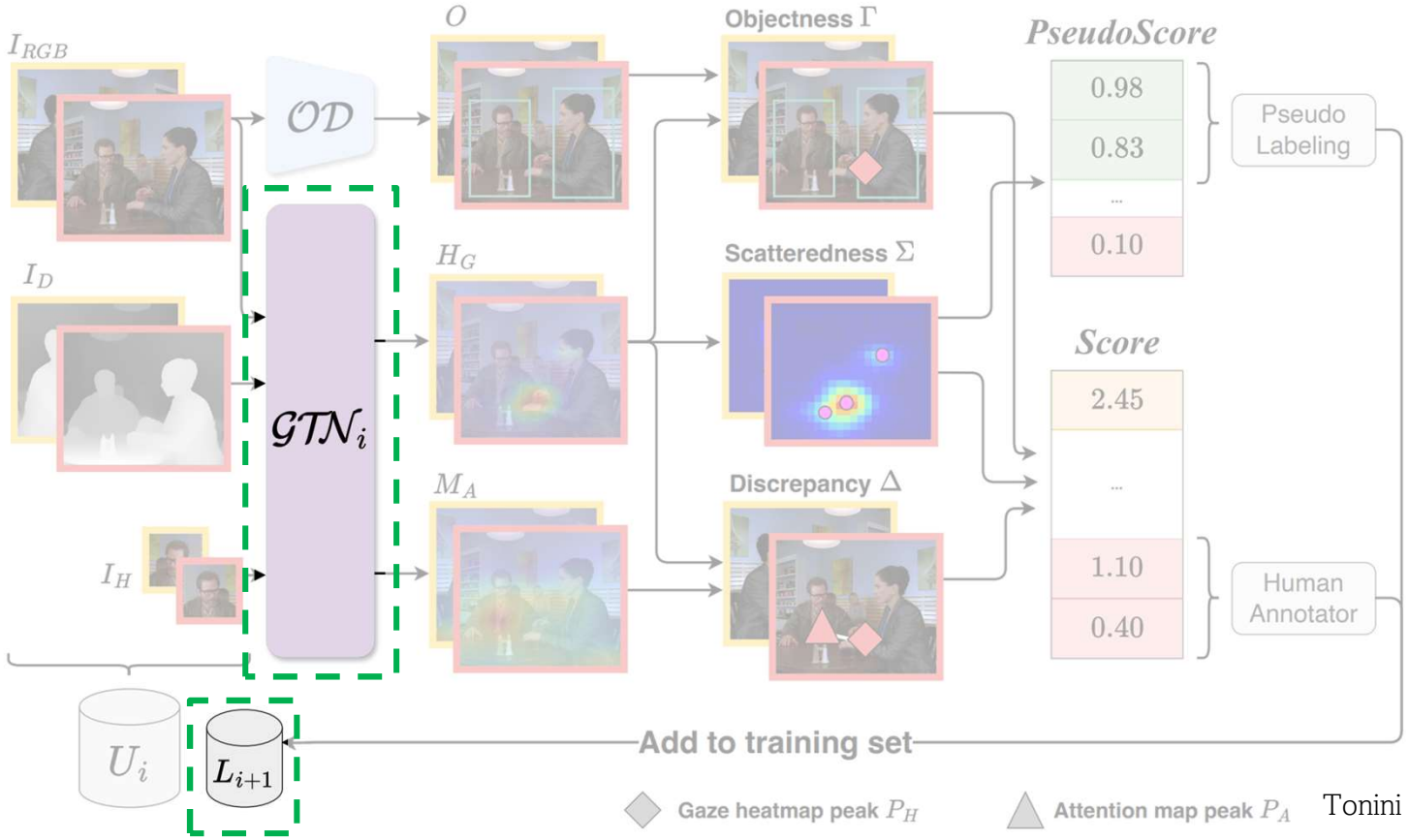
9 The newly labelled data are added to the training pool and used for the next AL cycle.



Tonini et al., ACM Multimedia, 2024

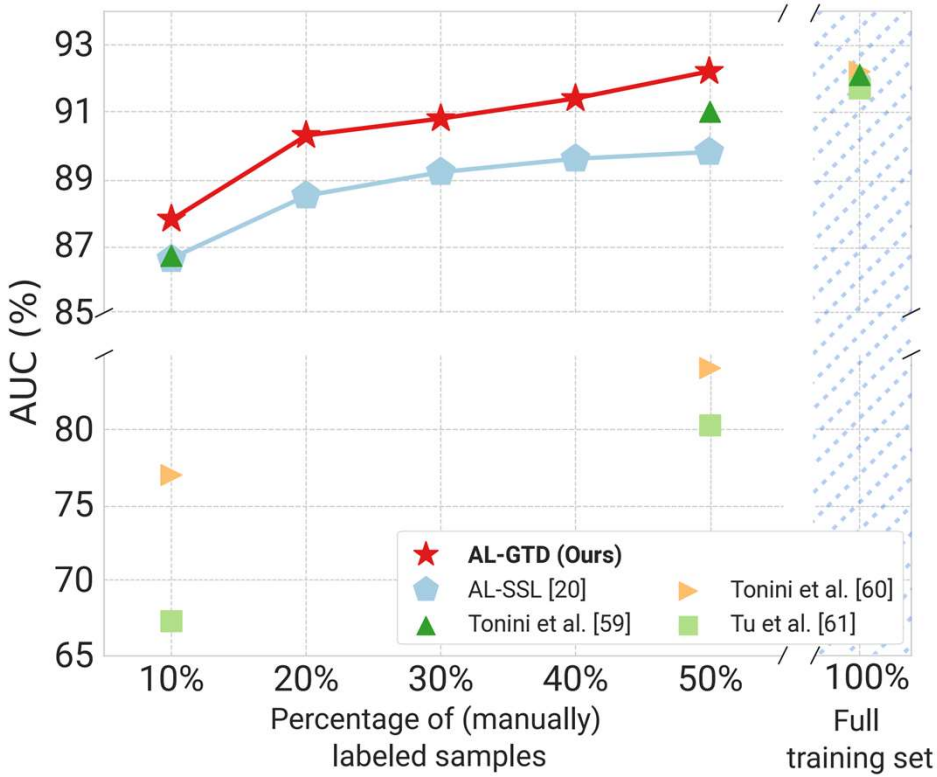
Our pipeline – Training the gaze model on new labeled data

(Re-)Train a gaze target detection method on a small random subset + newly labelled data and obtain predictions for unlabelled samples *U*.



Tonini et al., ACM Multimedia, 2024

Results

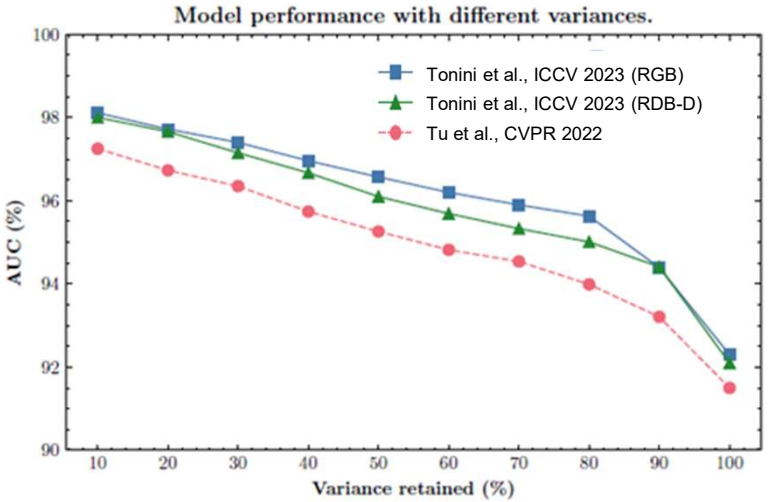


Our proposal achieves SOTA AUC performance with just **40-50%** of the usual training data and is competitive w.r.t. the current best-performing models even when trained on a mere 10-20% of the full training dataset.

Tonini et al., ACM Multimedia, 2024

Results

- Is the model suitable for complex scenes?
- Complex: the gaze points whose annotation has low reliability (high variance)



	Performance difference among complex and simple scenes			
Variance	5%	10%	50%	100%
10%	5.79%	5.92%	6.32%	5.79%
50%	8.74%	8.47%	9.44%	9.87%

No significant performance drop (AUC) on complex scenes with 5%, 10%, 50%, or 100% training data.

Tonini et al., ACM Multimedia, 2024

For further analysis

- Comparison with other Active Learning methods
- Comparison with several SOTA gaze target detector
- Ablation study
- ….

Code

Paper

Tonini et al., AL-GTD: Deep Active Learning for Gaze Target Detection, In Proc. of ACM Multimedia, 2024

The GAZE Team!

Francesco Tonini

Nicola Dall'Asen

Andrea Toaiari

Lorenzo Vaquero

Elisa Ricci

Marco Cristani

Vittorio Murino

THANK YOU!

Cigdem Beyan
Dept. of Computer Science
University of Verona

Slides with references can be found on my website!

I am hiring!

If you're interested in joining my team, feel free to connect ☺

beyancigdem | cbeyan.github.io | cbeyan | BeyanCigdem | Cigdem Beyan

UNIVERSITÀ di VERONA

References

- Beyan et al. 2023. Co-Located Human-Human Interaction Analysis using Nonverbal Cues: A Survey. Comput. Surveys 56, 5 (2023), 109:1–109:41.
- Tonini et al. 2023. Object-aware Gaze Target Detection. In Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Paris, France, 21860–21869.
- Recasens et al.. 2015. Where are they looking?. In Advances in Neural Information Processing Systems, Vol. 28. Curran Associates, Inc., Montreal, Quebec, Canada, 199–207.
- Chong et al. 2018. Connecting Gaze, Scene, and Attention: Generalized Attention Estimation via Joint Modeling of Gaze and Scene Saliency. In The European Conference on Computer Vision (ECCV).
- Lian et al. 2018. Believe it or not, we know what you are looking at!. In Asian Conference on Computer Vision. Springer, Springer, Perth, Australia, 35–50.
- Yi Fang et al. 2021. Dual attention guided gaze target detection in the wild. In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.11390–11399.
- Jun Bao et al. 2022. ESCNet: Gaze Target Detection with the Understanding of 3D Scenes. In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14126–14135.

UNIVERSITÀ di VERONA

References

- Tonini et al. 2022. Multimodal Across Domains Gaze Target Detection. In Proc. of the 2022 International Conference on Multimodal Interaction. ACM, 420–431.
- Miao et al. 2023. Patch-level Gaze Distribution Prediction for Gaze Following. In Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision. 880–889.
- Jin et al. 2022. Depth-aware gaze-following via auxiliary networks for robotics. Engineering Applications of Artificial Intelligence 113.
- Li et al. 2023. In the Eye of the Beholder: Gaze and Actions in First Person Video. IEEE Trans. Pattern Anal. Mach. Intell. 45, 6 (2023), 6731–6747.
- Gupta et al., Exploring the Zero-Shot Capabilities of Vision-Language Models for Improving Gaze Following, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops 2024.
- Toaiari et al. Upper-Body Pose-based Gaze Estimation for Privacy-Preserving 3D Gaze Target Detection, European Conference on Computer Vision 2024.

References

- Agtzidis et al. 2020. Two hours in Hollywood: A manually annotated ground truth data set of eye movements during movie clip watching. Journal of Eye Movement Research 13, 4 (2020), 1–12.
- Erwan et al.. 2018. A dataset of head and eye movements for 360 videos. In Proc. of the 9th ACM Multimedia Systems Conference. ACM, 432–437.
- Erel et al. 2023. iCatcher+: Robust and Automated Annotation of Infants' and Young Children's Gaze Behavior From Videos Collected in Laboratory, Field, and Online Studies. Advances in Methods and Practices in Psychological Science 6, 2 (2023).