# Object-aware Gaze Target Detection
# Supplementary Material

## A. The effect of variance in gaze annotations

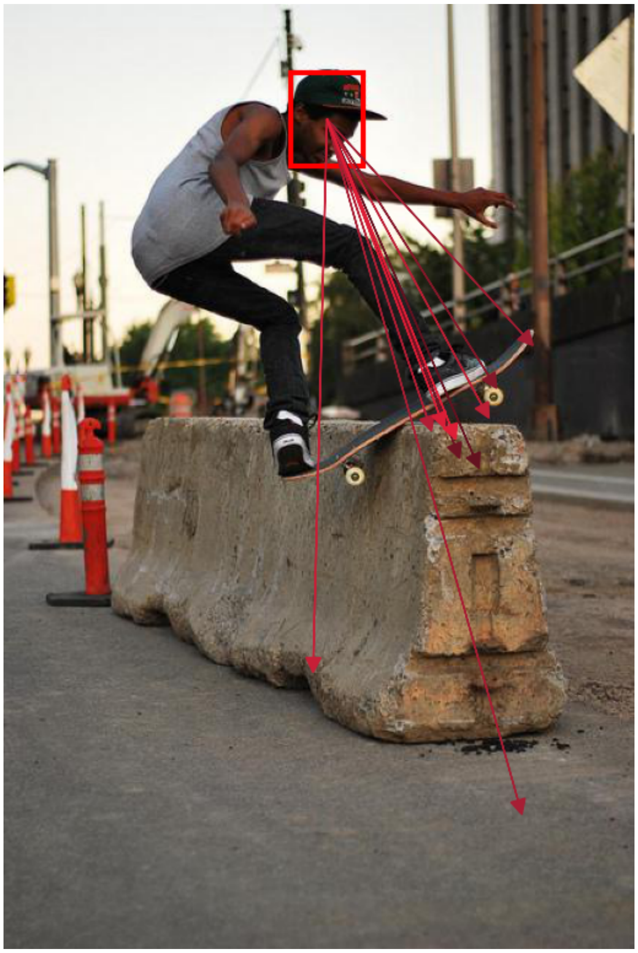

Figure 1: An image from the GazeFollow dataset's [10] test split which has several annotations for the gaze point with high annotation variance.

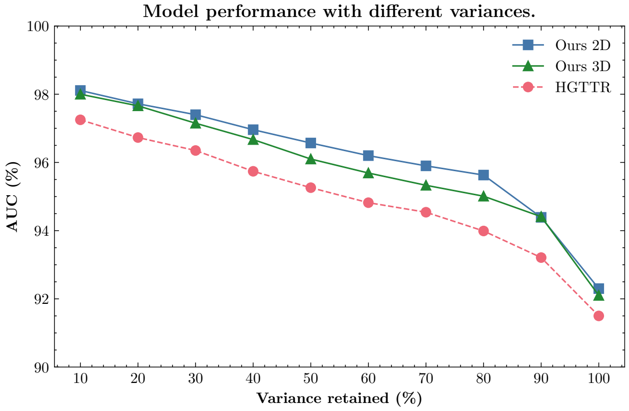

Figure 2: Performance of Our model (2D and 3D) and Tu et al. [12] *w.r.t.* the variance in the ground-truth annotations of the GazeFollow dataset [10].

As mentioned in the paper, the GazeFollow dataset [10] contains a single gaze point annotation for a single person in a scene in its training split. However, its test splits include several numbers of annotations with respect to a single person's gaze. The number of annotations can be varying up to 10 different gaze points for each person. Such an annotation procedure would not present an issue if all the annotators reached a consensus regarding the gaze point, however, as also shown in [8] the test split annotations of that dataset per person can vary remarkably. An example image and the corresponding gaze points demonstrating the variety across the annotations are given in Fig. 1.

On the other hand, the standard metric used to make evaluations on this dataset, aka AUC does not consider the (possible) varieties across the annotations. For this reason, herein as well as in the main paper, we present an additional evaluation procedure, which considers the multiple annotations that GazeFollow test split provides (see Fig. 2). That evaluation procedure can be described as follows. For each gaze point annotation corresponding to a single person, we compute its distance to the corresponding average gaze annotation point. We record all the distances for the whole test split and compose a distribution from them. Then, given this distribution, we keep the gaze points falling inside a certain threshold (shown as variance retained in the figure), and we opt for the deciles for easiness of computation. For each decile, we compute the AUC and report it in Fig. 2. As seen, our method is extremely effective when there is high annotation consensus, *i.e.* the distance from the average point falls in the first decile (*i.e.*, 0-10% variance retained in the figure); the performance slightly decreases until the eighth decile (*i.e*., 70-80% variance retained in the figure), with the last 20% representing high noise annotations (*i.e*., 80-100% variance retained in the figure) where the performance lowers at a faster rate. When we compare our performance against the state-of-the-art method of [12], one can observe a consistently higher performance for all cases both in 2D and 3D versions of our method. We speculate that the lower performance of Ours-3D *w.r.t.* Ours-2D can be since the human annotations were collected on 2D images.

## B. Additional evaluation on GazeFollow [10] and VideoAttentionTarget [2]

Table 1 reports the *Angular Error* [10] (i.e. the angle between predicted and ground-truth gaze vector) results and compare it with SOTA. Our method produces the best results out of all, while Ours (3D) is better than Ours (2D).

| | Ours (2D) | Ours (3D) | [12]$^\star$ | [11]$^\star$ | [5] | [1] | [4] | [2]$^\star$ | [7] | [10] |
|---|---|---|---|---|---|---|---|---|---|---|
| Min. ↓ | 4.0° | **3.5**° (−12.5%) | 6.6° | 8.1° | — | — | — | 9.1° | 8.8° | — |
| Avg. ↓ | 7.7° | **7.2**° (−6.2%) | 11.0° | 19.5° | 14.8° | 14.6° | 14.9° | 20.5° | 17.6° | 24.0° |
| Max. ↓ | 20.1° | **19.3**° (−3.9%) | 22.5° | 37.0° | — | — | — | 37.9° | — | — |

Table 1: Angular error on GazeFollow [10] $\star$ means our implementation. Improvements are w.r.t. "Ours (2D)".

## C. Implementation Details

We implemented our method in PyTorch and relied on the official code of DETR [13] as the backbone. The heads of DETR [13], *i.e.* the two MLPs for object classification and detection, were replaced by two larger MLPs that allow us to predict the location and classification of objects in the scene including the *heads*. Therefore, the number of classes of objects is adapted to accommodate the *head* class. We used a SOTA object detector, YOLOv8 [6], to pseudo-annotate objects in images that lack object annotations. This has been needed since the used datasets (except the COCO subset of the GazeFollow dataset) do not provide object annotations. We finetuned the *Object Detector Transformer* using head locations given in the used datasets as well as automatically obtained using an additional head detector, RetinaFace [3], and for other objects extracted from YOLOv8. RetinaFace was necessary (but other head detectors can be also adapted as shown in Tu et al. [12]) as we observed that both Tu et al. [12], and our method could not converge without head annotations of all heads in the image. The depth images were obtained by processing both datasets with a SOTA monocular depth estimation method called MIDAS [9].

## D. Qualitative Results

In this section, we provide additional qualitative results of the gaze heatmaps and the head bounding box of the gaze source (*i.e*., a person's head) and demonstrate the improved performance of our method *w.r.t.* the current state-of-the-art (SOTA) for both GazeFollow [10] and VideoAttentionTarget [2] datasets. Furthermore, we discuss some example cases in which our method has relatively lower performance ($AUC < 70\%$) *w.r.t.* ground-truth as well as Tu et al. [12]. Lastly, we compare our methods' versions in 2D and 3D and demonstrate the latter's effectiveness in challenging scenarios.

**Comparison with SOTA and ground-truth.** Fig. 3 and Fig. 4 compare our predictions with respect to the ground truth and the predictions of Tu et al. [12] on both datasets, GazeFollow [10] and VideoAttentionTarget [2]. As we can see, our model precisely predicts the gaze in many scenes where [12] is not able to. More importantly, we can see that predictions of both our method and Tu et al. [12] are in the field of view of the person whose gaze is to be predicted. However, [12] favors image regions closer to the gaze source (*i.e.* person's head).

**Relatively low-performing predictions.** Fig. 5 presents example images in which our method relatively performs worse. Notice that such images are highly challenging and most of the time also SOTA [12] underperform, *e.g.* in the second row of Fig. 5, where the head-pose makes it difficult to accurately predict the gaze. Conversely, when the face is not fully visible, e.g., in the third and fifth row of Fig. 5, we predict scattered heatmaps that cover the gaze point.

**The contribution of 3D gaze cone.** Fig. 6 demonstrates the results of our method with the 2D or 3D gaze cone (see main paper for additional details). This comparison aims to highlight the importance of the 3D cone particularly in challenging scenes. As the quantitative results in the main paper showed, the advantage of the 3D cone is especially visible in terms of the average and minimum distance between the ground-truth gaze point and the point of maximum confidence of the gaze heatmap. The example images in Fig. 6 demonstrate that in complex scenes, the 3D gaze vector and the corresponding 3D gaze cone help to decipher which object the person is looking at. Moreover, when the predictions with 2D cone are already high (*e.g.* the first row of Fig. 6), the 3D cone counterpart further consolidates the center of the heatmap towards the object, resulting in better performance.

## References

[1] Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14126–14135, 2022. 2

[2] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020. 2, 5

[3] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[4] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11390–11399, 2021. 2

[5] Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924, 2022. 2

[6] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 2023. 2

[7] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018. 2

[8] Qiaomu Miao, Minh Hoai, and Dimitris Samaras. Patch-level gaze distribution prediction for gaze following. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 880–889, 2023. 1

[9] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 2

[10] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 1, 2, 4, 6, 7

[11] Francesco Tonini, Cigdem Beyan, and Elisa Ricci. Multimodal across domains gaze target detection. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 420–431, 2022. 2

[12] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-end human-gaze-target detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2192–2200. IEEE, 2022. 1, 2, 4, 5, 6

[13] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021*, 2021. 2

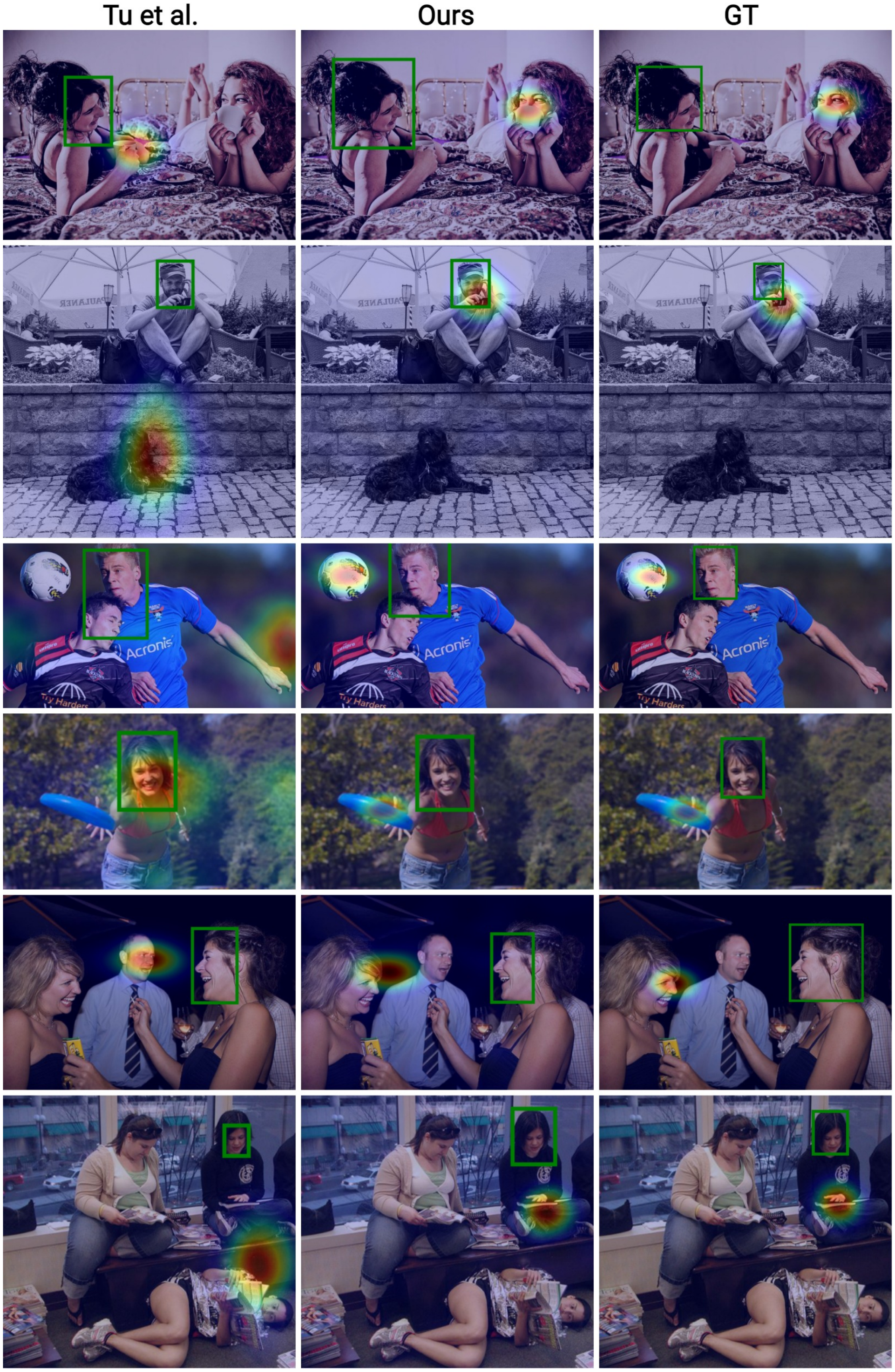

Figure 3: Qualitative results of our method (center) and Tu et al. [12] (left) *w.r.t.* the ground-truth annotations of the GazeFollow [10] dataset (right). Please note that we only show gaze predictions for the person whose gaze is included in the ground truth. The green boxes show the person-in-interest for the ground truth while they are the detected head for Our and Tu et al. [12]. Even though our method can detect the gaze of multiple persons in the scene simultaneously, for better visualization, we plot the predicted heatmaps and head locations per person.

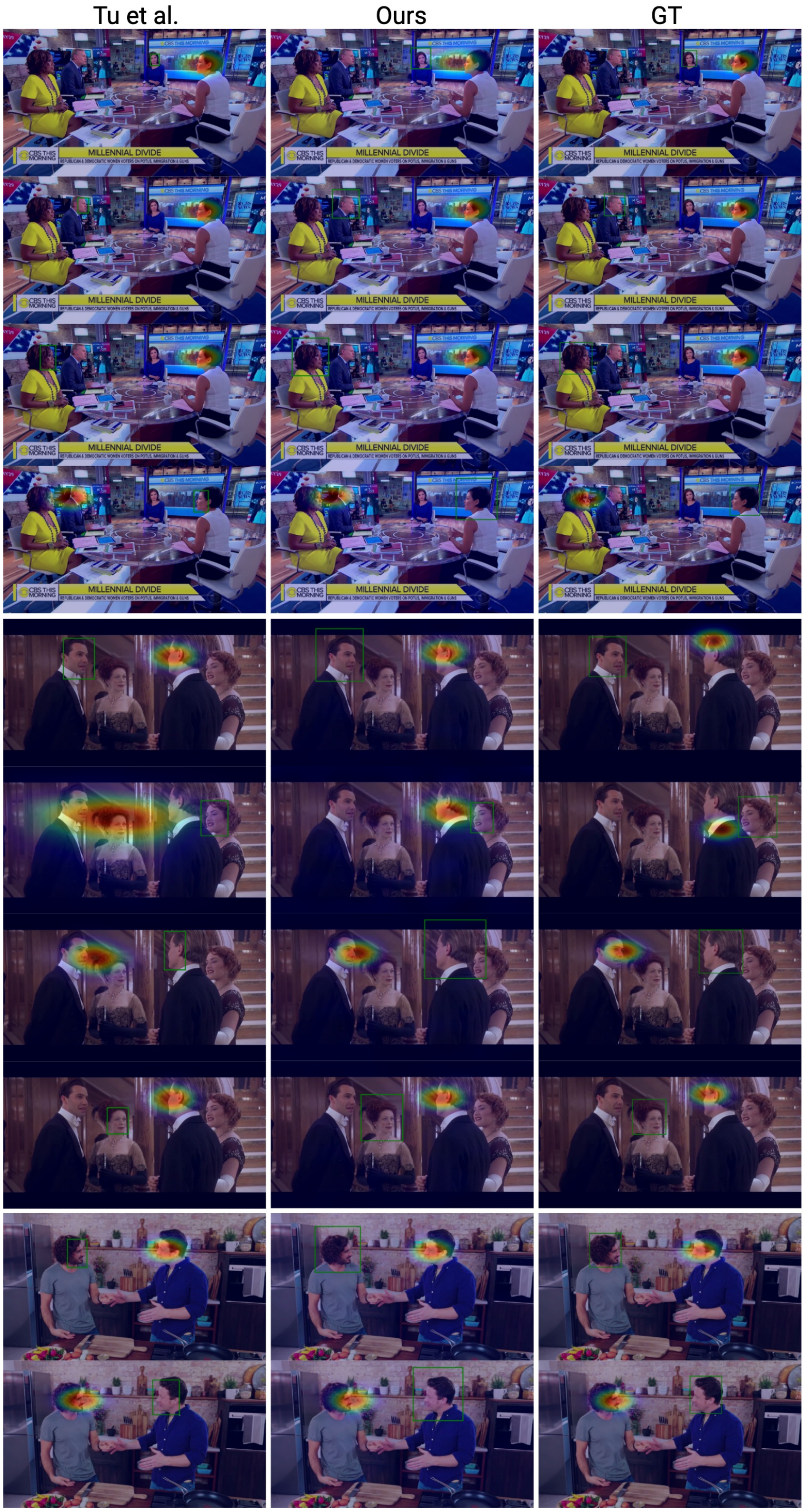

Figure 4: Qualitative results of our method (center) and Tu et al. [12] (left) *w.r.t.* the ground-truth annotations of the VideoAttentionTarget [2] dataset (right). Please note that we only show gaze predictions for the person whose gaze is included in the ground truth. The green boxes show the person-in-interest for the ground truth while they are the detected head for Our and Tu et al. [12]. Even though our method can detect the gaze of multiple persons in the scene simultaneously, for better visualization, we plot a single person's predicted heatmaps and head location.

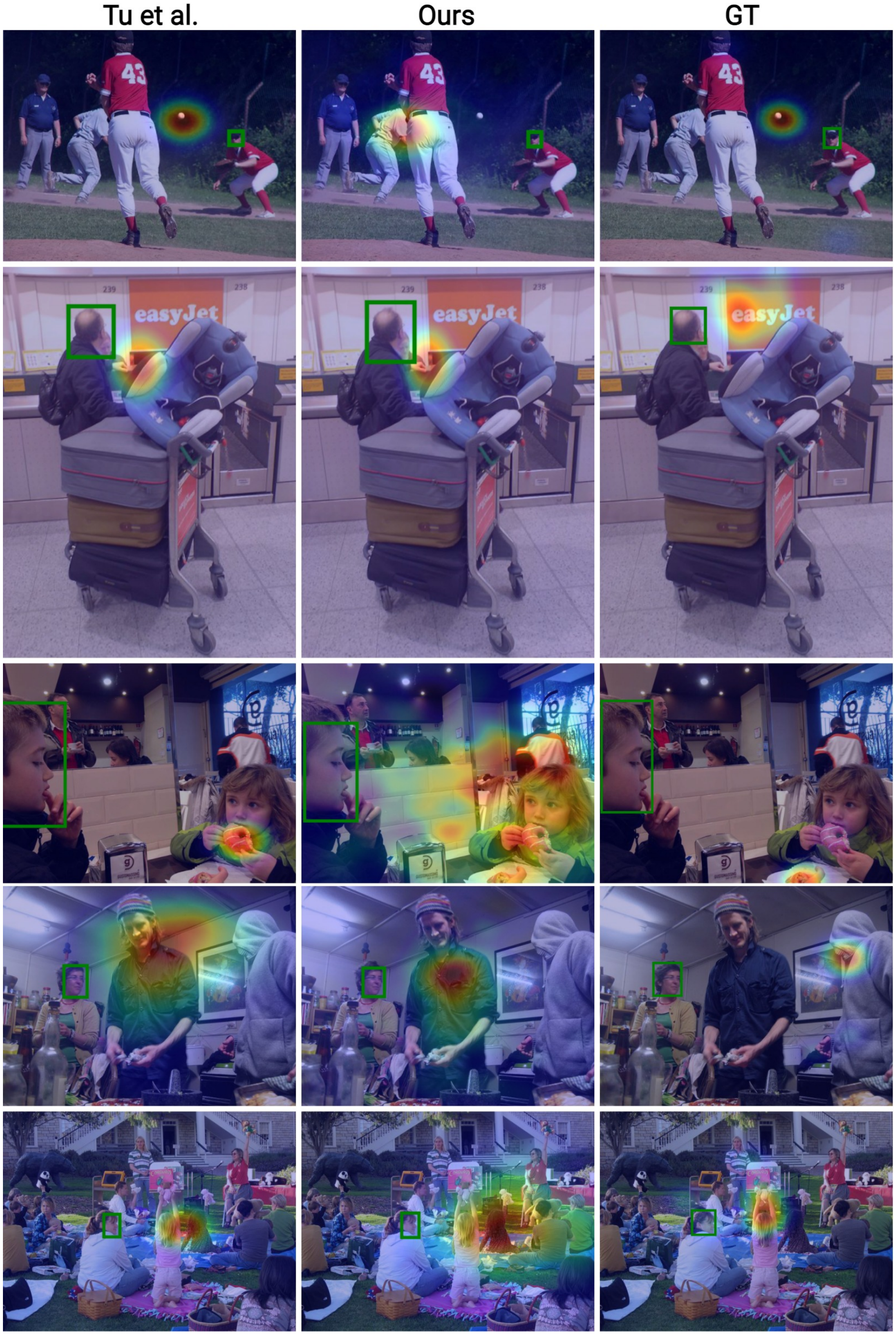

Figure 5: Qualitative results in which Our method performs relatively lower since the images are highly challenging due to several reasons (see text for details). Our method (center) and Tu et al. [12] (left) *w.r.t.* the ground-truth annotations of the GazeFollow [10] dataset (right). The green boxes show the person-in-interest for the ground truth while they are the detected head for Our and Tu et al. [12]. Even though our method can detect the gaze of multiple persons in the scene simultaneously, for better visualization, we plot a single person's predicted heatmaps and head location.

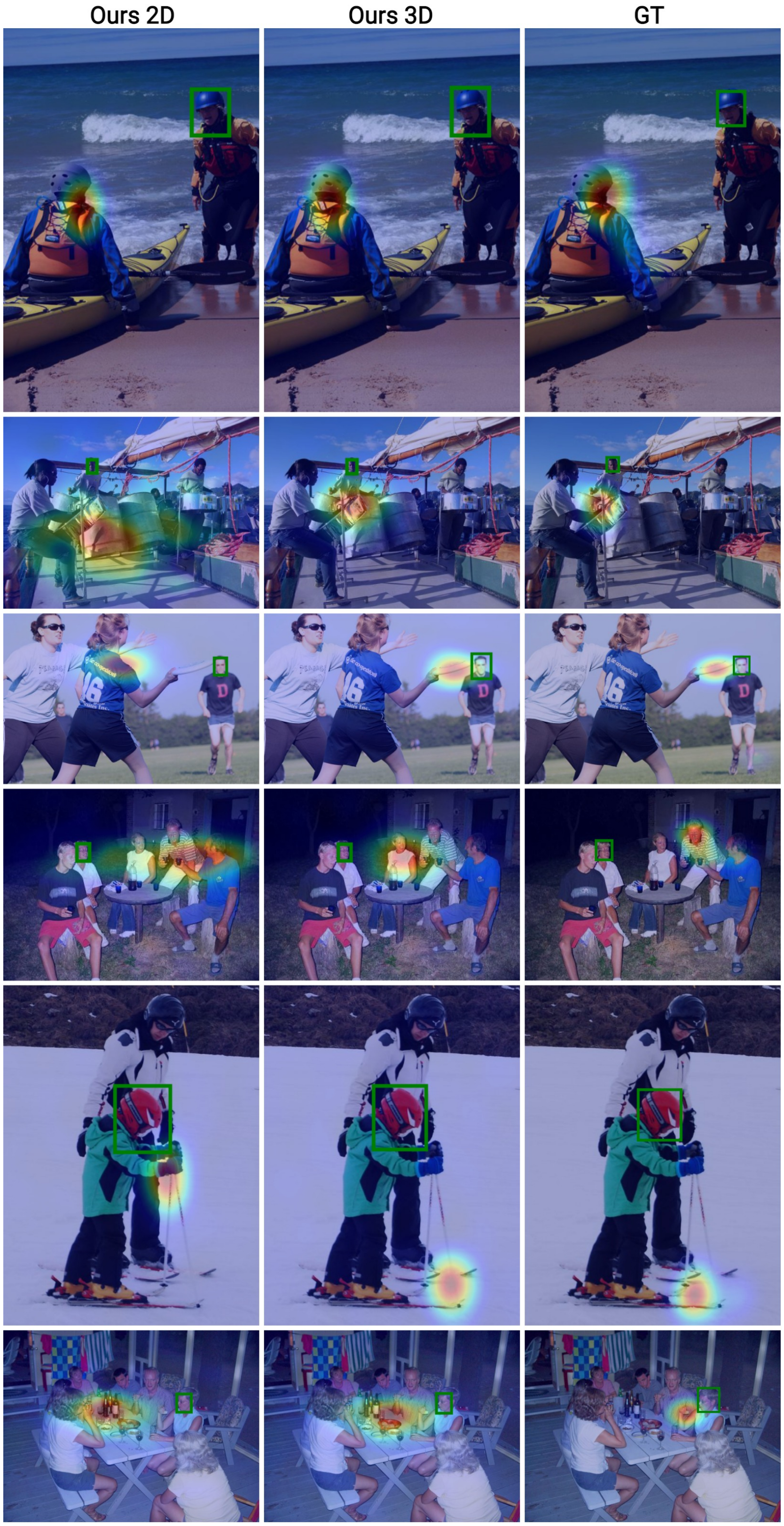

Figure 6: Qualitative results of our 2D (left) and 3D (center) method *w.r.t.* the ground-truth annotations of the GazeFollow [10] dataset (right), showing the importance of 3D-gaze cone building. Even though our method can detect the gaze of multiple persons in the scene simultaneously, for better visualization, we plot a single person's predicted heatmaps and head location.