Sharingan: A Transformer Architecture for Multi-Person Gaze Following

Supplementary Material

7. More Experiments

7.1. Qualitative Evaluation

We show multiple qualitative samples generated by our model in Figures 7 and 8, where we use our head detector to localize people. We can see that Sharingan performs well in different challenging situations and for all people in the scene (not only the annotated ones, which are often easier and in the foreground). This includes cases that

- Require depth reasoning (*e.g.* Figure 7 rows 3, 5, 7)
- Require understanding gestures (*e.g.* Figure 8 row 6)
- Have unusual camera angles (*e.g.* Figure 8 row 1)
- Involve complex social interactions (*e.g.* Figure 7 rows 2, 3, 4, and 6)
- Feature people seen from behind, where the face and the eyes are not visible (*e.g.* Figure 8 rows 1, 2, 5, 6)

The model is also able to capture social gaze behavior such as looking at people (*e.g.* Figure 8 rows 3, 5, and 7), and shared attention (*e.g.* Figure 7 rows 3 and 7). Finally, the heatmaps produced by our model successfully highlight other possible gaze targets in case of uncertainty (*e.g.* Figure 7 rows 2, 7 and Figure 8 rows 2, 5, 6).

Furthermore, we provide several examples of failure cases in Figure 9. We note that the model can fail at times in the presence of uncertainty: even if the heatmap captures the plausible targets, the $\arg\max$ might land on the wrong one (*e.g.* row 4). The model also seems to struggle with some unusual head poses and appearances. In row 2 for example, the gaze encoder only sees the hair from the top of the head, making it challenging to discern the body's orientation. In such cases, the predicted gaze vector is inaccurate, and so is the final prediction. This is also reflected in the heatmap which extends across half of the image. We believe that having access to the entire body pose of the person might prove useful in handling these situations. Moreover, the model might fail when the gaze target is completely occluded (*e.g.* row 5). This problem probably comes from the datasets themselves where annotated instances often correspond to visible targets. The authors of [35] proposed a gaze class to extend the traditional *in-vs-out* label, which incorporates a *gaze occluded* option. Having this prediction can help the user disregard these gaze instances, or deal with them separately (similar to the case when the person is looking outside the frame). Finally, Sharingan might fail when the gaze target selection requires complex reasoning, like when one person gazes at a distant object being pointed at by another person (*e.g.* row 3).

Method	AUC ↑	Avg. D. ↓	Min. D. ↓
Supervised	0.931	0.121	0.065
CLIP	0.923	0.139	0.080
MAE	0.931	**0.109**	**0.056**
MultiMAE	**0.944**	0.113	0.057

Table 6. Ablation results for the ViT pretraining.

7.2. ViT Pretraining

Given the limited size of the available benchmarks, all gaze following methods resort to pretraining instead of random initialization. In this section, we take a closer look at the influence of the pretraining strategy on the final performance of Sharingan. To this end, we compare different ViT initializations: 1. ImageNet-1k Supervised finetuning, 2. CLIP pretraining, 3. ImageNet-1k MAE, and 4. ImageNet-1k Multimodal MAE. The results are shown in Table 6. As expected, supervised classification doesn't translate as well to our dense prediction task compared to masked auto-encoding. Surprisingly, CLIP performs even worse. While the semantic information is useful to the task, we believe that the shortcoming of CLIP stems from its image-level representation while gaze following requires an object-level finer-grained understanding of the image. We also note that masked auto-encoding performs better overall, with the standard image-based MAE slightly outperforming its multimodal counterpart. However, MultiMAE seems to generalize better as evidenced by a cross-dataset evaluation on VideoAttentionTarget where we get a distance of 0.113 (MultiMAE) *vs* 0.117 (MAE).

7.3. Robustness to Inaccurate Head Boxes

As a two-stage approach, Sharingan requires access to head bounding boxes as input, typically obtained using off-the-shelf head detectors. However, the predicted head locations are naturally prone to inaccuracies. This raises the question of the model's robustness when provided with noisy head labels. To evaluate this aspect, we conducted an experiment where we jittered each head box coordinate in the test set of GazeFollow with *uniform* noise in $[-\alpha, \alpha]$ such that $\alpha = \beta.w_{\text{box}}$ and $\beta.h_{\text{box}}$ for x_i and y_i respectively. We find that the Avg. Dist. (averaged over multiple runs) for $\beta \in \{10\%, 20\%, 30\%\}$ only increased by 0.2%, 1.4%, and

Figure 7. Predictions of Sharingan on the VideoAttentionTarget and (test set of) GazeFollow datasets. The first column is the image, the second shows point predictions of all people, and the third is the heatmap of a randomly selected person. The model is trained on GazeFollow.

Figure 8. Predictions of Sharingan on the ChildPlay dataset. The first column is the image, the second shows point predictions of all people, and the third is the heatmap of a randomly selected person. The model is trained on GazeFollow.

Figure 9. Failure cases of Sharingan on the ChildPlay dataset. The first column is the image, the second shows point predictions of all people, and the third displays the heatmap of the person with an incorrect prediction. The model is trained on GazeFollow.

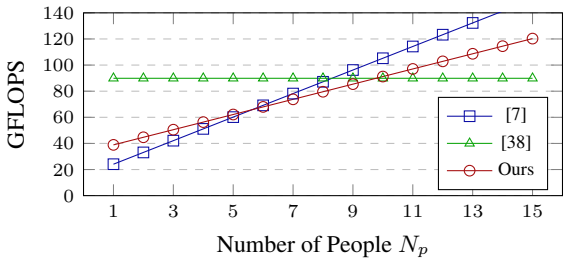

Figure 10. Comparative analysis of FLOPS *vs* number of people.

Method	Dist. ↓
Random	0.442
Chong [7]	0.138
Sharingan	**0.124**

Table 7. Cross-dataset performance on the DL Gaze dataset.

5.0% respectively. One important reason behind this robustness is our use of random noise to jitter head bounding box coordinates as a form of data augmentation during training.

7.4. Revisiting Model Efficiency

To further support our claim of efficiency, we provide a fair flops comparison with [7, 38] in Figure 10. Specifically, the flops count includes the head detection step for our model and [7], but not the depth extraction of [38]. We are better than [38] when $N_p <= 10$, and better than [7] when $N_p >= 5$.

7.5. Generalization

To assess the generalization robustness of our model, we tested it on other datasets and tasks related to gaze.

7.5.1 Gaze Following

First, we evaluate Sharingan (pretrained on GazeFollow) on the DL Gaze dataset [21], which records 16 volunteers performing several activities (*e.g.* talk, read, use a mobile phone) in 4 different indoor scenes (*i.e.* laboratory, working office, library, and corridor). The dataset contains 5526 frames annotated with 9481 gaze following instances. The images are generally very different from the ones in GazeFollow, and we use the distance metric for evaluation. The results, which are shown in Table 7, demonstrate our model's ability to generalize to other contexts.

7.5.2 Shared Attention

Next, we assess Sharingan's performance when the predicted heatmaps are processed to infer shared attention. To this end, we consider the test set of the VideoCoAtt dataset [10]. It contains 114100 test frames, 18101 (16%) of which contain shared attention instances. For each image, we predict the heatmaps of all people (*i.e.* annotated and automatically detected) and add them together. This image-based shared attention heatmap is used to evaluate two tasks: shared attention detection and shared attention localization.

For shared attention detection, the goal is to determine whether there is a shared attention instance happening in the frame. To do so, we simply find the maximum intensity value and consider it a positive prediction when it is above a certain threshold. The rationale is that if two or more people are looking at the same area, their cumulated heatmaps will result in a large peak. Since the heatmaps have a maximum value of 1, a perfectly predicted shared attention between 2 people means a maximum value of 2. In practice, it will be less than 2 because the points of maximum intensity of the two heatmaps will not perfectly align. Consequently, we report precision, recall, and f-score at a threshold of 1.6. We also vary the threshold between 1 and 2 to compute both the AUC and AP.

In terms of localization, the goal is to assess the distance between the predicted shared attention point (*i.e.* $\arg\max$ of the shared attention heatmap), and the ground truth (*i.e.* the center point of the annotated shared attention bounding box). In this case, we only consider the 18101 frames with a shared attention instance and use the standard distance metric computed at the original image resolution.

The results of this experiment are given in Table 8. Sharingan outperforms [7] on both tasks and all metrics except precision. Indeed, the model from [7] delivers slightly higher precision but performs significantly worse in terms of recall (*i.e.* 23-point difference).

Please note that a similar experiment was done in [7, 39], but we were not able to reproduce their results since the performance depends on the heads considered (*i.e.* [7] trained their own SSD head detector, and [39] predict both heads and gaze with their unified method). For a fair comparison, we tested both [7] and Sharingan using the same protocol outlined before. Unfortunately, the code and checkpoints from [39] are not available. Also, we chose to use AP, AUC, and F-score to evaluate shared attention detection because the dataset is heavily imbalanced (16-84 split) which makes the accuracy metric, as reported in [7, 39], not a suitable choice.

7.5.3 Mutual Gaze

Finally, we test the ability of our gaze following model to recognize mutual gaze behavior, *i.e.* whether two people are looking at each other. To this end, we use the test set of the UCO-LAEO dataset [25] which contains 2366 frames annotated with people's head bounding boxes and mutual

Method	Precision@1.6 ↑	Recall@1.6 ↑	F-score@1.6 ↑	AP ↑	AUC ↑	Dist. ↓
Random	—	—	—	—	—	186
Bias	—	—	—	—	—	108
Chong [7]	**54.50**	19.88	29.14	36.35	72.73	68
Sharingan	49.16	**43.56**	**46.19**	**42.96**	**81.20**	**55**

Table 8. Performance on the VideoCoAtt dataset for shared attention.

Method	Precision ↑	Recall ↑	F-score ↑
Random	45.76	49.90	47.74
Chong [7]	75.31	84.95	79.84
Sharingan	**78.45**	**92.23**	**84.79**

Table 9. Performance on the UCO-LAEO dataset for mutual gaze.

gaze instances. We predict gaze points for all annotated people in an image and consider pairwise instances between them. A predicted instance is considered positive if the gaze point of each person falls within the head bounding box of the other. We report the precision, recall, and f-score in Table 9. Once again, Sharingan outperforms the baselines by a significant margin across all metrics thereby marking its superiority.

Beyond the numbers, these experiments also serve to prove that Sharingan can be used to infer social gaze behavior simply by processing its output heatmaps according to the task. The qualitative results shown before also support this finding.

8. Discussion: One-Stage *vs* Two-Stage

Most previous works in gaze following solve the task using a two-stage approach where the first step is to detect people's heads and use them as input alongside the image to predict their gaze. Recently, authors from [39] and [38] attempted a one-stage end-to-end approach where the model takes only the image as input and regresses both people's head bounding boxes and their gaze heatmaps (among other things). The authors claim that this formulation is better, using efficiency and robustness as their main arguments. Aside from the difficulty of evaluating such methods through available benchmarks, we argue that multi-person two-stage approaches are more advantageous. First, we believe that person head detection is a solved task, so attempting to learn this is nothing short of reinventing the wheel. Incidentally, we found the head detector used in this paper to be extremely accurate, robust, and, even suitable for real-time applications (Yolo family). The only instances it seemed to miss were small background heads in low-quality images and uncommon head poses (*e.g.* child lying on the ground). Second, real-world gaze applications are often part of a larger system to analyze people's behaviors. For example, in the context of social robots interacting with individuals, people are typically already detected and tracked. The ability to exert control over the selection and presentation of subjects to the gaze model simplifies subsequent analysis and processing. In contrast, one-stage gaze methods require a matching step that is prone to errors and adds computation overhead. Moreover, implementations such as [38, 39] come with a hyperparameter for the maximum number of people they can handle, with a need for re-training to modify [39]. Instead, Sharingan can effortlessly accommodate a variable number of people without any changes. Finally, Sharingan is much easier and faster to train (*i.e.* 20 epochs on a single GPU for $\sim$10 hours *vs* 80 epochs on 8 GPUs for [39]).