# Believe It or Not, We Know What You Are Looking at!

Dongze Lian *, Zehao Yu*, and Shenghua Gao^
ShanghaiTech University

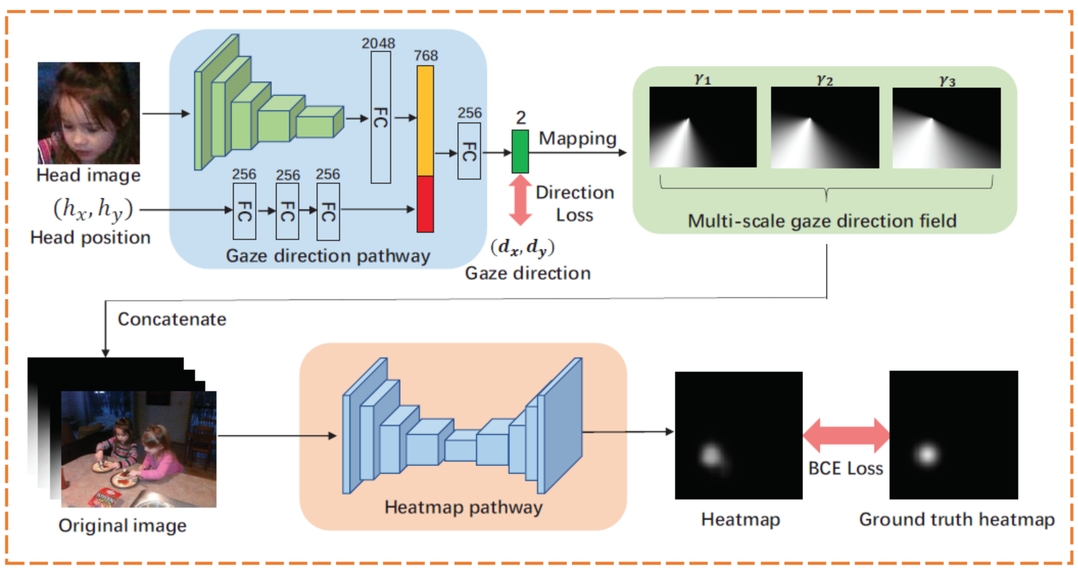
## Introduction:

- **Gaze Following:** Follow other people's gaze in a scene and infer where they are looking [figure (a) (b)].
- **Potential Applications:**
  - Understand the behavior of human;
  - New retailing scenario.
- **Challenges:**
  - Occluded head [figure (c)];
  - Ambiguity of gaze point [figure (d)].
- **Contributions:**
  - A psychological plausible two-stage solution;
  - Multi-scale gaze direction fields for different sizes of objects and various head positions;
  - A new video-based gaze following dataset;
  - State-of-the-art performance.

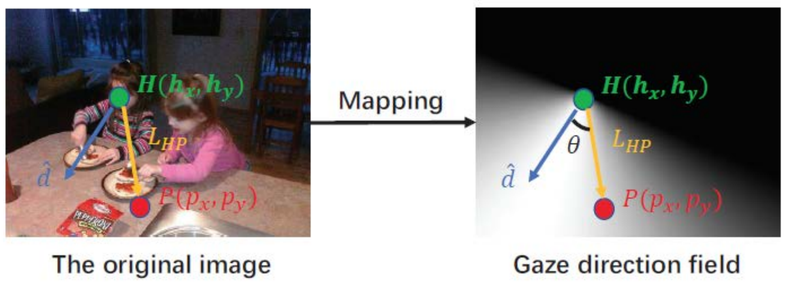

## Our approach:

- **Motivation:**
  - Mimic the behavior of a third-view person for gaze following.
- **Method:**
  - Stage-I: predict the gaze direction based on the head image and head position;
  - Stage-II: estimate heatmap based on the content information along the gaze direction.

## Network architecture:




- **Differentiable gaze direction field:**
  - Head position: $H = (h_x, h_y)$ , any point $P = (p_x, p_y)$ in image.
  - Direction from $H$ to $P$ : $G = (p_x - h_x, p_y - h_y)$;
  - The probability of point $P$ being the gaze point: $Sim(P) = \max(\frac{<G, \hat{d}>}{|G||\hat{d}|}, 0)$;
  - Multi-scale gaze direction field: $Sim(P, \gamma) = [Sim(P)]^{\gamma}$




- **Gaze direction loss:** $L_d = 1 - \frac{<d, \hat{d}>}{|d||\hat{d}|}$.
  $d$: ground truth gaze direction, $\hat{d}$: prediction.
- **Heatmap loss:** $L_h = -\frac{1}{N}\sum_{i=1}^{N} H_i \log(\widehat{H_i}) + (1 - H_i)\log(1 - \widehat{H_i})$,
  $H_i$: ground truth heatmap, $\widehat{H_i}$ : prediction.
- **Loss function:** $L = L_d + \lambda L_h$.

## Dataset:

- **Our Daily Life Gaze dataset (DL Gaze):**
  - 16 volunteers in 4 scenes (working office, laboratory, library and corridor in the building);
  - Gaze point is annotated by participants themselves;
  - 86 videos, 95K frames in total for evaluation.

## Experiments:

Dist: L2 distance, MDist: Minimum L2 distance, Ang: Angular error, MAng: Minimum Angular error.

**Table 1.** Performance comparison with existing methods on the GazeFollow dataset. One-scale and multi-scale correspond to the number of gaze direction fields in our model. For one-scale model, $\gamma = 1$.

| Methods | AUC | Dist | MDist | Ang | MAng |
|---|---|---|---|---|---|
| Center [22] | 0.633 | 0.313 | 0.230 | 49.0° | - |
| Random [22] | 0.504 | 0.484 | 0.391 | 69.0° | - |
| Fixed bias [22] | 0.674 | 0.306 | 0.219 | 48.0° | - |
| SVM + one grid [22] | 0.758 | 0.276 | 0.193 | 43.0° | - |
| SVM + shift grid [22] | 0.788 | 0.268 | 0.186 | 40.0° | - |
| Judd *et al.* [9] | 0.711 | 0.337 | 0.250 | 54.0° | - |
| SalGAN [19] | 0.848 | 0.238 | 0.192 | 36.7° | 22.4° |
| SalGAN for heatmap | 0.890 | 0.181 | 0.107 | 19.6° | 9.9° |
| Recasens *et al.* [22] | 0.878 | 0.190 | 0.113 | 24.0° | - |
| Recasens *et al.** [22] | 0.881 | 0.175 | 0.101 | 22.5° | 11.6° |
| One human [22] | 0.924 | 0.096 | 0.040 | 11.0° | - |
| Ours (one-scale) | 0.903 | 0.156 | 0.088 | 18.2° | 9.2° |
| Ours (multi-scale) | **0.906** | **0.145** | **0.081** | **17.6°** | **8.8°** |

**Table 2.** Performance comparison with existing methods on our dataset. Each frame only contains one gaze point, so only Dist and Ang are used for performance evaluation.

| Methods | Dist | Ang |
|---|---|---|
| Recasens *et al.* [22] | 0.203 | 26.9° |
| Recasens *et al.** [22] | 0.169 | 21.4° |
| Ours (multi-scale) | 0.157 | 18.7° |

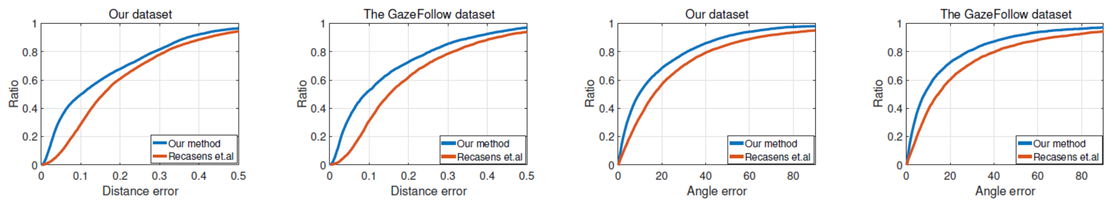
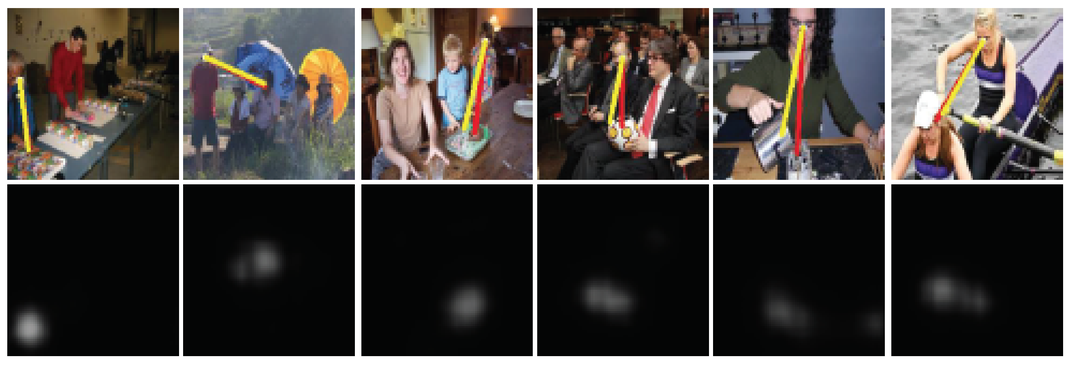



**Fig. 4.** Accumulative error curves of different methods on both datasets.

## Visualization of predicted results:

- **Predicted results:**

- **Predicted heatmaps:**



- **Some failures:**
  - Ambiguity and multimodal;
  - Small head or head occlusion, which is hard even for human;

**Reference:**
[1] Recasens*, A., Khosla*, A., Vondrick, C., Torralba, A.: Where are they looking? In: Advances in Neural Information Processing Systems (NIPS) (2015).

[2] Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1913-1921 (2015)

Code & dataset: https://github.com/svip-lab/GazeFollowing