

# Connecting Gaze, Scene, and Attention: Generalized Attention Estimation via Joint modeling of Gaze and Scene Saliency

Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg.

Georgia Institute of Technology. USA.

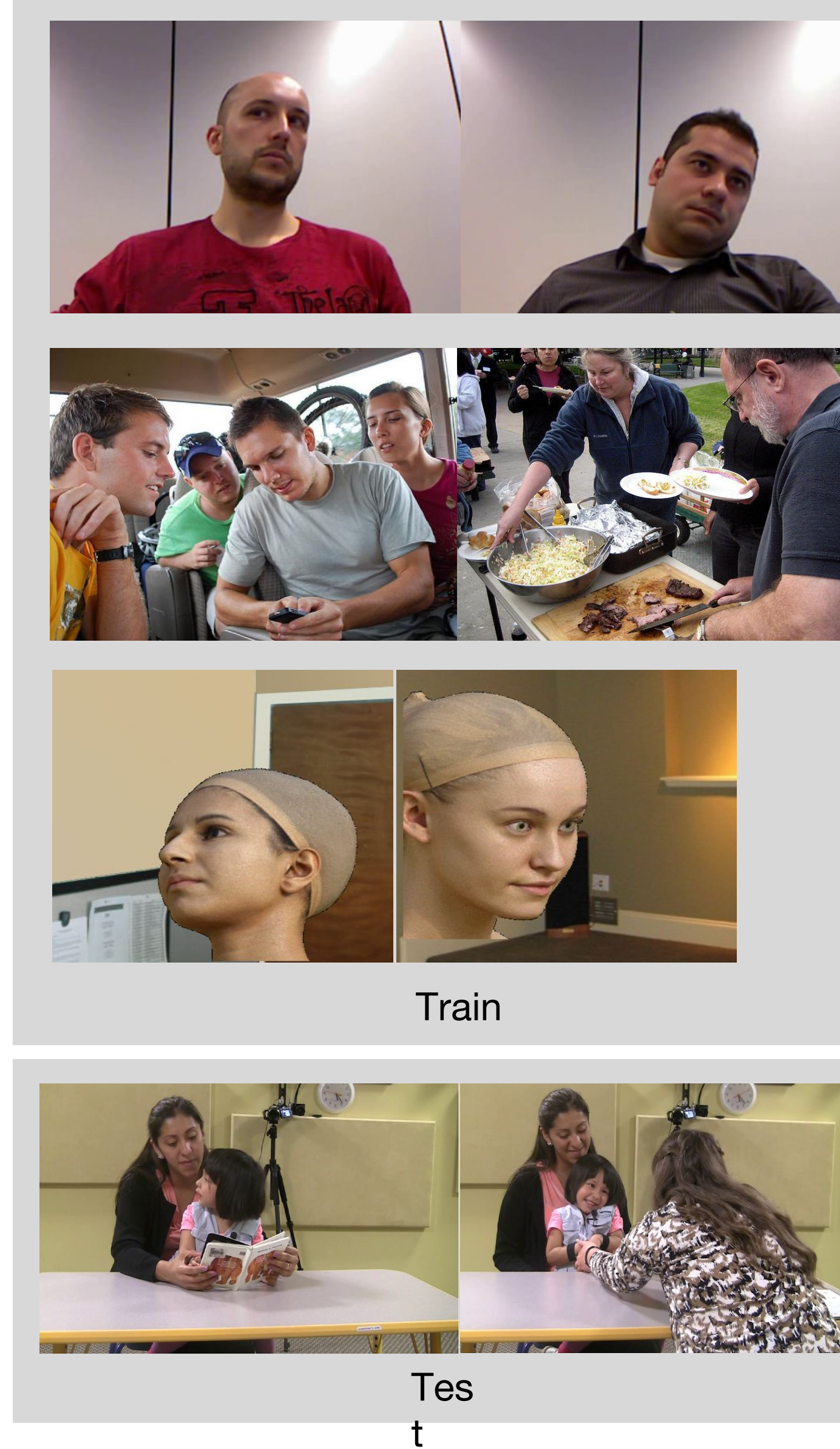
## Problem

- Human gaze behaviors are complex in the real world. **Automatically detecting and quantifying various types of visual attention from images** remains an open challenge.
- Current systems have focused on constrained versions of this problem in predetermined contexts.

## Our Approach

- We propose the new problem of **“generalized attention estimation”** and **design a system that can model the visual attention of subjects in unconstrained scenarios** which works across most natural scenarios.
- We exploit three public datasets that have been originally collected for different tasks to solve this problem.

## Dataset



- EYEDIAP dataset [1]
- Gaze and head pose variance
  - Target outside, clean background
  - Learn precise gaze angle representation
- GazeFollow dataset [2]
- Real world pictures
  - Target inside (with our additional annotation)
  - Learn gaze-relevant scene saliency representation
- SynHead dataset [3]
- Large head pose variation
  - Target outside, arbitrary background
  - Complement the other datasets
- MMDB dataset [4]
- Naturalistic social interactions
  - Frame-level annotations of subject's visual targets (with our additional annotation) among many other nonverbal behaviors

## Result



GazeFollow dataset (test split)

MMDB dataset

Table 1 Gaze-saliency evaluation on the GazeFollow test set

| Method          | AUC          | L2 Distance  | Min Distance |
|-----------------|--------------|--------------|--------------|
| Random          | 0.504        | 0.484        | 0.391        |
| Center          | 0.633        | 0.313        | 0.230        |
| Judd [17]       | 0.711        | 0.337        | 0.250        |
| GazeFollow [23] | 0.878        | 0.190        | 0.113        |
| Our             | <i>0.896</i> | <i>0.187</i> | <i>0.112</i> |

Table 3 Evaluation on MMDB - gaze target grid classification

| Grid Size | Method          | Precision | Recall |
|-----------|-----------------|-----------|--------|
| 2x2       | GazeFollow [23] | 0.344     | 0.715  |
|           | Our             | 0.744     | 0.851  |
| 5x5       | GazeFollow [23] | 0.210     | 0.437  |
|           | Our             | 0.614     | 0.683  |

Table 4 Evaluation of fixation likelihood on MMDB

| Method                                                    | Average Precision |
|-----------------------------------------------------------|-------------------|
| SVM with GazeFollow [23]                                  | 0.311             |
| SVM with GazeFollow [23]+gaze [32]                        | 0.531             |
| SVM with GazeFollow [23]+headpose [2]                     | 0.620             |
| SVM with gaze [32]+headpose [2]                           | 0.405             |
| SVM with GazeFollow [23]+gaze [32]+headpose [2]           | 0.624             |
| Random Forest with GazeFollow [23]                        | 0.707             |
| Random Forest with GazeFollow [23]+gaze [32]              | 0.727             |
| Random Forest with GazeFollow [23]+headpose [2]           | 0.785             |
| Random Forest with gaze [32]+headpose [2]                 | 0.512             |
| Random Forest with GazeFollow [23]+gaze [32]+headpose [2] | 0.773             |
| Our, trained only with GazeFollow dataset                 | 0.737             |
| Our, trained only with GazeFollow and EYEDIAP dataset     | 0.820             |
| Our final                                                 | <i>0.902</i>      |

- Our model achieves state-of-the-art result on the gaze following task, which consists in identifying the location of the scene the subject is looking at.

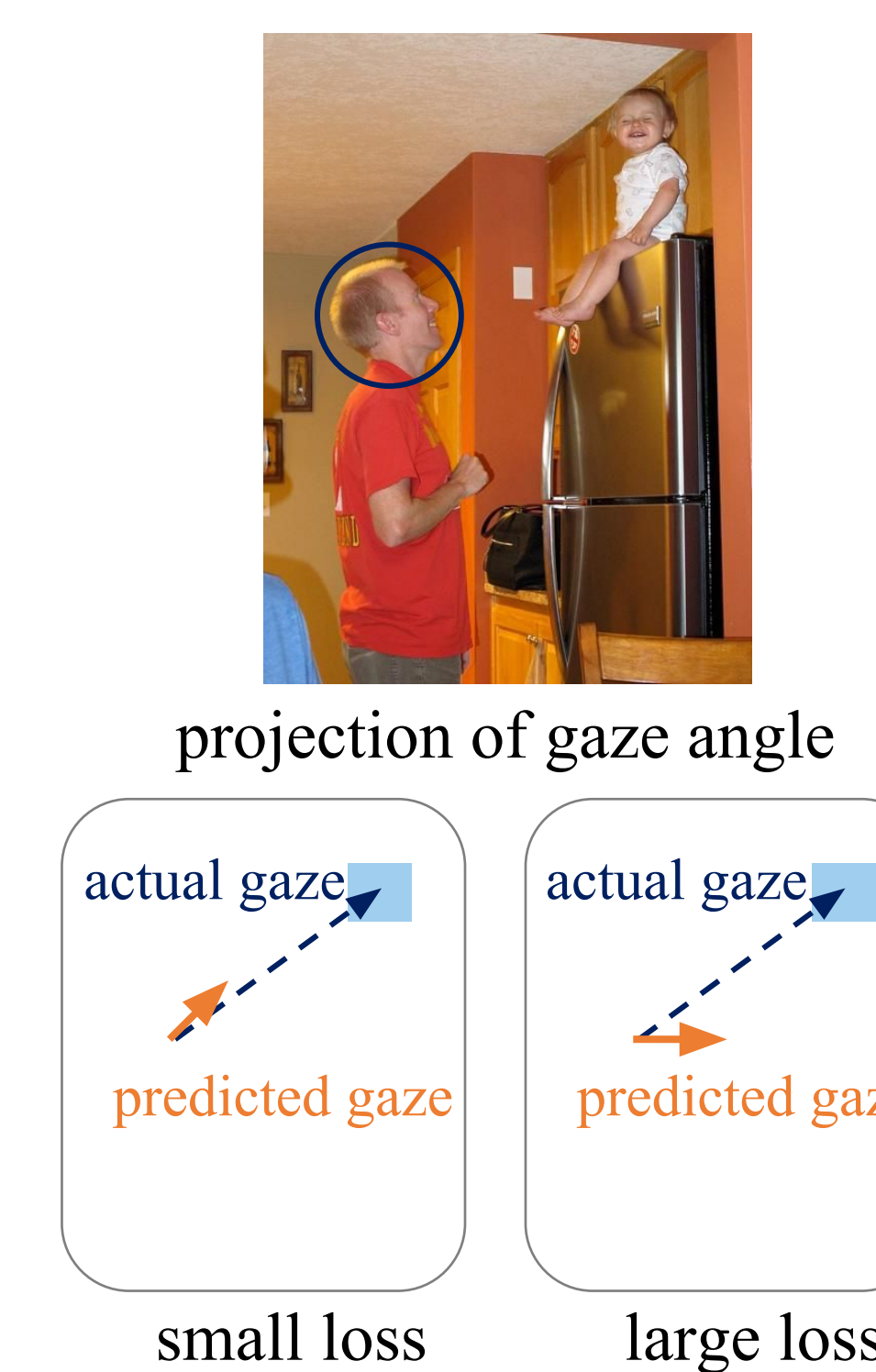
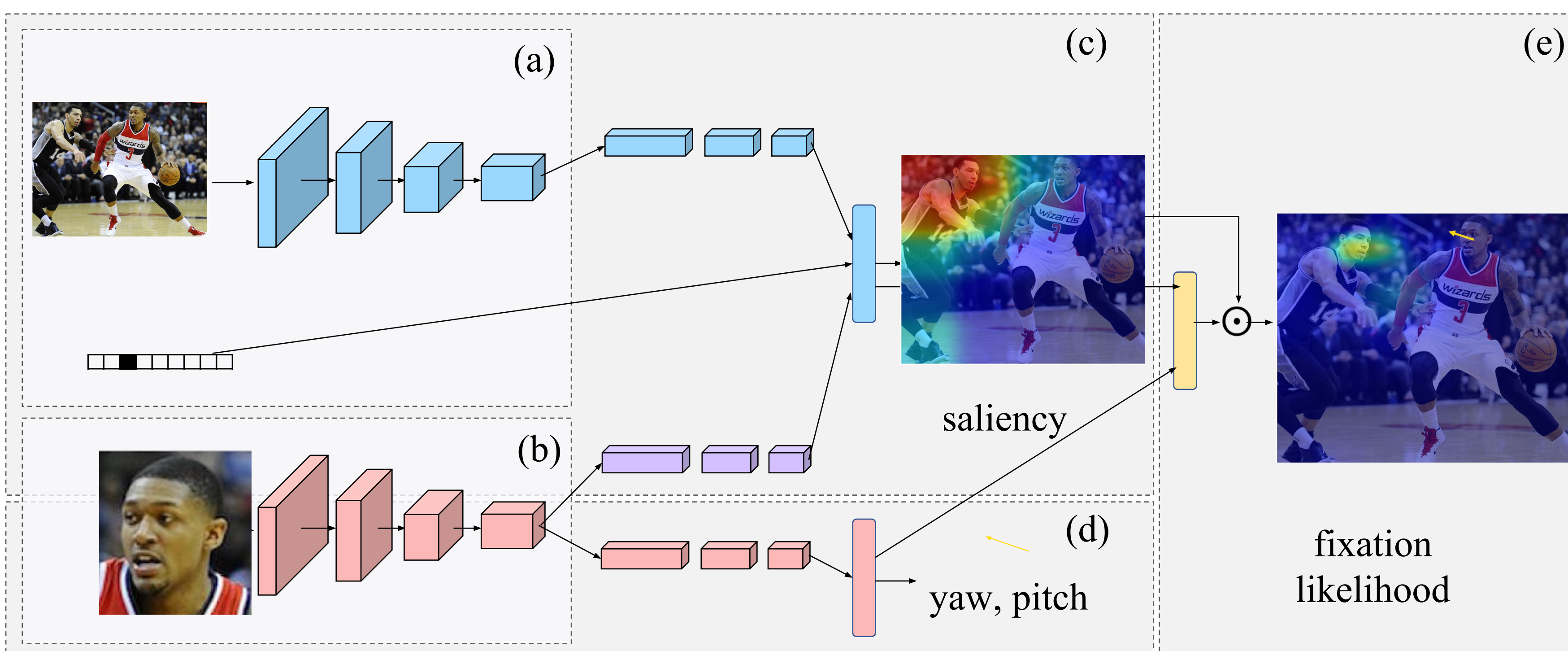
Table 2 Gaze angle evaluation on EYEDIAP

| Method        | Angular Error (degree) |
|---------------|------------------------|
| Wood [29]     | 11.3°                  |
| iTracker [18] | 8.3°                   |
| Zhang [32]    | 6.0°                   |
| Our           | 6.4°                   |

- Our model also competes with the state-of-the-art on the 3D gaze estimation task

- We evaluate our full model on a new challenging task on the MMDB dataset.
- We are the first to report attention estimation results on this problem. We compare our results to a variety of baseline tests.

## Method



- Project-and-Compare
- When angle is not directly available, a projection of its predicted angle is used as a measure of loss.

- Input = full scene image, a person's face location whose visual attention we want to predict, and the close-up face image.
- Scene and face images go through separate convolutional layers in such a way that (a) (b) and (c) contribute to saliency, and (b) and (d) contribute to gaze angle prediction. In the last layer, the final feature vectors for two tasks are combined to estimate how likely the person is actually fixating at a gaze target in the frame.
- Loss = Cross Entropy + Euclidean + Project-and-Compare.

## References

- [1] Mora, et al. EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras. ETRA 14.
- [2] Recasens, et al. Where are they looking? NIPS 15.
- [3] Gu, et al. Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network. CVPR 17.
- [4] Rehg, et al. Decoding Children's Social Behavior. CVPR 13.