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
School of Interactive Computing, Georgia Institute of Technology

In ECCV 2018

2019.08.19
Hanyang univ. AILAB 정지은

Introduction

Introduction

- Gaze-following task

Where are they looking?(NIPS 2015)

Introduction

fixation on an in-frame object | looking outside the frame | looking at the camera

Related work

Related work

- Gaze Estimation
    - Gaze estimation aims to predict the gaze of a human subject

Related work

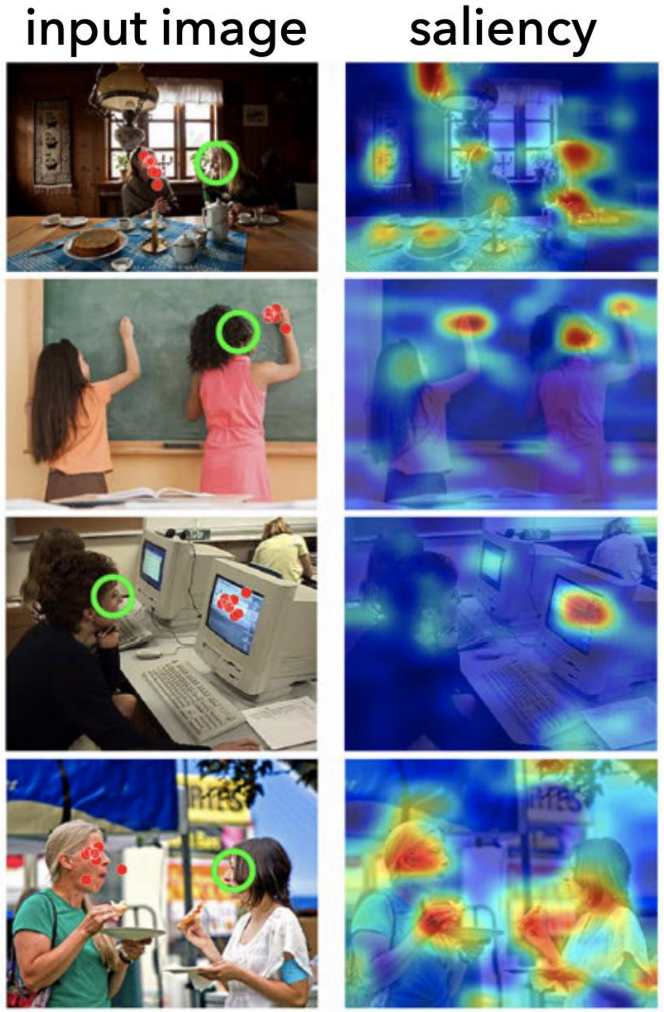
- Gaze Estimation
  - Gaze estimation aims to predict the gaze of a human subject

gaze vector using OpenFace2.0

Related work

- Gaze Estimation
  - Gaze estimation aims to predict the gaze of a human subject
- Visual Saliency
  - The objective of visual saliency prediction is to estimate locations in an image which attract the attention of humans looking at the image

Related work



Where are they looking?(NIPS 2015)

Related work

- Gaze Estimation
  - Gaze estimation aims to predict the gaze of a human subject
- Visual Saliency
  - The objective of visual saliency prediction is to estimate locations in an image which attract the attention of humans looking at the image
- Gaze Following
  - Given a single image containing one or more people, predict the location that each person in the scene is looking at

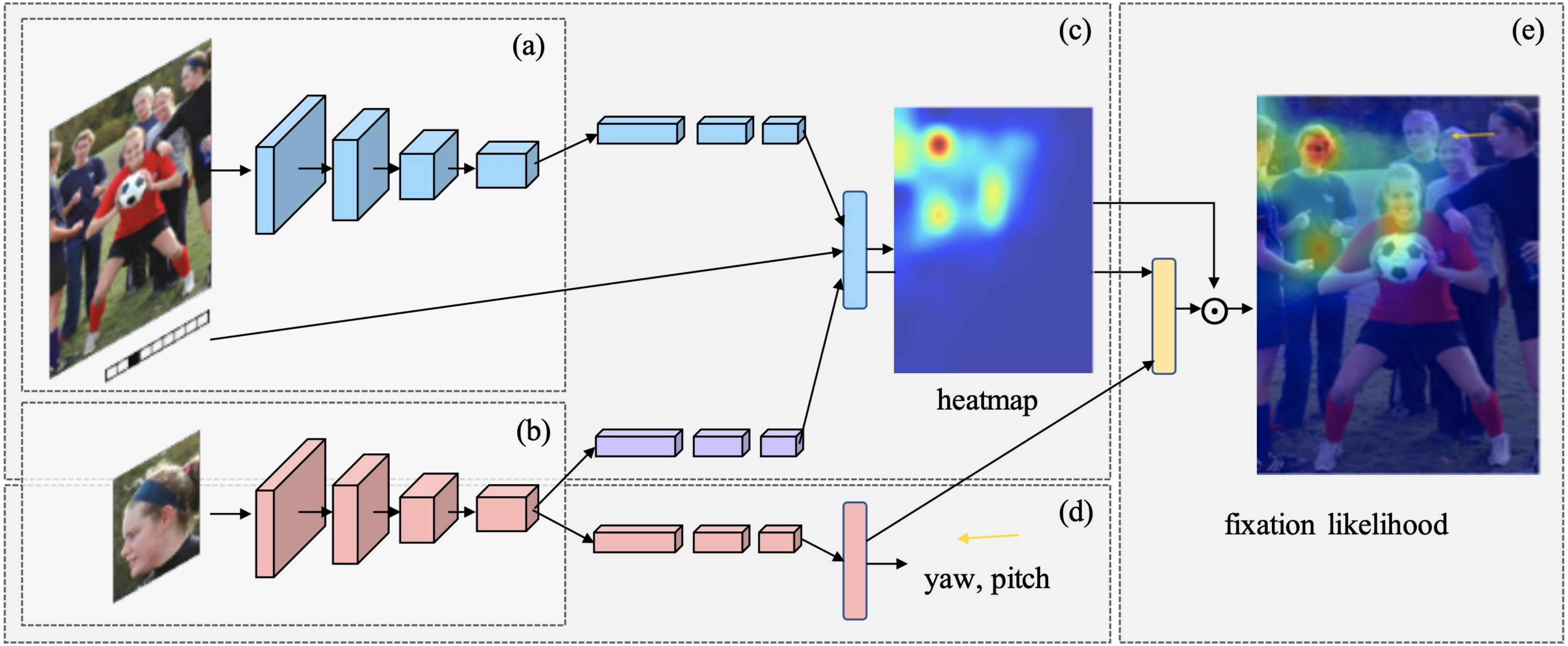
Related work

- Gaze Estimation
  - Gaze estimation aims to predict the gaze of a human subject
- Visual Saliency
  - The objective of visual saliency prediction is to estimate locations in an image which attract the attention of humans looking at the image
- Gaze Following
  - Given a single image containing one or more people, predict the location that each person in the scene is looking at
- Attention Modeling
  - We explicitly consider the gaze of the subject.

Method

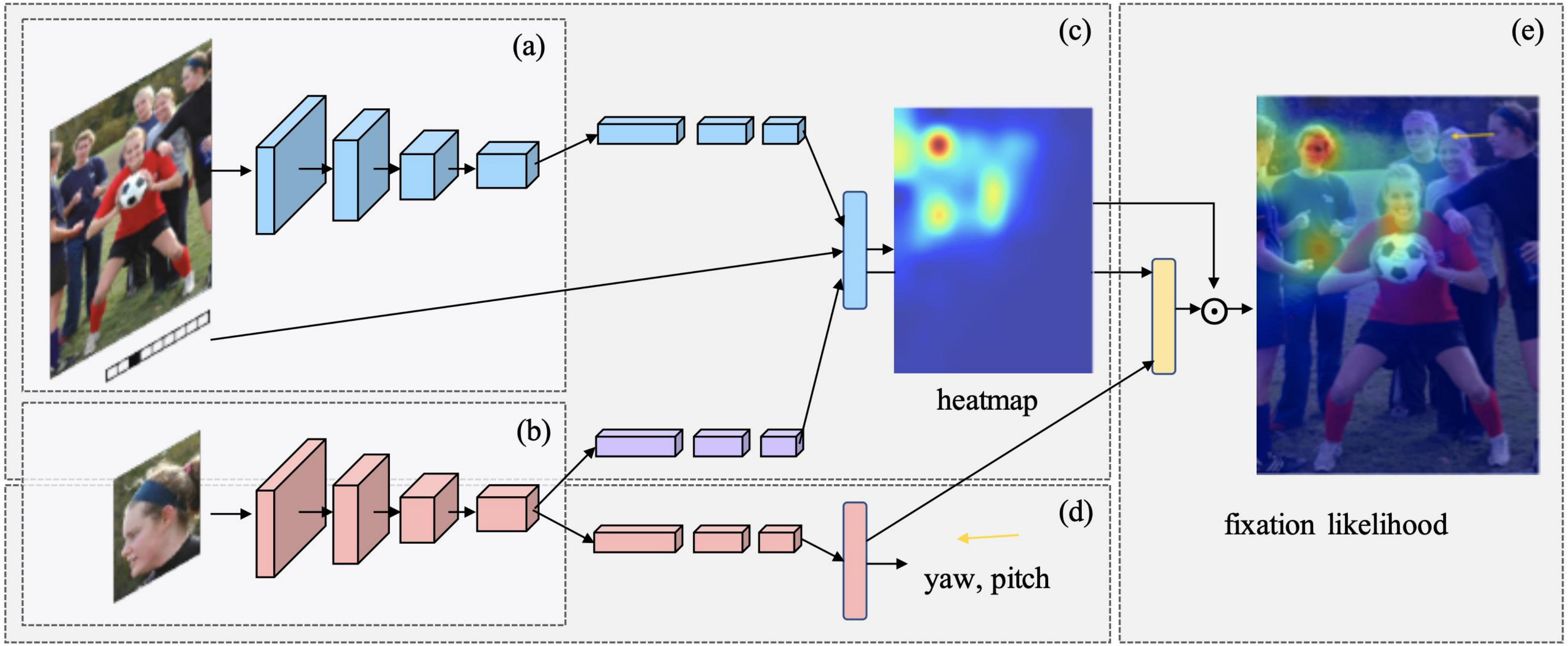
Method : Why separate the path?

- When we interpret a person's attention from an image, we infer their gaze direction and consider whether there are any salient objects in the image along the estimated direction



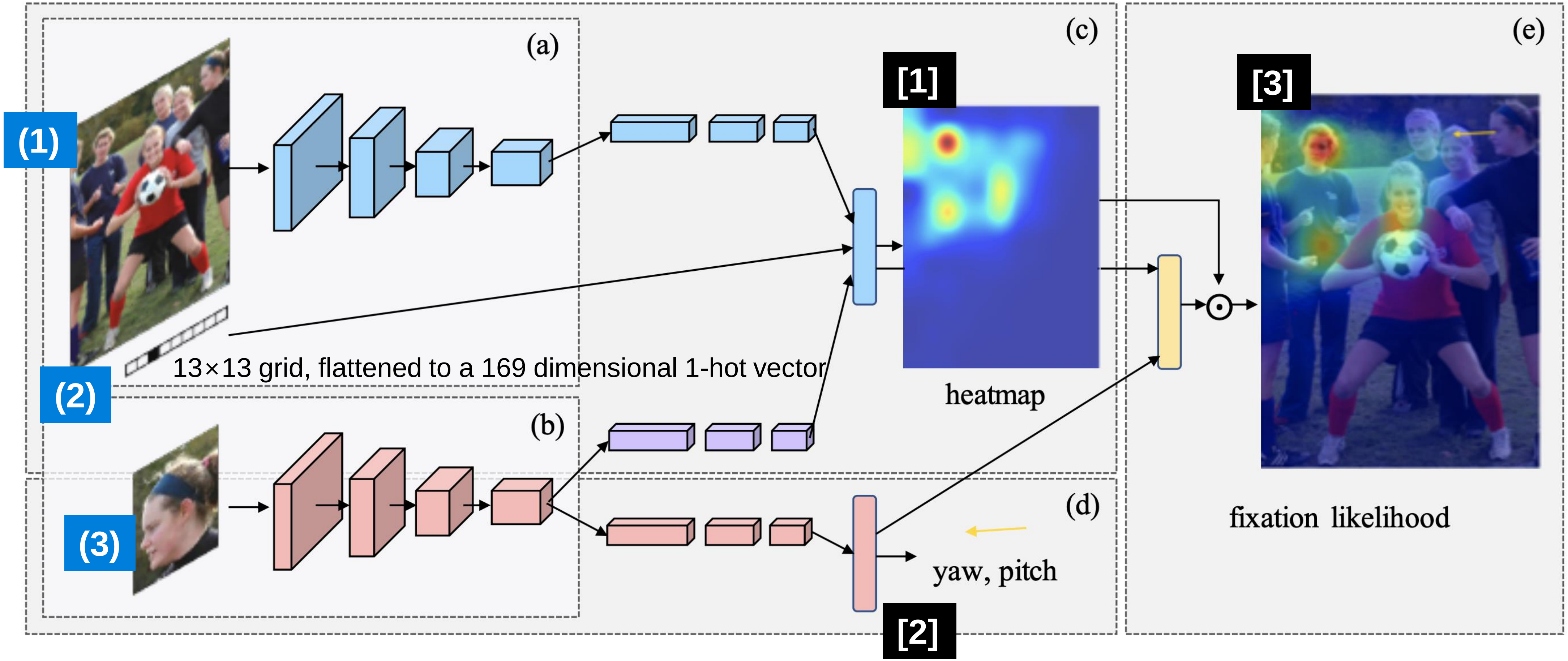
Method : Paths

- Path (a), (b) : Resnet50 (pre-trained on the ImageNet classification task)
- Path (c) : 2 conv pathways learn the heatmap
- Path (d) : training for the gaze angle
- Path (e) : learn “strength” of visual attention



Method : INPUTs & OUTPUTs

- 3 inputs : (1) the whole image (2) the location of the subject's face (3) a crop of the subject's face
- 3 outputs : [1] person-centric saliency map [2] gaze estimation (yaw, pitch) [3] fixation likelihood



Method : Cross-Domain Datasets

- No single dataset contains all of the information that we need to train the full model
- Leverage three different datasets, GazeFollow , EYEDIAP , and SynHead

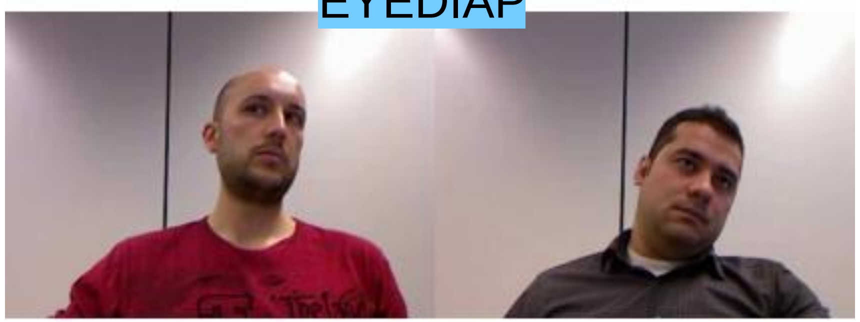
GazeFollow

- a real-world image dataset with manual annotations of the locations where people are looking
- collected <u>10 gaze annotations</u> per person for the test set
- BUT <u>actual 3D gaze angles are not available</u>
- we added additional annotations to this dataset in the form of a binary indicator label for “looking inside” or “looking outside” for every image.

Method : Cross-Domain Datasets

- No single dataset contains all of the information that we need to train the full model
- Leverage three different datasets, GazeFollow , EYEDIAP , and SynHead

EYEDIAP

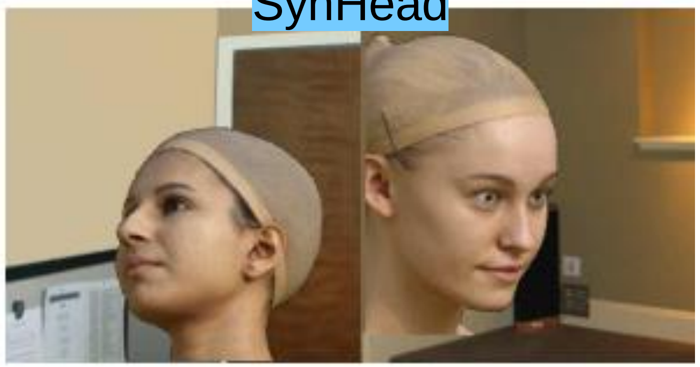


- for the evaluation of the gaze estimation task
- measured gaze angles range between -40° to 40°

Method : Cross-Domain Datasets

- No single dataset contains all of the information that we need to train the full model
- Leverage three different datasets, GazeFollow , EYEDIAP , and SynHead

SynHead



- for the head pose estimation task
- we use the labeled 3D head pose as the gaze angle ground truth
- the angle ranges are larger (between −90◦ and 90◦)
- include more diverse backgrounds
- SynHead entirely for training (head pose is not our task)

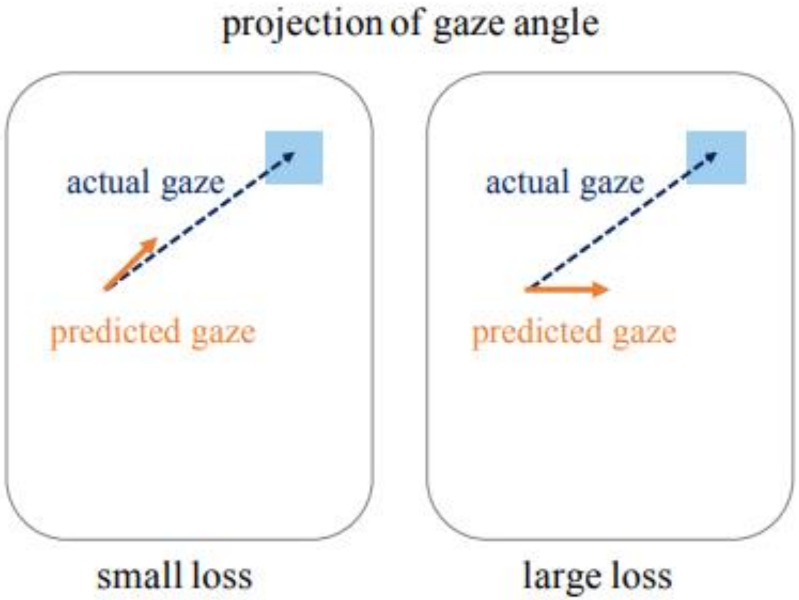
Method : Cross-Domain Datasets

Dataset	Training set	in vs out	Test set	in vs out
GazeFollow [23]	125,557	88.4% vs 11.6%	4,782	100% vs 0%
EYEDIAP [11]	72,613	0% vs 100%	18,153	0% vs 100%
SynHead [13]	75,400	0% vs 100%	-	-
MMDB [25]	-	-	4,965	41.4% vs 58.6%

MMDB(for test)

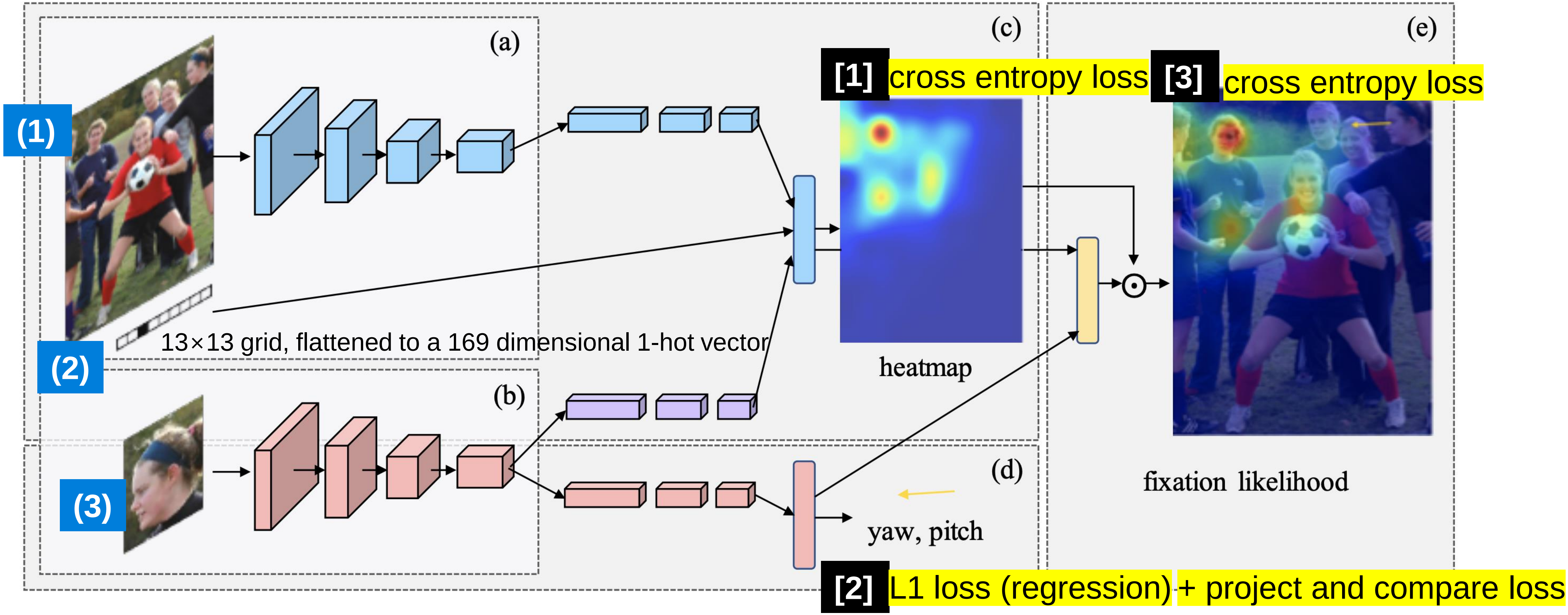
Method : Loss

- **Gaze angle regression** : L1 loss
- **Heat map & Fixation likelihood** : cross entropy loss
- **Project and Compare Loss :** cosine distance



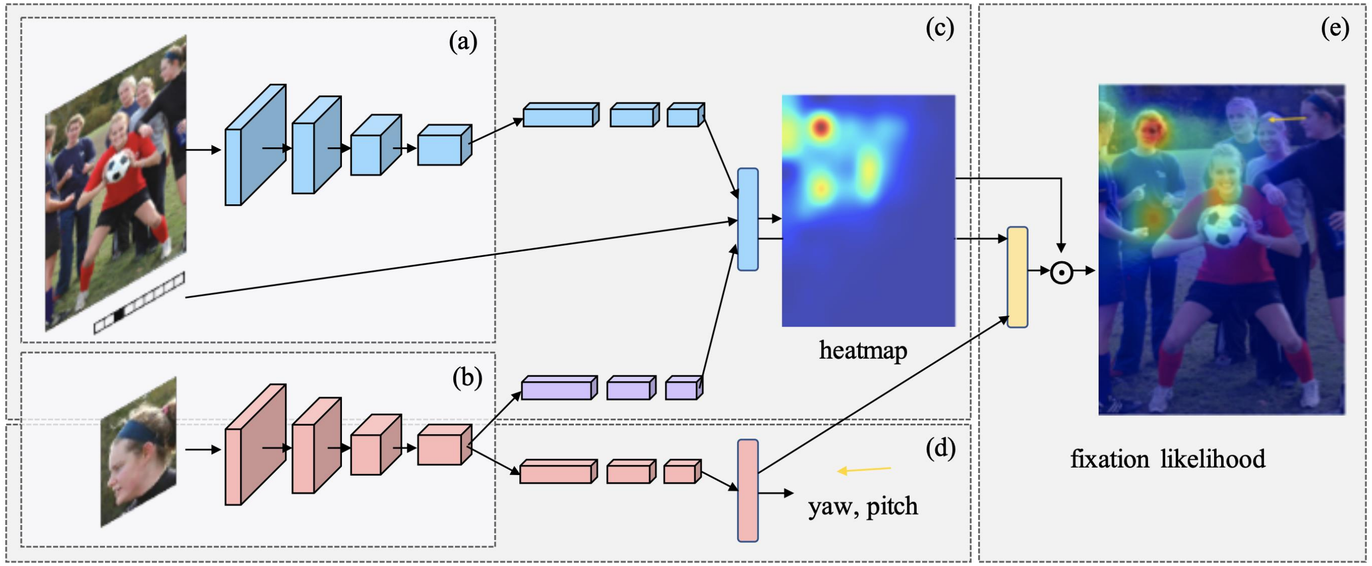
Method : INPUTs & OUTPUTs & Loss

- 3 inputs : (1) the whole image (2) the location of the subject's face (3) a crop of the subject's face
- 3 outputs : [1] person-centric saliency map [2] gaze estimation (yaw, pitch) [3] fixation likelihood



Method : Training Procedure

- **Only update the relevant parts** of the network based on which dataset the training sample is from, while freezing other irrelevant layers during back-propagation
  - When learning **gaze angle estimation**, only update the angle pathway **(b) and (d)**
  - When learning **saliency**, update the scene pathway **(a), (b) and (c)** while freezing all other layers
  - When training **fixation likelihood**, only update the layer **(e)**



Evaluation

Evaluation

- (1) Evaluate the person-dependent saliency map
- (2) Gaze angle estimation(prediction)
- (3) General attention estimation
- (4) Evaluate our method by changing model architectures and training dataset

Evaluation : (1) person-dependent saliency map

Evaluation : (1) person-dependent saliency map

- Evaluating Saliency Maps is AUC Metric
- If our model behaves perfectly, the AUC will be 1 while chance performance is 0.5 (**Higher is better)**

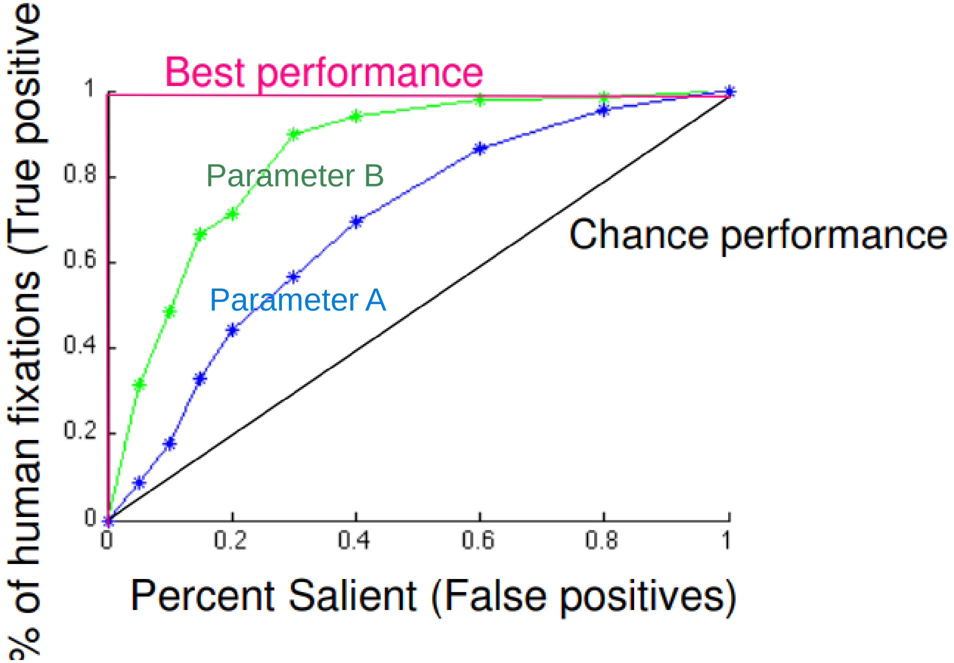
Table 2. Gaze-saliency evaluation on the GazeFollow test set

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	*0.896*	*0.187*	*0.112*

Supplement

- **AUC : Area under the ROC** (Receiver Operating Characteristic Curve)
- ROC shows the performance of the classification model

< ROC >



- Saliency map is thresholded to become a binary classifier
  - 1 if pixel at x,y over threshold, 0 otherwise
  - By varying the threshold we can get a ROC curve

0.2	0.1	0.4
0.2	0.7	0.3
0.2	0.7	0

Parameter threshold = 0.5

Critical Object

Evaluation : (2) Gaze Angle Prediction

- Yaw and pitch on the chosen EYEDIAP test split
- Our method is trained on multiple tasks
- All other methods are trained solely on the gaze angle prediction task

Table 3. Gaze angle evaluation on EYEDIAP

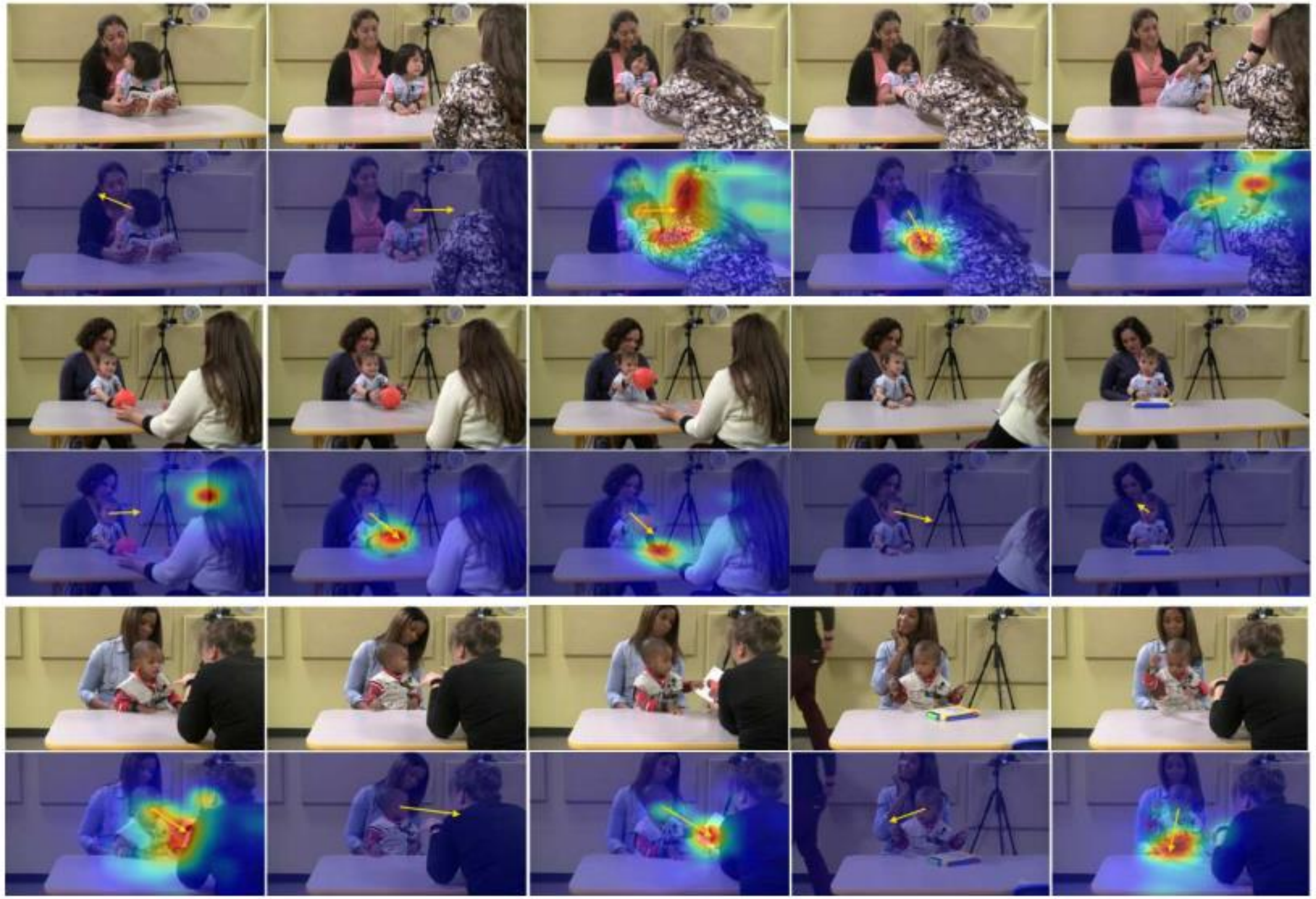
Method	Angular Error (degree)
Wood [29]	11.3°
iTracker [18]	8.3°
Zhang [32]	*6.0°*
Our	6.4°

Evaluation : (3) General attention estimation

- Limitations is the inability to correctly predict the “outside” case, where the subject is looking outside of the frame => Evaluate the method on the generalized attention prediction task
- **The MMDB dataset** is one of the largest datasets that contains children’s social and communicative behaviors
- Designed a gaze target grid classification task, where each test image is divided into N×N grids
- Using our method’s fixation likelihood map we predict the positive gaze grid square

MMDB(for test)

Evaluation : (3) General attention estimation



Evaluation : (3) General attention estimation

- GazeFollow [23] is base model

Table 4. 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

Evaluation : (4) Alternative Model and Diagnostics

- Omitting EYEDIAP or SynHead training dataset did not have much impact on the heatmap estimation
- Changing model architecture (Map resolution, ROI-pooling) considerably affected the scores

Table 6. Additional model evaluation and diagnostics on the GazeFollow test split

Method	AUC	L2 Distance
No EYEDIAP	0.887	0.197
No SynHead	0.895	0.191
No EYEDIAP and SynHead	0.891	0.194
No project-and-compare loss	0.895	0.189
Map resolution 15x15	0.778	0.194
ROI-pooling	0.700	0.325
Our final	*0.896*	*0.187*

Challenging cases

- When the target is within the frame but occluded by other object
- When the subject is closer to the camera than some salient object in the background

Fig. 7. Challenging cases due to occlusion and the lack of depth understanding.

Conclusion

Conclusion

- Proposed a multi-task learning approach and neural architecture leveraging three different datasets which tackles this problem and works across multiple naturalistic social scenarios
- Achieved state-of-the-art performance on the single-task gaze-saliency prediction
- Competed with state-of-the-art methods on gaze estimation benchmarks
- Achieved promising performance on the generalized attention prediction problem on the MMDB dataset

- **… We got to know about ‘Gaze-Following’ task**

Thank You