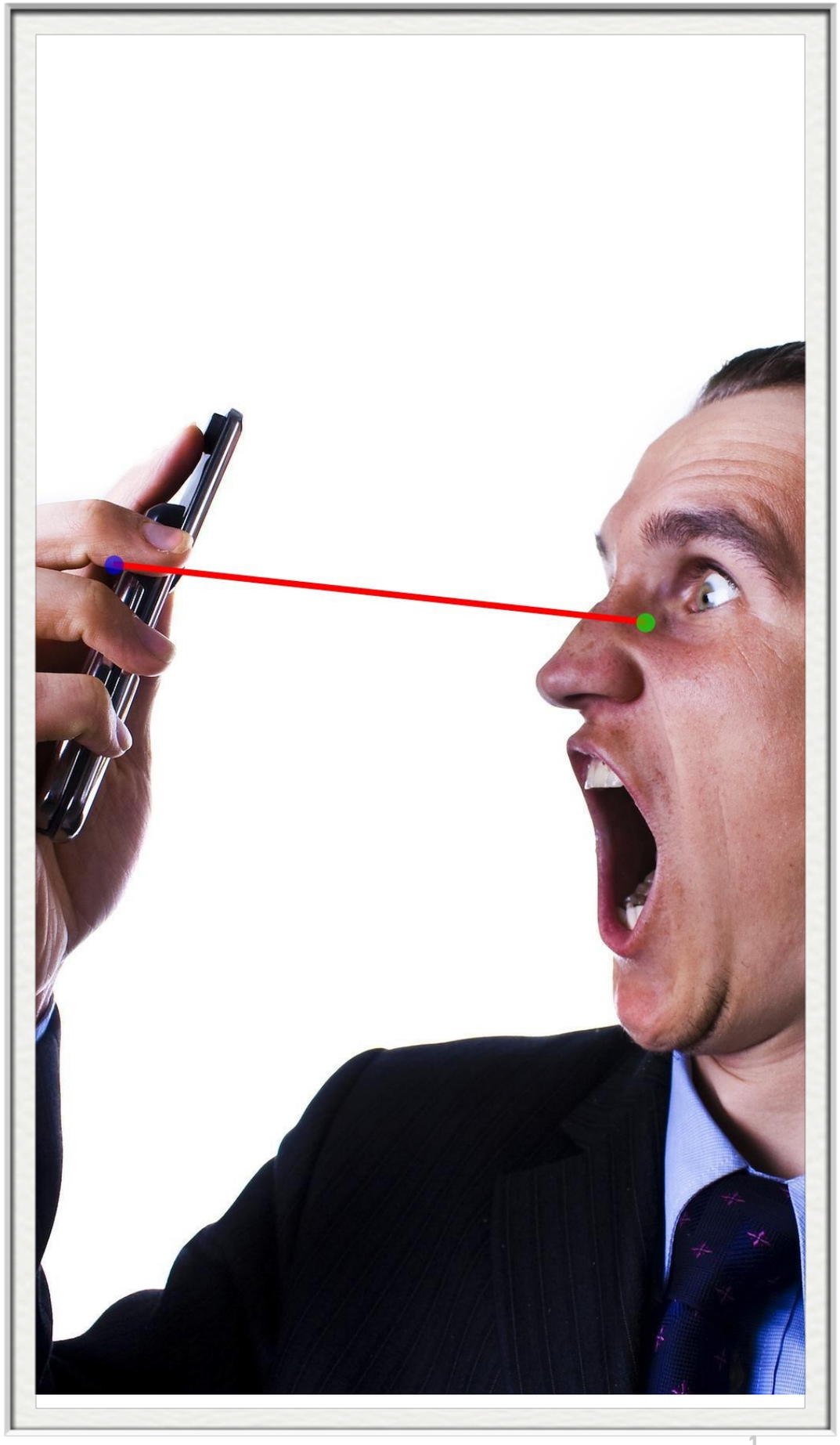
WHERE ARE THEY LOOKING?

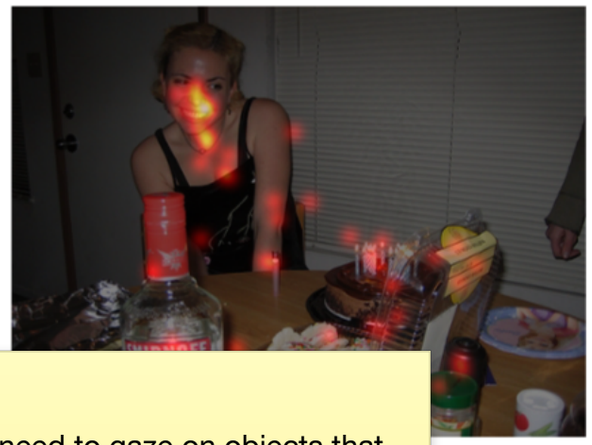
Adria recasens, MIT

Presenter:
Dongguang You



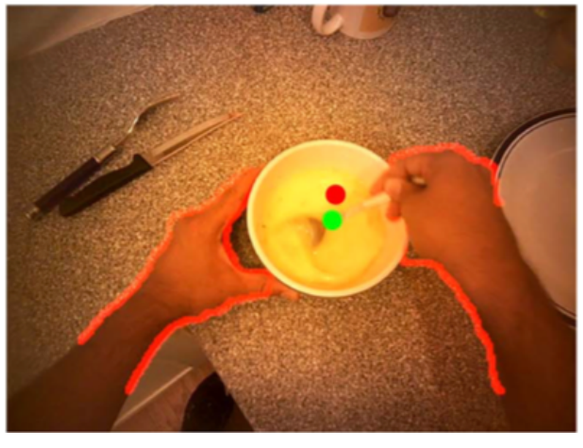
RELATED WORK

- The Secrets of Salient Object Segmentation
  - free-view saliency

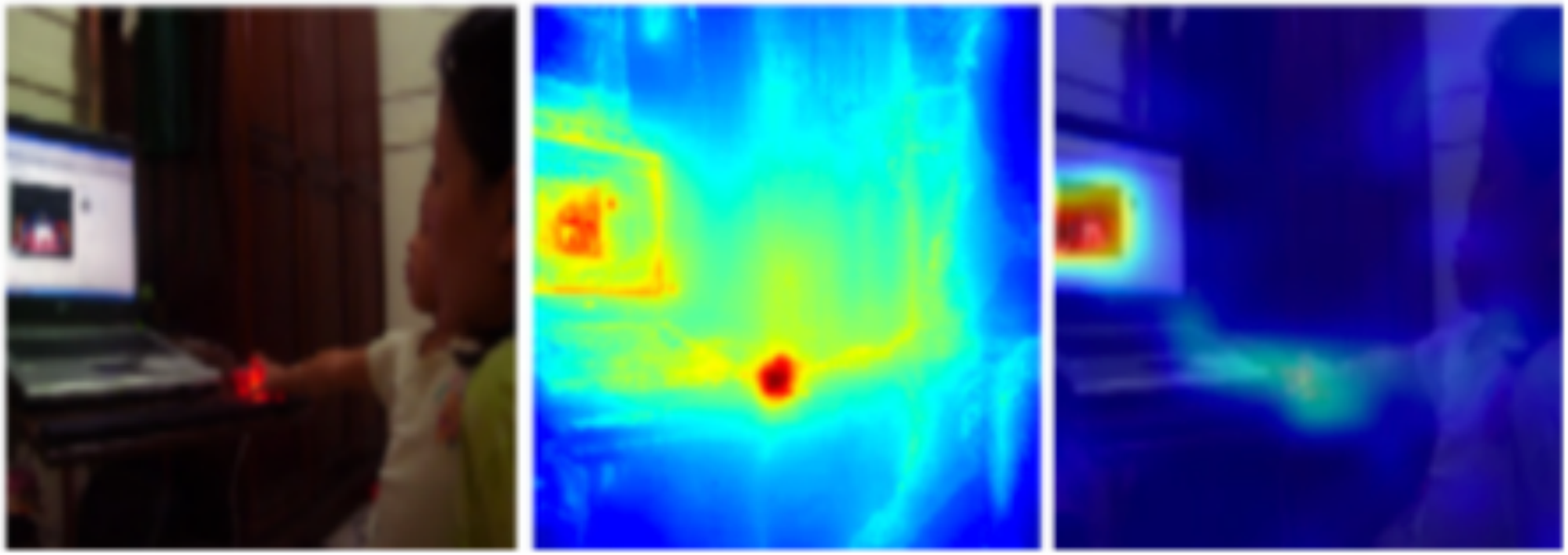


for gaze saliency, people may need to gaze on objects that are not visually salient to perform a task

- Learning to Predict Gaze in Egocentric Video
  - gaze saliency (gaze following)



THIRD PERSON VIEWPOINT



input

free-view saliency

gaze saliency

difference?

This illustrates how free-view saliency differs from gaze saliency.

Not only it doesn't consider gaze direction, but also it highlights some wrong objects, such as mouse with red light in the lower right image. People performing task in the picture may focus on objects that are not salient by free view

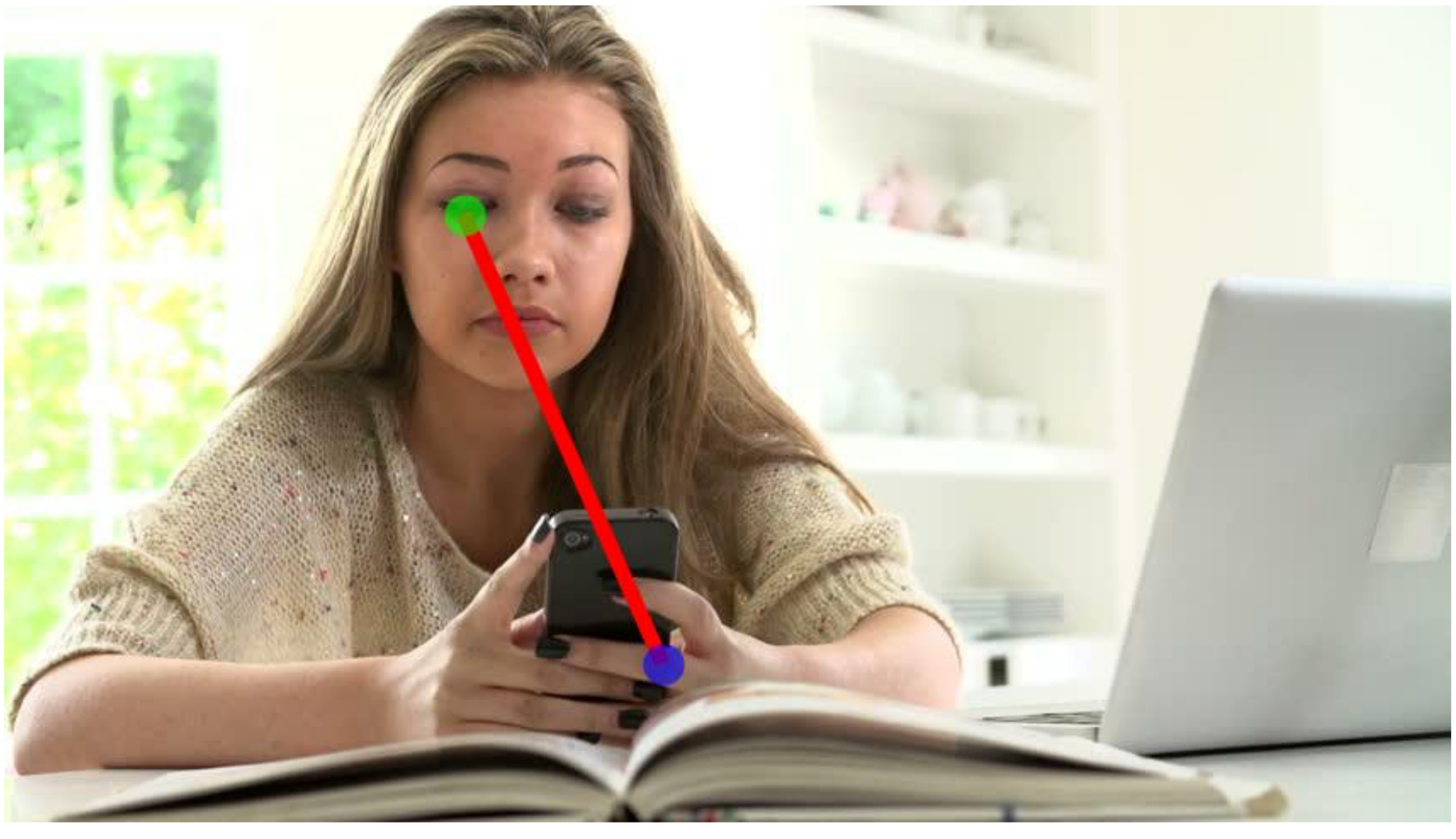
PROBLEM DEFINITION

- Predicting gaze saliency from a 3rd person viewpont
- Where are they looking at?

- Assumptions:
  - 2-D head positions are given
  - people are looking at objects inside the image

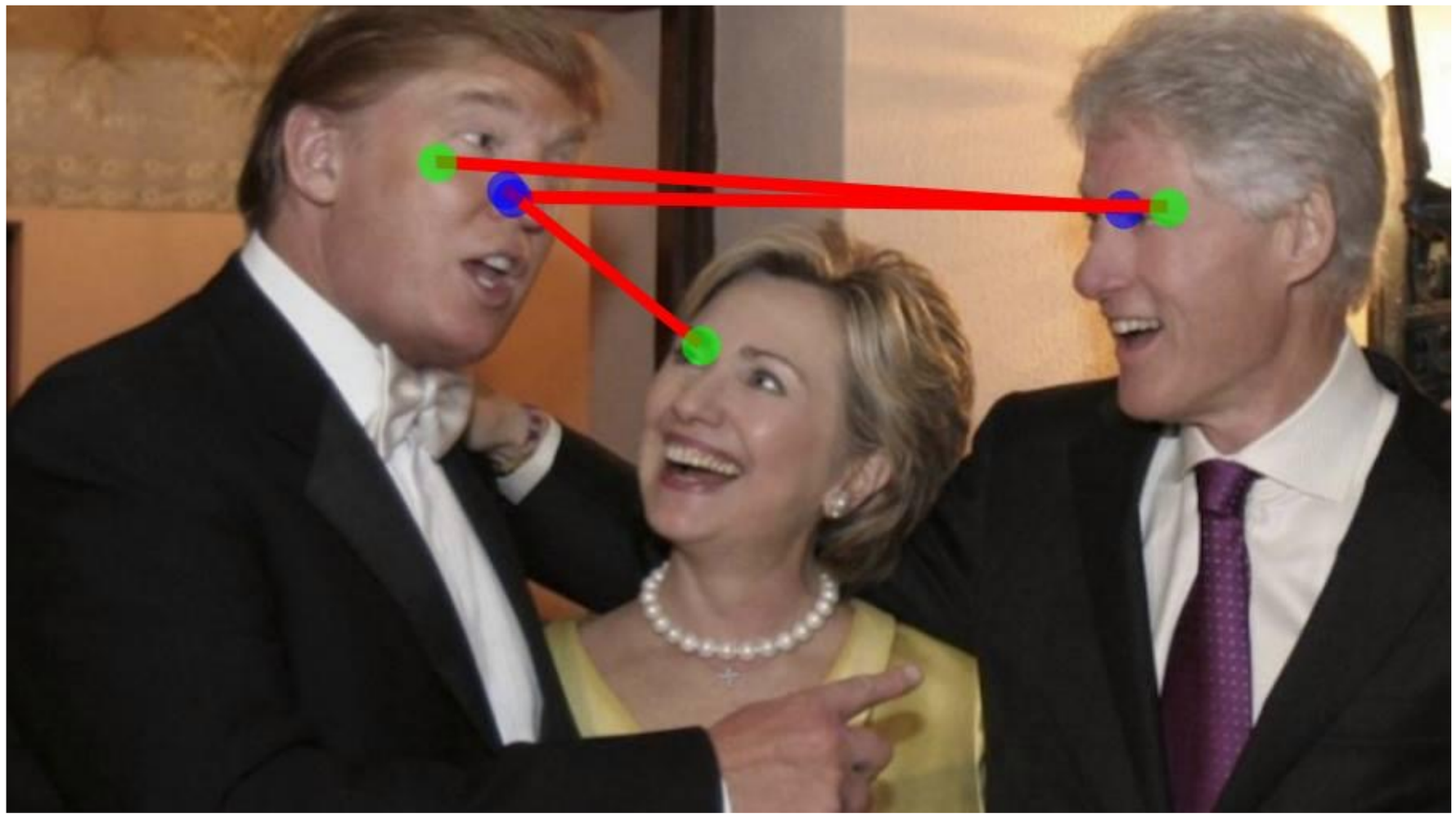
APPLICATIONS

- Behavior understanding



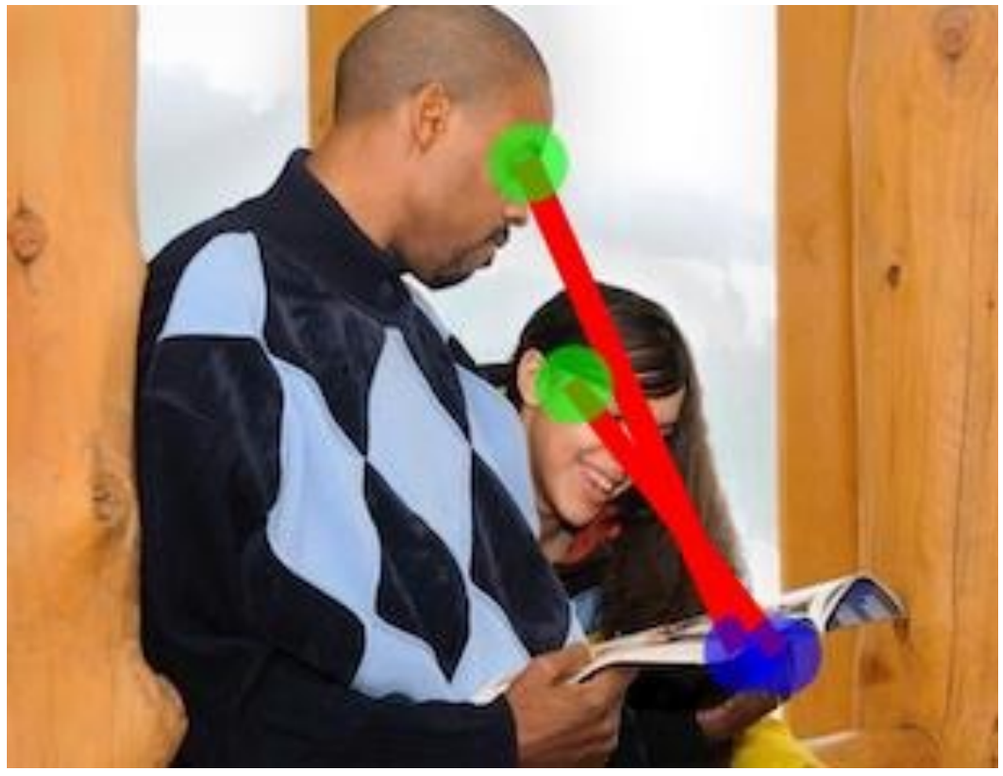
APPLICATIONS

- Behavior understanding
- Social situation understanding
  - do people know each other?



APPLICATIONS

- Behavior understanding
- Social situation understanding
  - do people know each other?
  - are people collaborating on the same task?



DIFFICULTY

CONTRIBUTION OF THIS WORK

- Solve gaze following in 3rd person view, instead of ego-centric

- Predict exact gaze location rather than just direction

- Do not require 3D location info of people in the scene

DATASET

1,548 from SUN

9,135 from Action 40

33,790 from MS COCO

508 from Imagenet

7,791 from Pascal

198,097 from Places

gaze annotation

9 more annotations per person in the scene

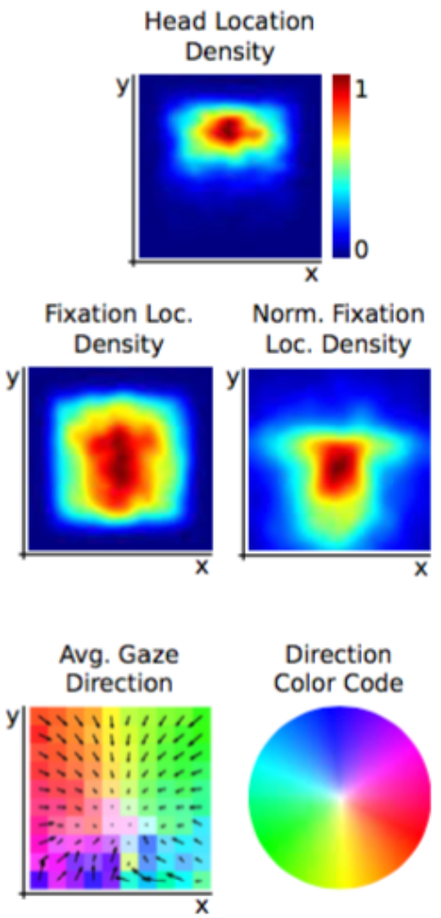
GazeFollow dataset:
130,339 people in 122,143 images

testing: 4,782 people

training: rest

TEST DATA EXAMPLES & STATS

(a) Example test images and annotations



(b) Test set statistics

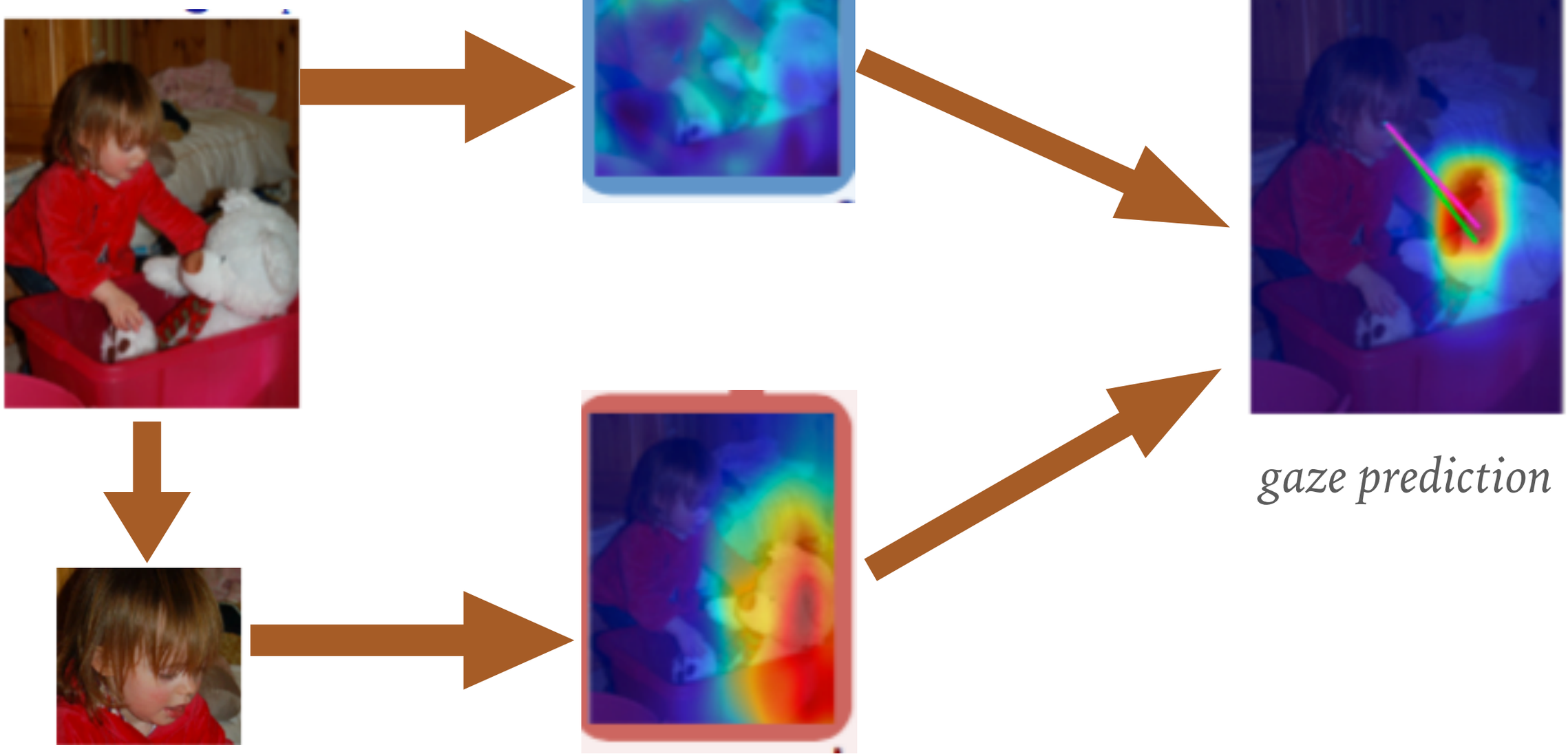
APPROACH

How do human predict where a person in a picture is looking at?

human first estimate the possible gaze directions based on head pose, then find the most salient objects in those directions

APPROACH

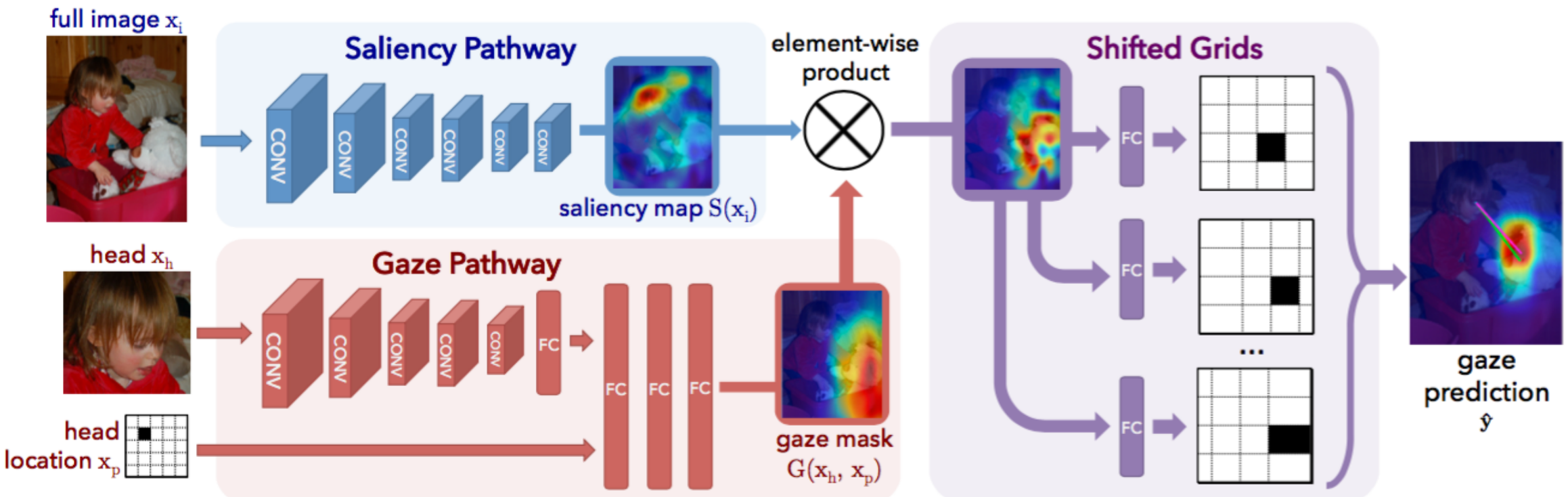
Saliency Map



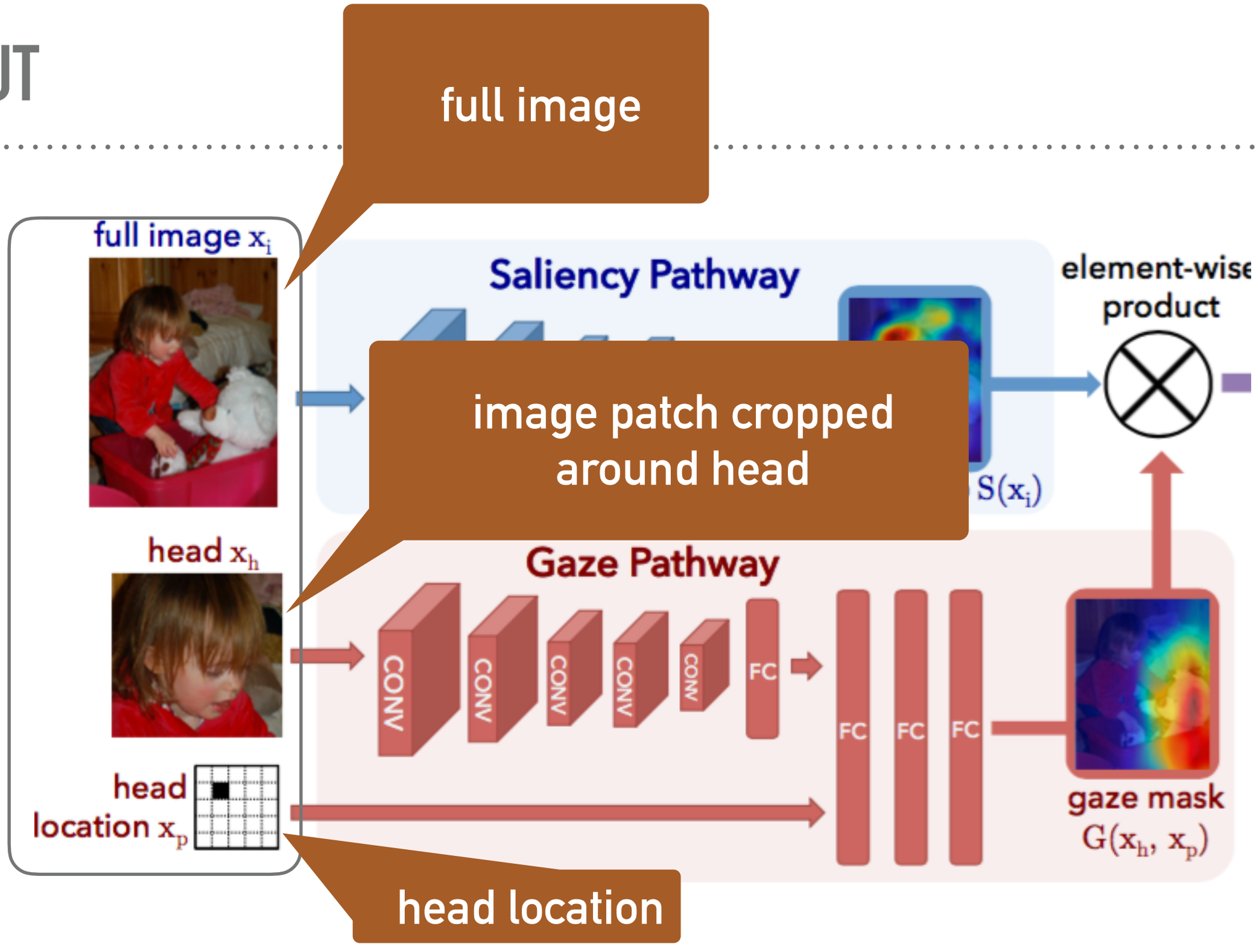
gaze prediction

Gaze direction map(Gaze mask)

MODEL

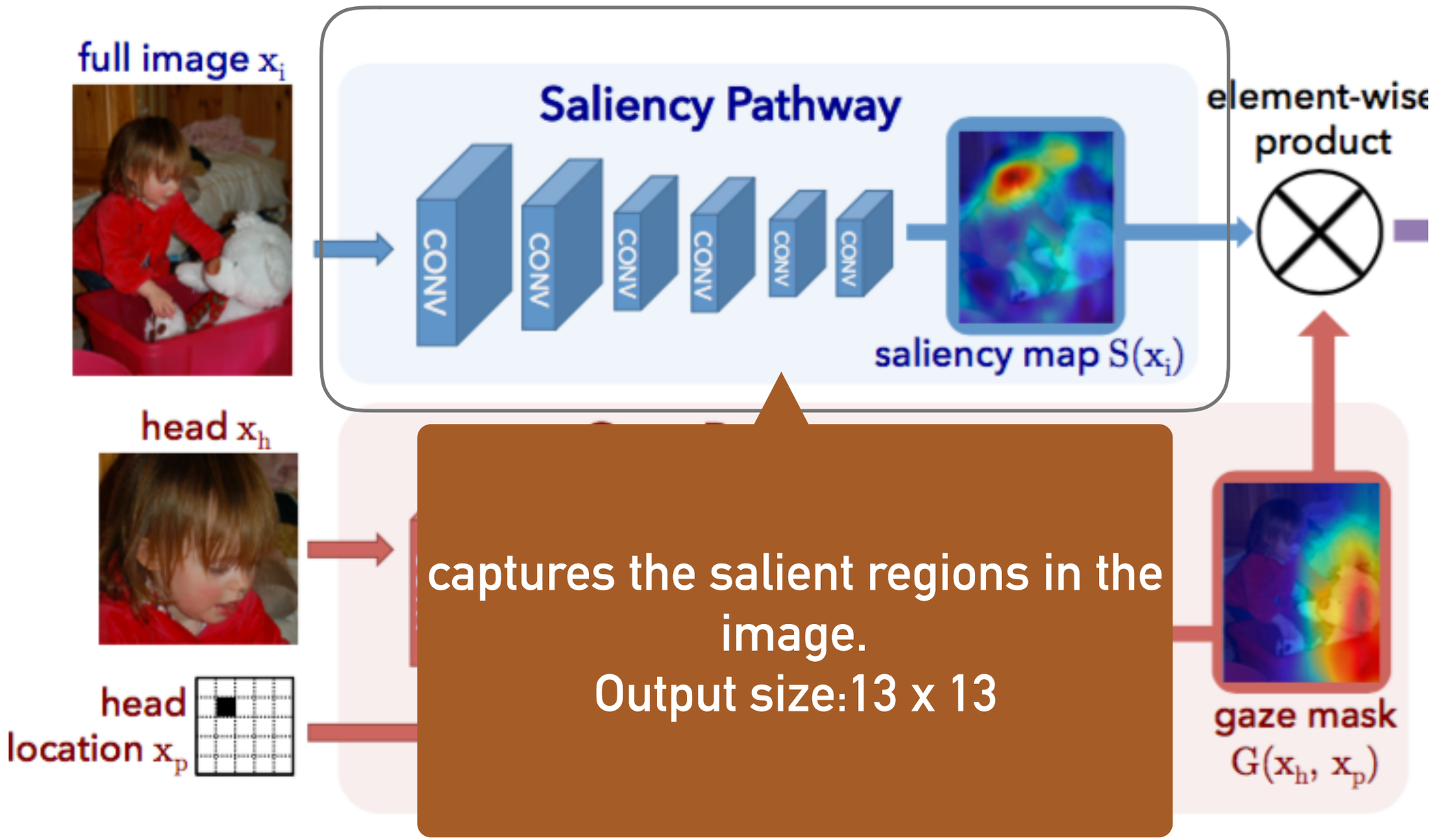

full image x_i
Saliency Pathway
CONV
CONV
CONV
CONV
CONV
CONV
saliency map $S(x_i)$
element-wise product
Shifted Grids
FC
FC
FC
...
head x_h
Gaze Pathway
CONV
CONV
CONV
CONV
CONV
FC
FC
FC
FC
gaze mask $G(x_h, x_p)$
head location x_p
gaze prediction $\hat{y}$

INPUT

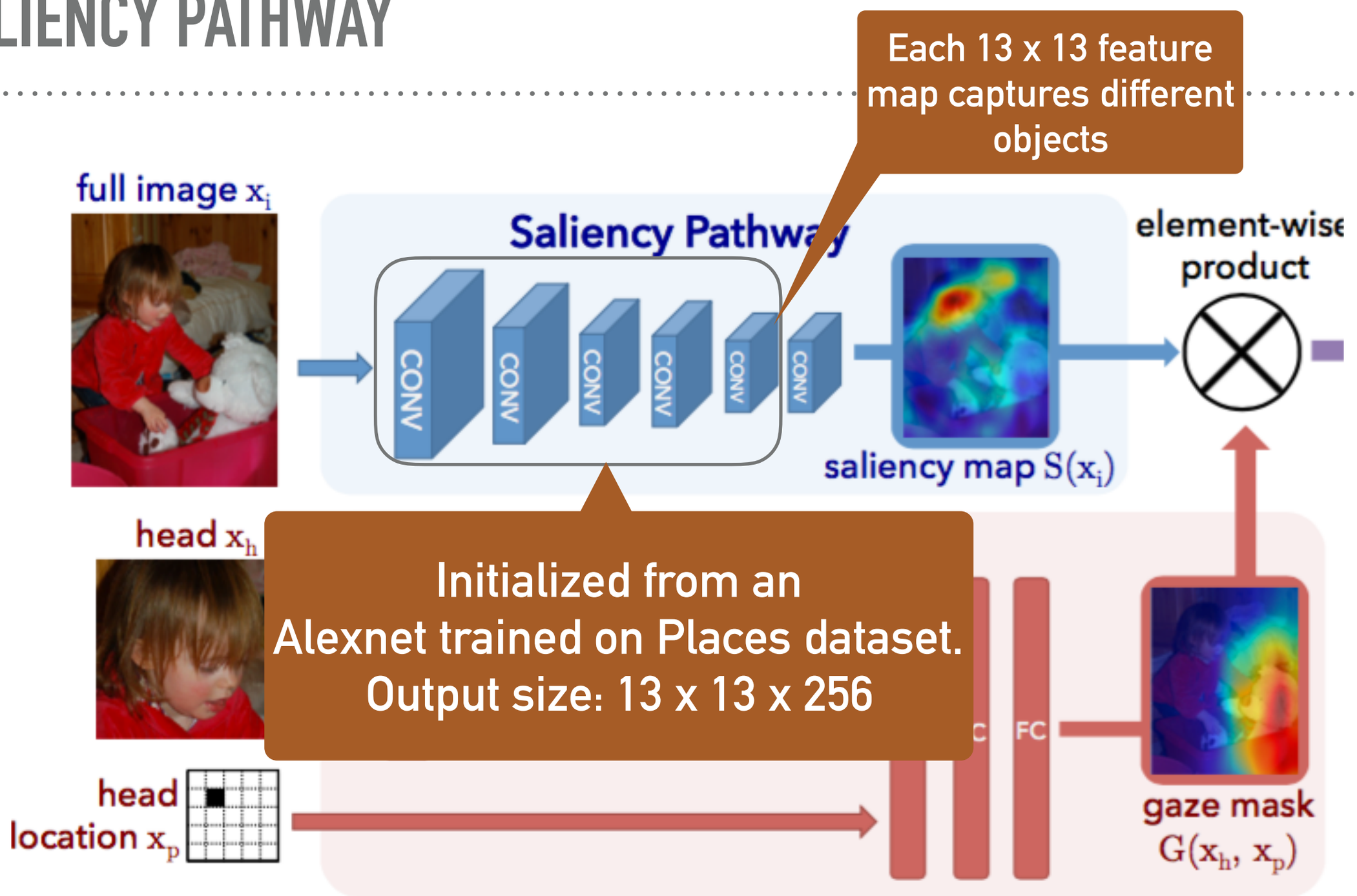


How does it force Saliency Pathway and Gaze Pathway to learn the saliency map and gaze mask respectively?

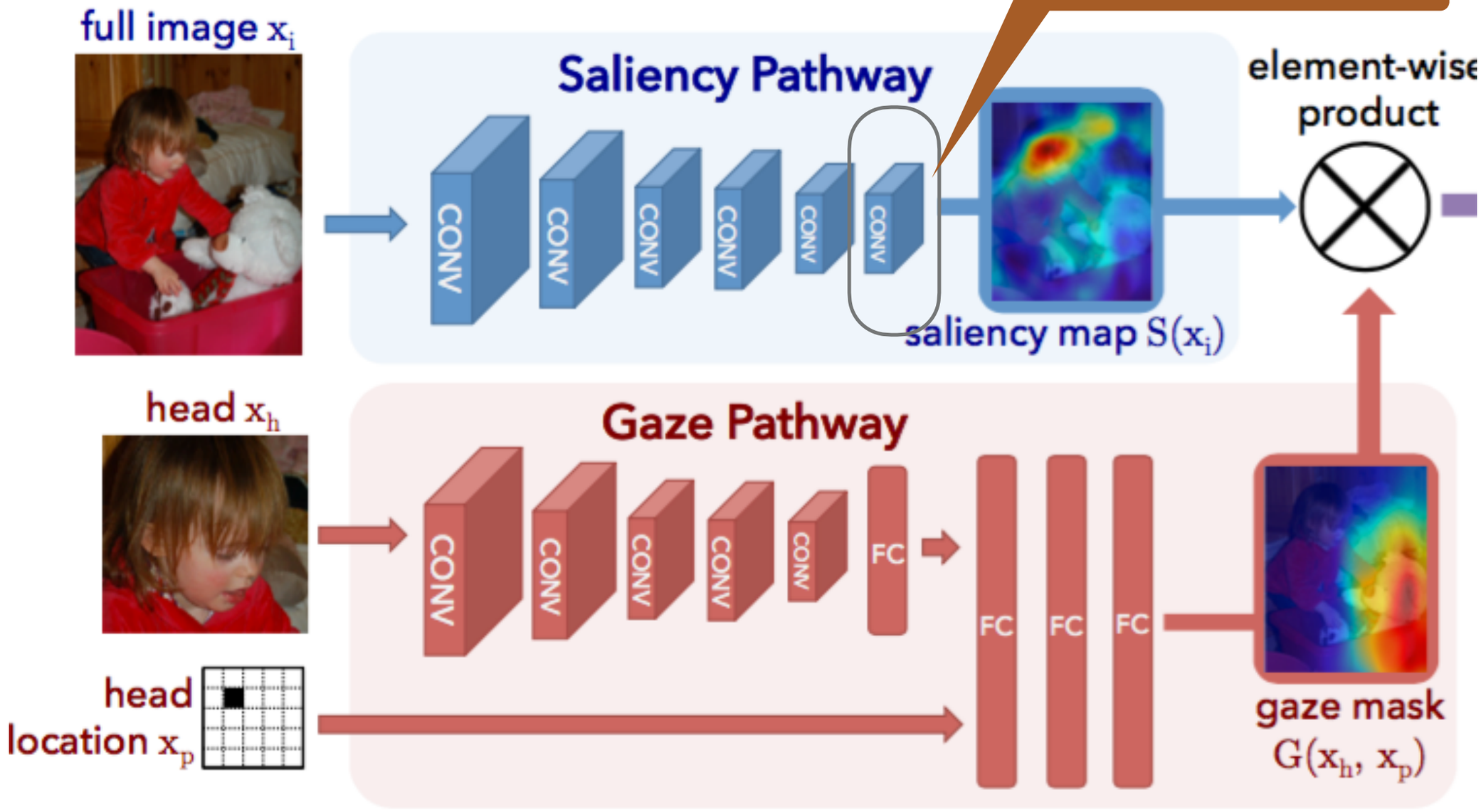
SALIENCY PATHWAY



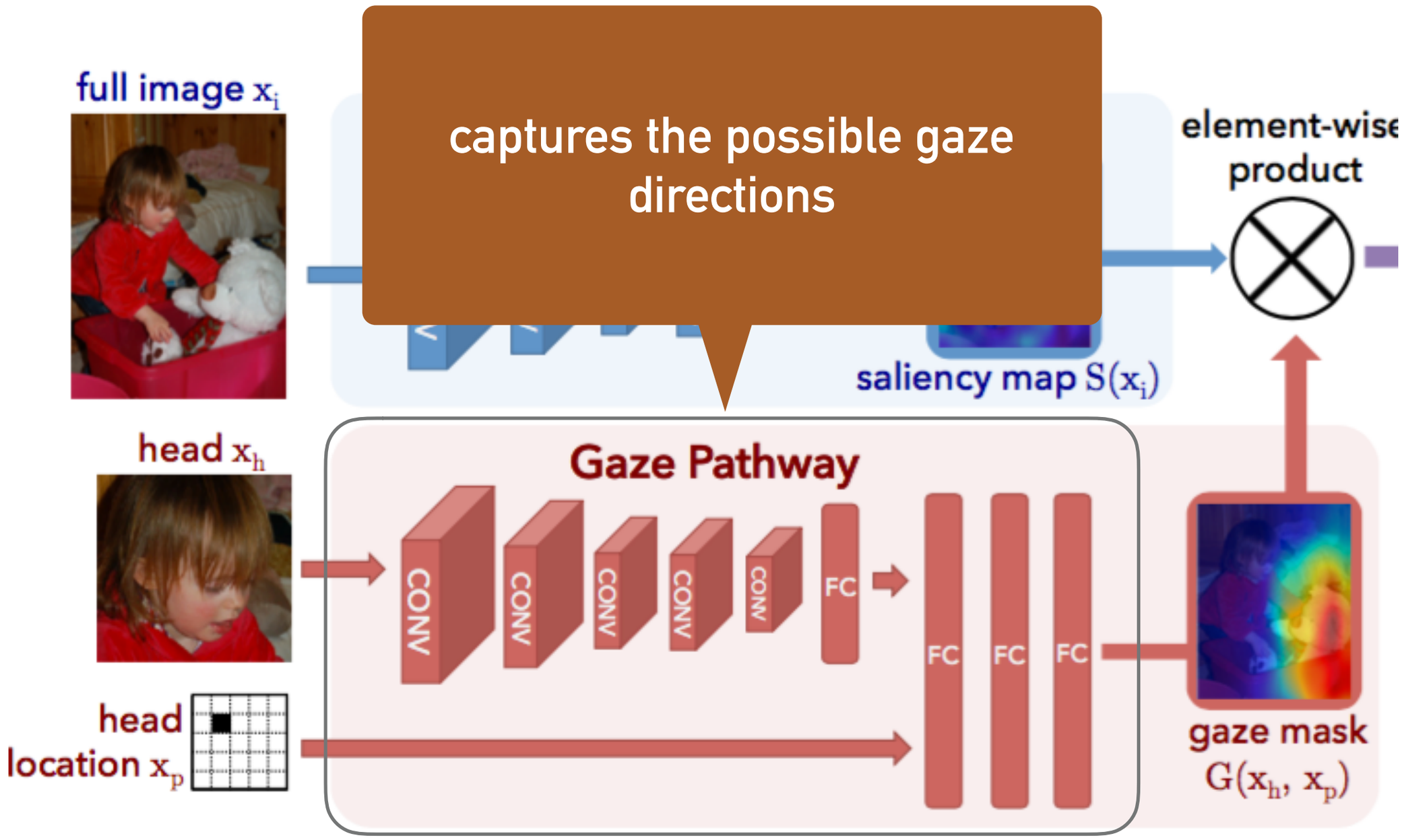
SALIENCY PATHWAY



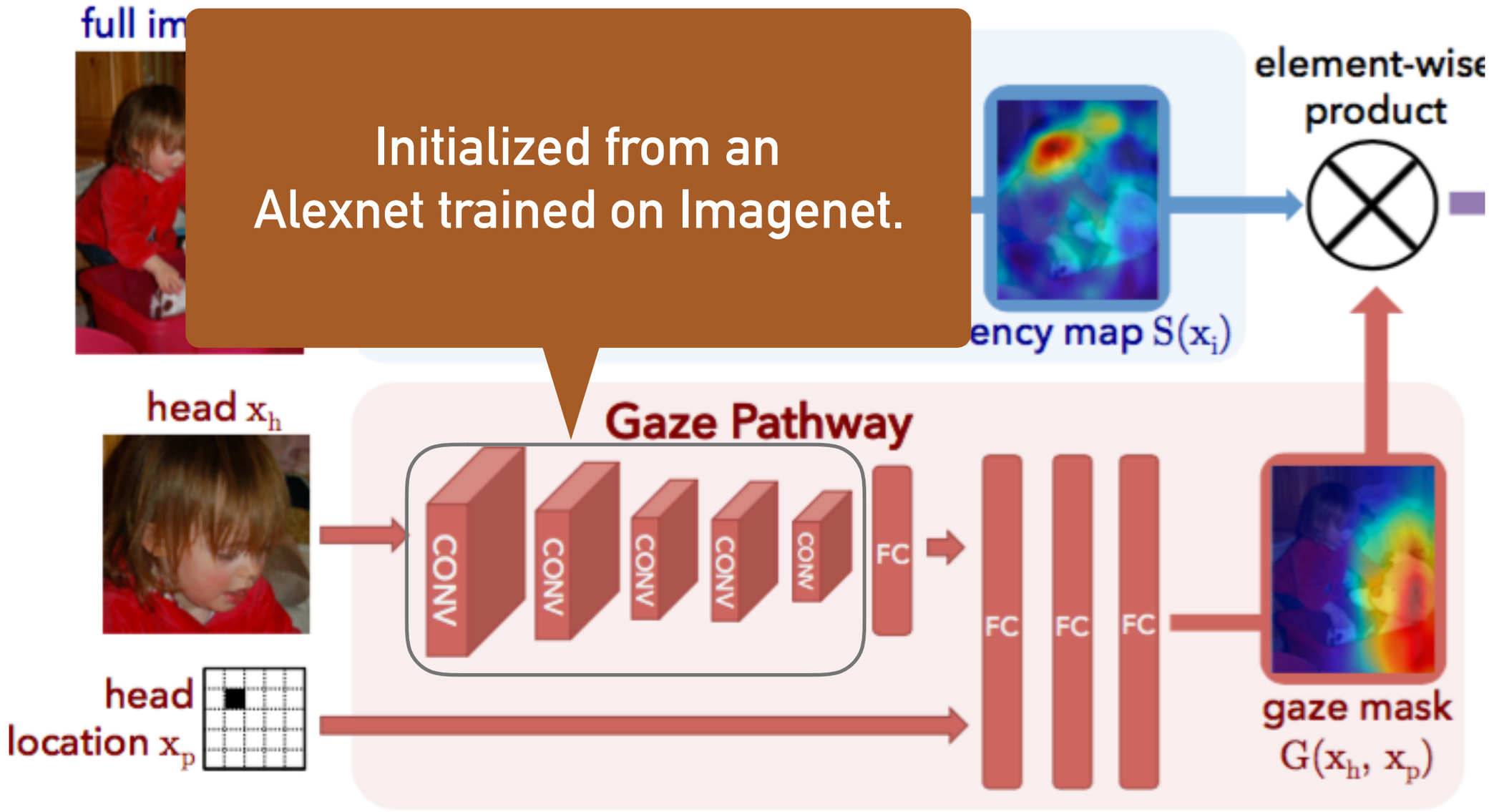
SALIENCY PATHWAY



GAZE PATHWAY

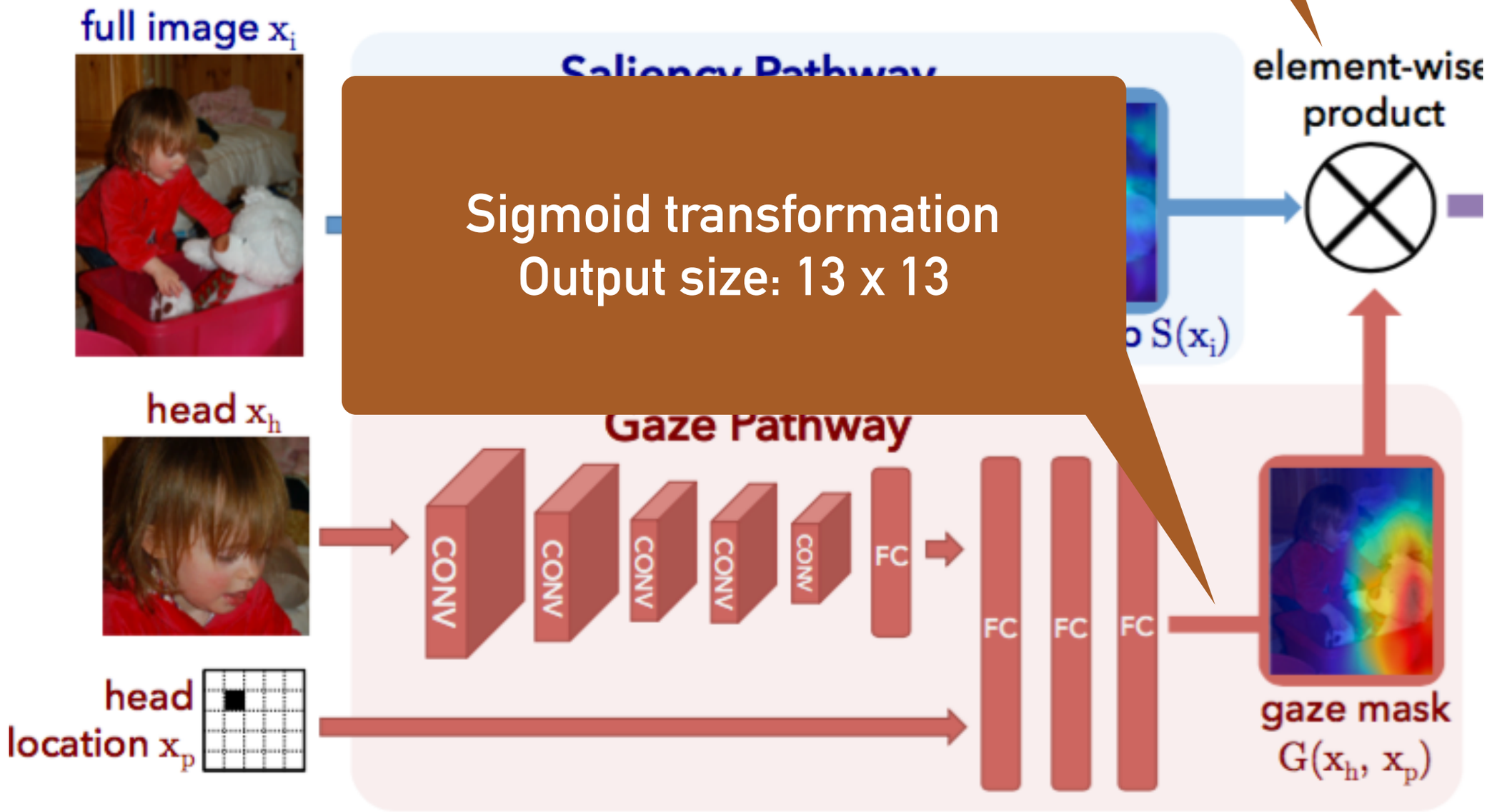


GAZE PATHWAY



GAZE PATHWAY

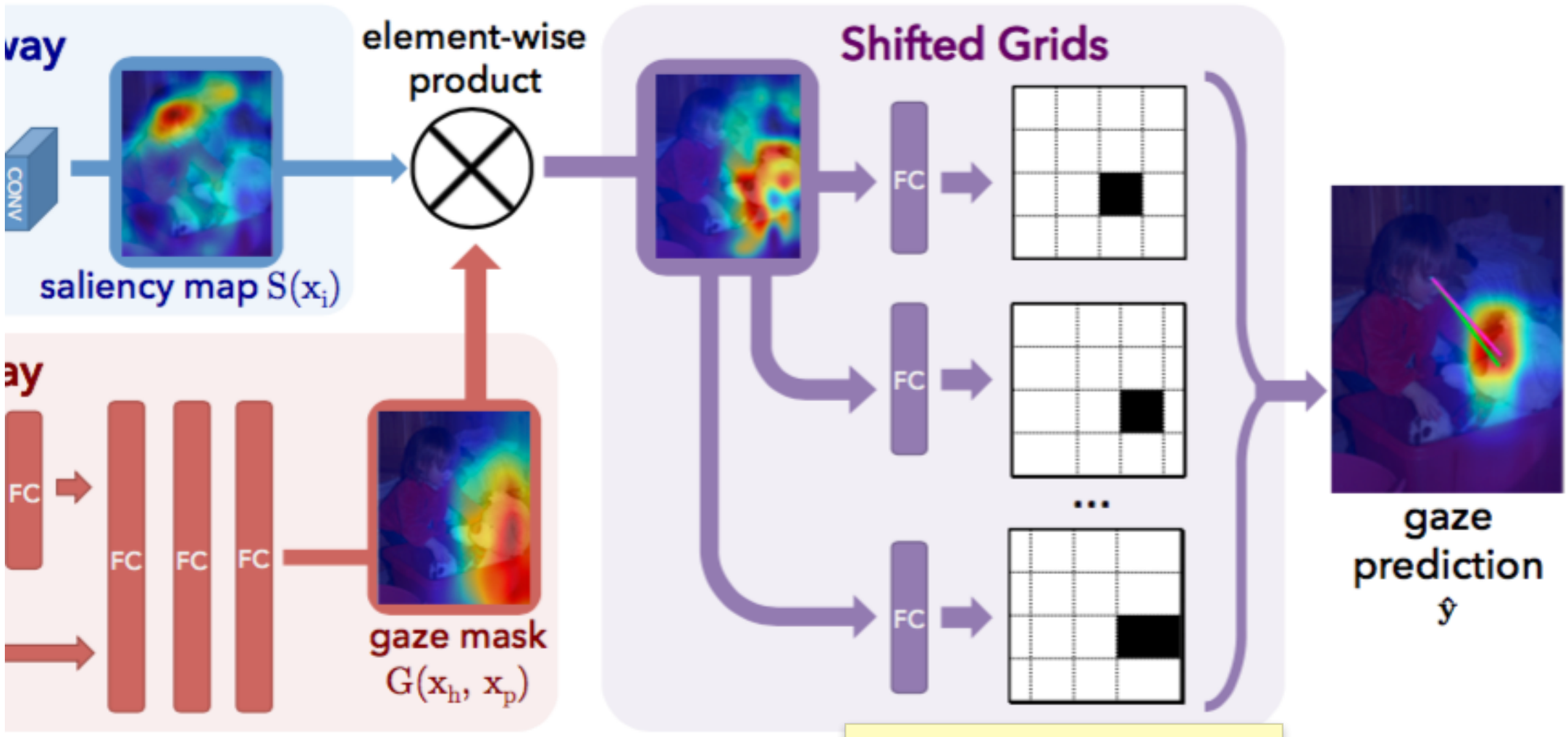
Combine the saliency map with the gaze mask



Sigmoid transformation
Output size: 13 x 13

MULTIMODAL PREDICTION WITH SHIF

They treated it as a multimodal classification problem instead of regression problem, because there are ambiguities in gaze location, and regression would just take the middle



The softmax loss penalizes all wrong grids uniformly, but we want it to penalize less on grids closer to the answer, so we compute loss on all shifted grids and take an average.

In their model they used shifted grids with size 5 x 5

QUANTITATIVE RESULT

Model	AUC	Dist.	Min Dist.	Ang.
Our	**0.878**	**0.190**	**0.113**	**24°**
SVM+shift grid	0.788	0.268	0.186	40°
SVM+one grid	0.758	0.276	0.193	43°
Judd [11]	0.711	0.337	0.250	54°
Fixed		0.306	0.219	48°
Cente		0.313	0.230	49°
Rand		0.484	0.391	69°
One h		0.096	0.040	11°

valuation

Recall that there are ten ground-truth gaze for each person in test images

AUC: rank the grid by their softmax prob and draw the ROC curve
Dist: distance to the average ground-truth
Min Dist: distance to closest ground-truth

Ang: angular distance between prediction and average ground-truth

Comparisons:
1. free-viewing saliency is different from gaze fixation. Also, free-viewing saliency doesn't consider gaze directions

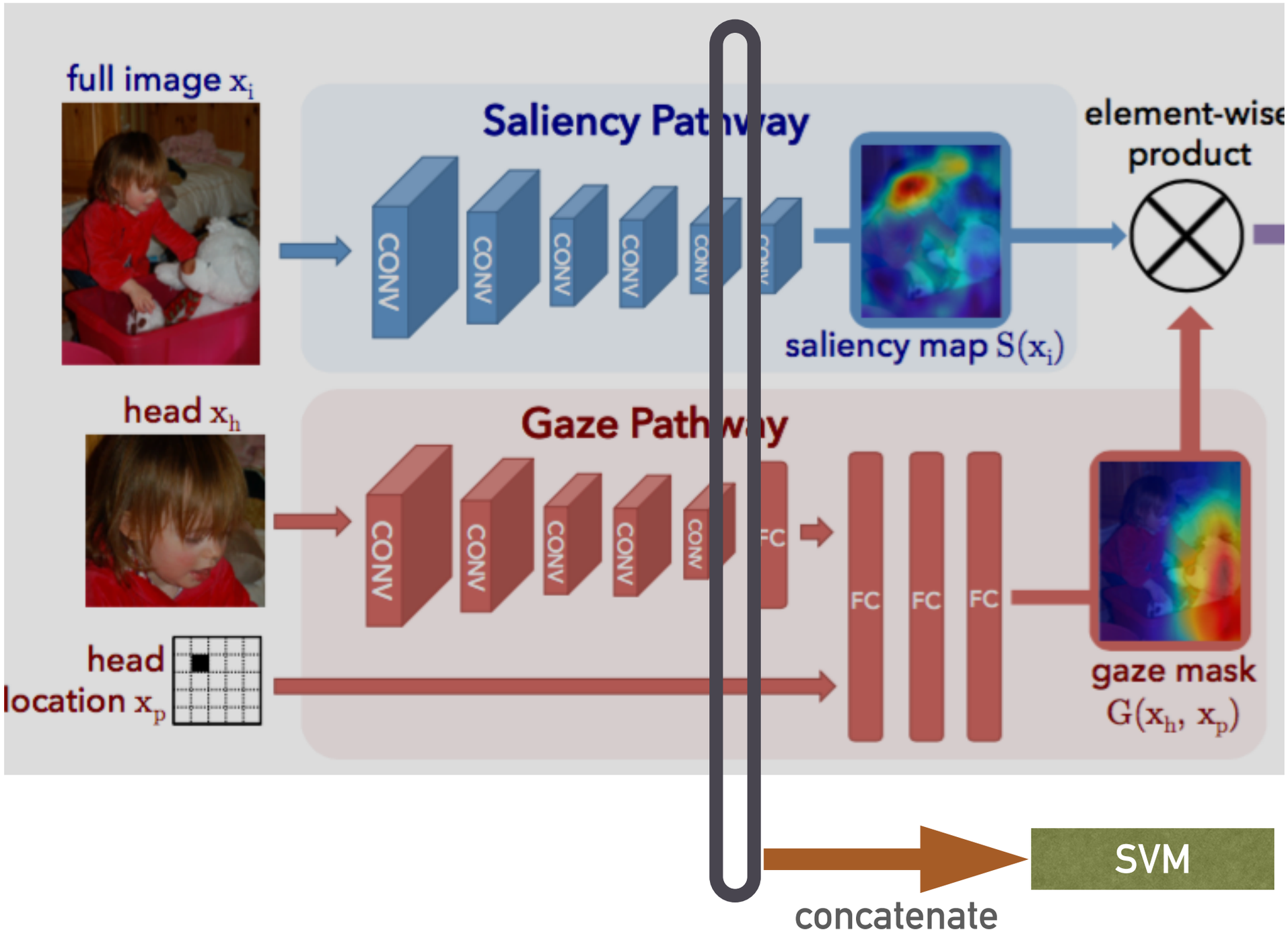
t

Center: The prediction is always the center of the image.

Fixed bias: The prediction is given by the average of fixations from the training set for heads in similar locations as the test image.

Judd[11]: We use a state-of-the-art free-viewing saliency model [11] as a predictor of gaze

SVM BASELINE



QUANTITATIVE RESULT

Model	AUC	Dist.	Min Dist.	Ang.
Our	**0.878**	**0.190**	**0.113**	**24°**
SVM+shift grid	0.788	0.268	0.186	40°
SVM+one grid	0.758	0.276	0.193	43°
Judd [11]	0.711	0.337	0.250	54°
Fixed		0.306	0.219	48°
Cente		0.313	0.230	49°
Rand		0.484	0.391	69°
One h		0.096	0.040	11°

valuation

Comparisons:

2. The model outperforms the SVM + shift grid baseline, SVM didn't have the learned weight in saliency pathway or the extra fully connected layers in gaze pathway. It also doesn't include the element wise multiplication. Therefore this decrease in performance suggests one or more of these components may play an important role. Later we will show that the element wise multiplication is actually not that important

3. Shifted grid improved the classification performance by a small margin

ABLATION STUDY

Model	AUC	Dist.	Min Dist.	Ang.
No image	0.821	0.221	0.142	27°
No position	0.837	0.238	0.158	32°
No head	0.822	0.264	0.179	41°
No eltwise	0.876	0.193	0.117	25°
5 × 5 grid	0.839	0.245	0.164	36°
10 × 10 grid	0.873	0.218	0.138	30°
L2 loss	0.768	0.245	0.169	34°
Our full	0.878	0.190	0.113	24°

(b) Model Diagnostics

1. Although the role of element wise multiplication All three input are important for this network. However, the full image doesn't affect the angular distance that much, which makes sense because the angular distance only depends on the correctness of gaze direction.

2. Elementwise multiplication of saliency map and gaze mask doesn't help that much

3. Their full model uses shifted grids with size 5 x 5. As can be seen, shifted grids did improve all measures by a large margin

4. The prediction of regression with L2 loss is much less accurate than classification result

QUALITATIV

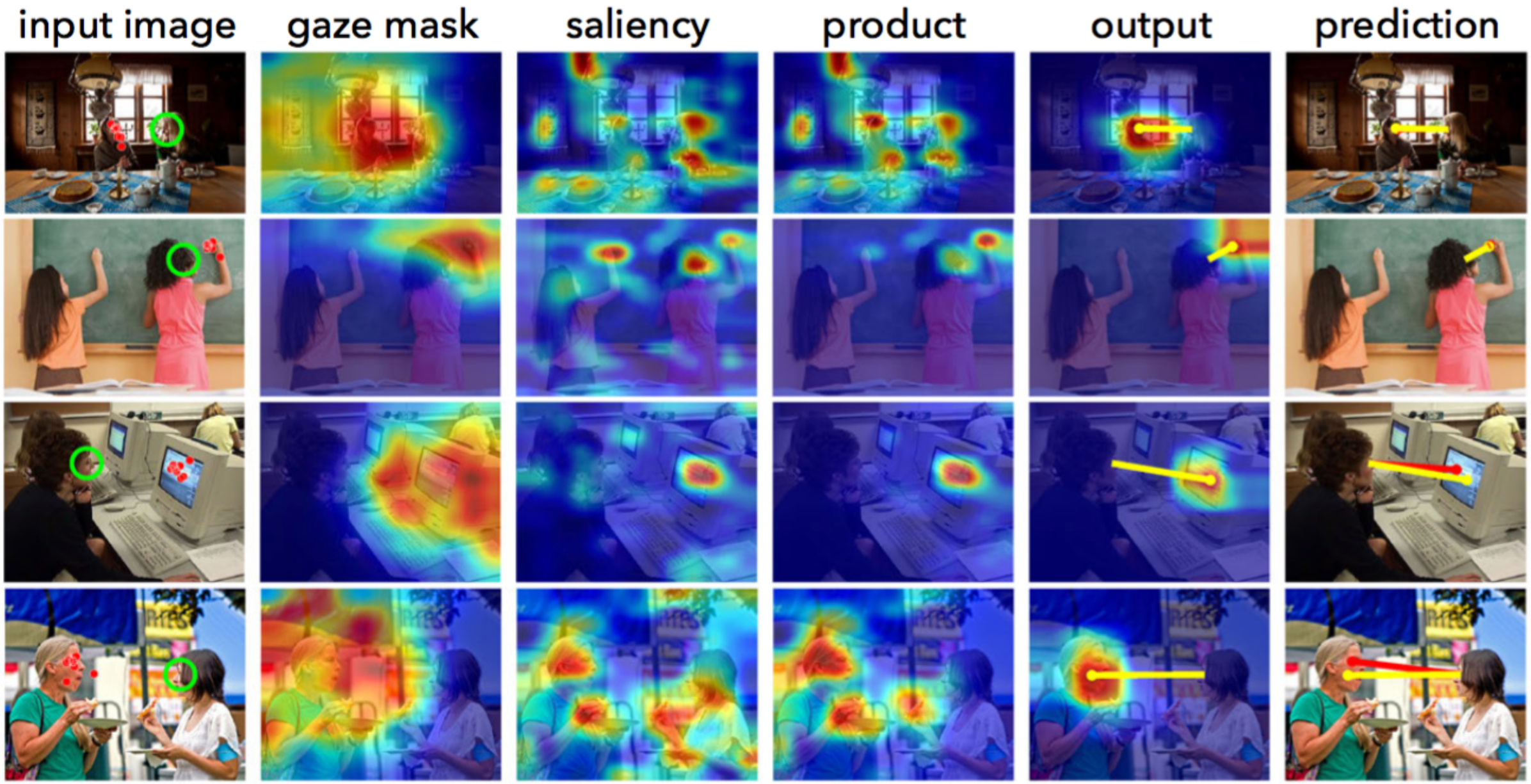
The model is able to find both reasonable gaze directions as well as salient objects on those directions

This proves that gaze mask is useful because the model is able to predict different gaze location for different people in the same picture

The first picture in the second row:

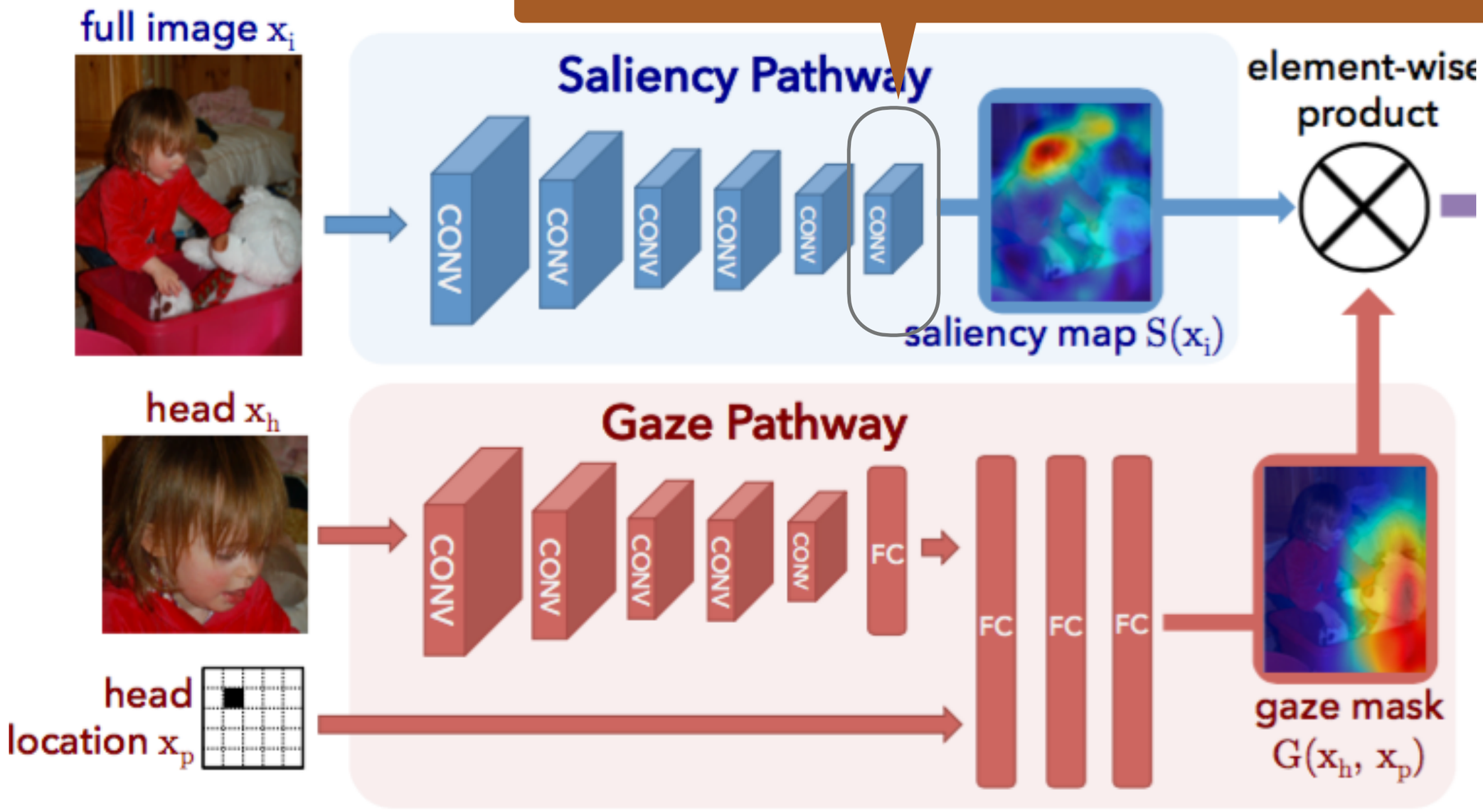
Figure 5: **Qualitative results:** We show several examples of successes and failures of our model. The red lines indicate **ground truth gaze**, and the yellow, our **predicted gaze**.

QUALITATIVE RESULT2



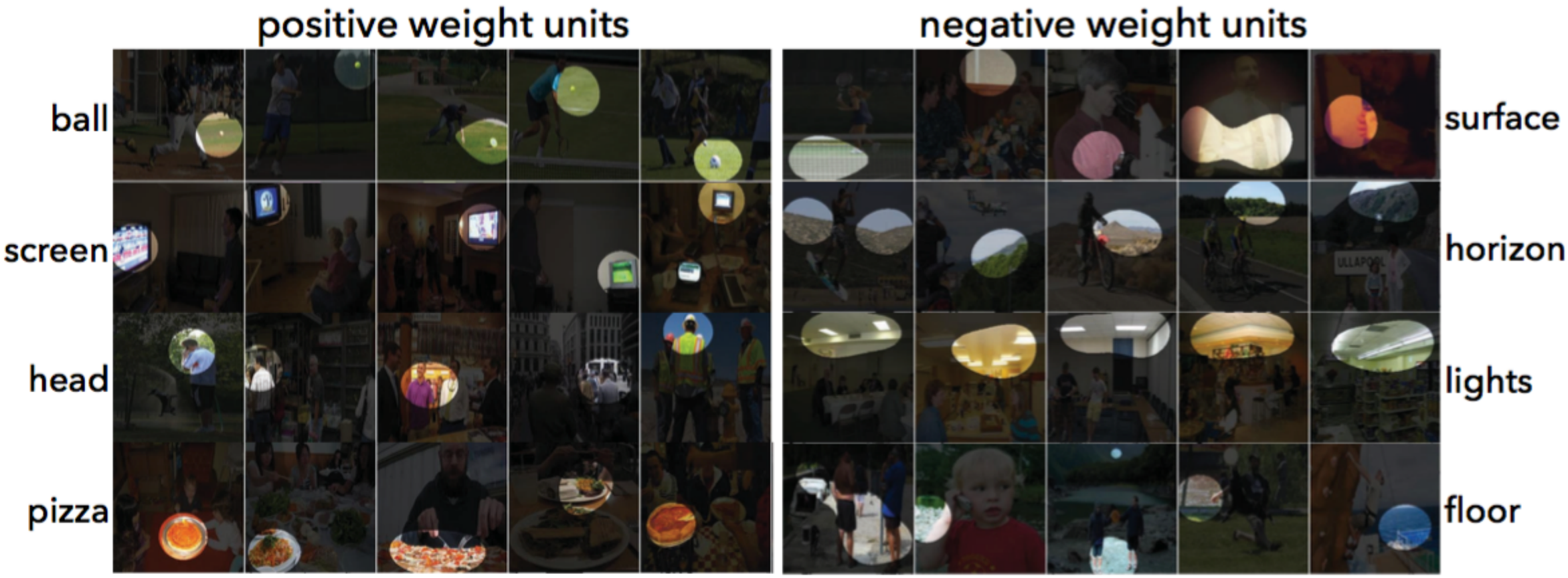
RECALL THAT:

Weighted sum of 256 feature maps.
Each feature map captures some object patterns.
Weights are learned such that objects people usually look at have higher(positive) weights.



QUALITATIVE RESULT3

- Top activation image regions for 8 conv5 neurons in Saliency pathway



EVALUATION

Strength:

- Combine gaze direction and visual saliency
- Good performance
- Use head position instead of face position
  - can handle the case where only the back is seen

Weakness:

- Ignoring depth -> unreasonable prediction
- Cross-entropy loss VS shifted grids?

DEMO

- http://gazefollow.csail.mit.edu/demo.html

- Photo with people appearing in their back
  - http://jessgibbsphotography.com/wp-content/uploads/2013/01/crowds_of_people_take_photos_of_flag_ceremony_outside_town_hall.jpg

- Photo where people are staring at objects outside the image
  - http://www.celwalls.com/wallpapers/large/7525.jpg