Who are they looking at?

Automatic Eye Gaze Following for Classroom Observation Video Analysis

Arkar Min Aung, Anand Ramakrishnan, and Jacob Whitehill

Worcester Polytechnic Institute (WPI)

Context

Classroom observation

- In the USA (and other countries), it is commonplace for administrators, researchers, and other teachers to make **classroom observations**:
  - Live
  - Video-based

Classroom observation

- These observation sessions are used for:
  - Professional development
  - Accountability
  - Educational research

Classroom observation protocols

- Classroom sessions are coded using one of several standard observation protocols to characterize different aspects of classroom instruction.

- One of the most commonly used protocols is the Classroom Assessment Scoring System (CLASS; Pianta, et al. 2008).

CLASS

Pianta, et al. (2008)

- An underlying assumption of the CLASS is that the quality of **teacher-student interactions** can be measured independently of the curriculum being taught.

- Significant evidence that CLASS scores predict children's downstream academic, cognitive, and emotional outcomes, e.g.:
  - Reading achievement (Ponitz, et al. 2009)
  - Engagement (Curby, et al. 2014)
  - Executive functioning (Weiland, et al. 2013)

CLASS

Pianta, et al. (2008)

Domain
Emotional support
Classroom organization
Instructional support

CLASS

Pianta, et al. (2008)

Domain	**Dimension**
Emotional support	Positive climate
	Negative climate
	Teacher sensitivity
	Regard for child perspectives
Classroom organization	Behavioral management
	Productivity
	Instructional learning formats
Instructional support	Concept development
	Quality of feedback
	Language modeling
	Literacy focus

CLASS

Pianta, et al. (2008)

Domain	**Dimension**	**Indicators**
Emotional support	Positive climate	
	Negative climate	
	Teacher sensitivity	**Awareness** Responsiveness Address problems
	Regard for child perspectives	
Classroom organization	Behavioral management	
	Productivity	
	Instructional learning formats	
Instructional support	Concept development	
	Quality of feedback	
	Language modeling	
	Literacy focus	

CLASS

Pianta, et al. (2008)

Domain	Dimension	Indicators	Behavioral markers
Emotional support	Positive climate		
	Negative climate		
	Teacher sensitivity	**Awareness** Responsiveness Address problems	… **Notices lack of understanding** …
	Regard for child perspectives		
Classroom organization	Behavioral management		
	Productivity		
	Instructional learning formats		
Instructional support	Concept development		
	Quality of feedback		
	Language modeling		
	Literacy focus		

Manual classroom observation

- With the CLASS, human annotators assign one number (1-7) to each dimension once every 15 minutes.
  - Sparse
  - Expensive
  - Non-specific (difficult to label *which* children/teachers were most important)

Automated classroom observation

- It could be useful to (partially) automate this process:
  - More frequent and specific feedback to teachers
  - Improved lens to estimate impact of educational interventions

Automated classroom observation: feasibility

- Some dimensions are likely more automatable than others.

- For some emotional support dimensions, the behavioral markers are related to:
  - Facial expression

Automated classroom observation: feasibility

- Some dimensions are likely more automatable than others.
- For some emotional support dimensions, the behavioral markers are related to:
  - Facial expression
  - Physical proximity

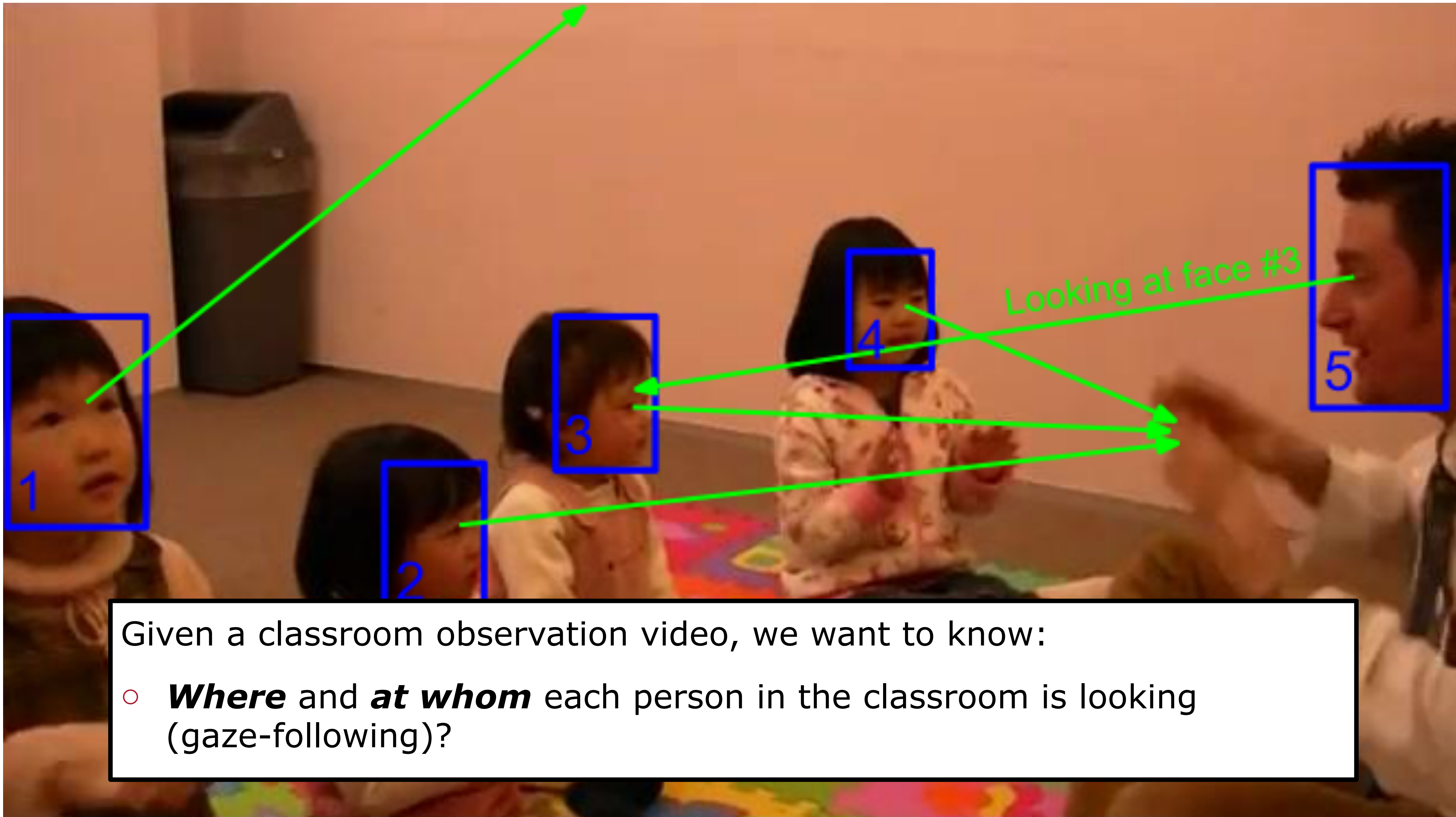
Automated classroom observation: feasibility

- Some dimensions are likely more automatable than others.

- For some emotional support dimensions, the behavioral markers are related to:
  - Facial expression
  - Physical proximity
  - **Mutual eye-gaze between students and teachers.**

Gaze following

Problem Statement



Given a classroom observation video, we want to know:

- ***Where*** and ***at whom*** each person in the classroom is looking (gaze-following)?

Gaze-following in 2-D Static Images

- Annotating gaze locations in 2-D images:
  - Can be ambiguous since 2-D images does not have depth information.
  - **Assumption:** Knowing gaze location in 2-D images can be informative for downstream processing.
- 2-D images are a lot easier to obtain than 3-D images (RGB-D images).

Classroom observation videos

Video taken from: https://www.youtube.com/watch?v=cjNv2dQCFEk&list=PLI4ATdTjD5eSPcj9YJe4I6kV5heH-s-58&index=4&t=1084s

Classroom observation videos

- Multiple students and teachers
- Highly cluttered
- Significant occlusion
- Extreme head poses (with faces sometimes pointing away from camera)

Video taken from: https://www.youtube.com/watch?v=cjNv2dQCFEk&list=PLI4ATdTjD5eSPcj9YJe4I6kV5heH-s-58&index=4&t=1084s

Differences in Datasets

MS COCO, SUN, Actions, Places, PASCAL Datasets

Classroom Observation Video Images

Data Collection

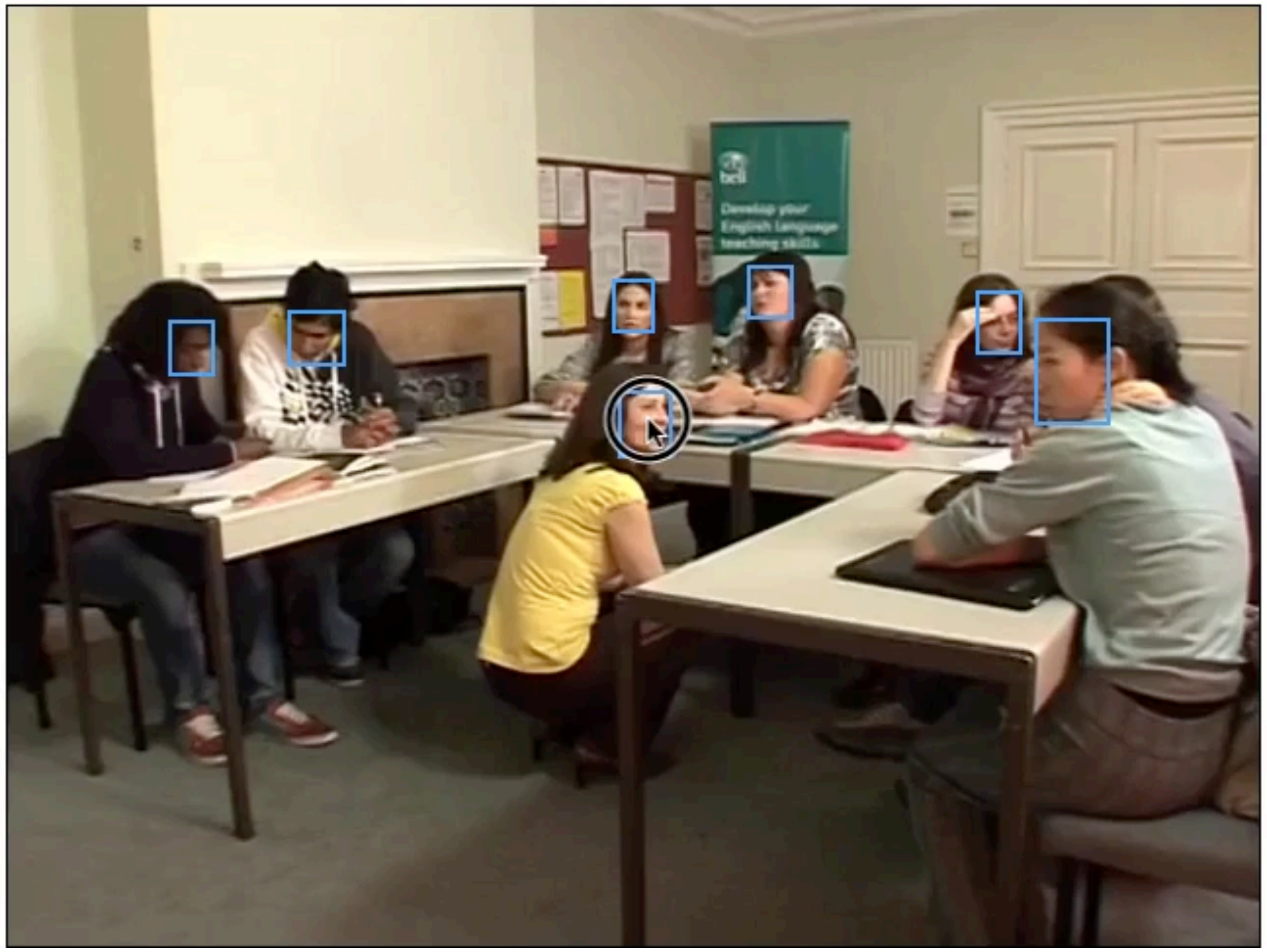
Data Sourcing

- Use 70 classroom observation videos[1] publicly available on YouTube.
- Extract 1 frame approximately every 10 seconds.
- Use Faster R-CNN for face detection[2] to obtain face bounding boxes in extracted frames.
- 7.85 faces per image on average (for the whole dataset)

[1] Ramakrishnan, A., and Whitehill, J. Youtube pre-school dataset, 2017.
[2] Jiang, H., and Learned-Miller, E. Face detection with the faster r-cnn. In IEEE Automatic Face & Gesture Recognition (2017).

Data Annotation

- Tool built with HTML5+Javascript and deployed on Amazon Mechanical Turk (AMT).
- Collects gaze location as well as binary indication of whether the gaze ends inside or outside the image.



Data Annotation

- 3 labelers per image on average on AMT to annotate the gaze of each face.
- 408 unique annotators.
- Collected three gaze annotations each for 17,758 faces in 2,263 images.
- After cleaning data, obtained a total of 48,907 gaze annotations.

Dataset

- Training data is augmented by flipping images and gazes left to right.
- Data split
  - 70% Training
  - 15% Validation
  - 15% Testing
- Sets of people in training, validation, and test don’t overlap.
- No image from the same video occurs in more than one data split.

Sample Annotations (for 3 labelers)

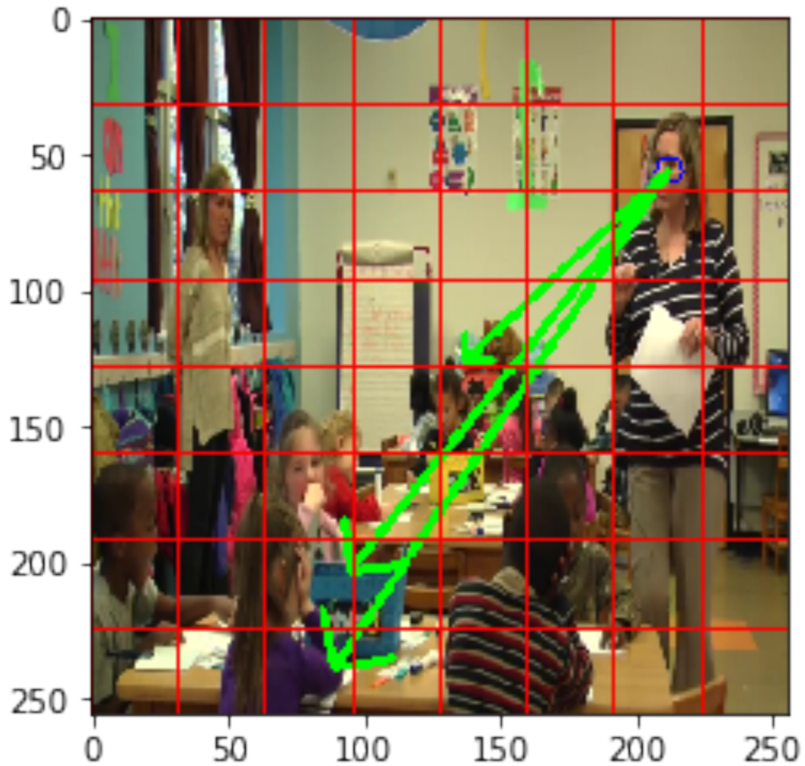
Network Design

To regress or to classify?

- The task of following the gaze of a person can be formulated as either:
  - A classification task
  - A regression task

(x,y) coordinates and soft labels

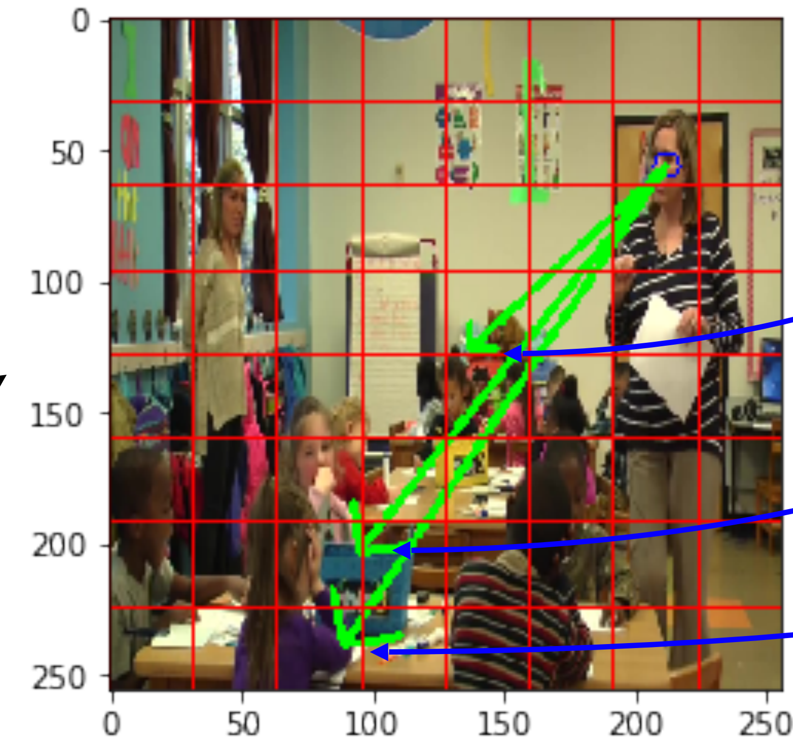
256x256 pixel image



Regression

Classification

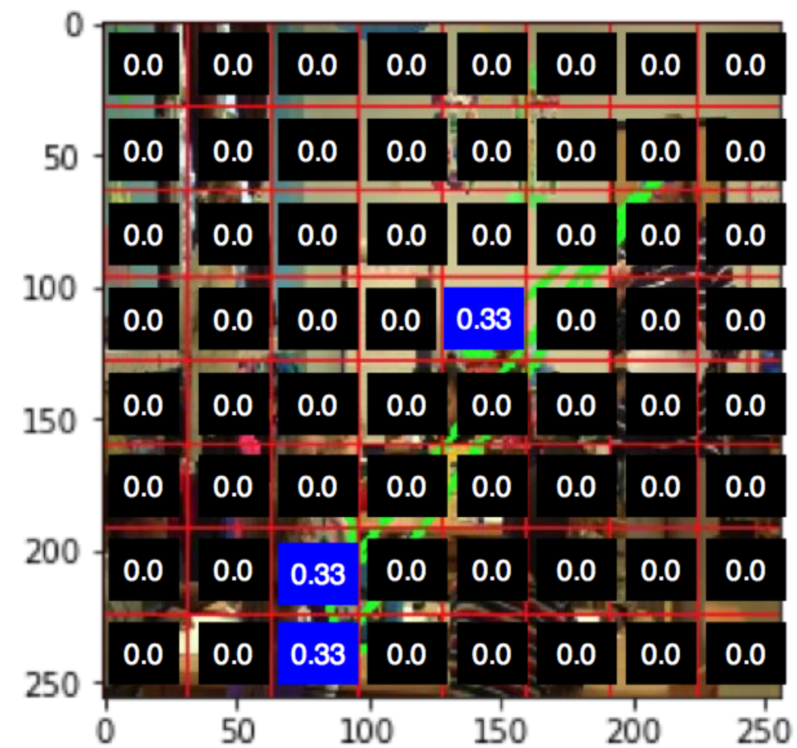
Separate training label



(140, 130)

(100, 200)

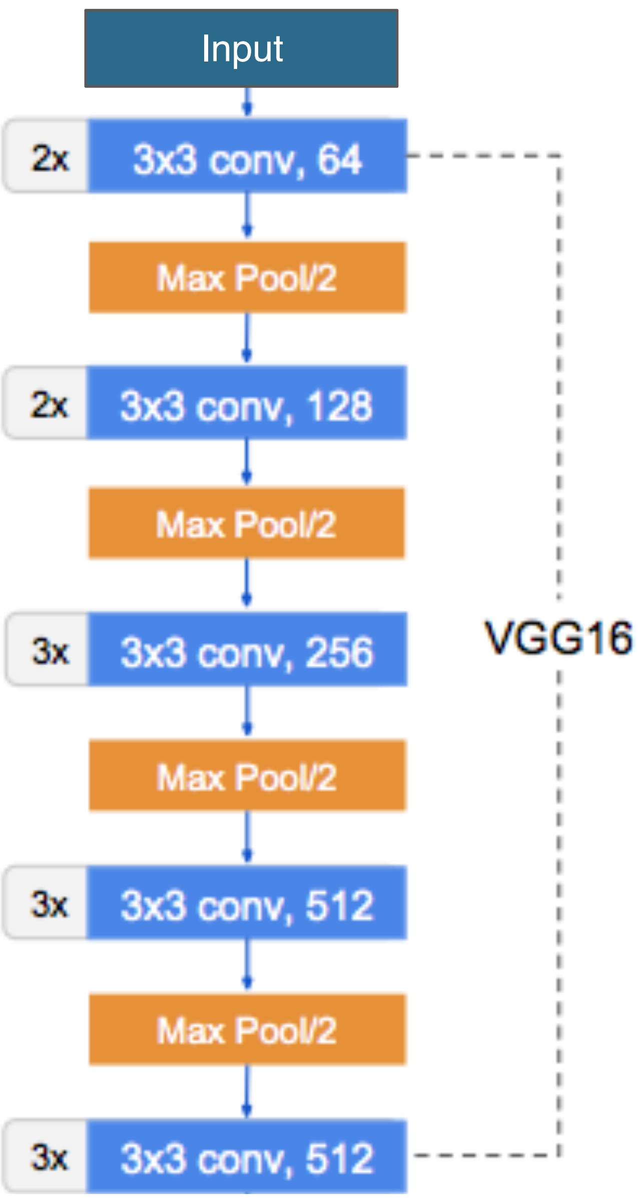
(90, 235)



8x8 Grid

Deep Learning Architecture

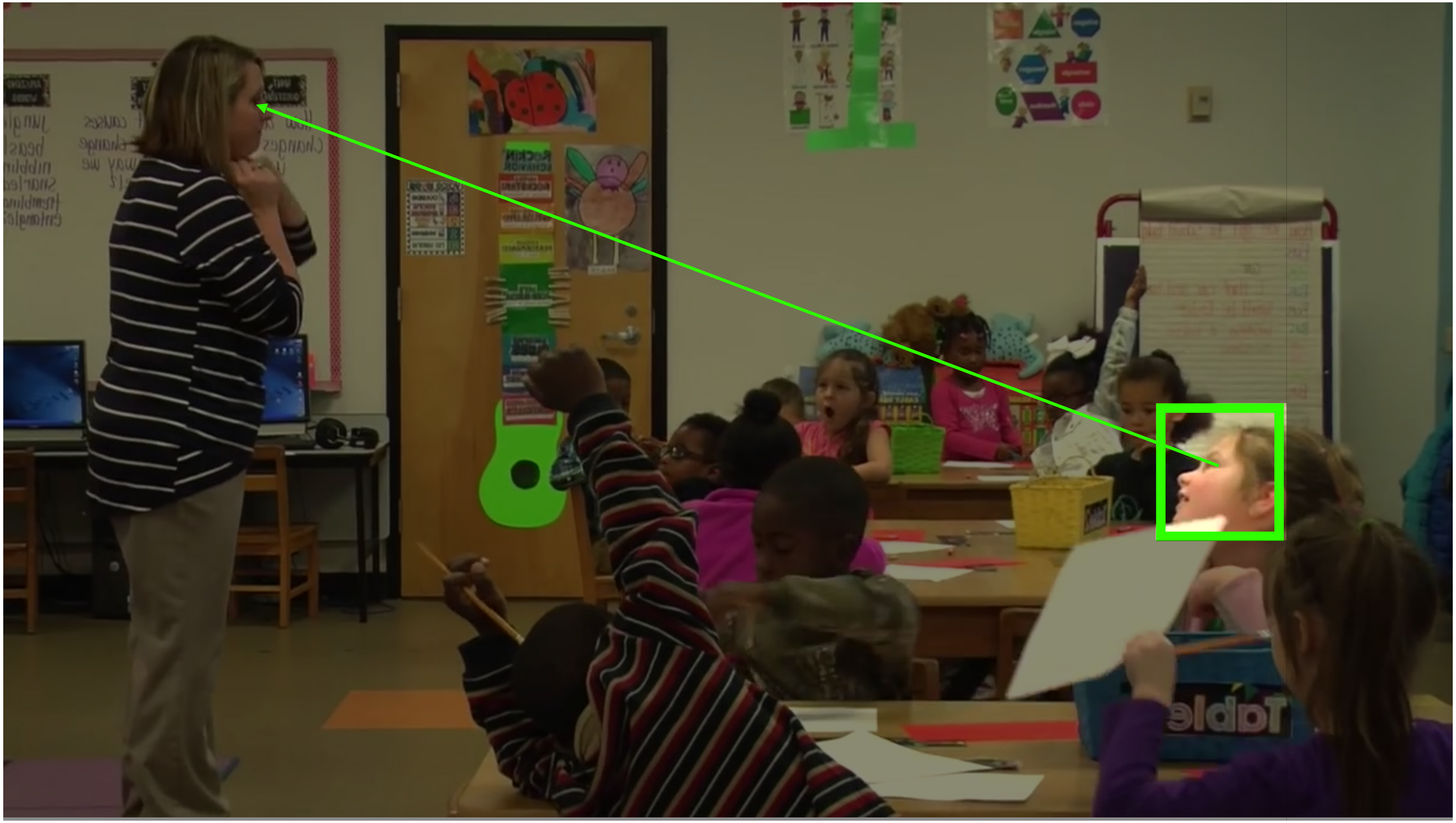
- Approach is inspired by Recasens, et al (2015) [1].
- We use VGG16[2] as the base architecture.
- We use different optimization techniques.
  - Transfer learning with fine tuning.
- Multiple-tasks
  - Predict the gaze location.
  - Predict whether the gaze ends inside or outside the image (In/Out gaze).



[1] Recasens, A., Khosla, A., Vondrick, C., and Torralba, A. Where are they looking? In Advances in Neural Information Processing Systems (2015).
[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

Take this image for an example

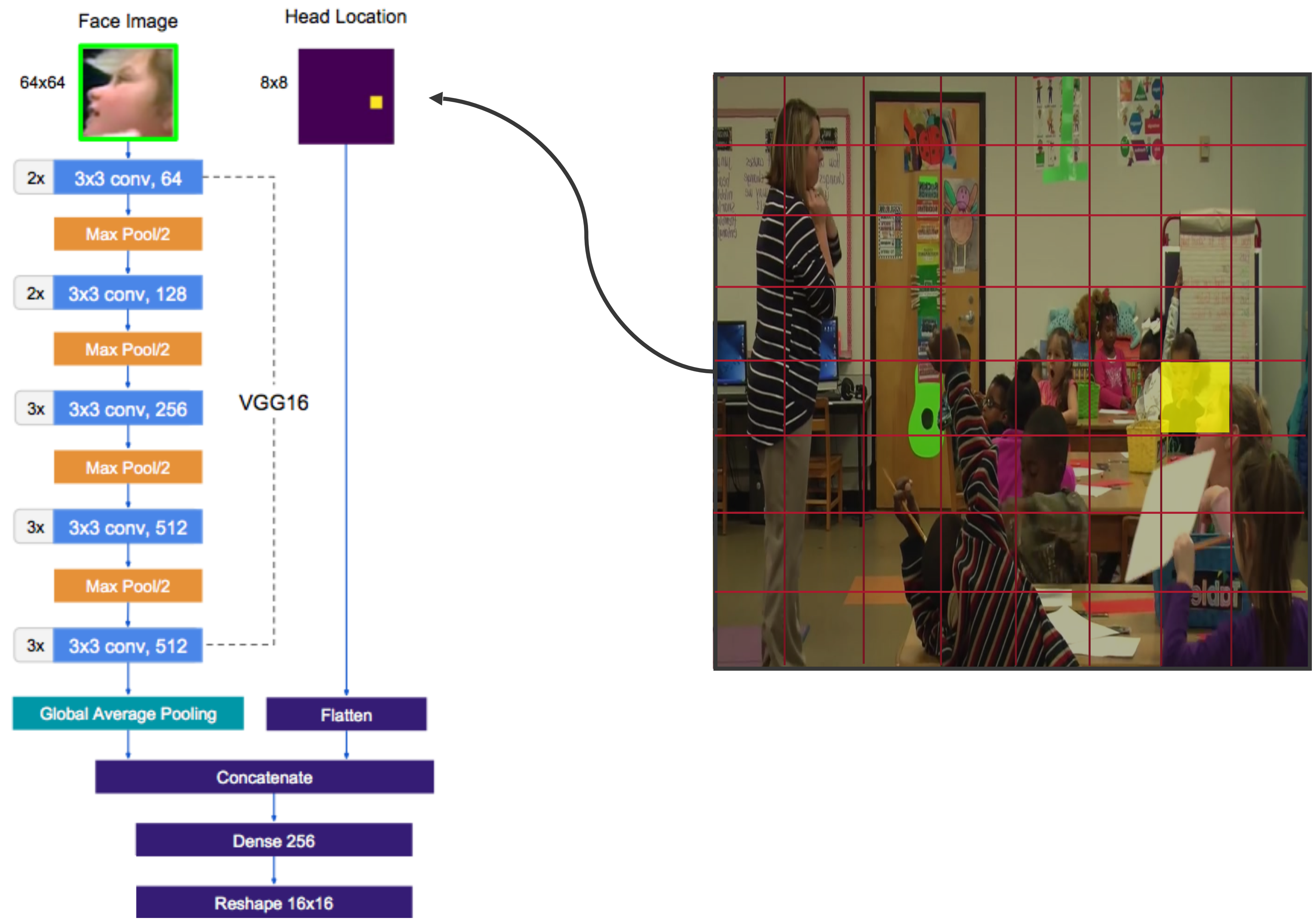
We want to know the gaze of this girl

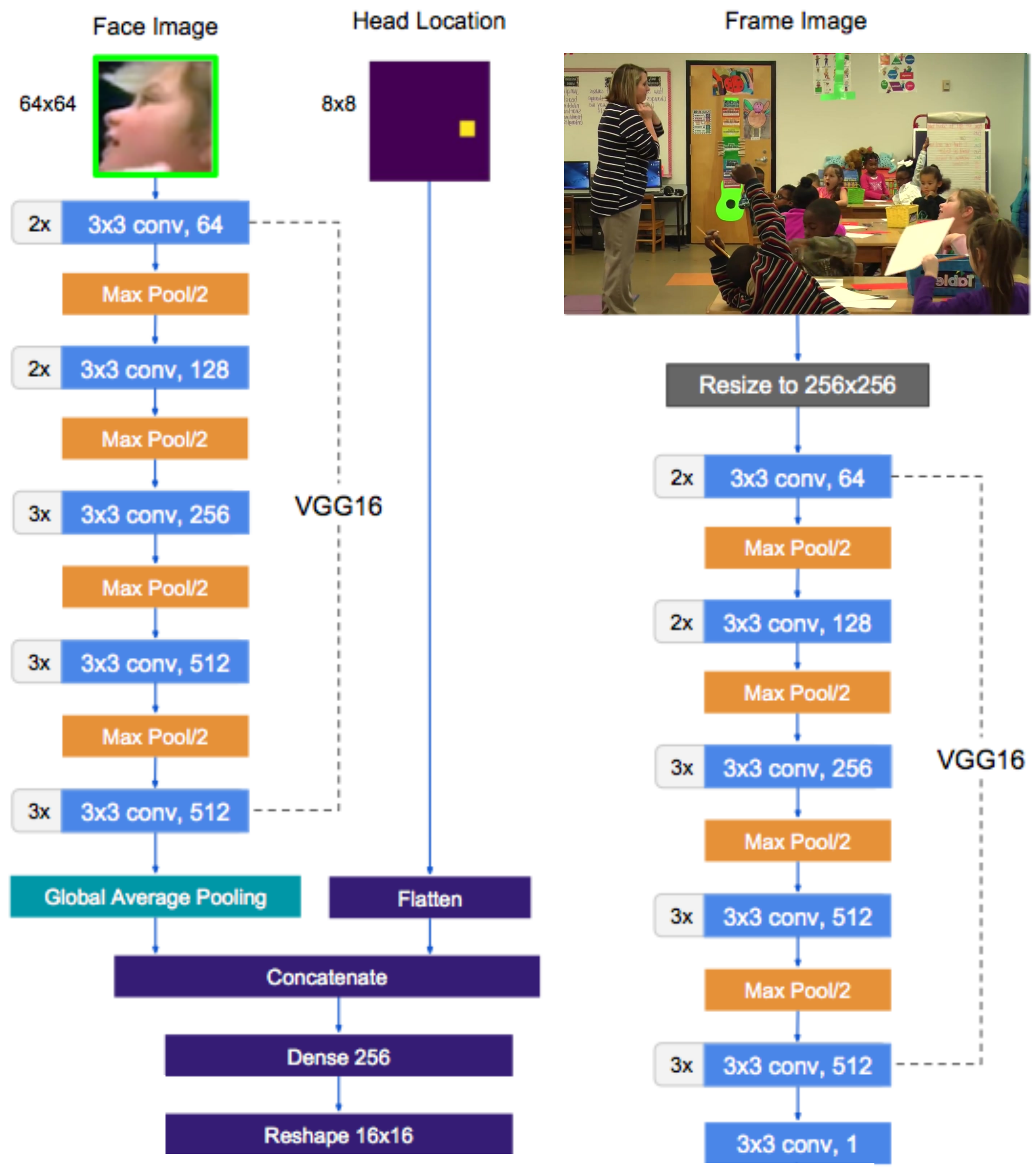


Face-to-Gaze pathway

Only have access to close-up face image and head location

Intuition:
1) Infer gaze from head pose



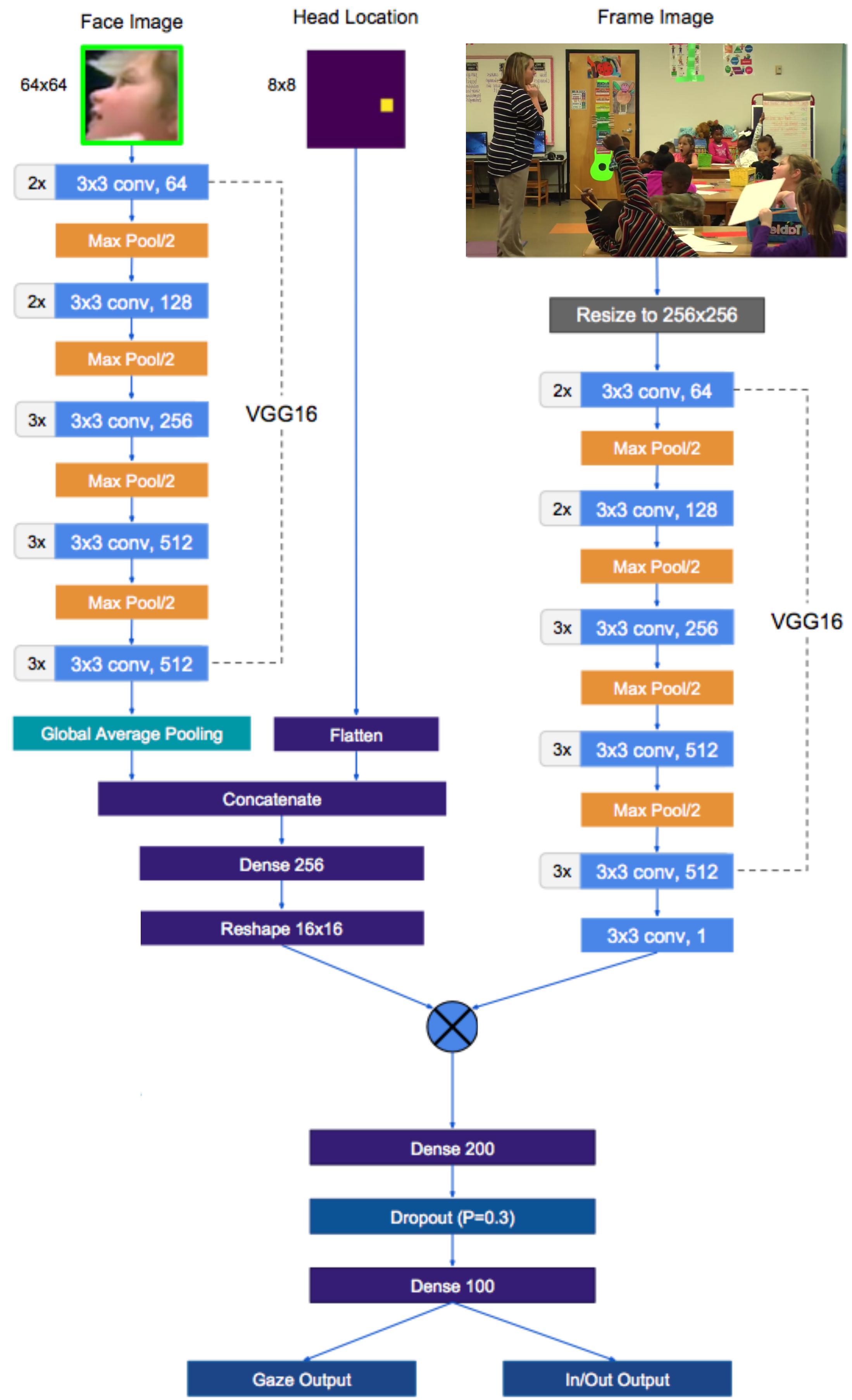


Frame pathway

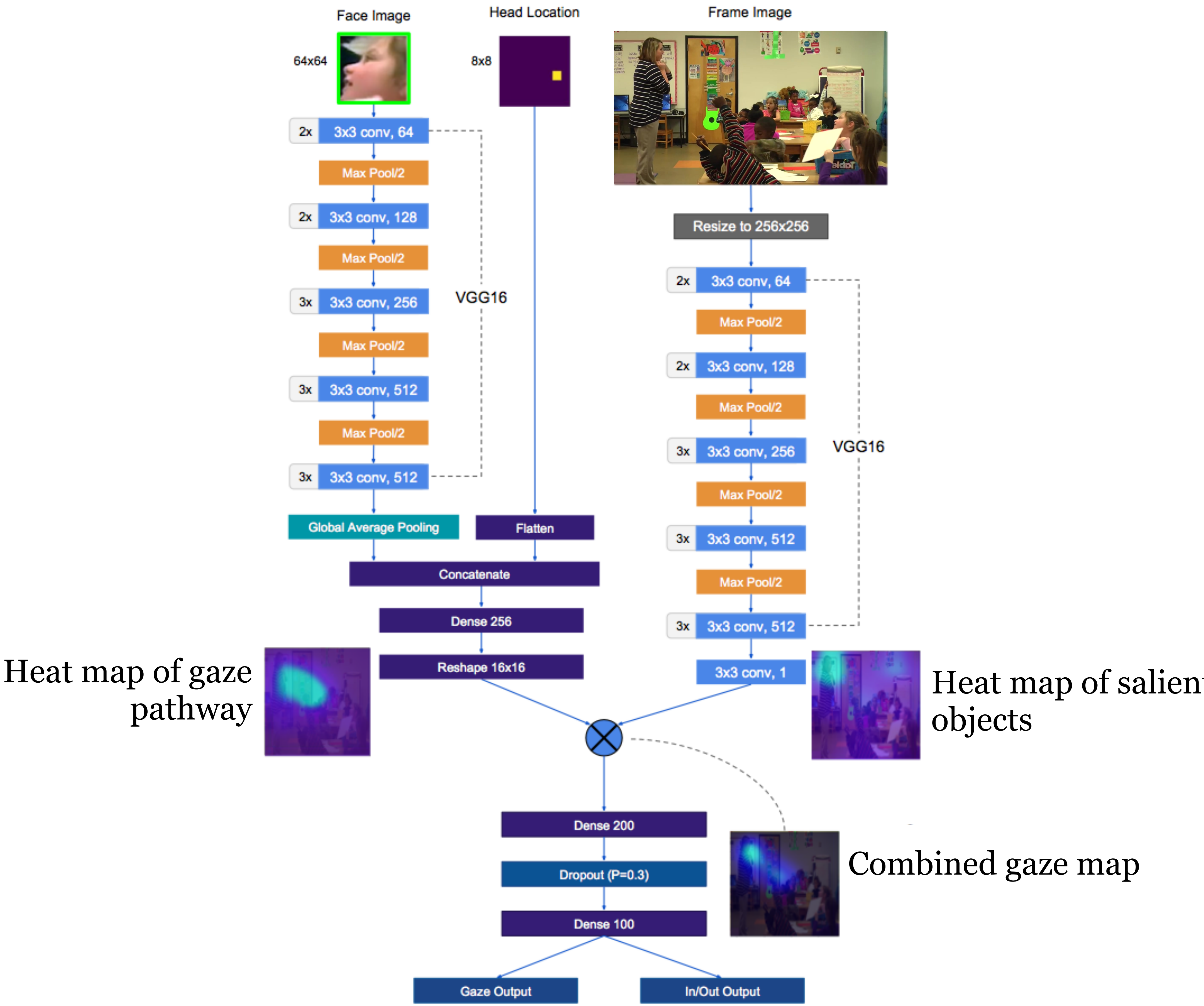
Only have access to image of the scene without knowing anything about where the subject of interest is

Intuition:

1) Learn to detect salient objects



Face Image
Head Location
Frame Image
64x64
8x8
2x 3x3 conv, 64
Max Pool/2
2x 3x3 conv, 128
Max Pool/2
3x 3x3 conv, 256
Max Pool/2
3x 3x3 conv, 512
Max Pool/2
3x 3x3 conv, 512
VGG16
Global Average Pooling
Flatten
Concatenate
Dense 256
Reshape 16x16
Resize to 256x256
2x 3x3 conv, 64
Max Pool/2
2x 3x3 conv, 128
Max Pool/2
3x 3x3 conv, 256
Max Pool/2
3x 3x3 conv, 512
Max Pool/2
3x 3x3 conv, 512
VGG16
3x3 conv, 1
Dense 200
Dropout (P=0.3)
Dense 100
Gaze Output
In/Out Output


Face Image
64x64
Head Location
8x8
Frame Image
2x 3x3 conv, 64
Max Pool/2
2x 3x3 conv, 128
Max Pool/2
3x 3x3 conv, 256
Max Pool/2
3x 3x3 conv, 512
Max Pool/2
3x 3x3 conv, 512
VGG16
Global Average Pooling
Flatten
Concatenate
Dense 256
Reshape 16x16
Resize to 256x256
2x 3x3 conv, 64
Max Pool/2
2x 3x3 conv, 128
Max Pool/2
3x 3x3 conv, 256
Max Pool/2
3x 3x3 conv, 512
Max Pool/2
3x 3x3 conv, 512
VGG16
3x3 conv, 1
Heat map of gaze pathway
Heat map of salient objects
Dense 200
Dropout (P=0.3)
Dense 100
Combined gaze map
Gaze Output
In/Out Output

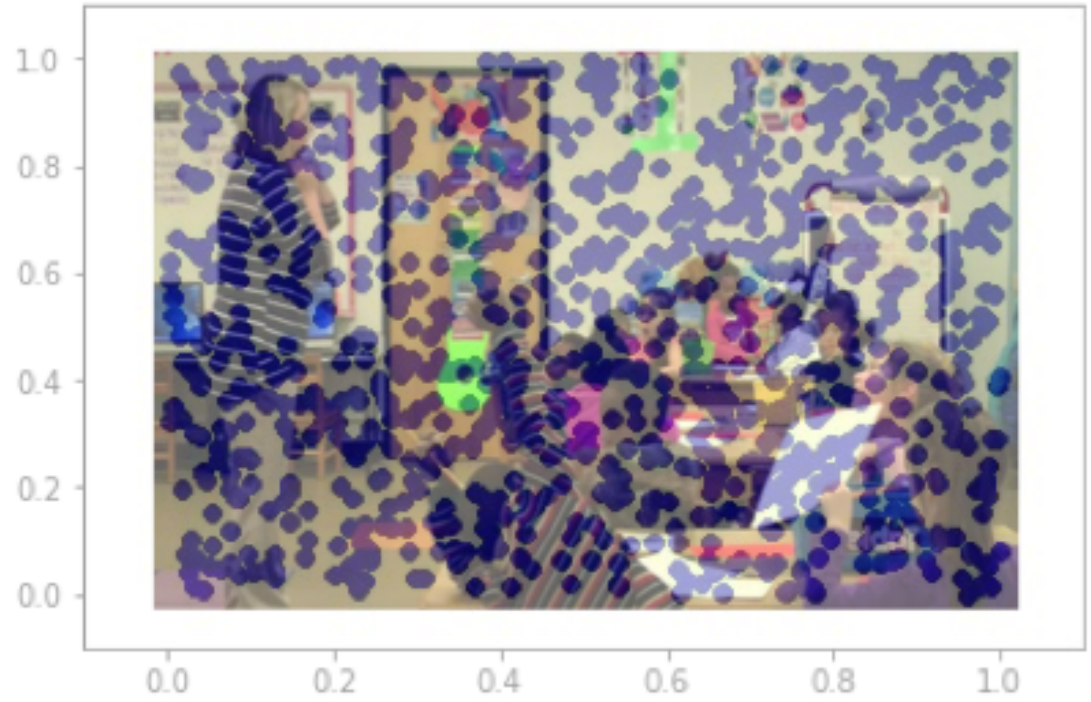
Research Questions

1. How accurately can the Merged Model predict gaze locations?
2. Can our Merged Model predict whom the person is looking at?

Results

Regression Baselines

- **Random Gaze**: Random location over the whole image.



[3] Judd, T., Ehinger, K., Durand, F., and Torralba, A. Learning to predict where humans look. In International Conference on Computer Vision (2009).

Regression Baselines

- **Random Gaze**: Random location over the whole image.
- **Center Region**: Random gaze constrained to center 10% of the image. Motivated by Judd, et al[3].

[3] Judd, T., Ehinger, K., Durand, F., and Torralba, A. Learning to predict where humans look. In International Conference on Computer Vision (2009).

Regression Baselines

- **Random Gaze**: Random location over the whole image.
- **Center Region**: Random gaze constrained to center 10% of the image. Motivated by Judd, et al[3].
- **Linear regression**: use shallow network to predict (x,y) from close-up cropped face and head location.

[3] Judd, T., Ehinger, K., Durand, F., and Torralba, A. Learning to predict where humans look. In International Conference on Computer Vision (2009).

Regression Baselines

- **Random Gaze**: Random location over the whole image.
- **Center Region**: Random gaze constrained to center 10% of the image. Motivated by Judd, et al[3].
- **Linear regression**: use shallow network to predict (x,y) from close-up cropped face and head location.
- **Face-to-Gaze**: Left half of **Merged Model**. Only have access to close-up cropped face and head location.

[3] Judd, T., Ehinger, K., Durand, F., and Torralba, A. Learning to predict where humans look. In International Conference on Computer Vision (2009).

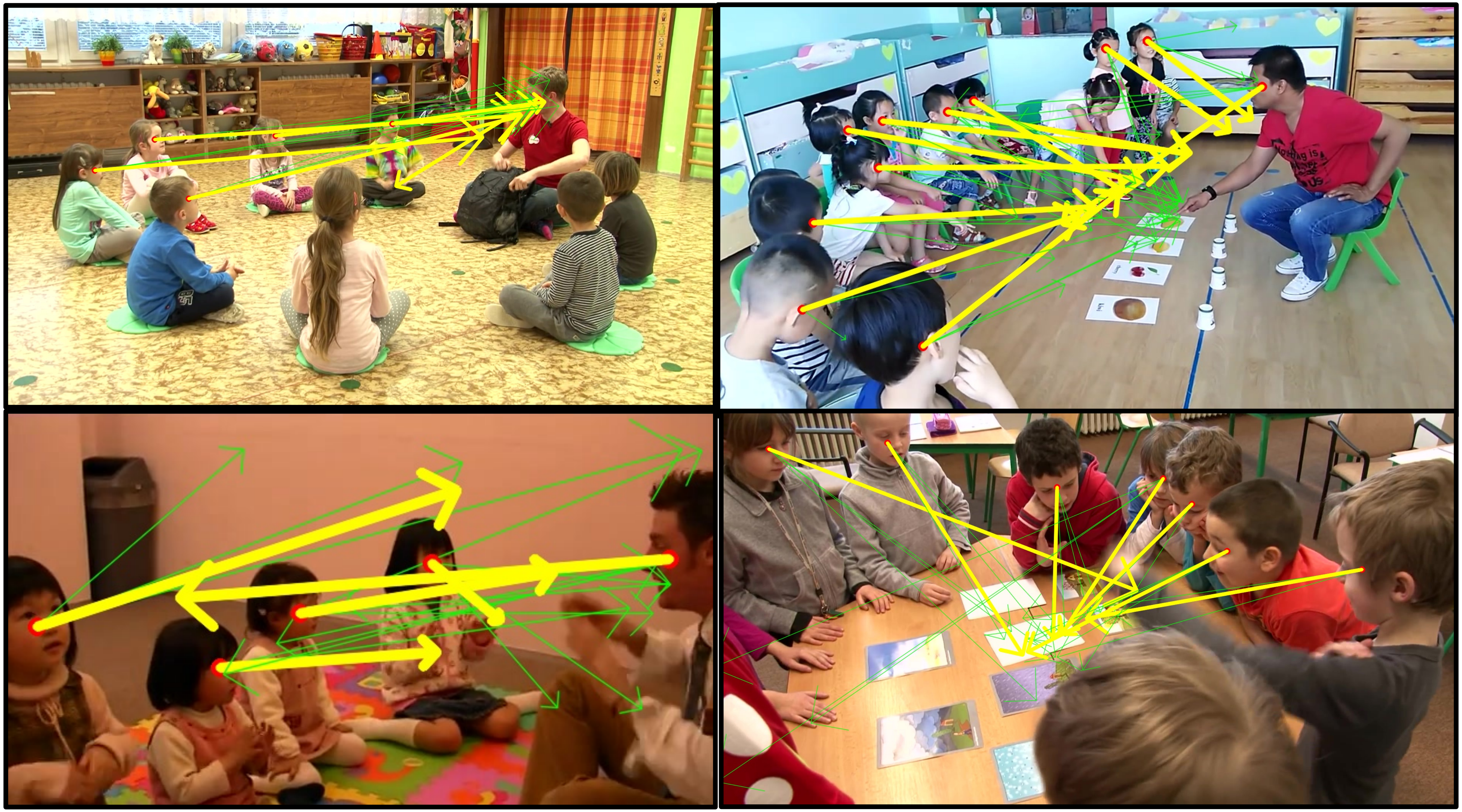
Regression Results

Regression results (within 256x256 pixel image)

	MAE*	Mean Euclidean Distance*	Mean Absolute Angular Error	AUC for In/Out
Random Gaze	79.74	124.15	67.24°	-
Center Region	52.76	82.11	48.36°	-
Linear Regression	49.63	77.34	55.21°	-
Face-to-Gaze	45.74	71.53	39.91°	0.54
Merged Model	**44.49**	**69.82**	**38.30°**	**0.62**
Human	25.91	41.04	18.38°	0.70

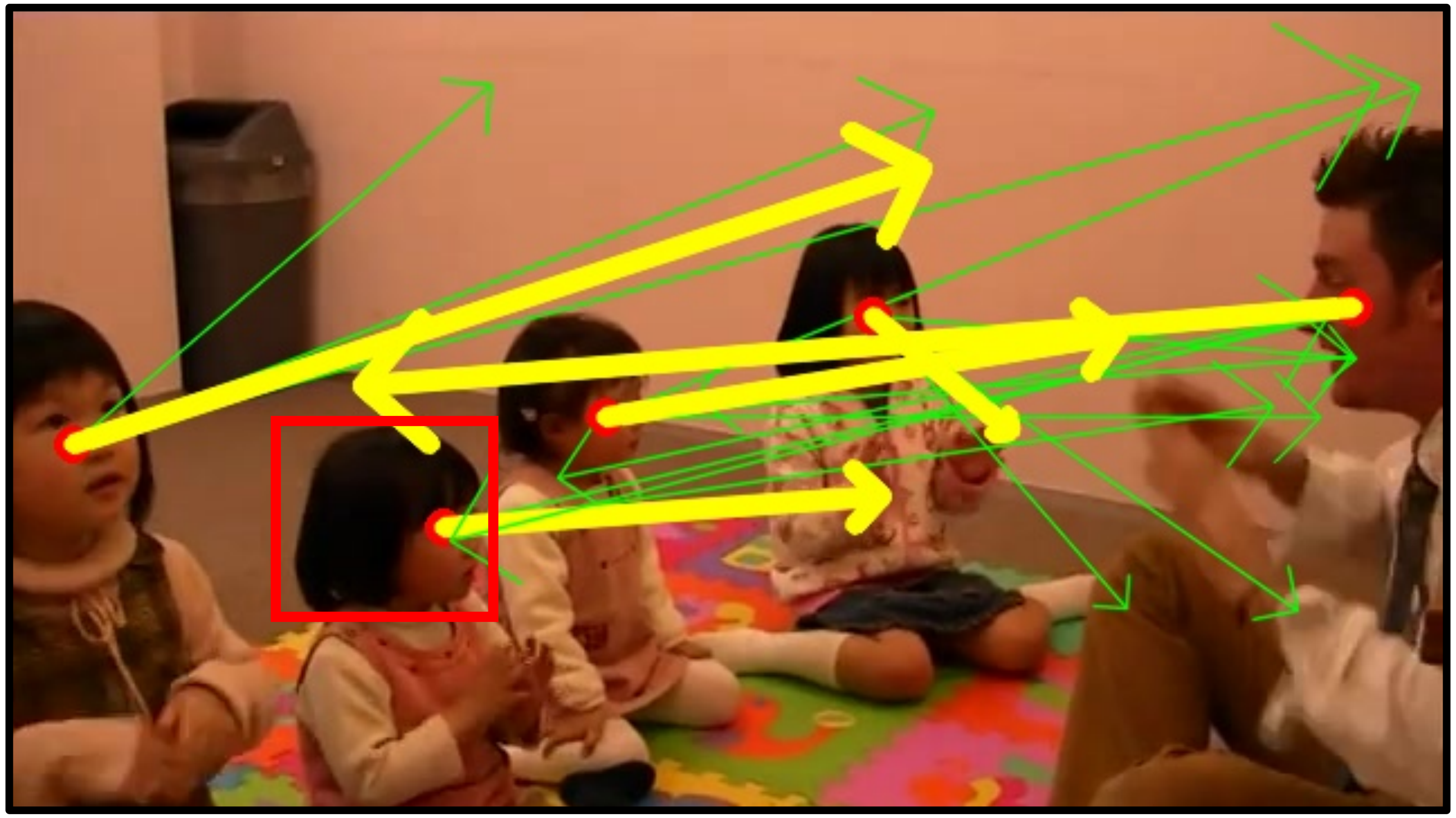
*Distance in pixels

Qualitative Results (Regression)



Qualitative Results (Regression)

- The merged model sometimes accurately estimates the direction, but not the distance, of the gaze.
- E.g., the girl in red box is looking at teacher’s hands but the gaze end-point stops before getting to the hands.



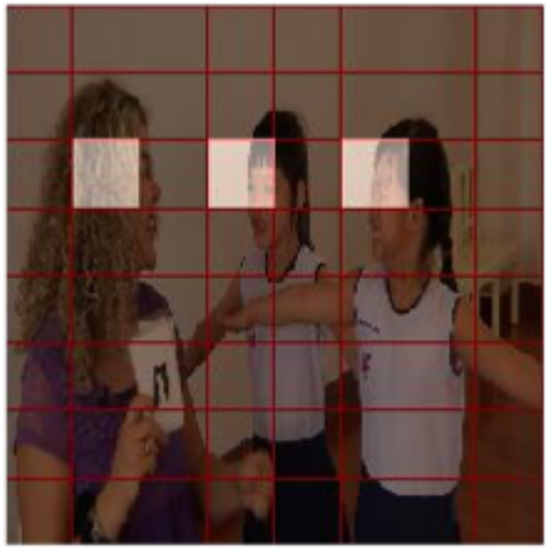
Who are they looking at?

Who are they looking at?

- Analyze subset of faces s.t. all annotators agree he/she is looking at another *face* (not just any other object).
- Prediction task: *given* that the person is looking at a face, *whose face* is he/she looking at?

Merged Model Predictions on faces

- Start with the network's predictions on 8x8 grid.
- Remove any cells containing no faces.
- Find top *k*=1 cells with highest predicted gaze probability.
- Predict the face contained within that cell.

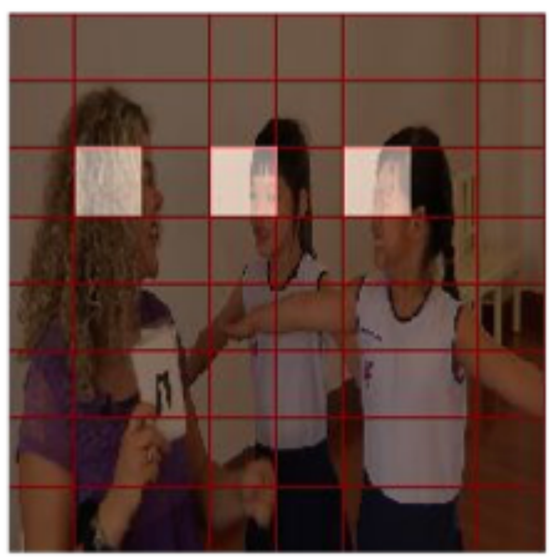


Face cells on 8x8 grid

Merged model predictions in color
(Top 1 face – 3
Top 2 face – 2 or 3
Top 3 faces – 1,2 or 3)

Merged Model Predictions on faces

- Start with the network's predictions on 8x8 grid.
- Remove any cells containing no faces.
- Find top k=1 cells with highest predicted gaze probability.
- Predict the face contained within that cell.
- Can also consider top k=1,2,3 faces (c.f. object detection literature).



Face cells on 8x8 grid

Merged model predictions in color
(Top 1 face – 3
Top 2 face – 2 or 3
Top 3 faces – 1,2 or 3)

Results for “Who are they looking at?”

Probability of correctly identifying which face a person is looking at on 8 × 8 grid.

Top k faces	$k = 1$	$k = 2$	$k = 3$
Random Face	0.15	0.30	0.45
Merged Model	**0.47**	**0.65**	**0.79**
Human	0.82		

- 6.87 faces per image on average (for test set)

Results for “Who are they looking at?”

Probability of correctly identifying which face a person is looking at on 8 × 8 grid.

Top k faces	$k = 1$	$k = 2$	$k = 3$
Random Face	0.15	0.30	0.45
Merged Model	**0.47**	**0.65**	**0.79**
Human	0.82		

- 6.87 faces per image on average (for test set)

- 79% of the time, NN can correctly “narrow down” the gazed-at face to a set of 3 people.

Summary

Summary

- With a modest-sized (70 classroom observation videos) dataset, we can train a NN to predict eye gaze (where & whom) from 2-D images.
  - **Whom**: 79% of the time, NN can correctly "narrow down" the possible gaze targets to < 1/2 the number of classroom participants.

Summary

- With a modest-sized (70 classroom observation videos) dataset, we can train a NN to predict eye gaze (where & whom) from 2-D images.
  - **Whom**: 79% of the time, NN can correctly "narrow down" the possible gaze targets to < 1/2 the number of classroom participants.
- Eye gaze is just one of many behavioral markers that could be useful for classroom observation.

Summary

- With a modest-sized (70 classroom observation videos) dataset, we can train a NN to predict eye gaze (where & whom) from 2-D images.
  - **Whom**: 79% of the time, NN can correctly "narrow down" the possible gaze targets to < 1/2 the number of classroom participants.
- Eye gaze is just one of many behavioral markers that could be useful for classroom observation.
- Long-term goal is to integrate many (noisy) predictors into an automated — or hybrid — classroom observation system.

End

spencer

NSF