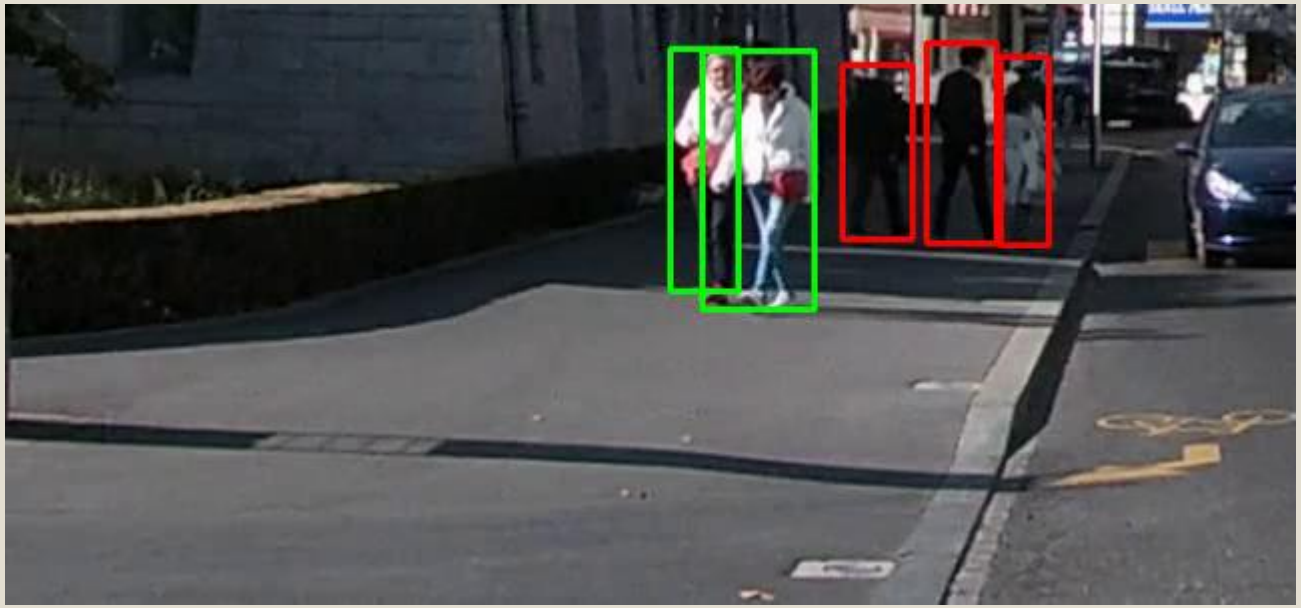
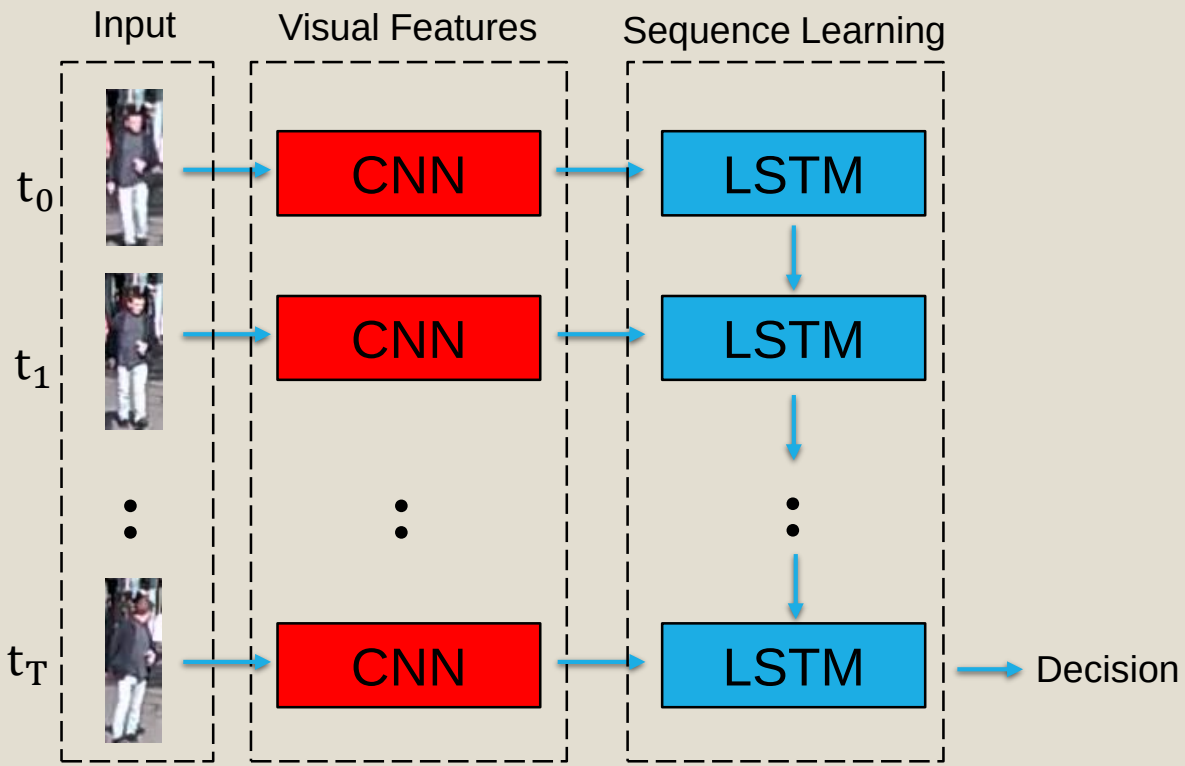


Pedestrian Intention Prediction

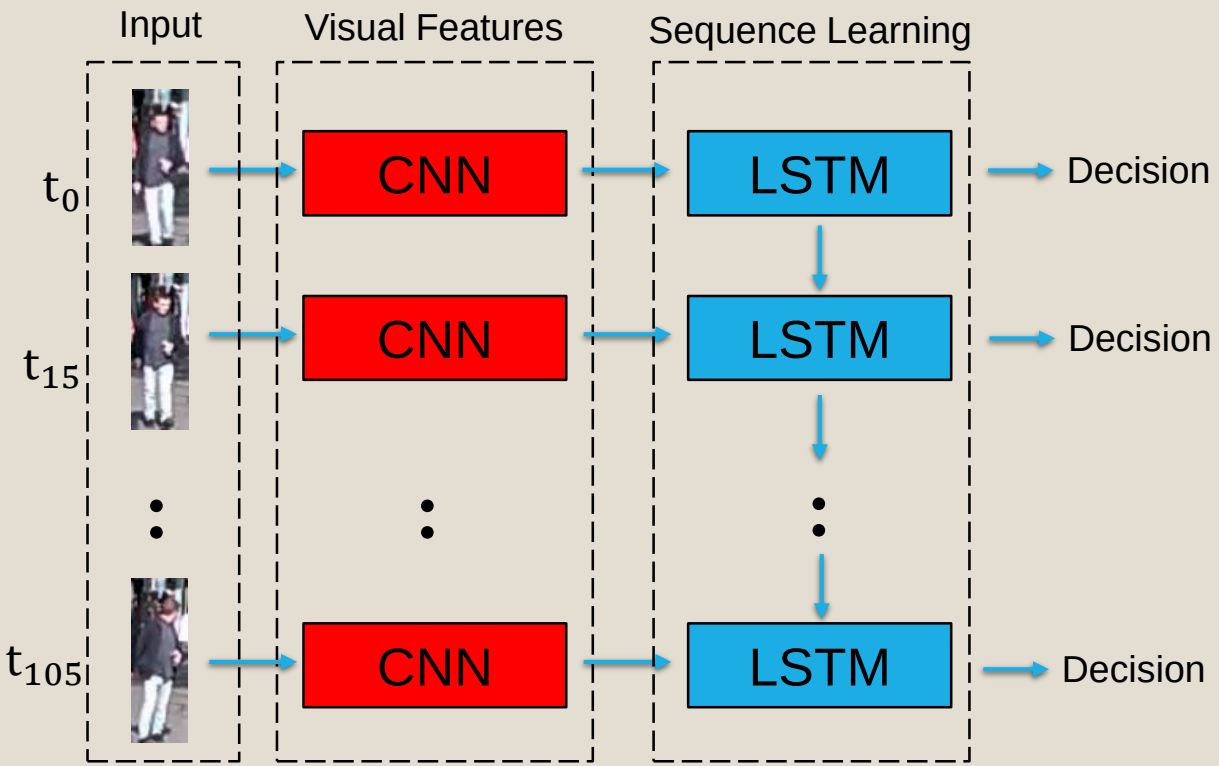
Muhammad Haziq Bin Razali
Supervisor: Professor Alexandre Alahi

Baseline (Training)

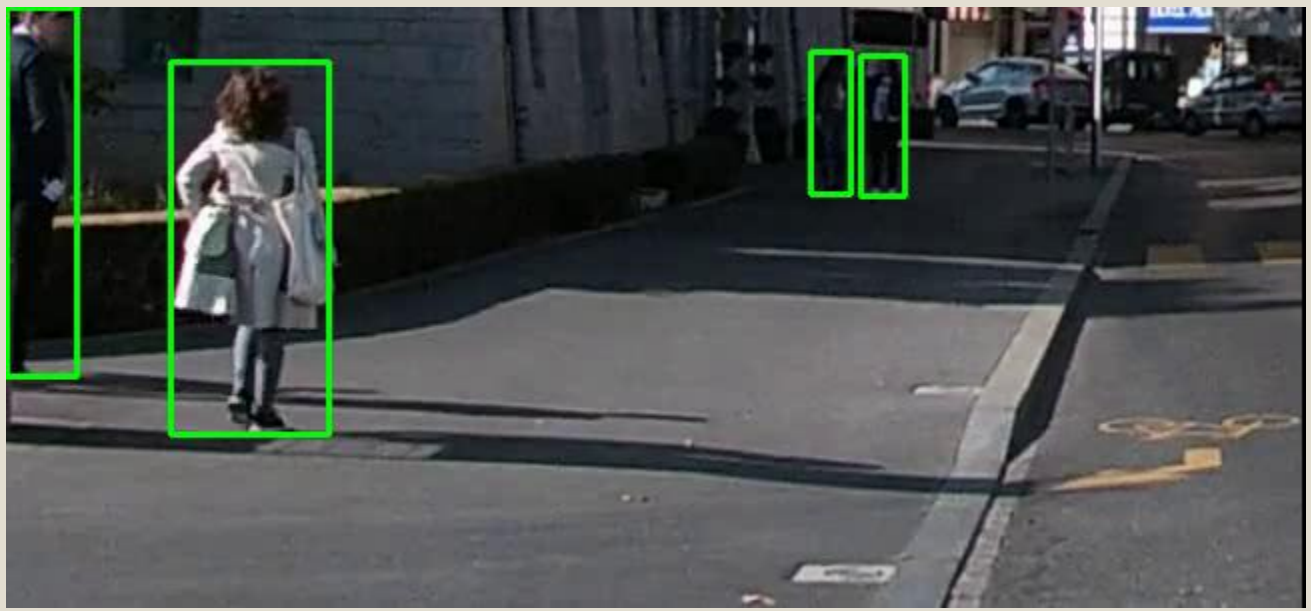


- Model should learn to detect changes in orientation
- Trained on final 8 crops spaced 15 frames apart
- Only used crops that are not in the crossing
- Classifier at the end

Baseline (Testing)



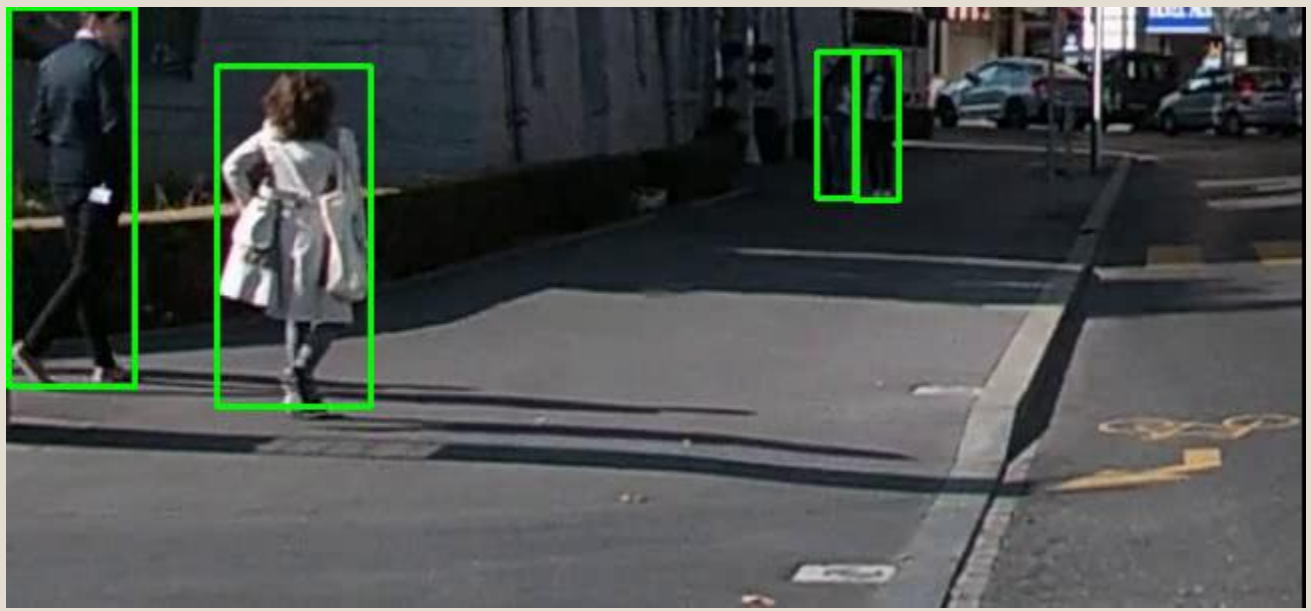
- Classifier at every timestep
- Model to detect changes in orientation



Not crossing

Crossing

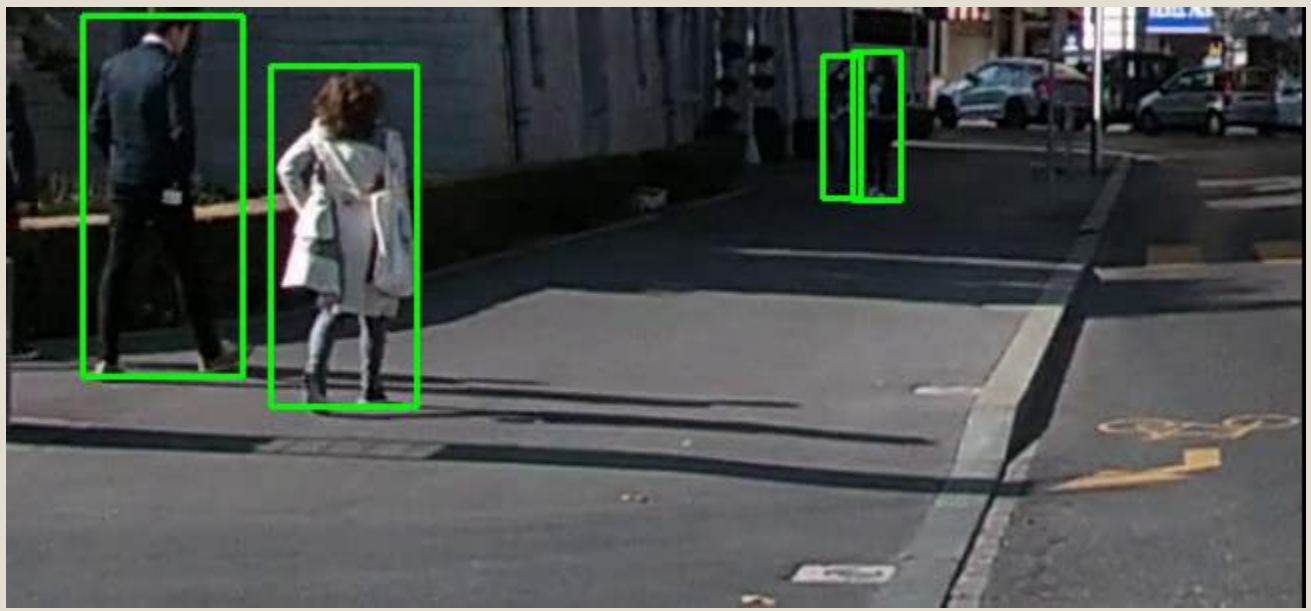
- Input processed every 15 frames



Not crossing

Crossing

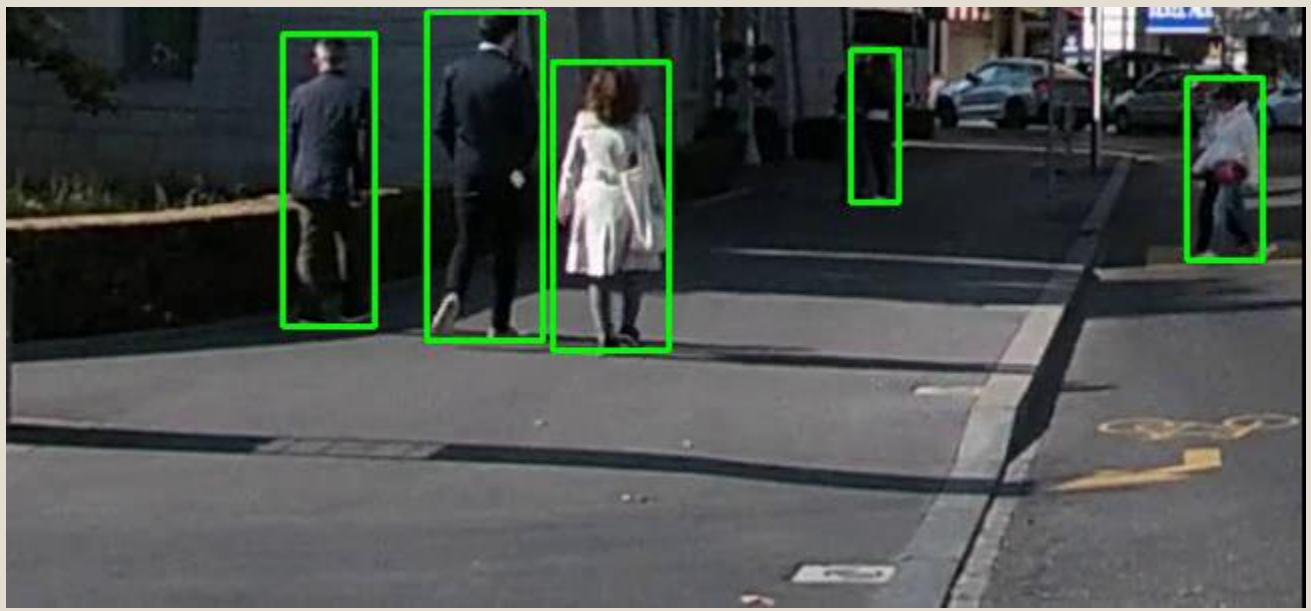
- Input processed every 15 frames



Not crossing

Crossing

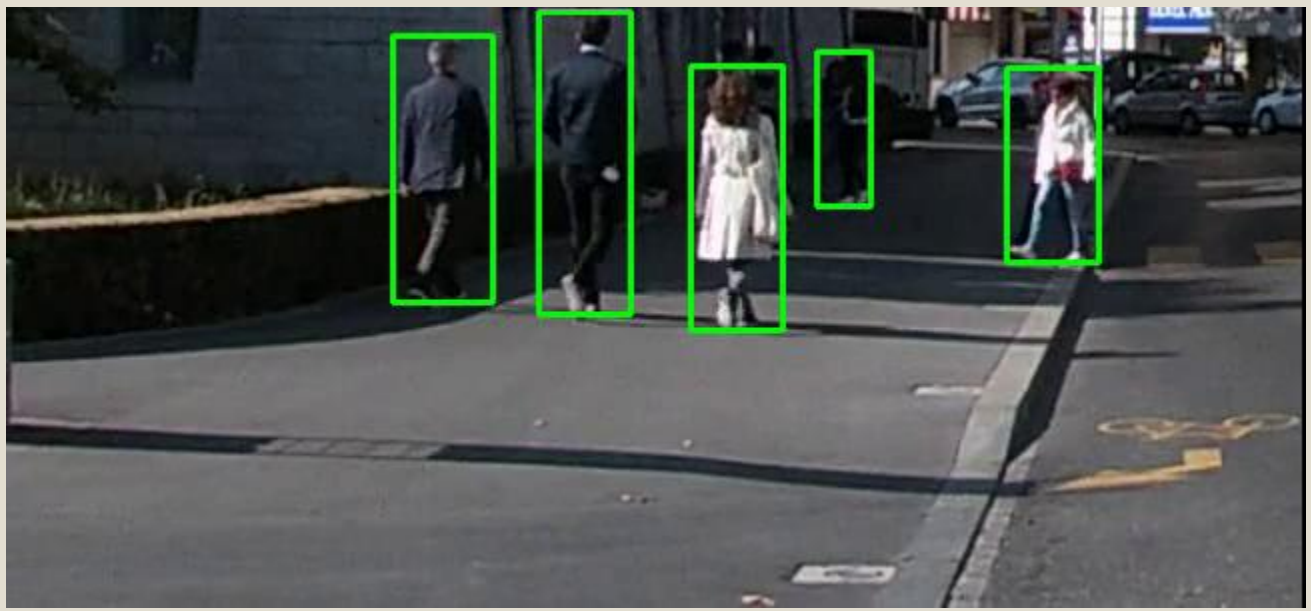
- Input processed every 15 frames



Not crossing

Crossing

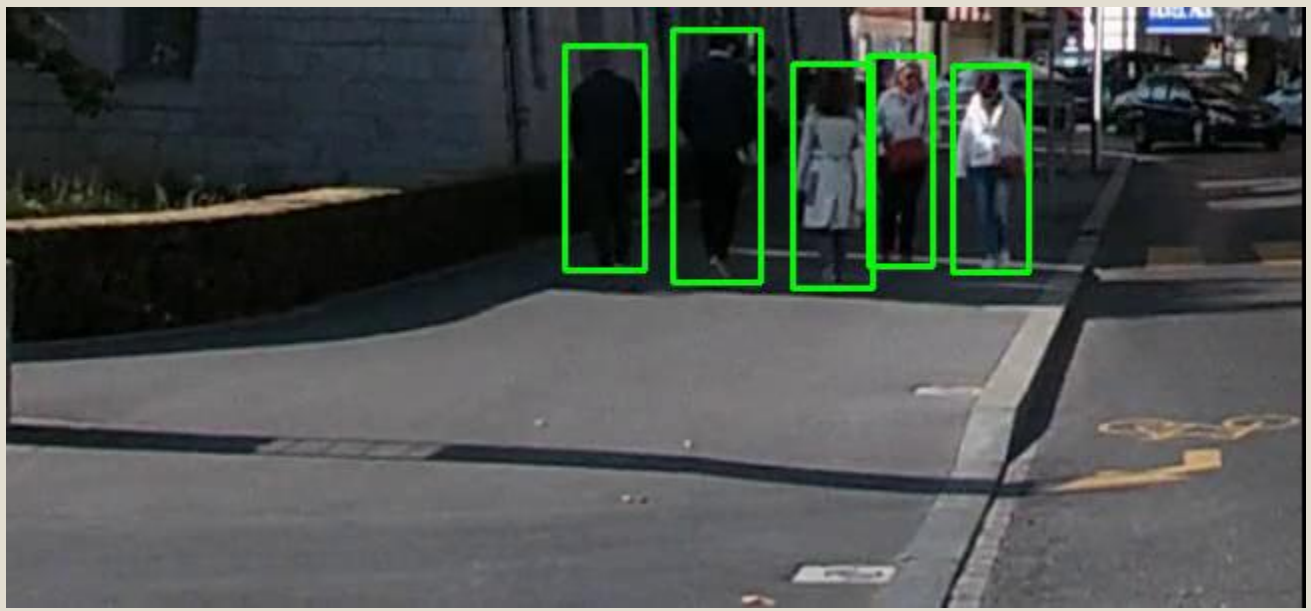
- Input processed every 15 frames



Not crossing

Crossing

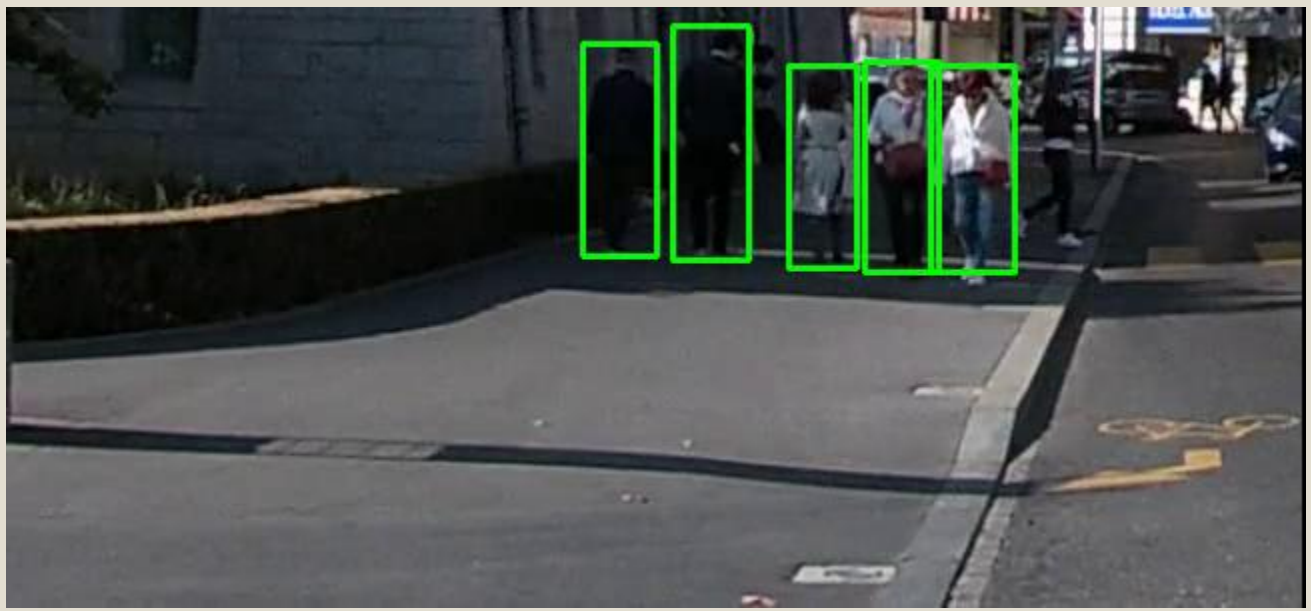
- Input processed every 15 frames



Not crossing

Crossing

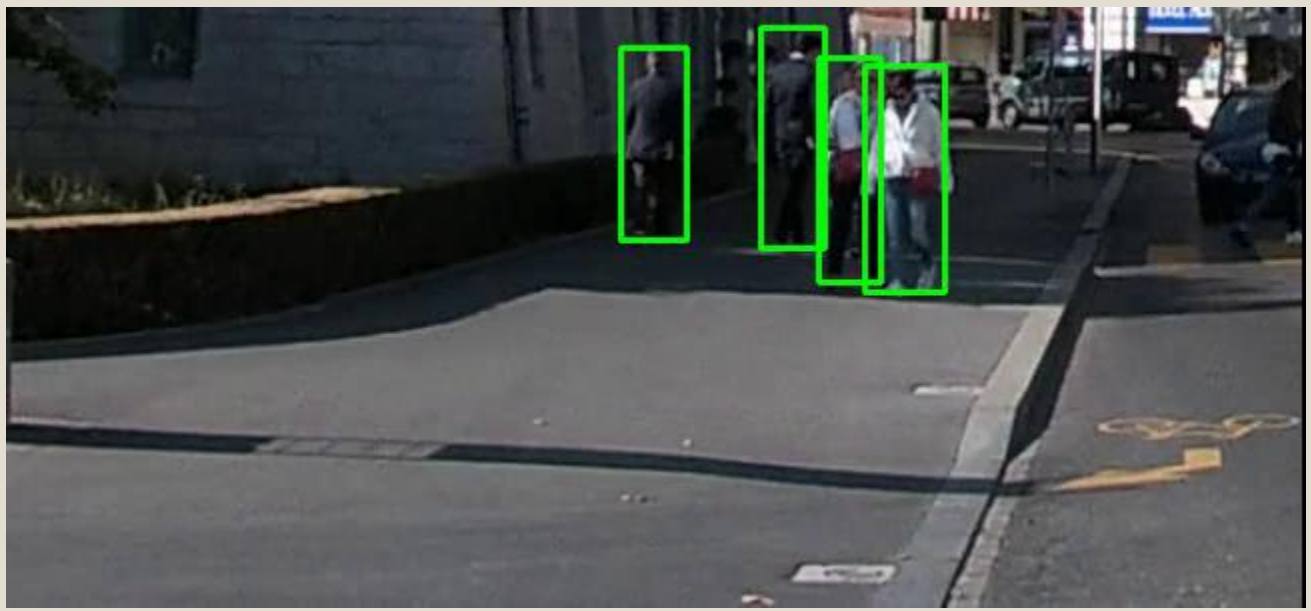
- Input processed every 15 frames



Not crossing

Crossing

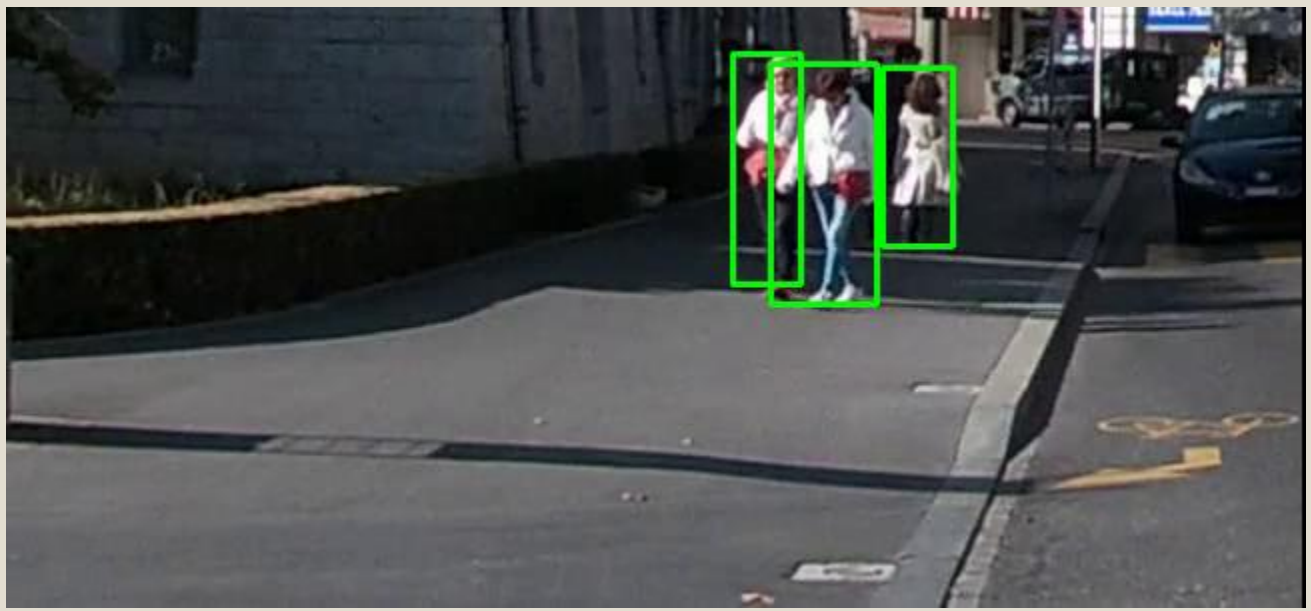
- Input processed every 15 frames



Not crossing

Crossing

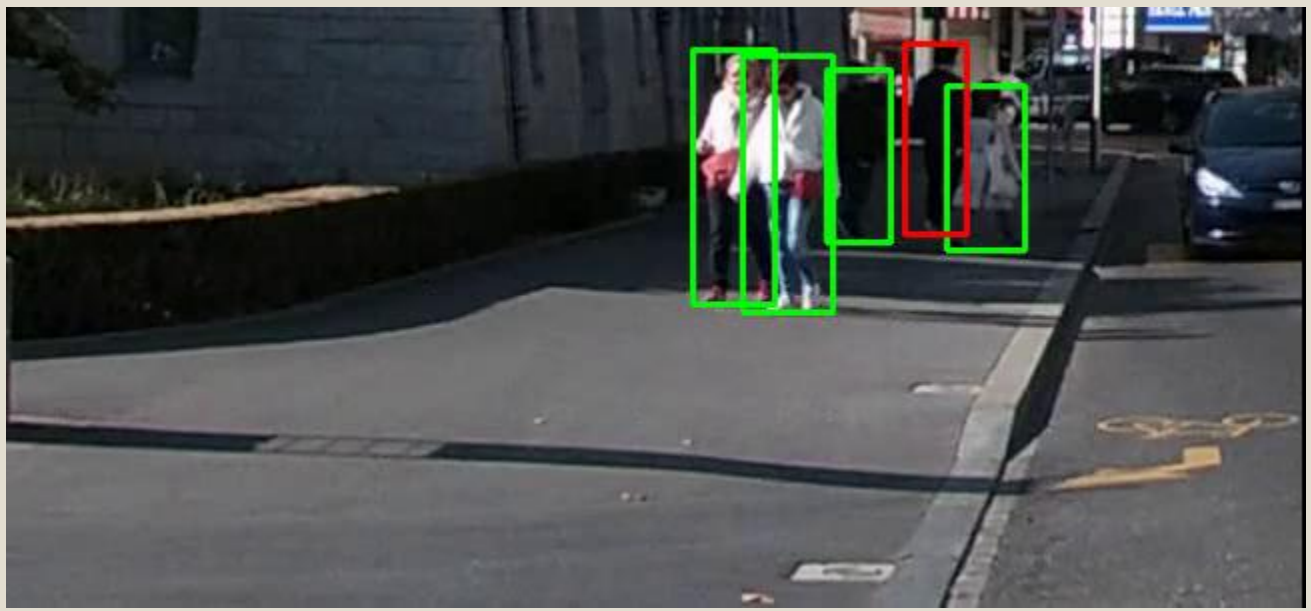
- Input processed every 15 frames



Not crossing

Crossing

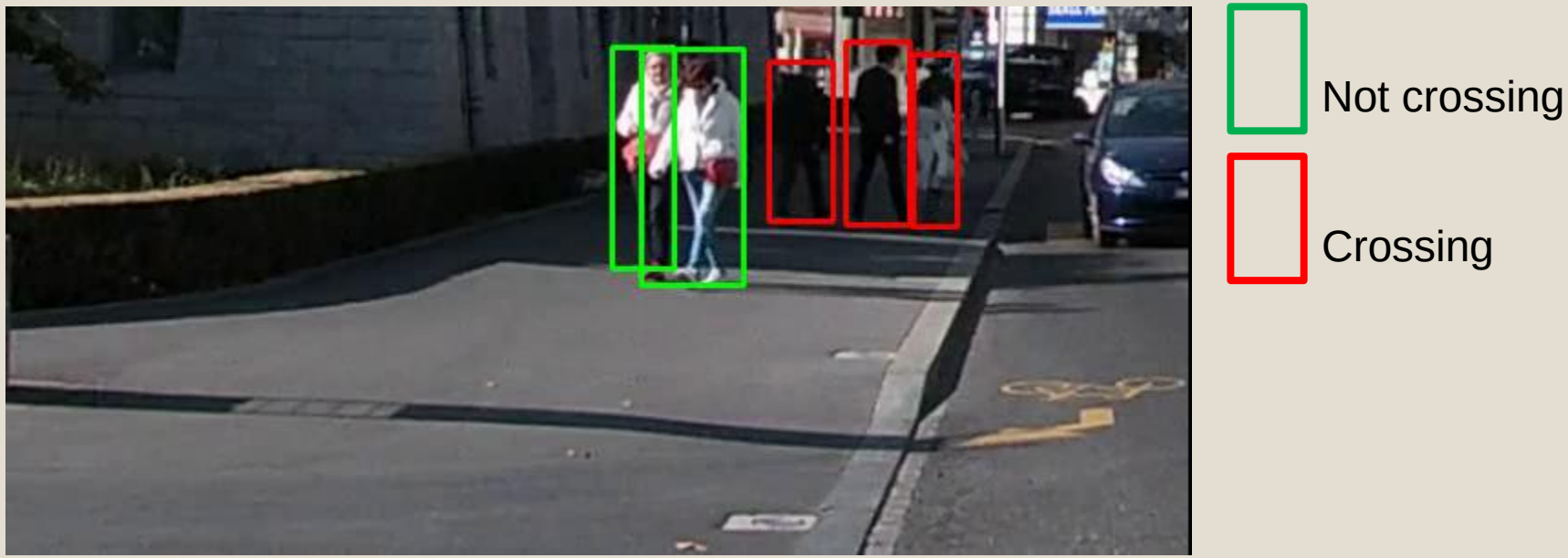
- Input processed every 15 frames



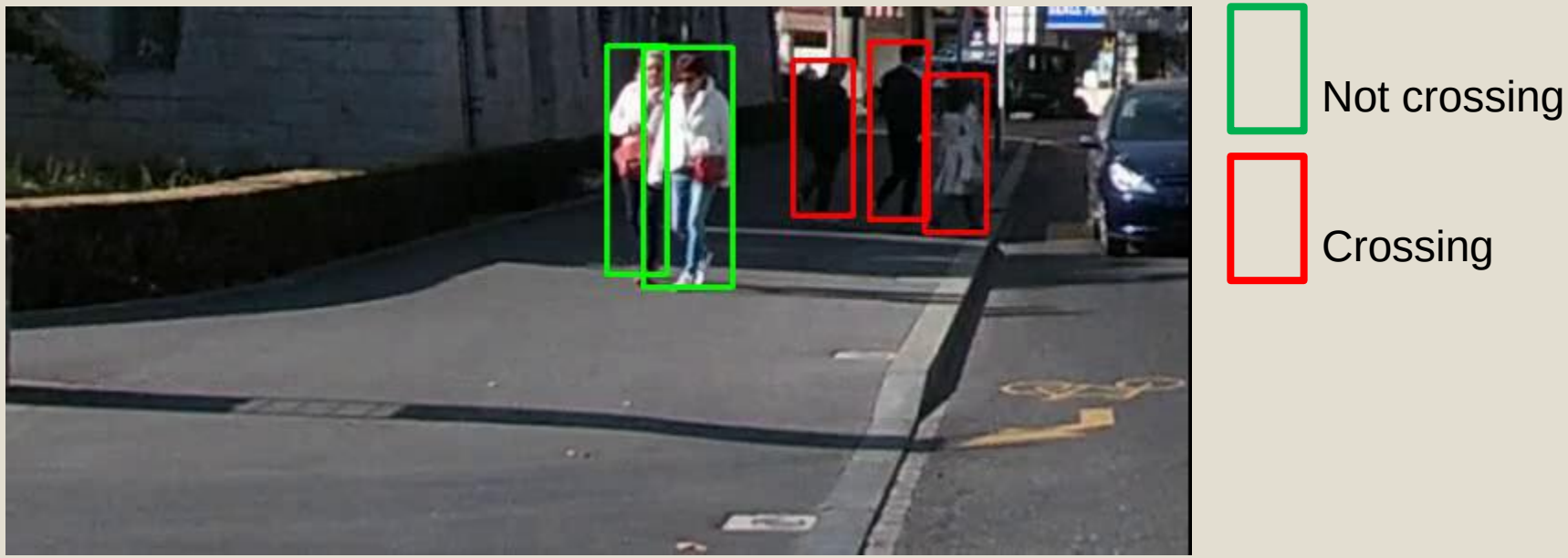
Not crossing

Crossing

- Classifier detects change in orientation as they turn towards the crossing



- Classifier detects change in orientation as they turn towards the crossing

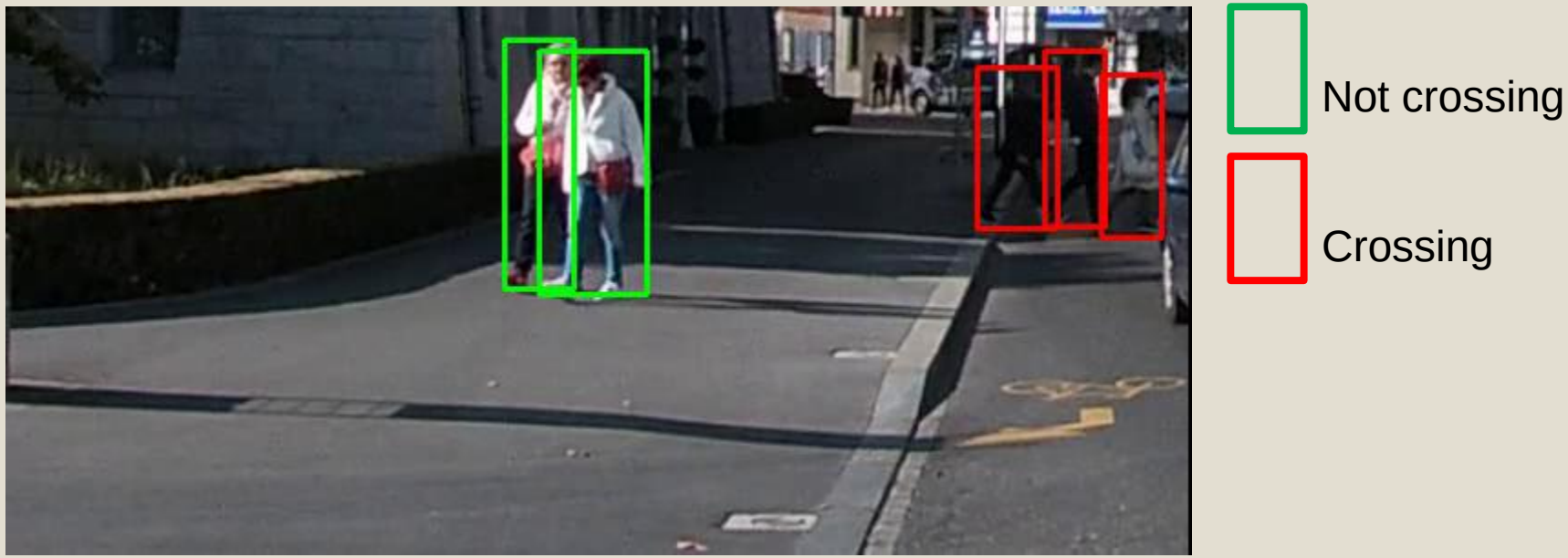


- Classifier detects change in orientation as they turn towards the crossing

Not crossing

Crossing

- Classifier predicts “not crossing"
- Architecture only trained to predict at the final 8’th timestep
- But made to continuously predict for testing



- Eventually switches back to “crossing”

Not crossing

Crossing

Not crossing

Crossing

Not crossing

Crossing

Not crossing

Crossing

Not crossing

Crossing

Not crossing

Crossing

Not crossing

Crossing

Not crossing

Crossing

Not crossing

Crossing

Not crossing

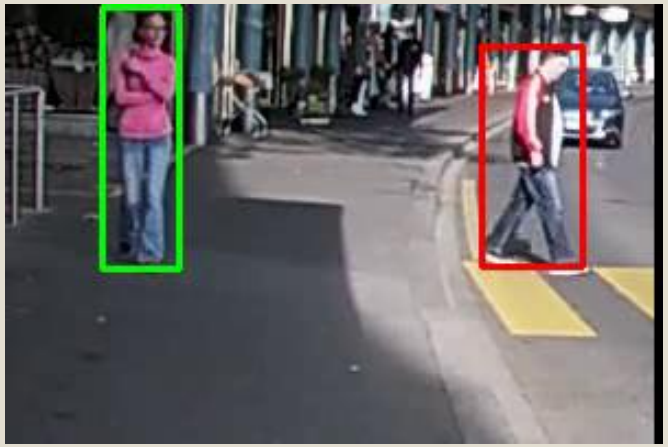
Crossing

Not crossing

Crossing

Not crossing

Crossing



Not crossing

Crossing

Not crossing

Crossing

Not crossing

Crossing

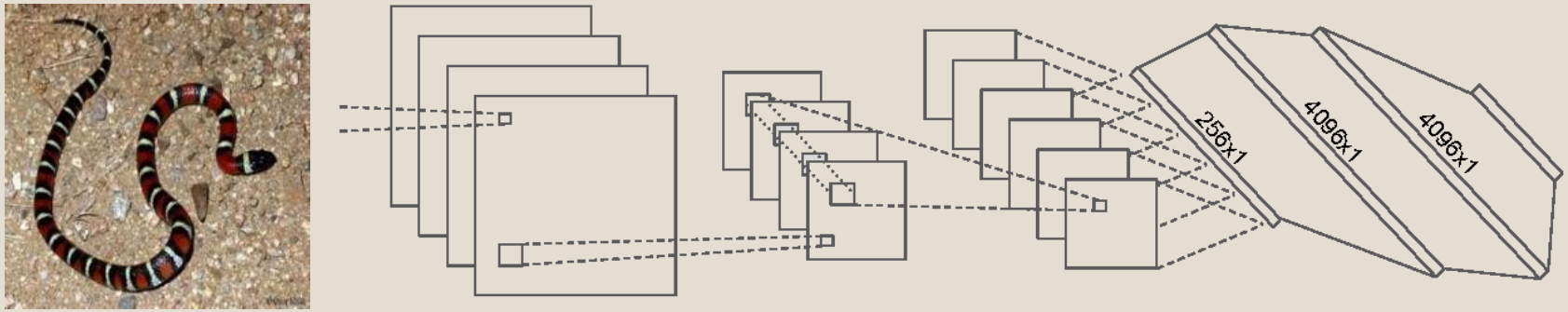
Not crossing

Crossing

Guided Backpropagation

Forward Pass

Which pixels are important for classification ?

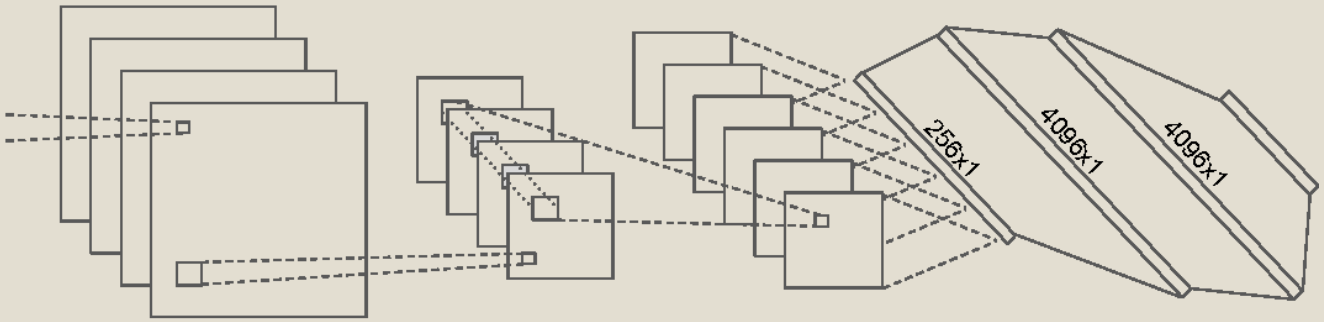


Snake

Guided Backpropagation

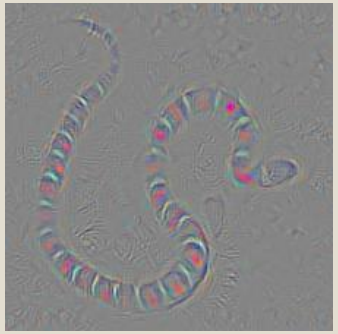
Forward Pass

Which pixels are important for classification ?



Snake

Backward Pass

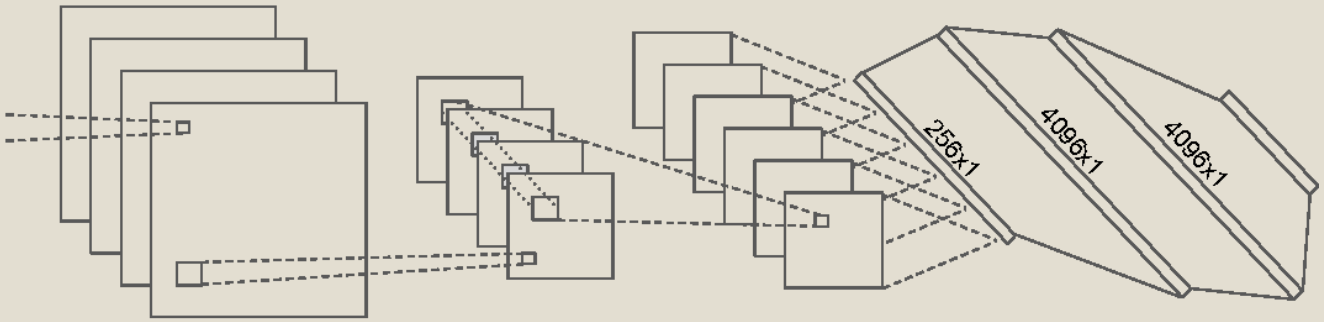


- Compute gradient of **<u>class score</u>** with respect to **<u>input</u>**

Guided Backpropagation

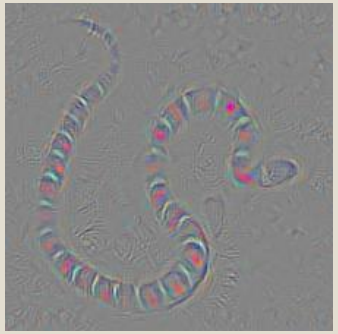
Which pixels are important for classification ?

Forward Pass



Snake

Backward Pass

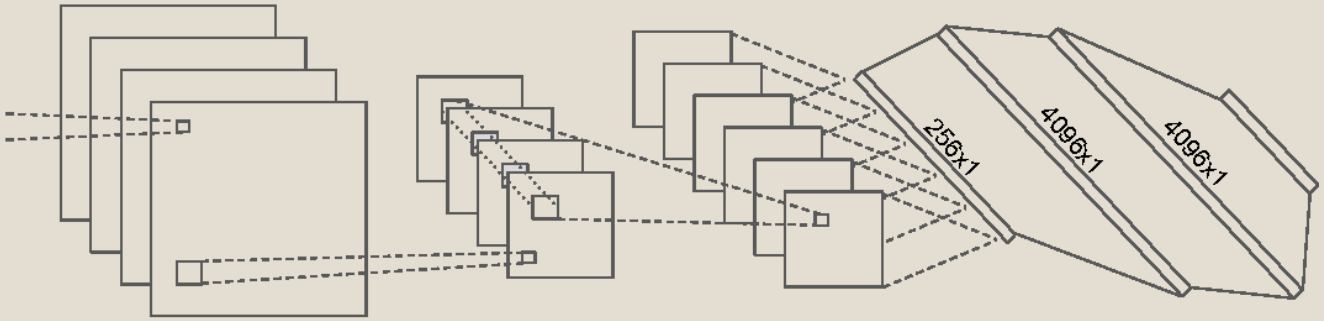


- Compute gradient of **<u>class score</u>** with respect to **<u>input</u>**
- ‘Gradient image’ illustrates pixels that positively affect the output class

Guided Backpropagation

Which pixels are important for classification ?

Forward Pass



Dog

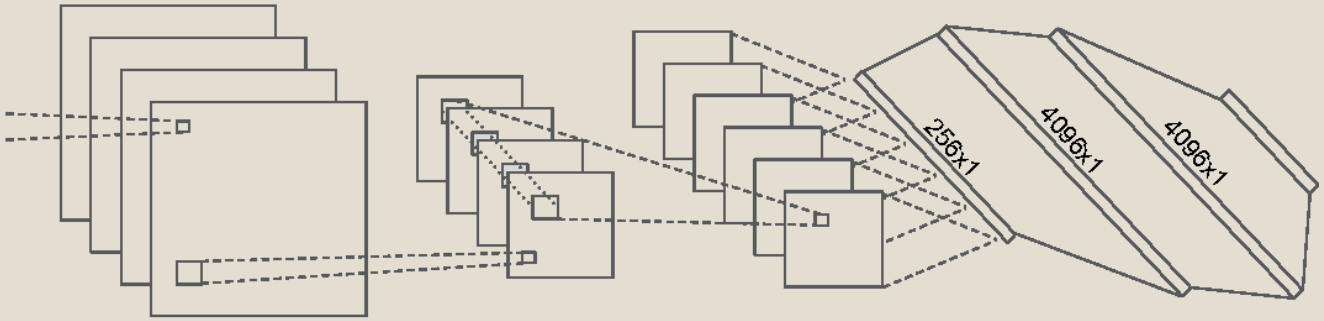
Backward Pass

- Compute gradient of **<u>class score</u>** with respect to **<u>input</u>**
- ‘Gradient image’ illustrates pixels that positively affect the output class

Guided Backpropagation

Which pixels are important for classification ?

Forward Pass



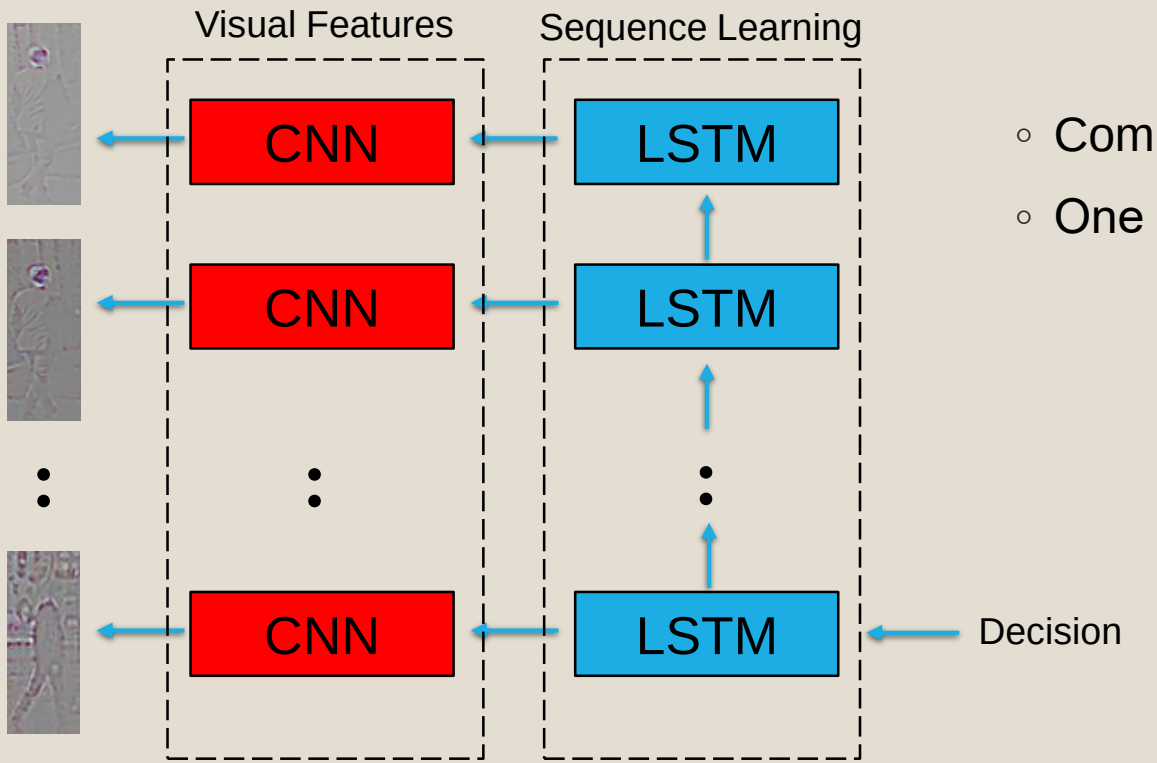
Dog

Backward Pass

- Compute gradient of **<u>class score</u>** with respect to **<u>input</u>**
- ‘Gradient image’ illustrates pixels that positively affect the output class
- Shows us what the architecture looks at during the forward pass

Guided Backpropagation

Which pixels are important for classification ?

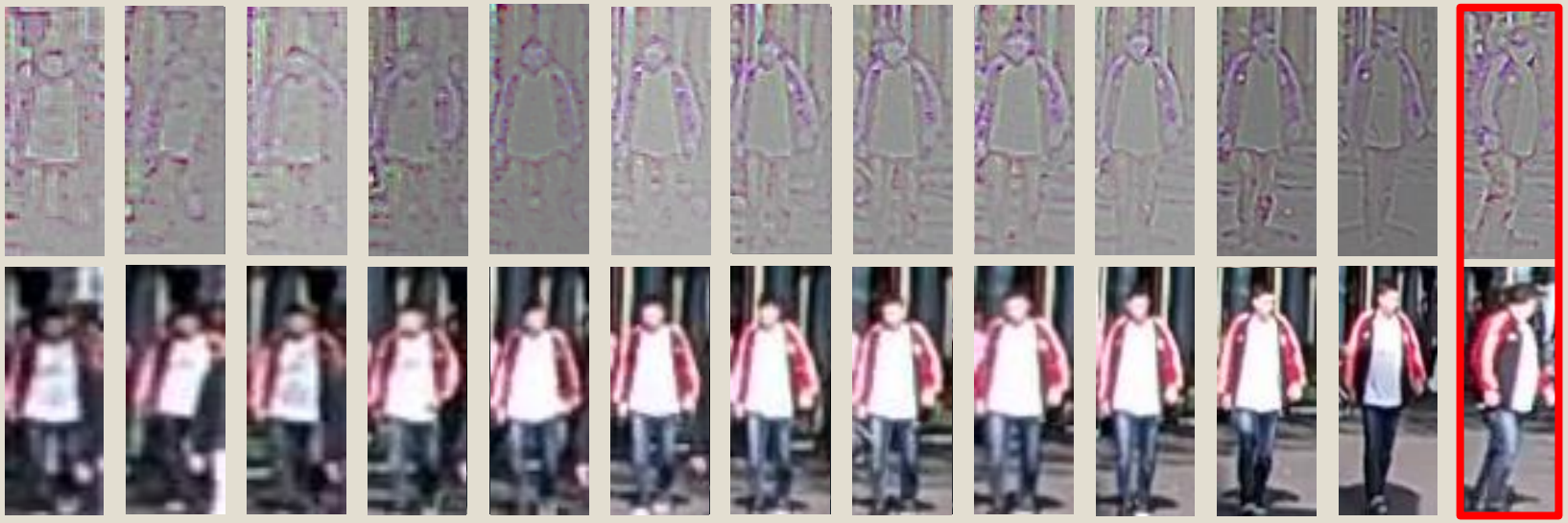


- Compute gradient of **decision** with respect to **input**
- One ‘gradient image’ for each input

Guided Backpropagation

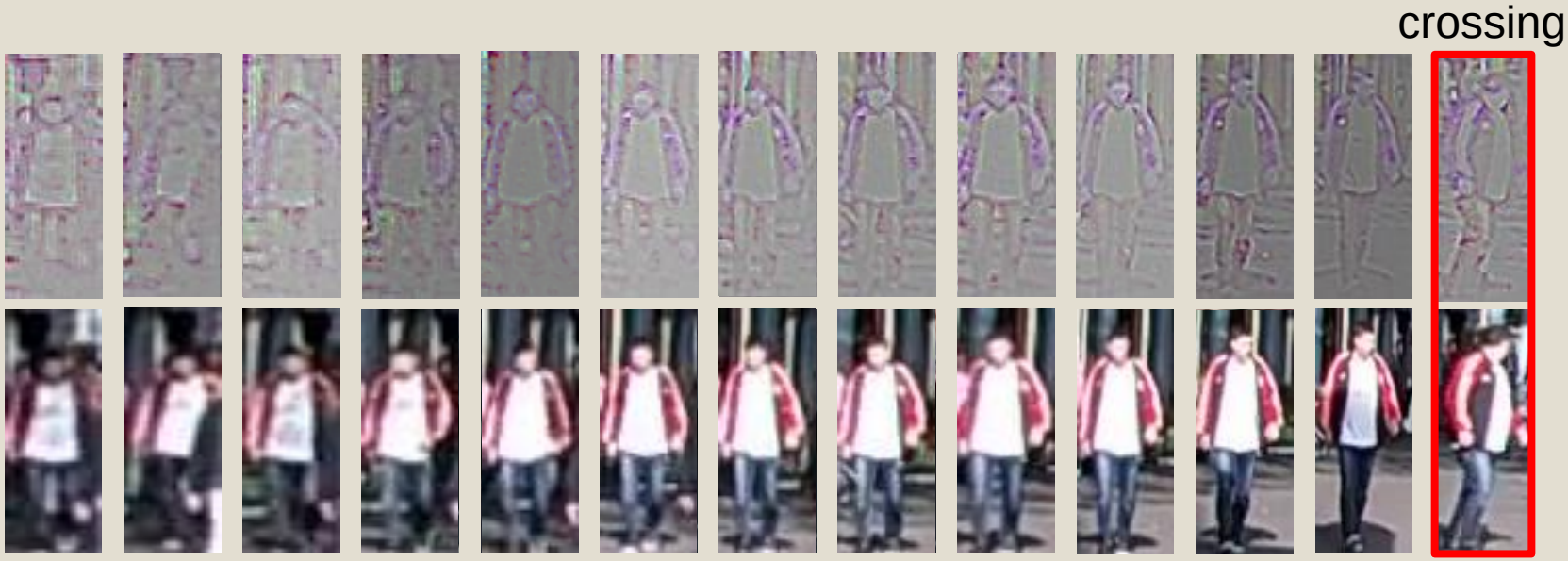
Backprop when classifier predicts a crossing

crossing



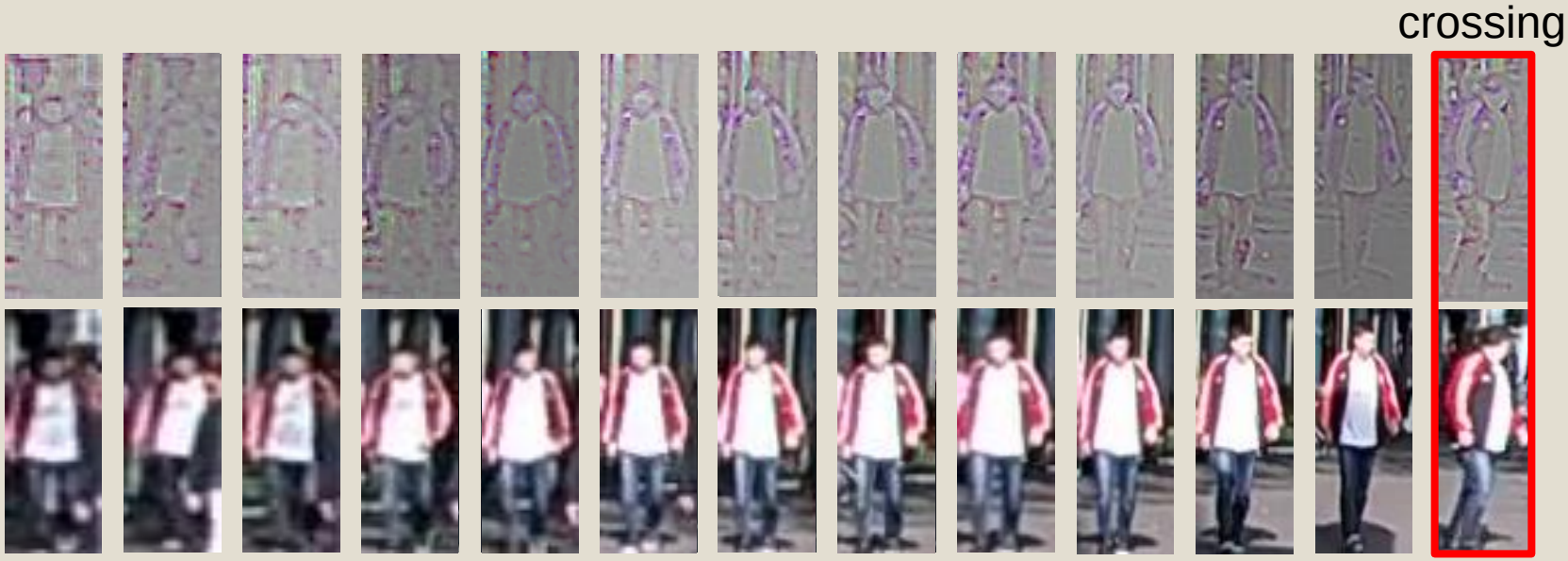
Guided Backpropagation

- Architecture is looking at the pedestrian and the head



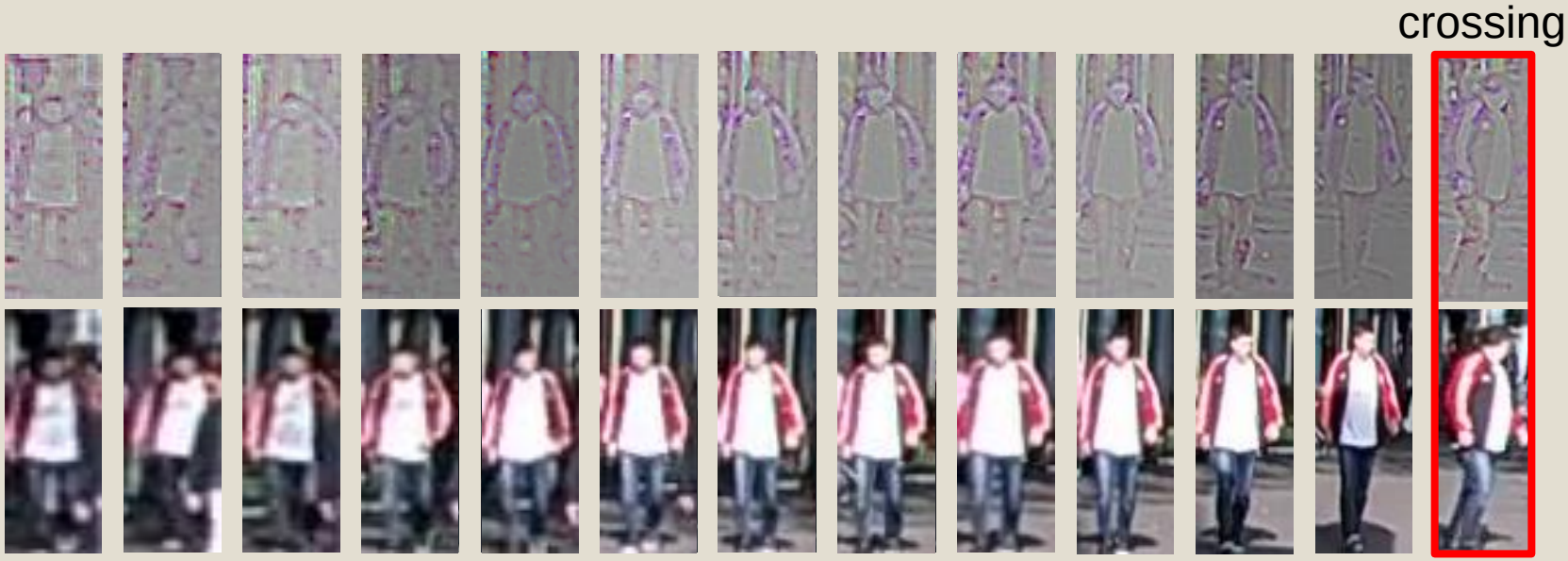
Guided Backpropagation

- Architecture is looking at the pedestrian and the head
- Architecture is looking at the background (can resolve via instance segmentation)



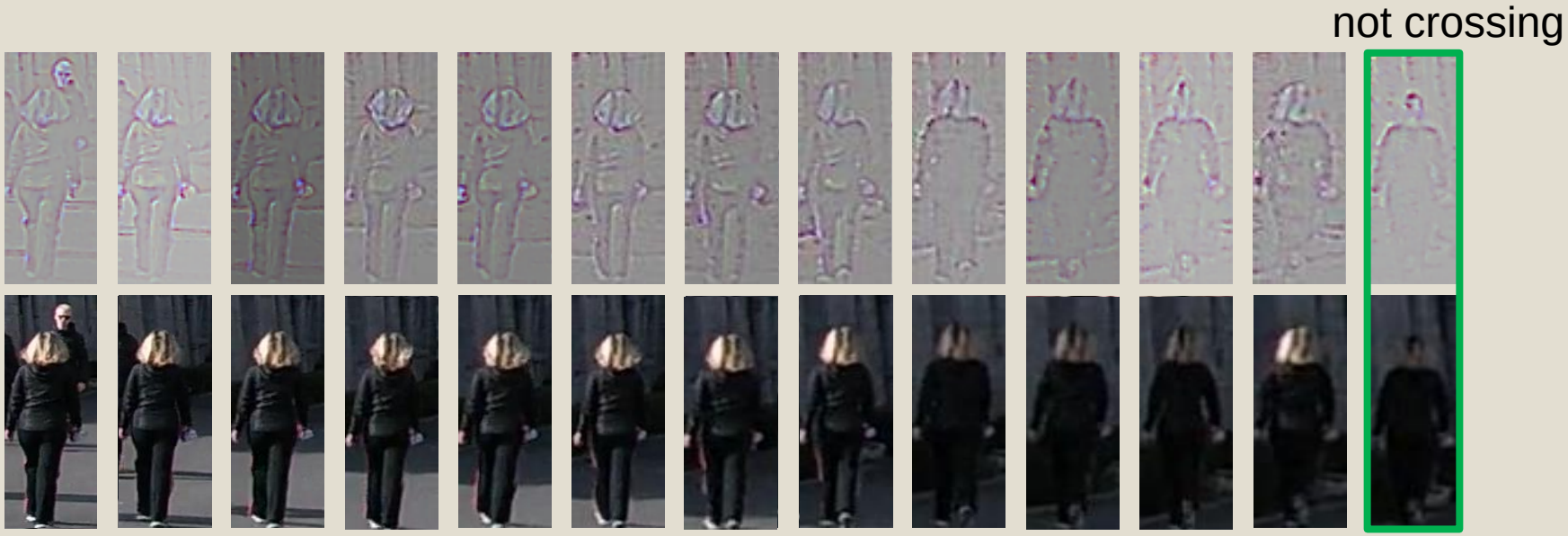
Guided Backpropagation

- Architecture is looking at the pedestrian and the head
- Architecture is looking at the background (can resolve via instance segmentation)
- Architecture uses every frame to make its prediction (can resolve via attentive framework?)



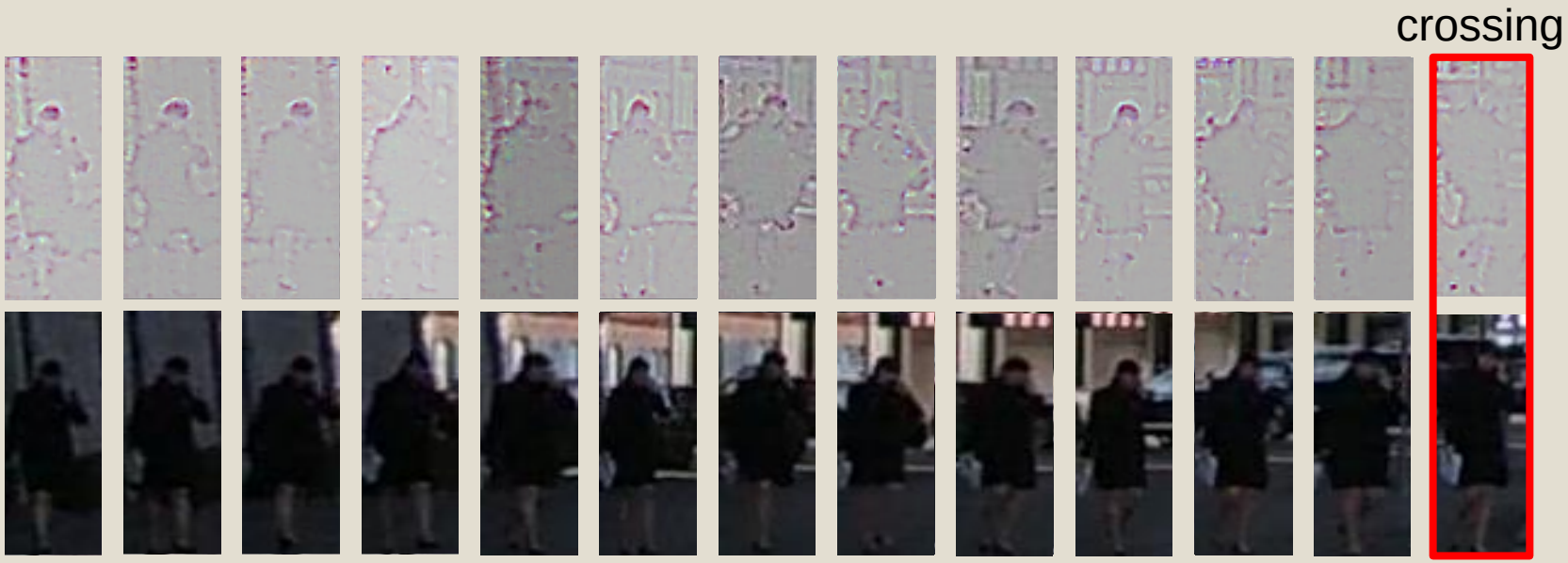
Guided Backpropagation

- Architecture is looking at the pedestrian and the head
- Architecture is looking at the background (can resolve via instance segmentation)
- Architecture uses every frame to make its prediction (can resolve via attentive framework?)

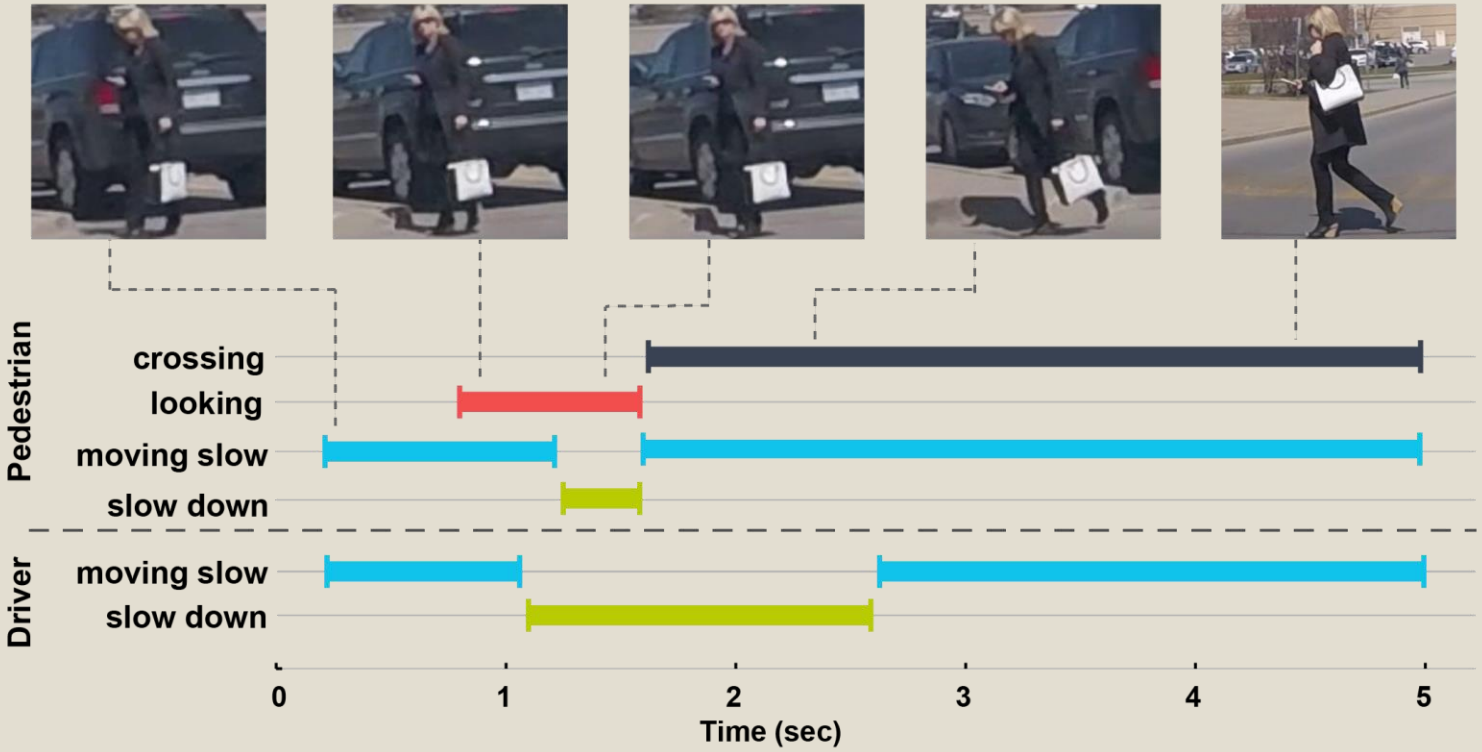


Guided Backpropagation

- Architecture is looking at the pedestrian and the head
- Architecture is looking at the background (can resolve via instance segmentation)
- Architecture uses every frame to make its prediction (can resolve via attentive framework?)

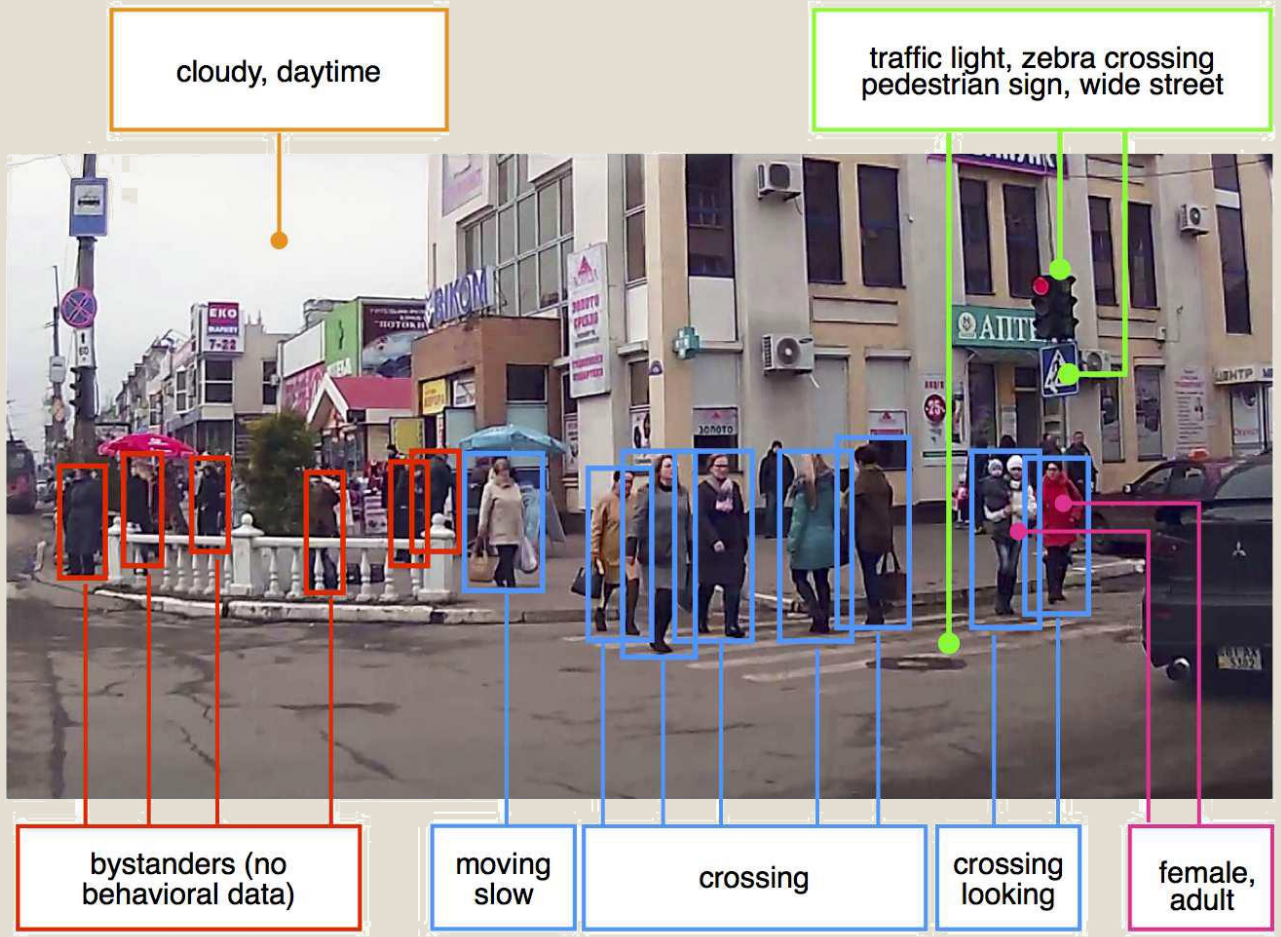


JAAD Dataset



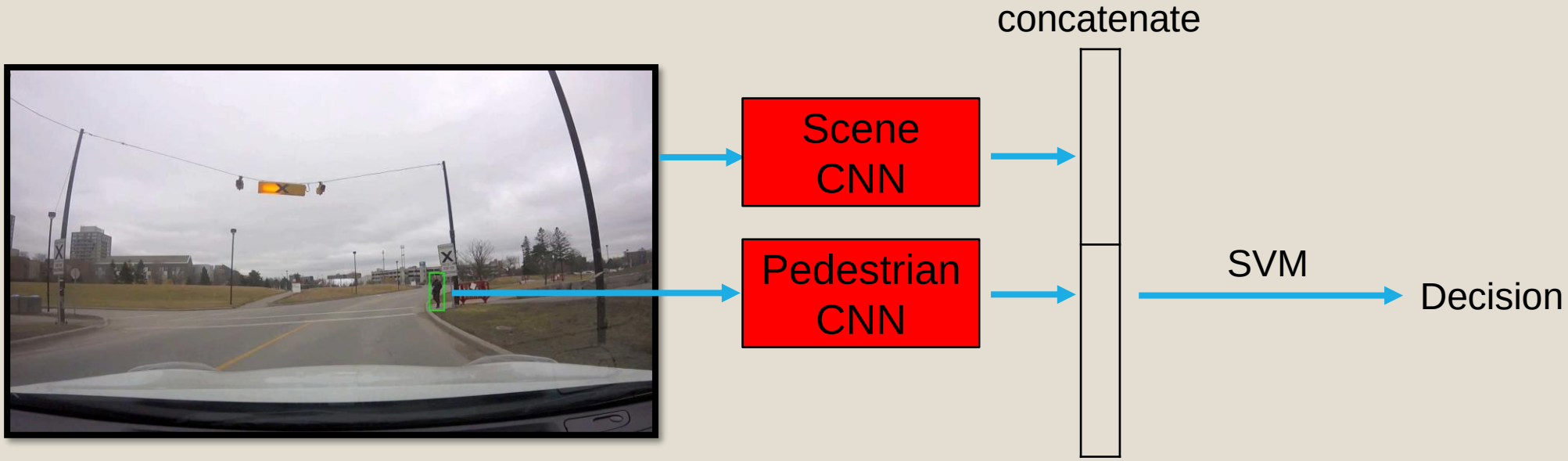
- Recorded from a moving vehicle
- Behavioural annotations
- Scene annotations
- 346 videos, each 5-10 seconds

JAAD Dataset



- Recorded from a moving vehicle
- Behavioural annotations
- Scene annotations
- 346 videos, each 5-10 seconds

Baseline



- Extract features describing pedestrian's action and scene
- Linear SVM
- Architecture uses only 1 frame instead of a sequence to make its prediction

Baseline

Method	Precision (%)	No samples
Pedestrian CNN	39.24	3324
Pedestrian CNN + Scene CNN	62.73	3324
CNN LSTM *	50	~ 350

* Training precision of 89%

Thank You