TU Graz

Pedestrian Detection in RGB-D Images from an Elevated Viewpoint

C. Ertler, H. Possegger, M. Opitz and H. Bischof, Institute for Computer Graphics and Vision

7th February 2017

Motivation

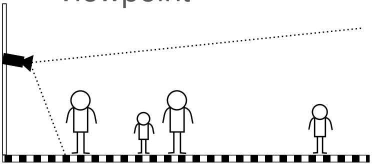
- Traffic light control system
- Predict intent of pedestrians
  - Want to cross the road?
  - Direction?
- Pedestrian detection as pre-processing step

Camera Setup

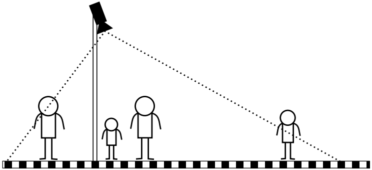
- Stereo cameras mounted on traffic light filming downwards

  $\rightarrow$ disparity data

- Overhead viewpoint *vs.* classical surveillance viewpoint



Surveillance viewpoint



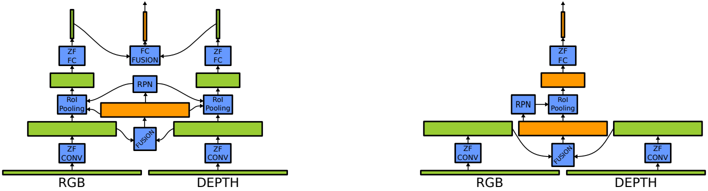
Overhead viewpoint

Viewpoint Challenges

VS.

Main Contributions

- Pedestrian detector for overhead views
- Faster R-CNN for RGB-D images
    - Two modality fusion architecturess
    - Several modality fusion layers



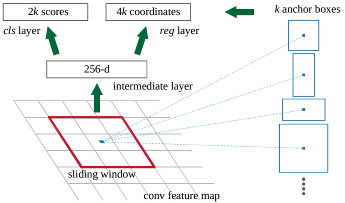
- Improve results with trained non-maximum suppression (NMS)

Faster R-CNN (Ren et al. 2015)

- Use classification networks for detection
- CNN features are used in two stages
  - Region Proposal Network (RPN)
  - Region Pooling $\rightarrow$ Region Classification

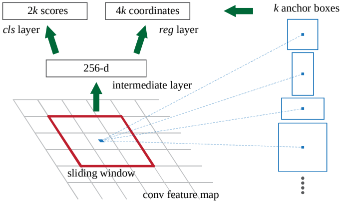
Faster R-CNN (Ren et al. 2015)

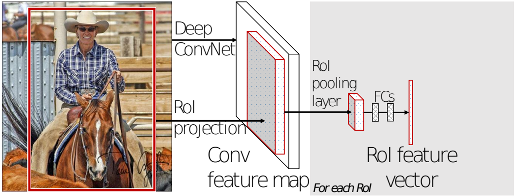
- Use classification networks for detection
- CNN features are used in two stages
  - Region Proposal Network (RPN)
  - Region Pooling → Region Classification



Faster R-CNN (Ren et al. 2015)

- Use classification networks for detection
- CNN features are used in two stages
  - Region Proposal Network (RPN)
  - Region Pooling → Region Classification





Incorporating Disparity Data

- Transfer learning between modalities
  - Zeiler & Fergus network
- Disparity depends on position relative to camera

  $\rightarrow$ Data variation
- Solution: Height above ground (HAG) encoding
  - Estimate ground plane of the scene
  - Compute HAG
  - Apply colormap to HAG data

Incorporating Disparity Data

- Transfer learning between modalities
  - Zeiler & Fergus network

- Disparity depends on position relative to camera

  $\rightarrow$ Data variation

- Solution: Height above ground (HAG) encoding
  - Estimate ground plane of the scene
  - Compute HAG
  - Apply colormap to HAG data

Incorporating Disparity Data

- Transfer learning between modalities
  - Zeiler & Fergus network

- Disparity depends on position relative to camera

  $\rightarrow$ Data variation

- Solution: Height above ground (HAG) encoding
  - Estimate ground plane of the scene
  - Compute HAG
  - Apply colormap to HAG data

Height above Ground Encoding

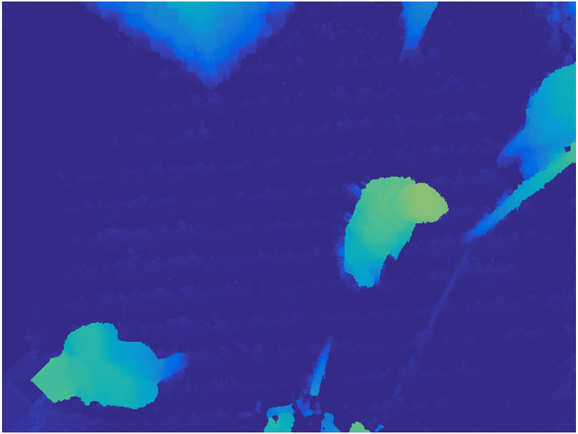
Stereo Images

Height above Ground Encoding

Disparity Map

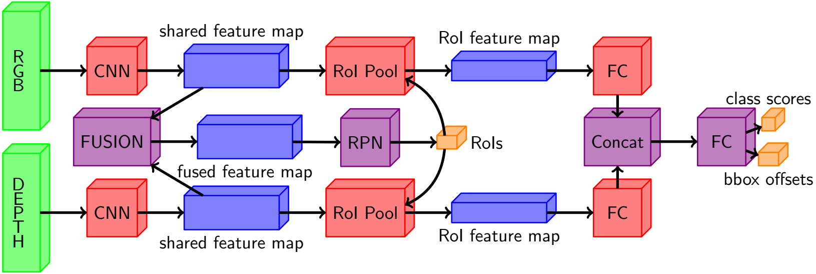
Height above Ground Encoding

Colored HAG



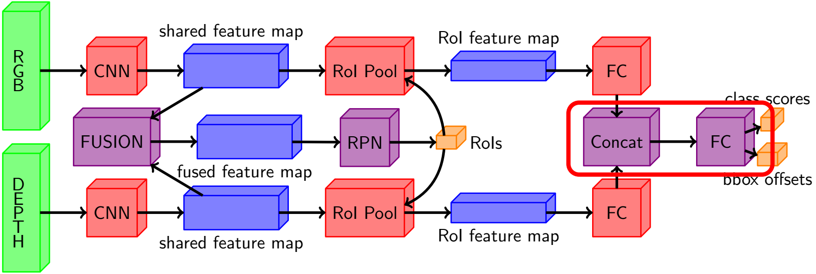
Late Fusion

- 2 independent network streams
- Fusion after last hidden layer
- Concatenate feature maps and learn additional fully-connected fusion layer



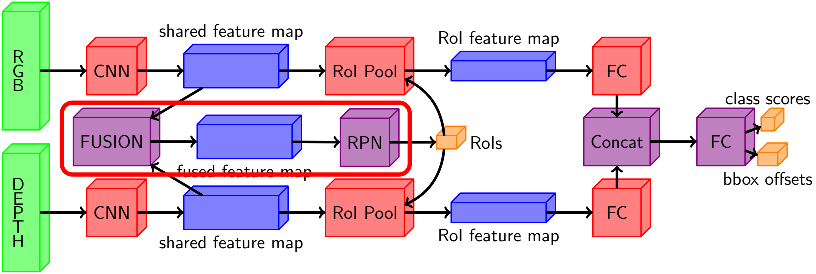
Late Fusion

- 2 independent network streams
- Fusion after last hidden layer
- Concatenate feature maps and learn additional fully-connected fusion layer



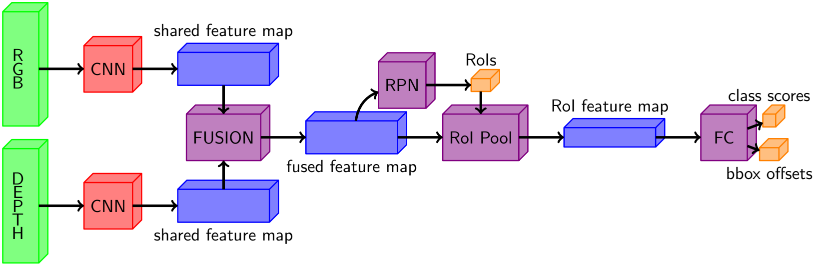
Late Fusion

- 2 independent network streams
- Fusion after last hidden layer
- Concatenate feature maps and learn additional fully-connected fusion layer

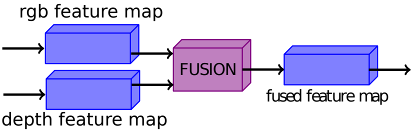


Mid-layer Fusion

- Fusion of mid-layer representations
- Single stream after convolutional layers
- Number of parameters significantly reduced
  - 117 M *vs.* 45 M



Modality Fusion Layers



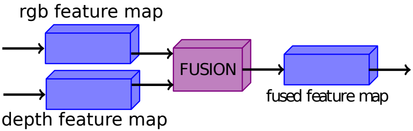
- Parameterless fusion
  - Average
  - Sum
  - Max

  element-wise

- Parametrized fusion
  - 1×1 Convolution
  - Inception tower

  concatenated features

Modality Fusion Layers



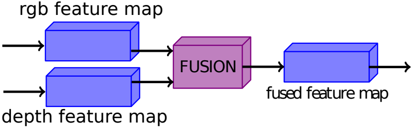
- Parameterless fusion
  - Average
  - Sum
  - Max

  element-wise

- Parametrized fusion
  - 1×1 Convolution
  - Inception tower

  concatenated features

Modality Fusion Layers



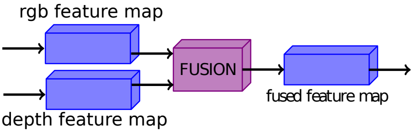
- Parameterless fusion
  - Average
  - Sum
  - Max

  } element-wise

- Parametrized fusion
  - 1×1 Convolution
  - Inception tower

  } concatenated features

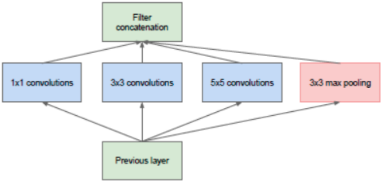
Modality Fusion Layers



- Parameterless fusion
  - Average
  - Sum
  - Max

  $\Big\}$ element-wise

- Parametrized fusion
  - 1×1 Convolution
  - Inception tower



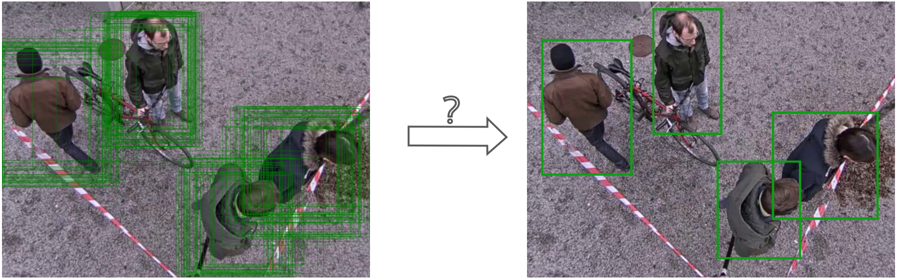
Learning Non-Maximum Suppression

Greedy NMS

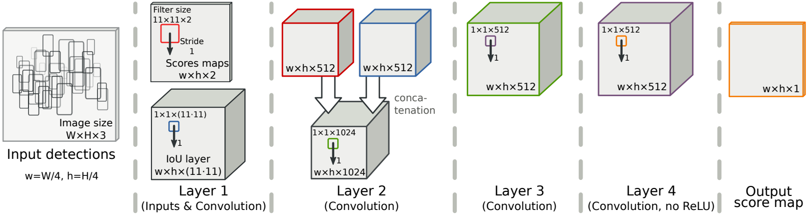
- De-facto standard in object detection
- Need to choose constant overlap threshold

  $\rightarrow$ heavily tuned to validation set

- Trade-off between recall and precision



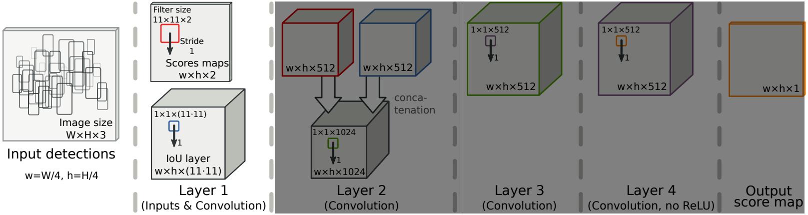
Tnet (Hosang et al. 2016)



- Fully convolutional network
- Inputs are detection boxes encoded as
  - Score maps
  - IoU of the boxes
- Output is final score map after suppression

  → No post-processing needed

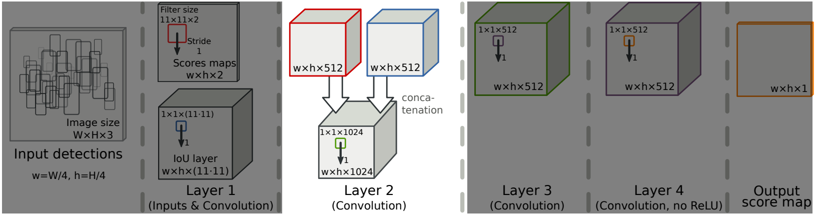
Tnet (Hosang et al. 2016)



- Fully convolutional network
- Inputs are detection boxes encoded as
  - Score maps
  - IoU of the boxes
- Output is final score map after suppression

  → No post-processing needed

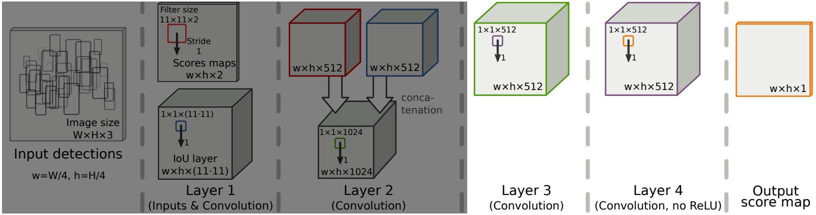
Tnet (Hosang et al. 2016)



- Fully convolutional network
- Inputs are detection boxes encoded as
  - Score maps
  - IoU of the boxes
- Output is final score map after suppression

  → No post-processing needed

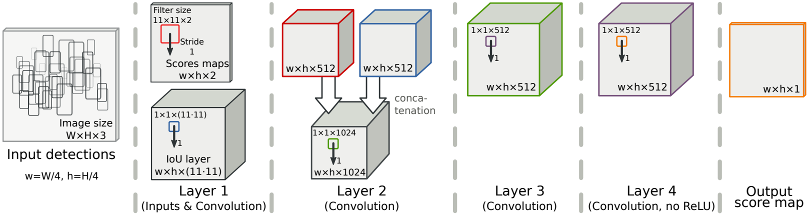
Tnet (Hosang et al. 2016)



- Fully convolutional network
- Inputs are detection boxes encoded as
    - Score maps
    - IoU of the boxes
- Output is final score map after suppression

  → No post-processing needed

Tnet (Hosang et al. 2016)



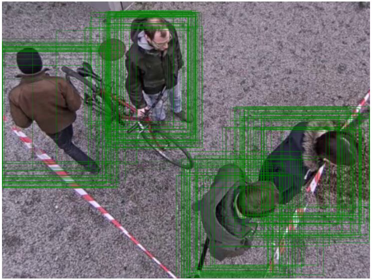
- Fully convolutional network
- Inputs are detection boxes encoded as
  - Score maps
  - IoU of the boxes
- Output is final score map after suppression

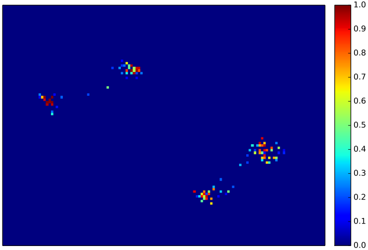
$\rightarrow$ No post-processing needed

Sparse Score Maps

- Detection scores in 2D grid
- Sparse detections from Faster R-CNN

→ Zero loss weights in empty regions

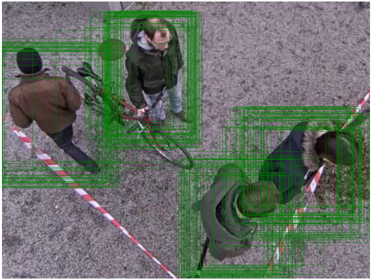


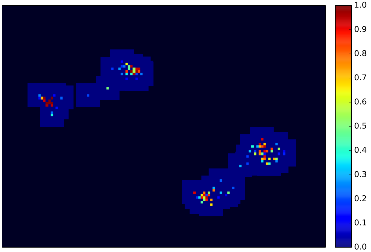


Sparse Score Maps

- Detection scores in 2D grid
- Sparse detections from Faster R-CNN

$\rightarrow$ Zero loss weights in empty regions





Evaluation

Fusion Evaluation

- Training set recorded on a public site (VIENNA)
  - 447 images, 1194 annotations
- Test set recorded on the campus (CAMPUS)
  - 321 images, 832 annotations

Fusion Evaluation

- Compare RGB network with different fusion networks

Model	**AP**	
	Mid-layer	**Late**
RGB-only	81.95 % (0.35)	
HAG-only	52.05 % (3.55)	
Sum fusion	88.60 % (1.00)	87.55 % (0.65)
Average fusion	87.00 % (0.00)	87.70 % (0.90)
Max fusion	89.89 % (0.20)	87.65 % (0.75)
Conv fusion	86.35 % (0.55)	85.60 % (1.10)
Inception fusion	88.85 % (0.85)	—

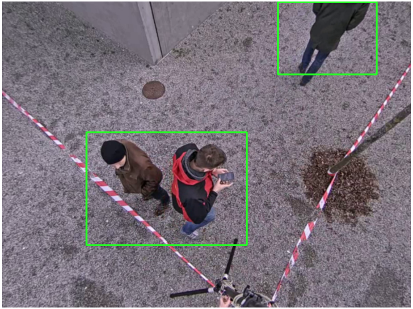
NMS Evaluation

- Compare Tnet with different greedy NMS thresholds
- Test set is split into samples with and without overlapping ground truth boxes

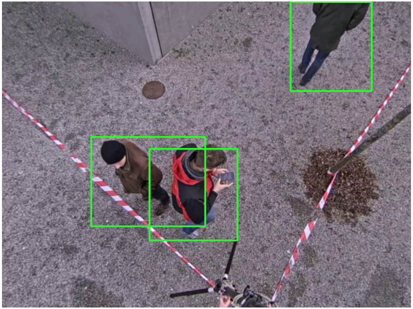
Model	AP All	AP Overlapping	AP Non-overlapping
Tnet	90.10 %	87.00 %	95.90 %
NMS 0.9	41.20 %	37.30 %	49.40 %
NMS 0.8	67.80 %	61.80 %	76.40 %
NMS 0.7	85.60 %	78.10 %	93.40 %
NMS 0.6	89.70 %	82.30 %	95.40 %
NMS 0.5	88.30 %	81.00 %	95.90 %
NMS 0.4	87.10 %	79.30 %	95.30 %

Fusion Evaluation — Qualitative Results

Nearby pedestrians



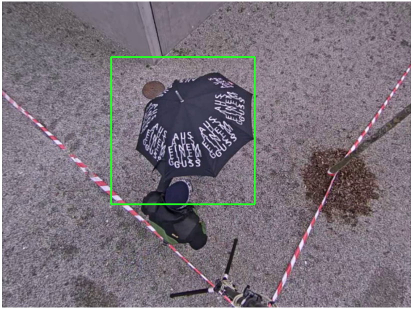
RGB-only



Mid-layer Max Fusion

Fusion Evaluation — Qualitative Results

Generalization



RGB-only

Mid-layer Max Fusion

Fusion Evaluation — Qualitative Results

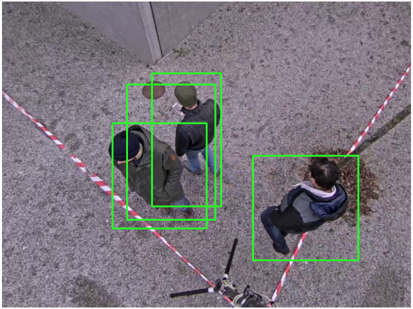
Bounding box regression

RGB-only

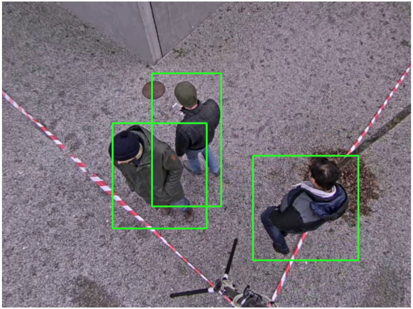
Mid-layer Max Fusion

NMS Evaluation — Qualitative Results

False positives



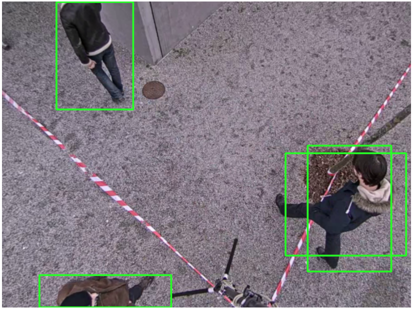
NMS 0.6



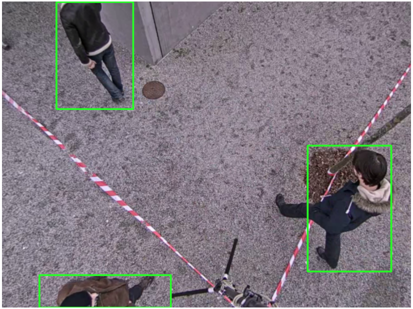
Tnet

NMS Evaluation — Qualitative Results

Double detections



NMS 0.6



Tnet

NMS Evaluation — Qualitative Results

False negatives

NMS 0.6

Tnet

Conclusion

- Modality fusion in Faster R-CNN model
- Mid-layer fusion has better performance and is less complex than late fusion

- Replace Greedy NMS by learned model
- Eliminates the constant threshold

Questions

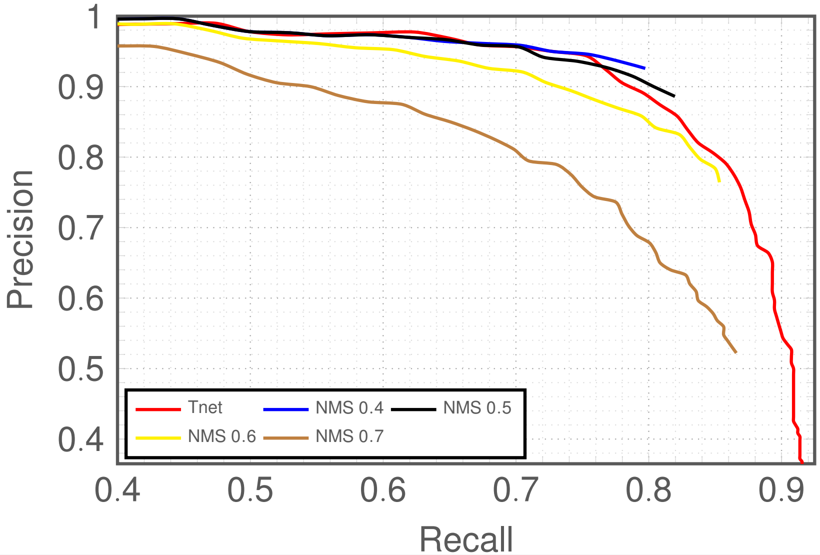
Thank You

Bibliography I

[1] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. “A Convnet for Non-Maximum Suppression”. In: Proceedings of the German Conference on Pattern Recognition (GCPR). 2016.

[2] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS). 2015.

NMS Evaluation — Precision *vs.* Recall



Runtime Performance

- Experiments on NVIDIA GTX 970 with 4GB

- RGB: 67 ms
- Mid-layer fusion: 87 ms
- Late fusion: 119 ms

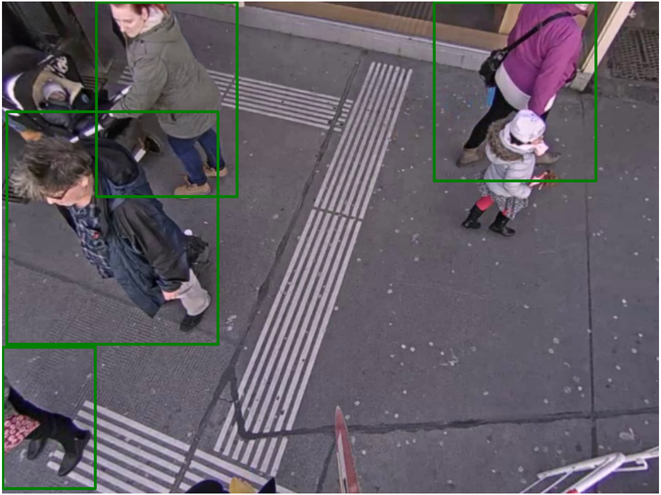
$\rightarrow$ Mid-layer fusion only 20 ms slower

- Greedy NMS: 14 ms
- Tnet: 28 ms

Additional Qualitative Examples

Additional Qualitative Examples

Additional Qualitative Examples



Additional Qualitative Examples

Additional Qualitative Examples