Cooperative Multi-Scale Convolutional Neural Networks for Person Detection

Markus Eisenbach

Ilmenau University of Technology (Germany)

Neuroinformatics and Cognitive Robotics Lab
markus.eisenbach@tu-ilmenau.de
www.tu-ilmenau.de/neurob

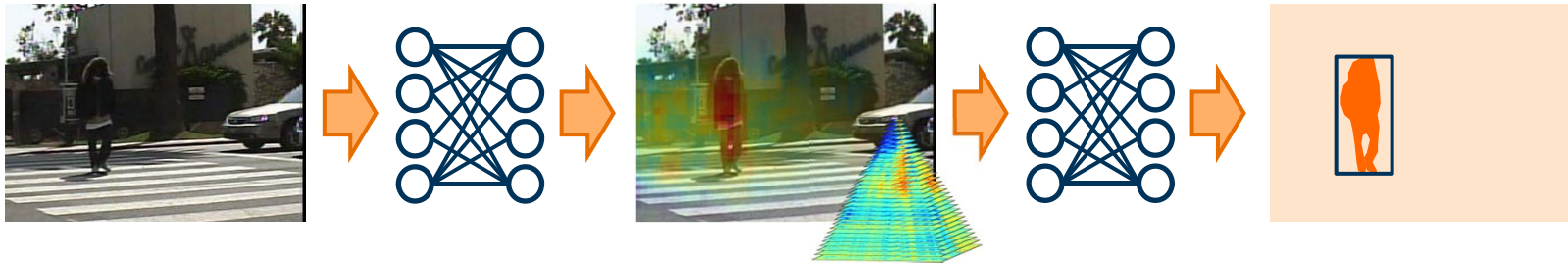
D. Seichter, T. Wengefeld, H.-M. Gross
Ilmenau University of Technology
Germany

Outline

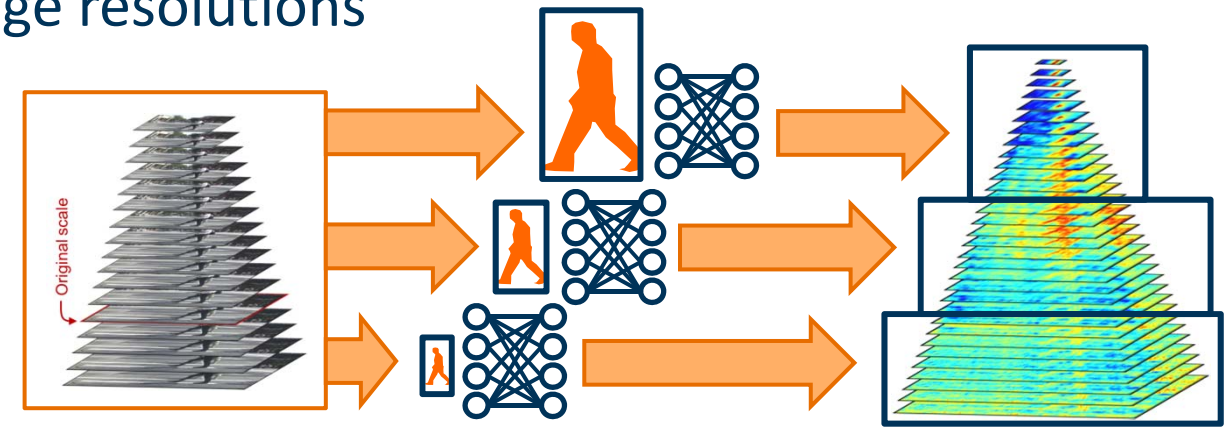
What can you expect?

Outline – What can you expect?

- **Fully neural approach**
  - Starting from raw pixels
  - To bounding boxes of detected persons



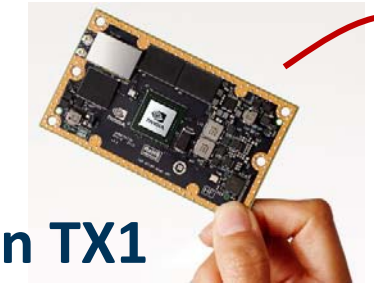
- **Hybrid multi-scale detection**
  - Multiple classifier detection window sizes
  - Multiple image resolutions



Outline – What can you expect?

- **Relatively fast approach** (on special hardware)

TITAN X



Jetson TX1

- **State of the art performance without domain specific training**
  - Good generality for person detection domains

Training

Test

Motivation

Why yet another person detector?

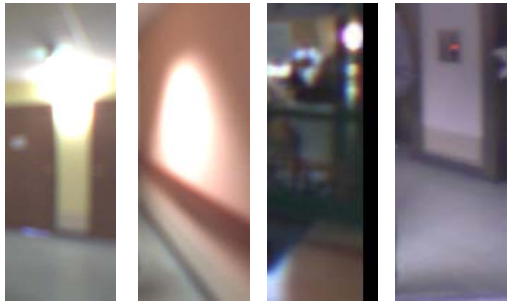
Why yet another person detector?
– Motivation from a robotic viewpoint

Do state of the art person detectors perform well in real-world applications?

Example 1: User (re-)identification

- Often **misaligned** bounding boxes
  → **erroneous feature extraction**

- Background objects (**false detections**) stored in user model
  → **Mismatches** with other users if their models include these false detections, too

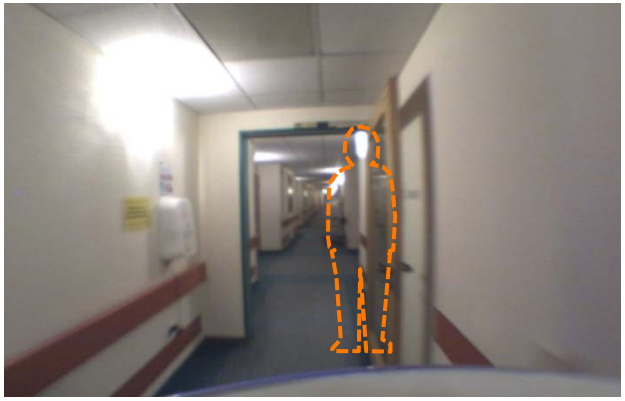


Why yet another person detector?
– Motivation from a robotic viewpoint

Do state of the art person detectors perform well in real-world applications?

Example 2: Navigation

- Many **false detections near doors** (vertical gradients)
  → **Robot does not pass** door due to possible personal space violation

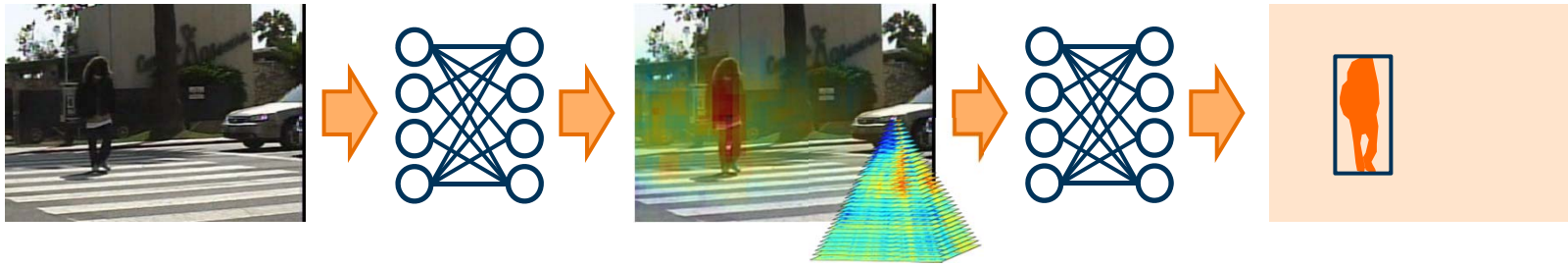


Conclusion:

- Either poor performance (**many false positives**)
- **Or** more sophisticated but **bad run-time** (and hard to parallelize)

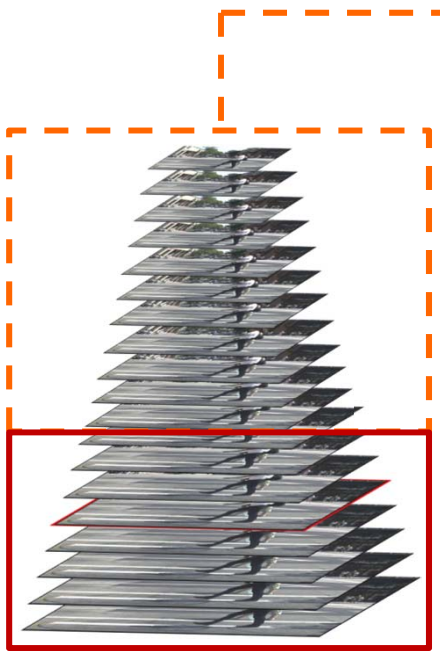
Our Approach

What is the basic idea?



Our Approach

Original scale

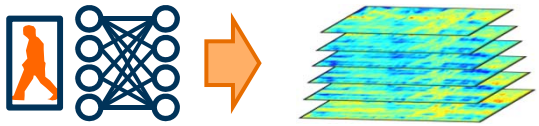


Scaling factor:
7 x downscale
= half image size

Near
128 x 48

80 x 30

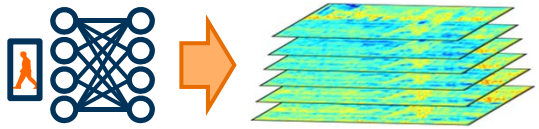
Medium
64 x 24



6

40 x 15

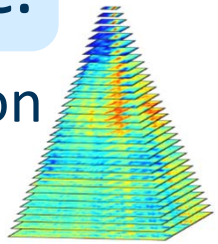
Far
32 x 12



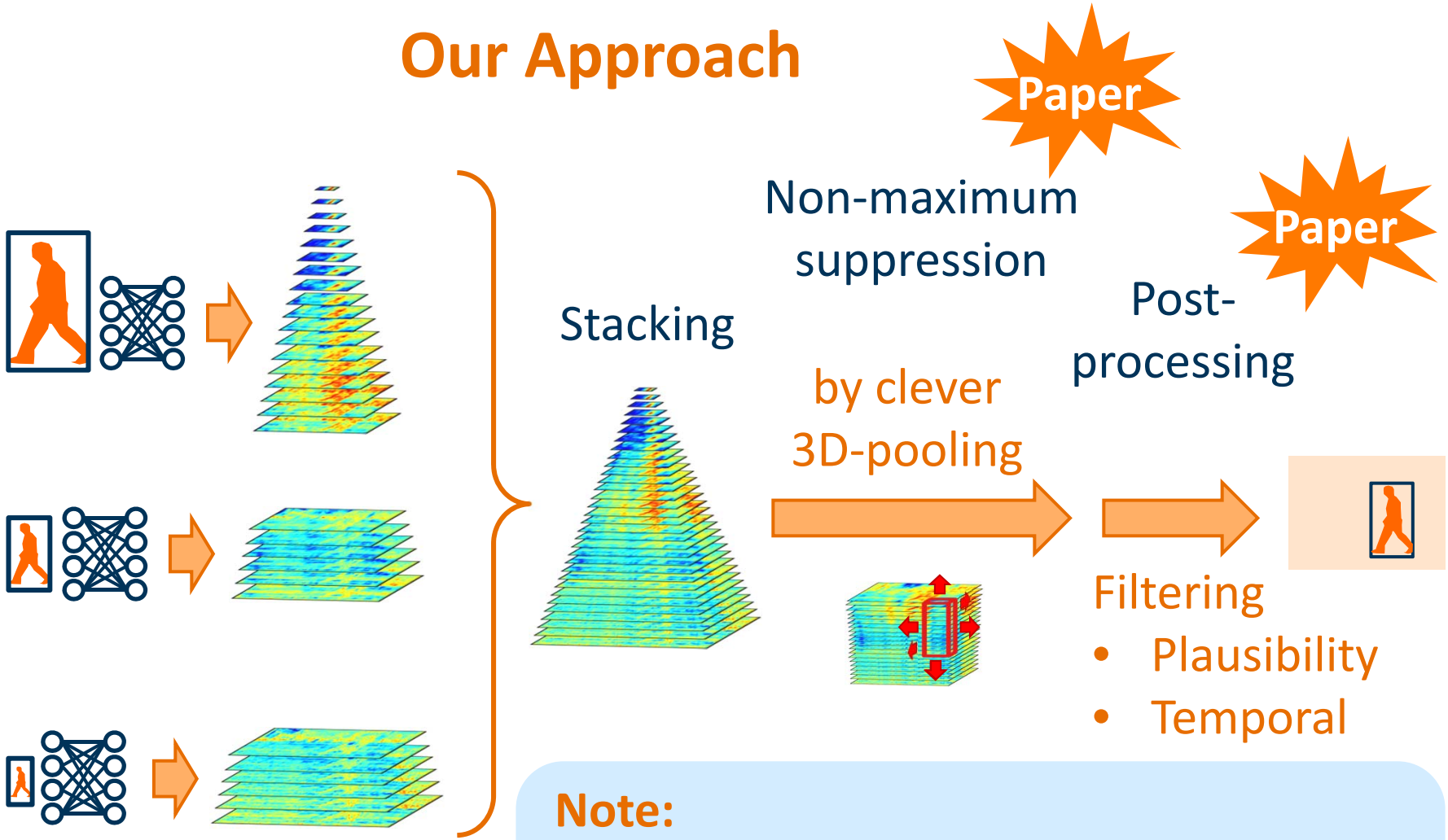
6

Objective:

- Position
- Scale



Our Approach



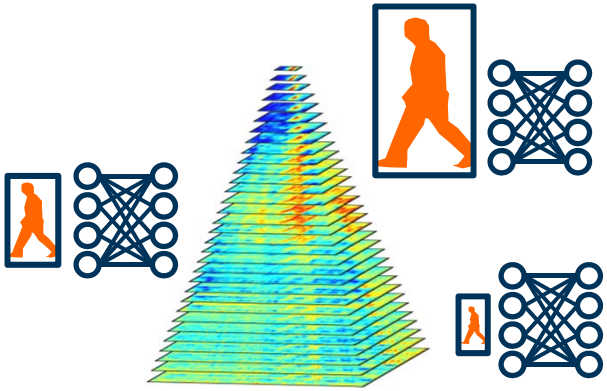
Note:

Stacking only possible due to

- Inter-coordinated network topologies
- Comparable outputs

Our Approach

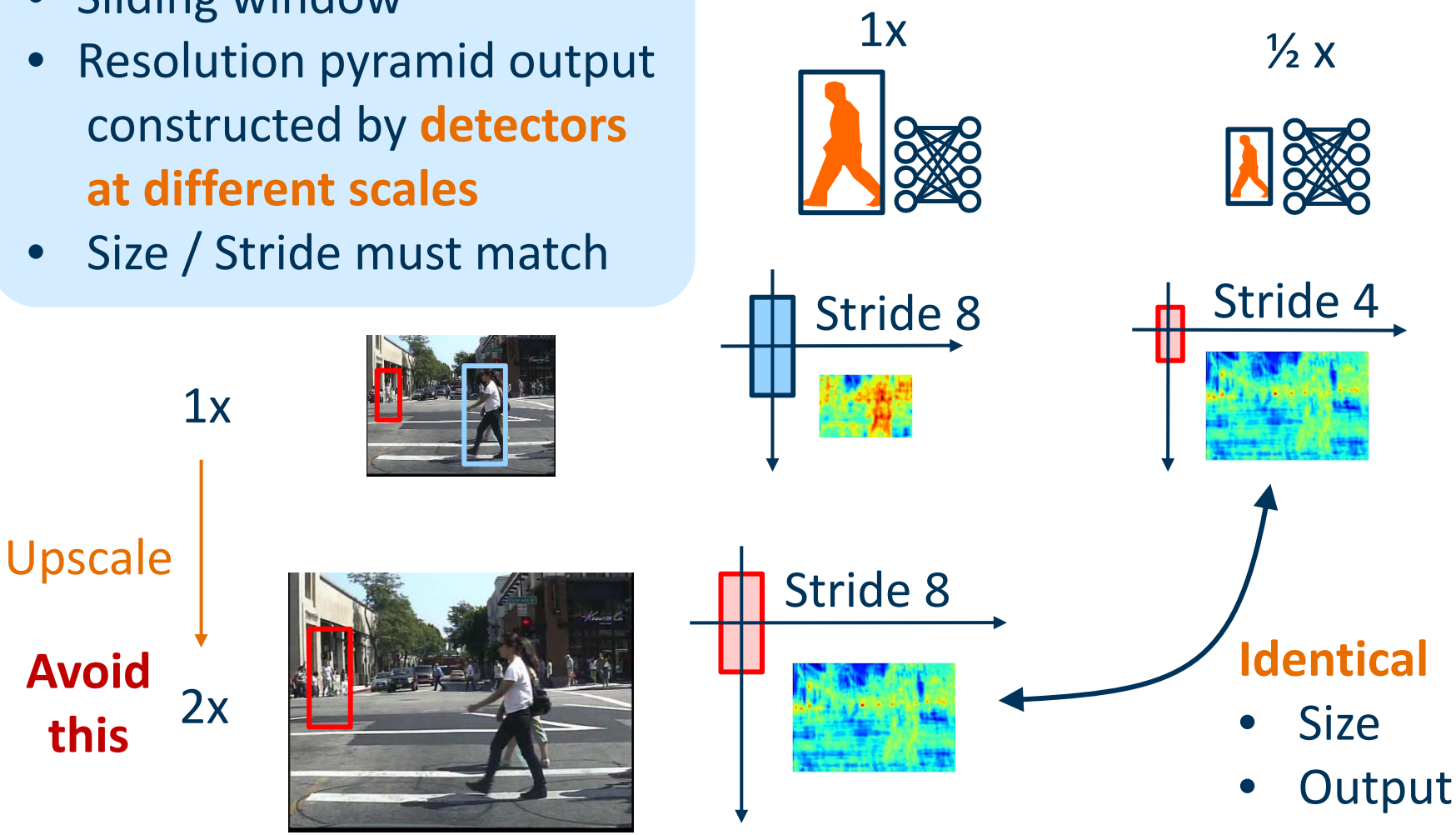
How to configure the networks?



How to configure the networks?

Objective

- Sliding window
- Resolution pyramid output constructed by **detectors at different scales**
- Size / Stride must match



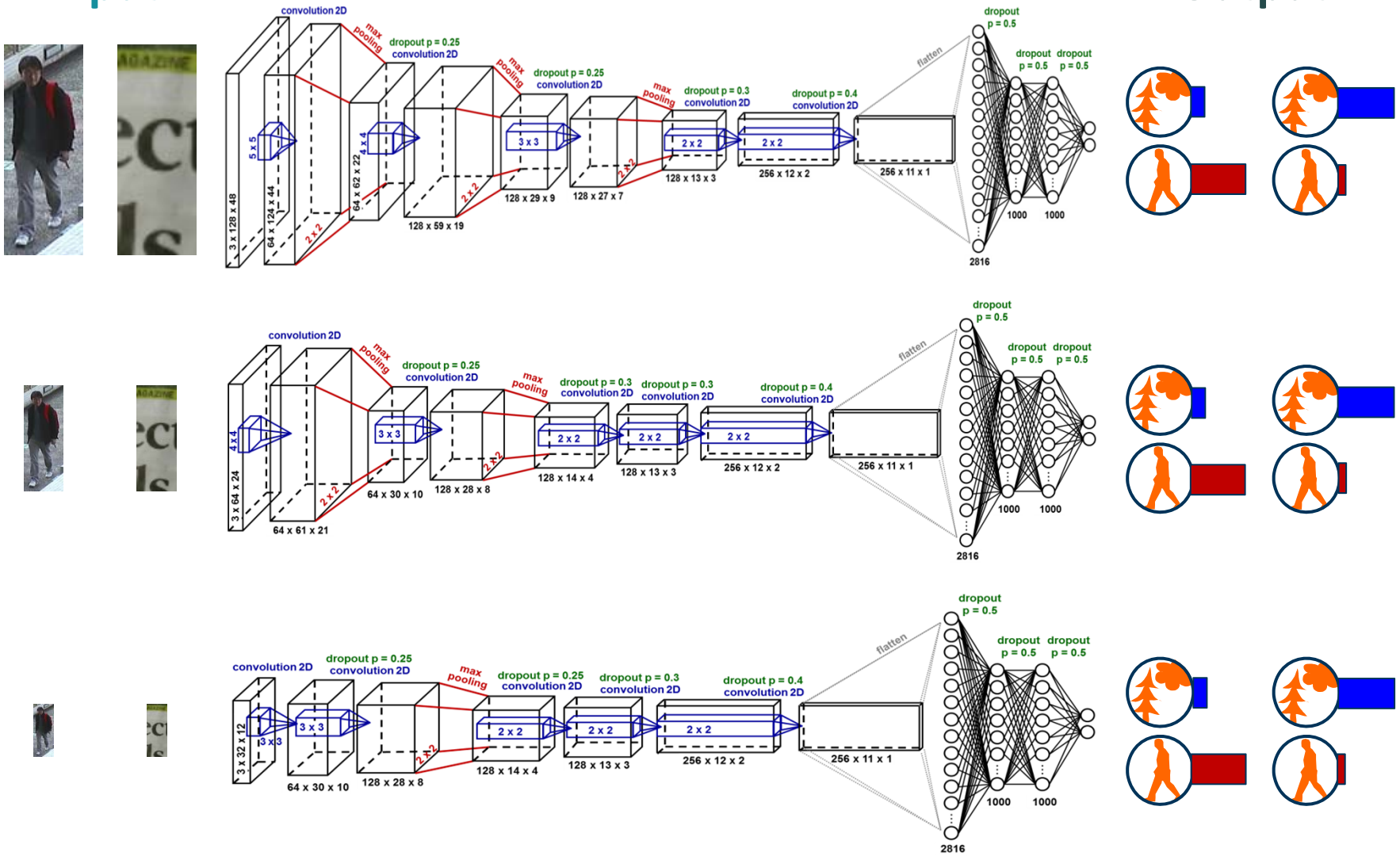
Network Topology

Training phase

Network Topology – Training

Input

Output

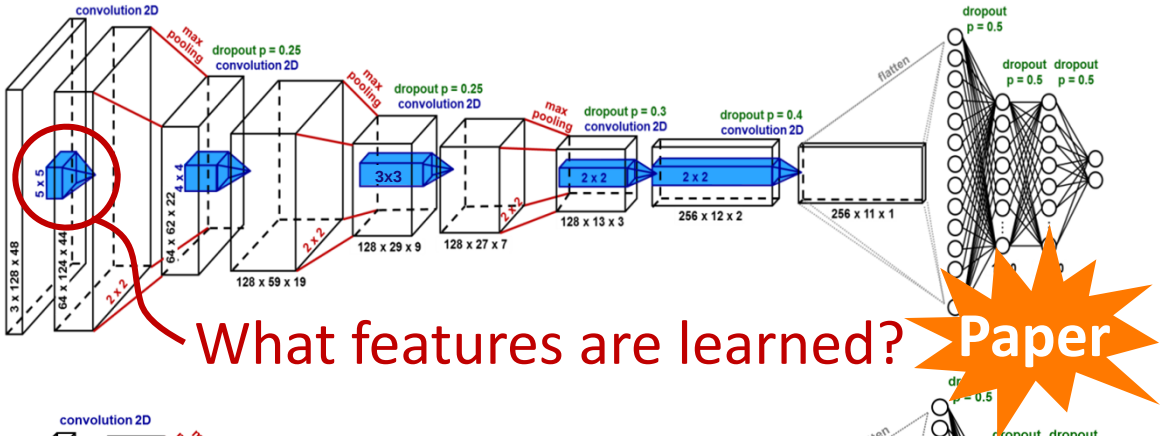


Network Topology – Training

Input

Near

128 x 48 x 3

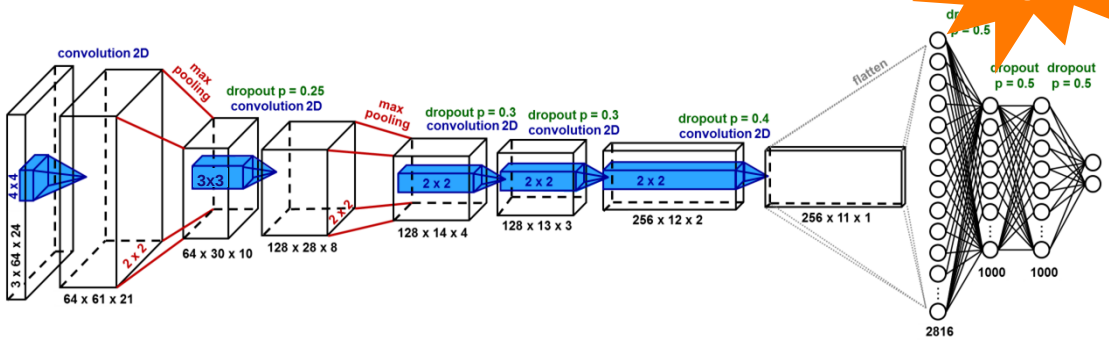


What features are learned?

Paper

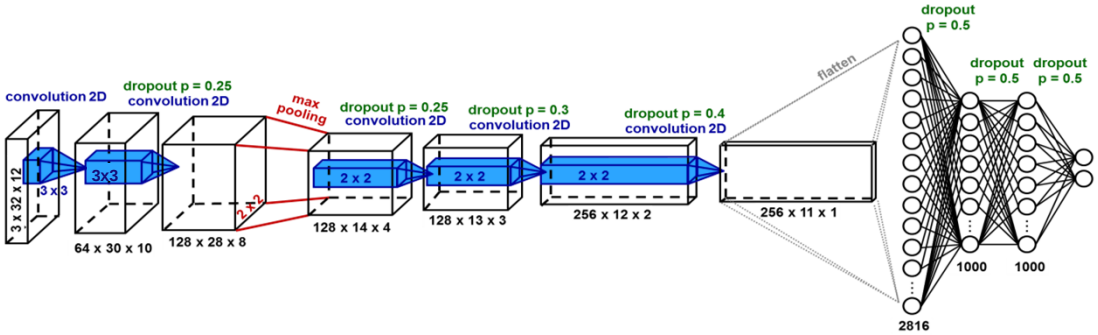
Medium

64 x 24 x 3



Far

32 x 12 x 3



EACH

5 Conv. layers

Filter sizes

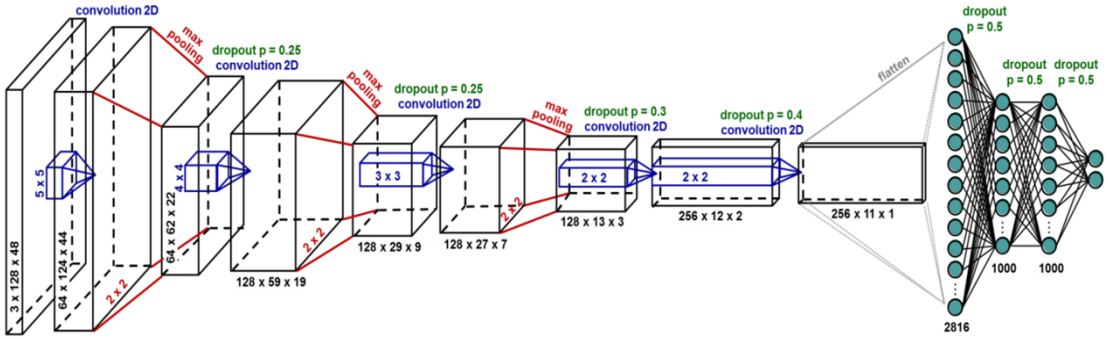
between **2 x 2** and **5 x 5**

Network Topology – Training

Input

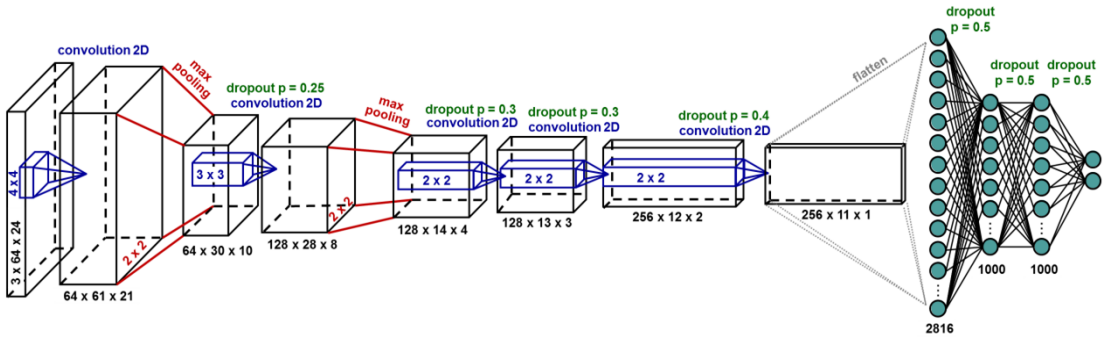
Near

128 x 48 x 3



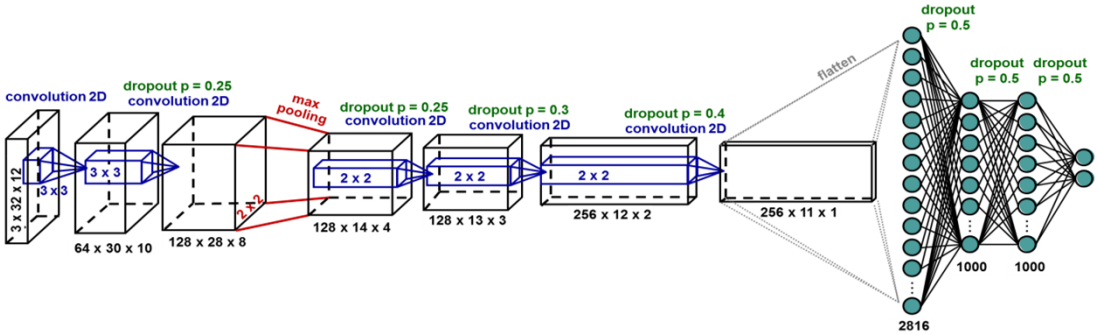
Medium

64 x 24 x 3



Far

32 x 12 x 3



EACH

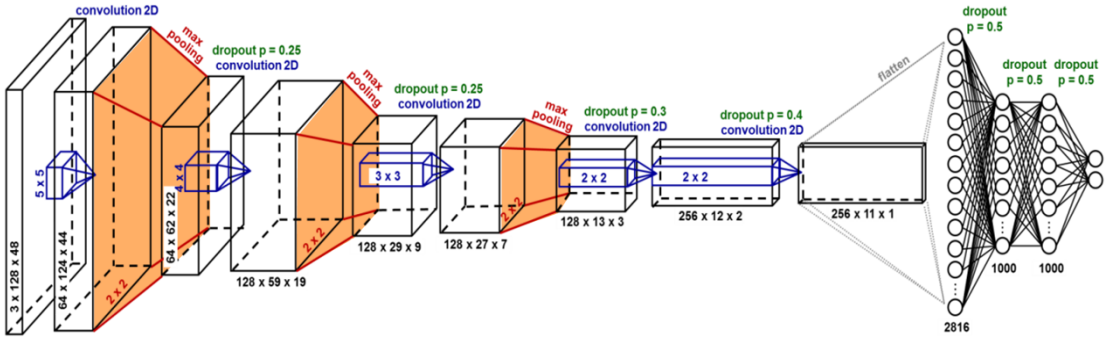
2 Fully connected layers

+ Softmax output layer

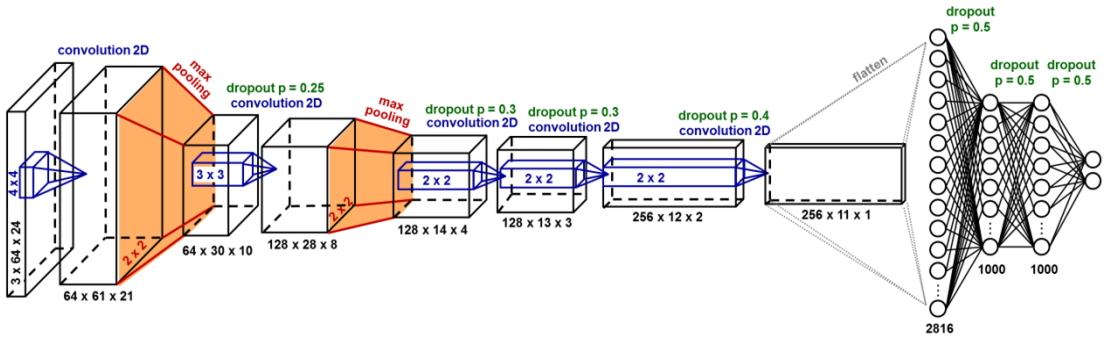
Network Topology – Training

Pooling

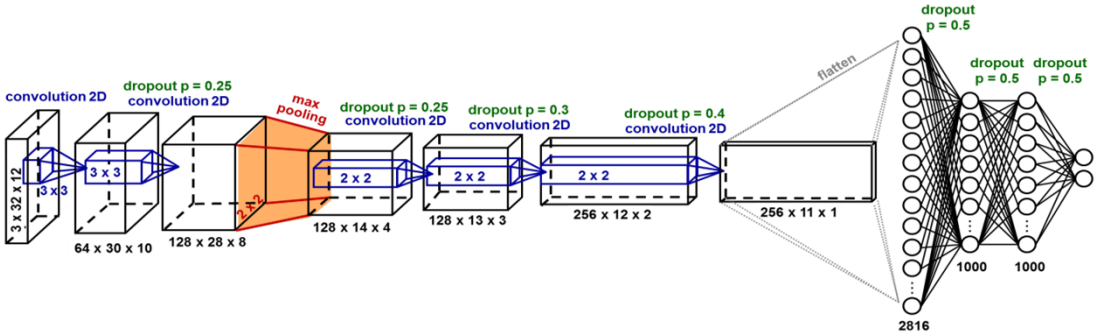
3 Layers



2 Layers



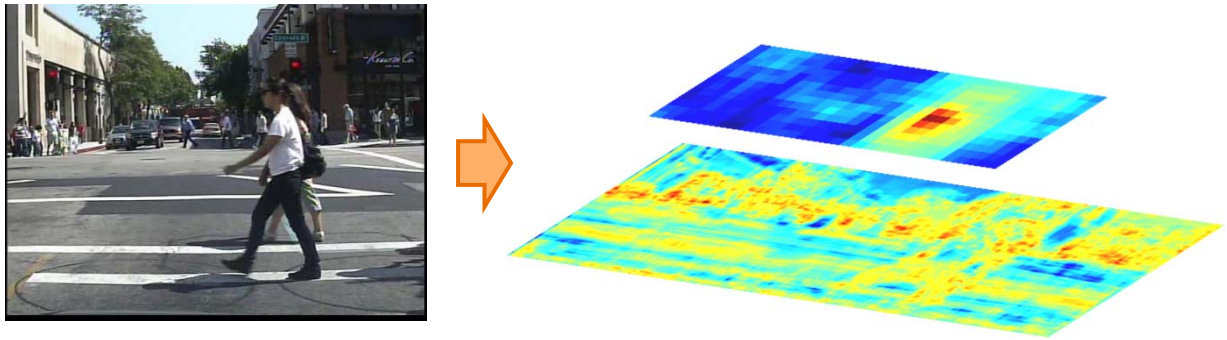
1 Layer



2 x 2
Max
pooling
with
stride 2

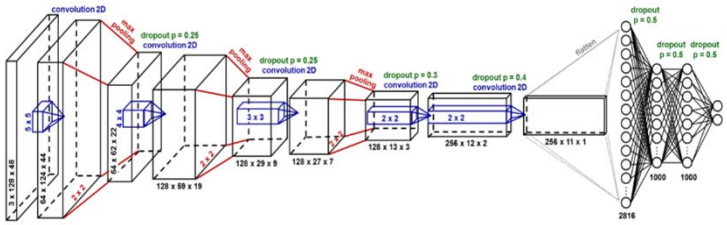
Network Topology

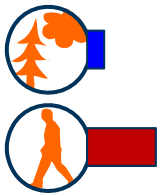
Application phase



Network Topology – Application

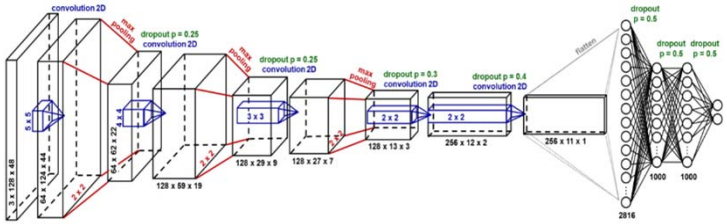
What we have:

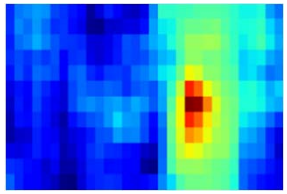


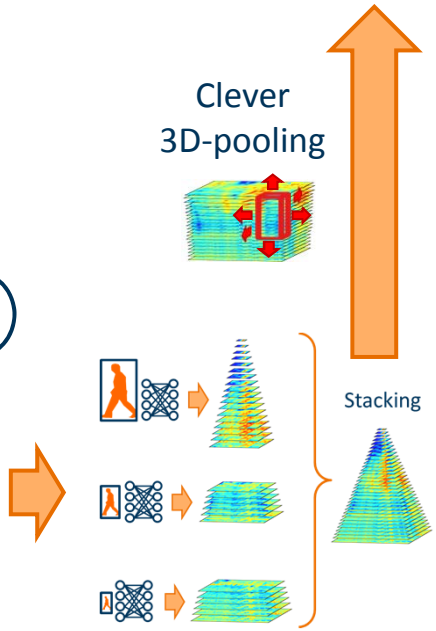


What we need this for:

What we want:



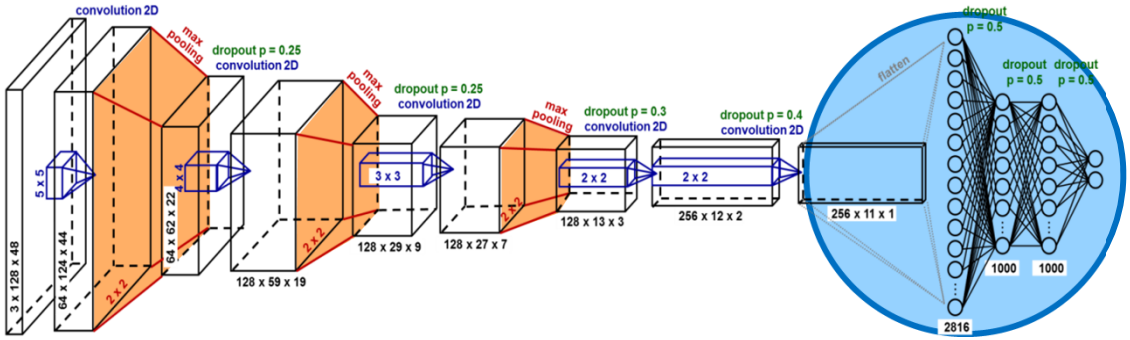


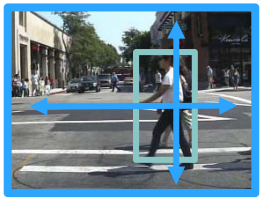


Network Topology – Application

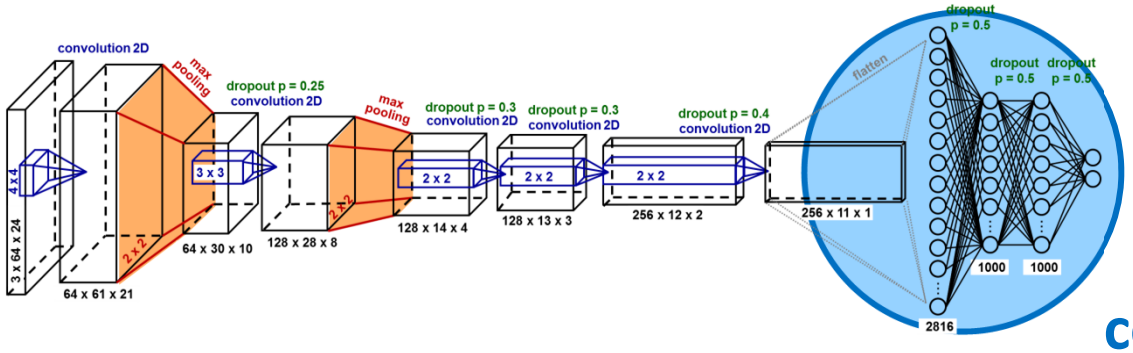
Pooling

3 Layers
= 3 x Scale
= **Stride 8**



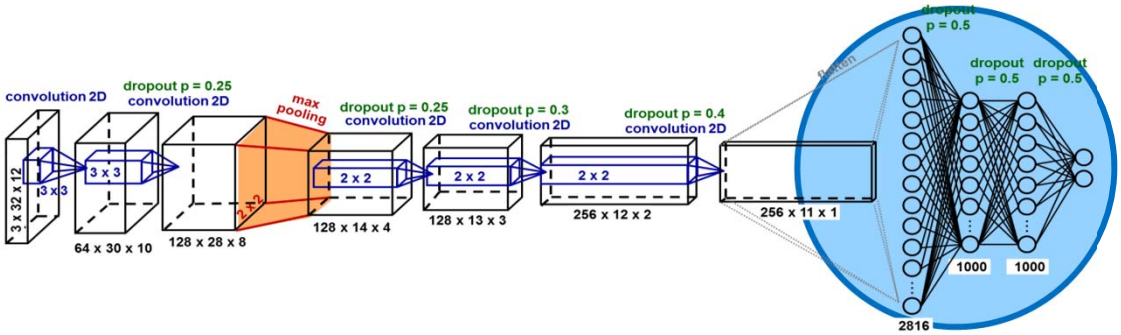


2 Layers
= 2 x Scale
= **Stride 4**



Rearrange fully connected layers to convolutional layers

1 Layer
= 1 x Scale
= **Stride 2**

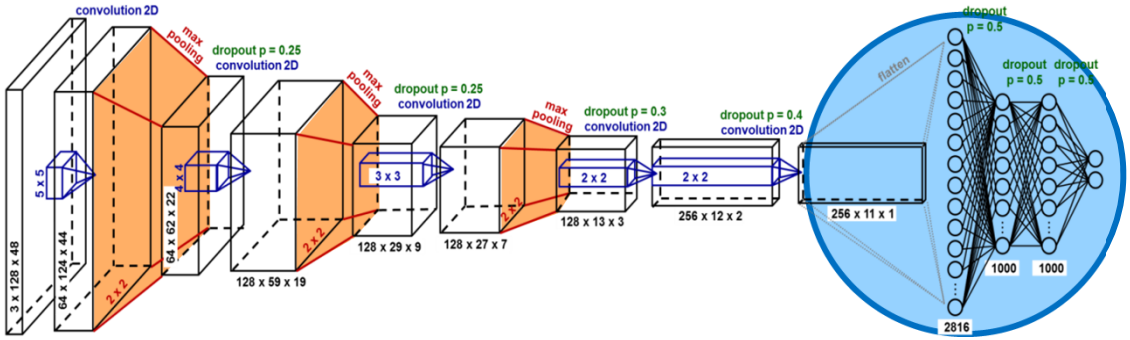


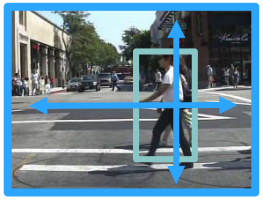
Paper

Network Topology – Application

Pooling

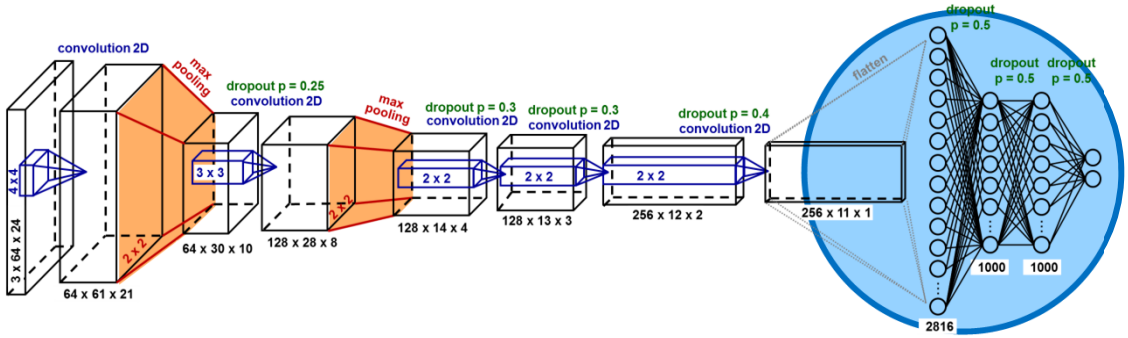
3 Layers
= 3 x Scale
= **Stride 8**



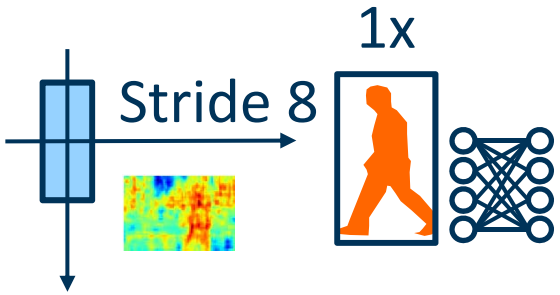


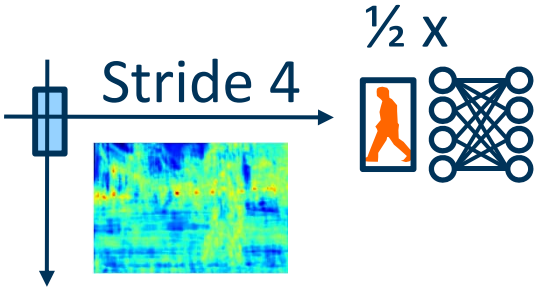
Rearrange fully connected layers to conv. layers

2 Layers
= 2 x Scale
= **Stride 4**



Objective:

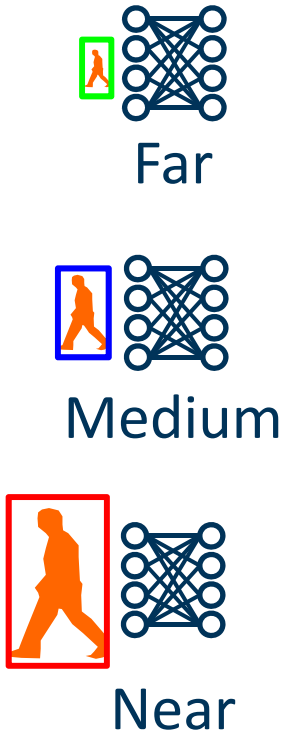


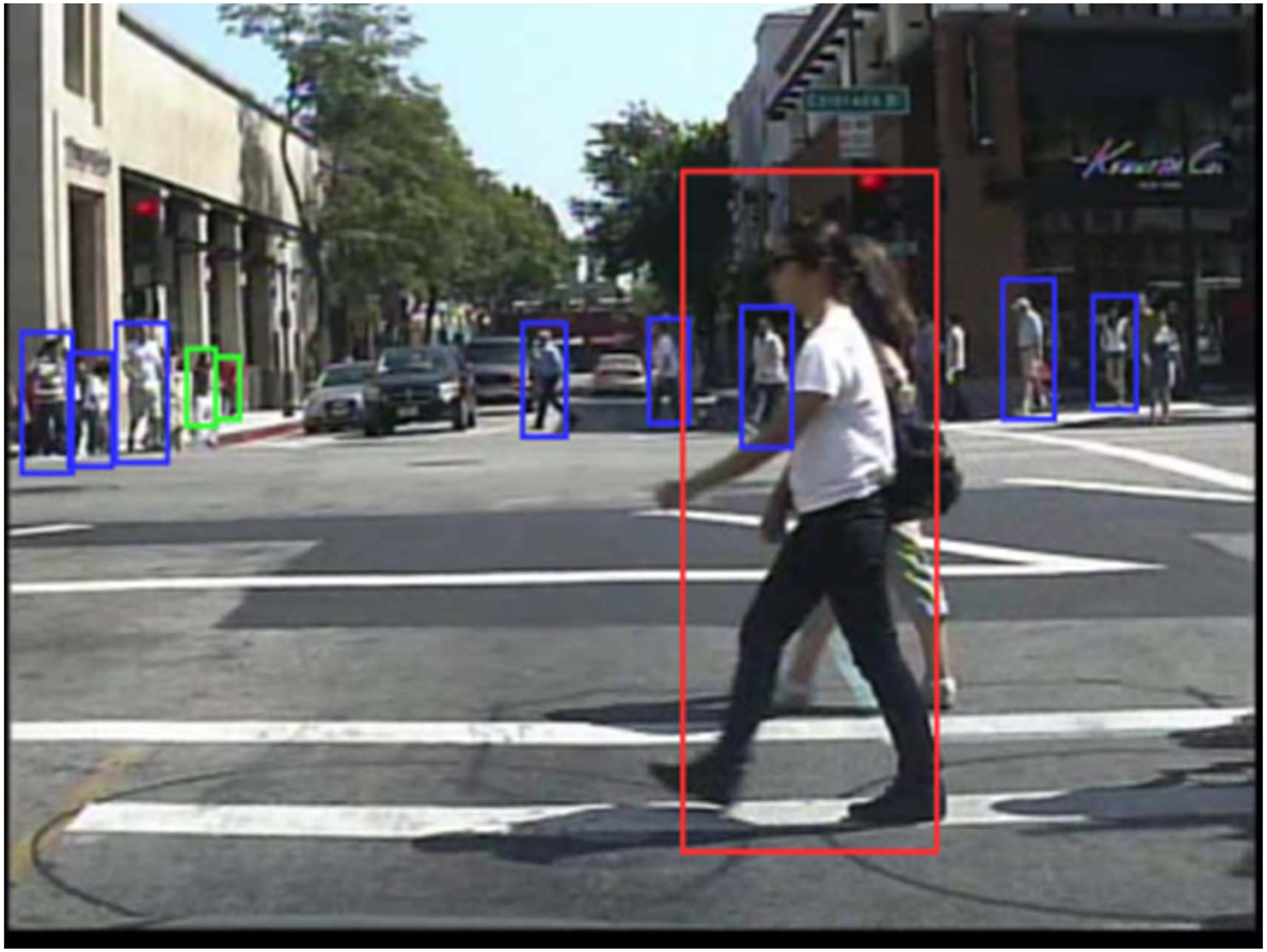


Results

Visual Examples

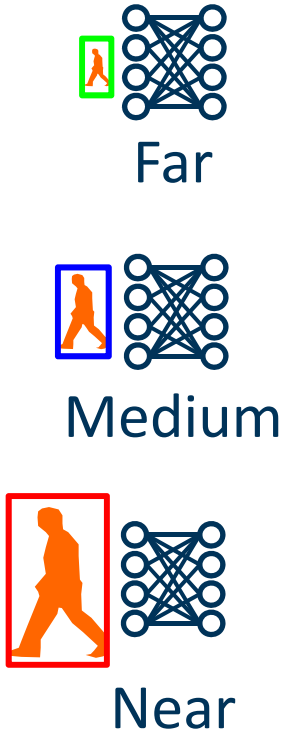
Results – Visual Examples

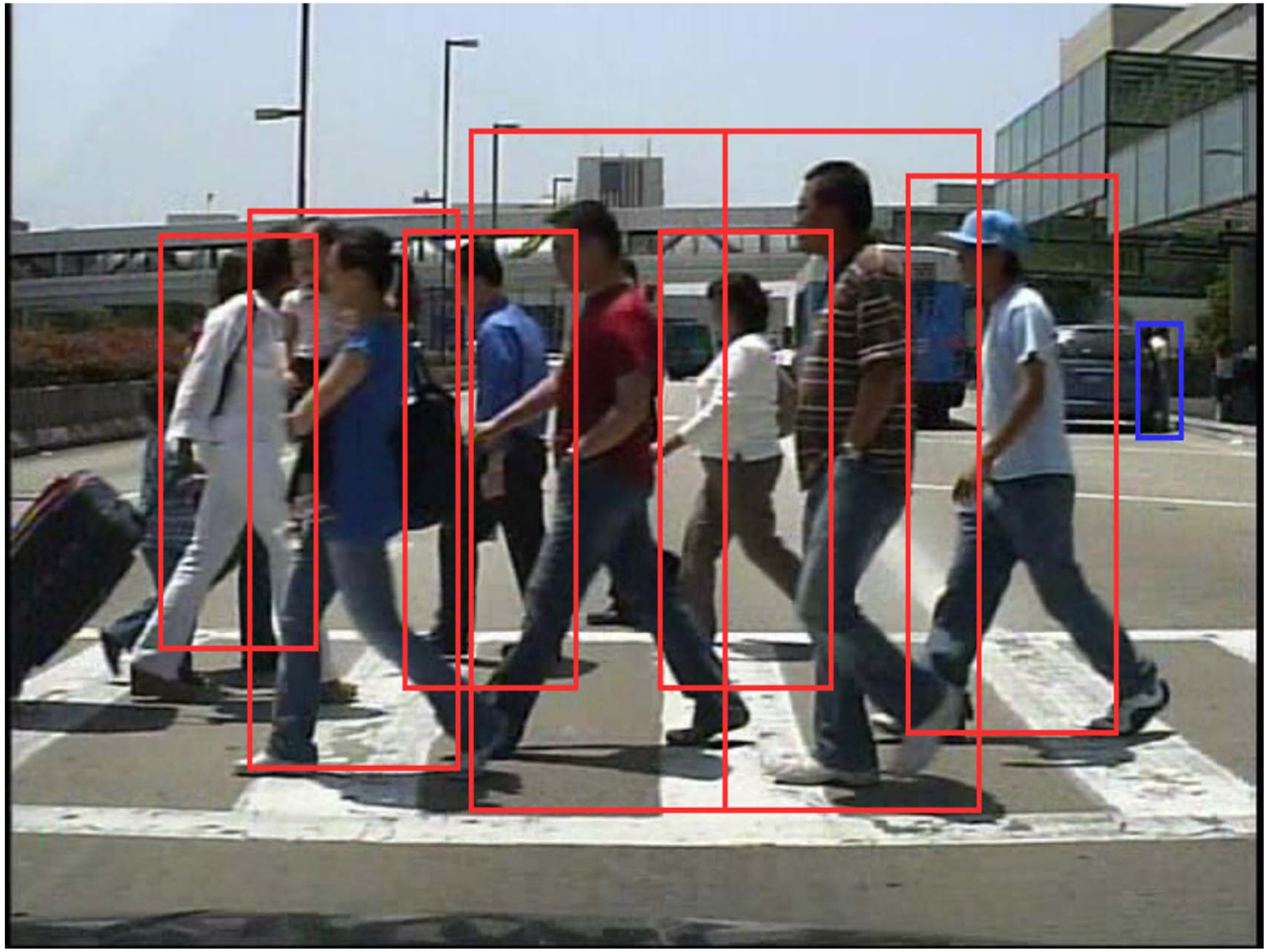




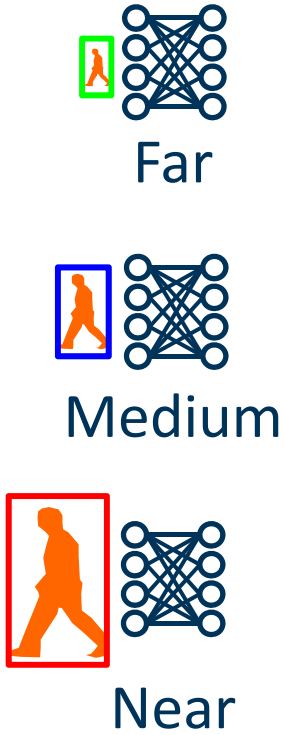
Results – Visual Examples

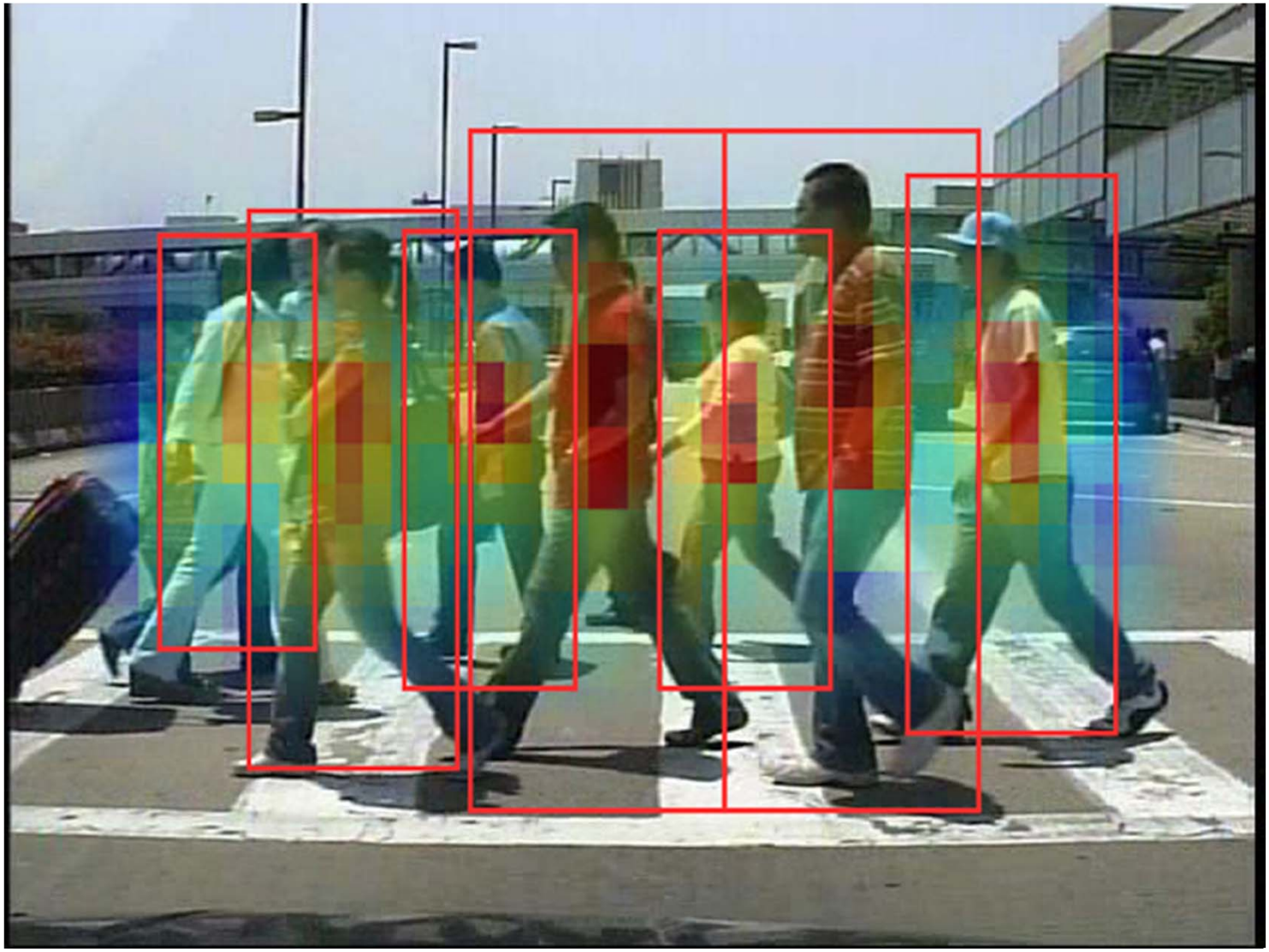
Results – Visual Examples





Results – Visual Examples





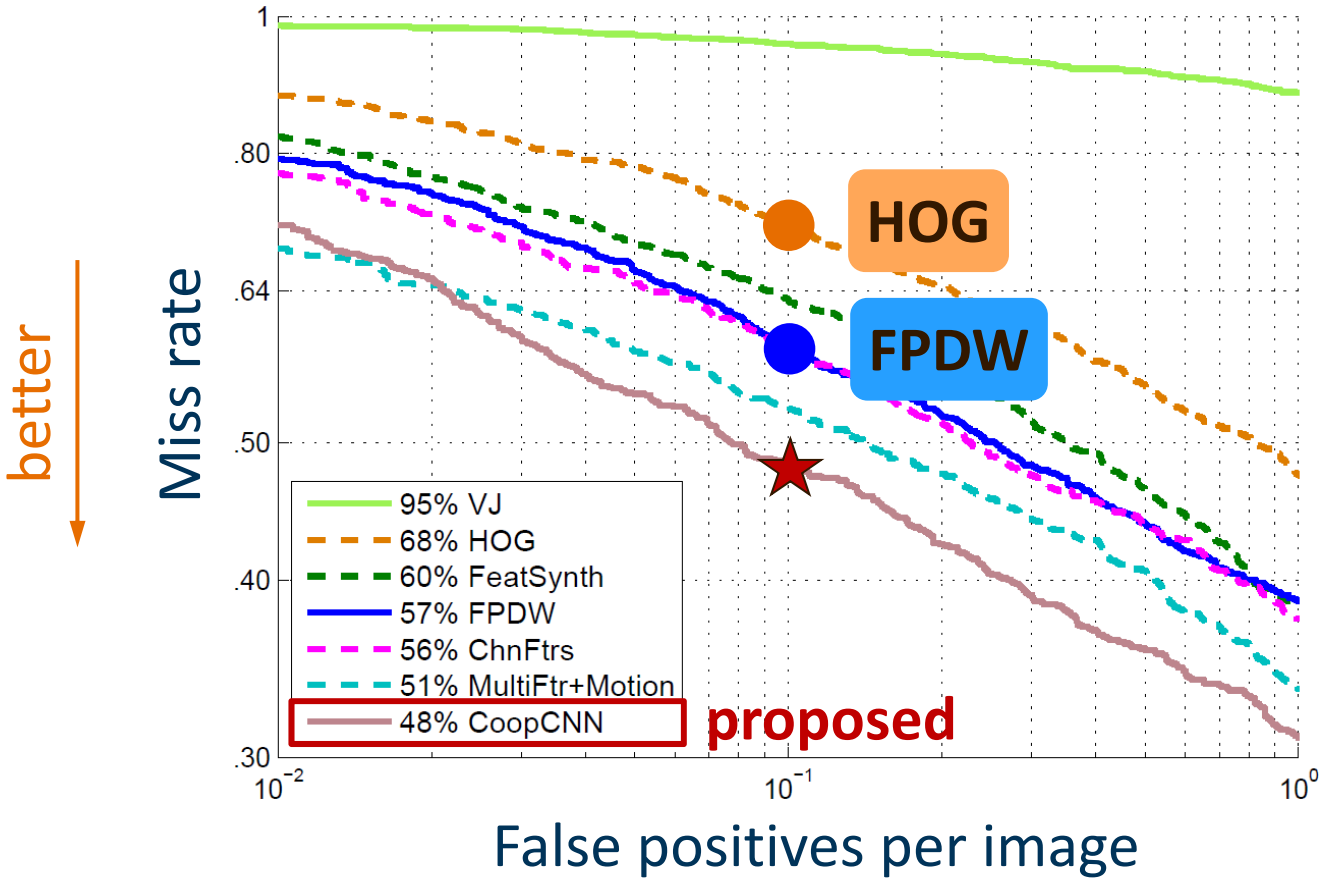
Evaluation

Comparison to state of the art

Evaluation

Comparison to state of the art

- DET Curve (→ ROC on logarithmic axes)
- All detectors not trained on CALTECH



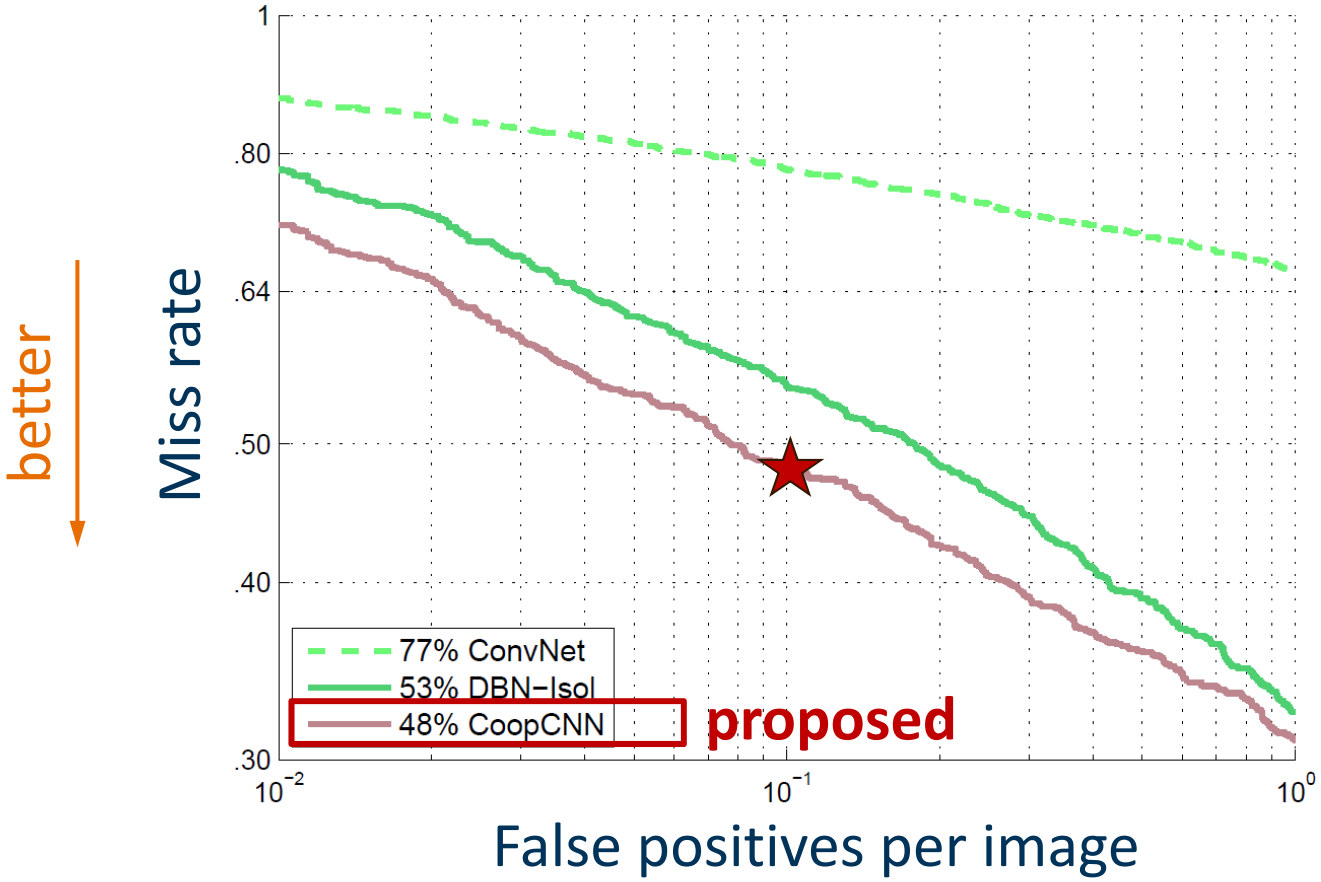
Details

Paper

Evaluation

Comparison to state of the art

- DET Curve (→ ROC on logarithmic axes)
- **Deep learning** detectors not trained on CALTECH



Details

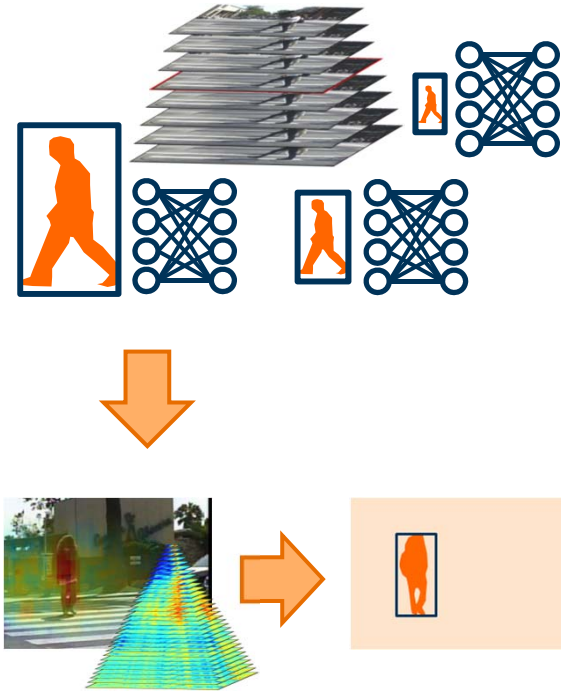
Paper

Conclusion

Conclusion

Cooperative Multi-Scale Convolutional Neural Networks for Person Detection

- **Motivation**: State of the art detectors perform poor in real world scenarios
- **Own Approach**: Fully neural, Hybrid multi-scale detection (detectors / resolution pyramid)
- **Evaluation**: State of the art performance without domain specific training
- **Performance**: Relatively fast on special hardware
- **Future work**: Transfer learning



Clever 3D pooling

Appendix

Training Dataset

Training dataset

Paper

Positive data (persons)

100,107 Patches

from 22 datasets

- Surveillance
- Pedestrian zone
- Sport scenes
- Robot in rehab clinic
- Car traffic scenes (5.6%)

Negative data (non-persons)

628,636 Patches

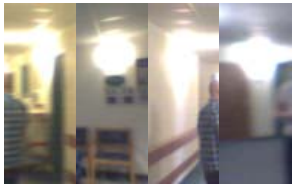
usual

- Landscapes
- Urban scenes
- Robot

- No car traffic scenes!

special cases

- Typical errors



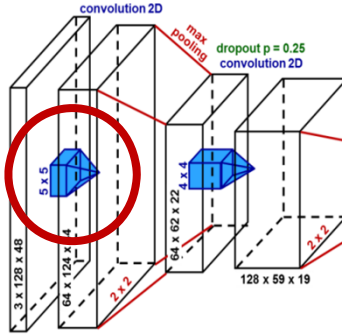
- Misaligned

No training on CALTECH → Test set

Evaluation

What did the nets learn?

1st Conv. layer

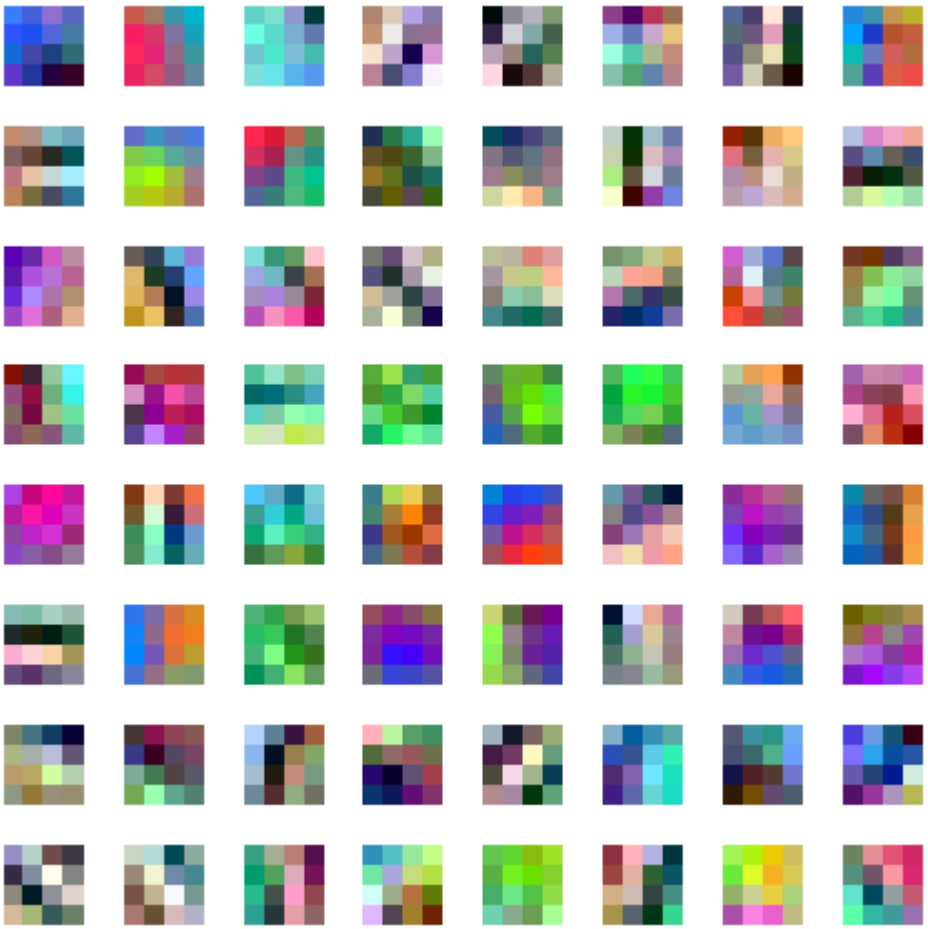


Evaluation – Learned filters

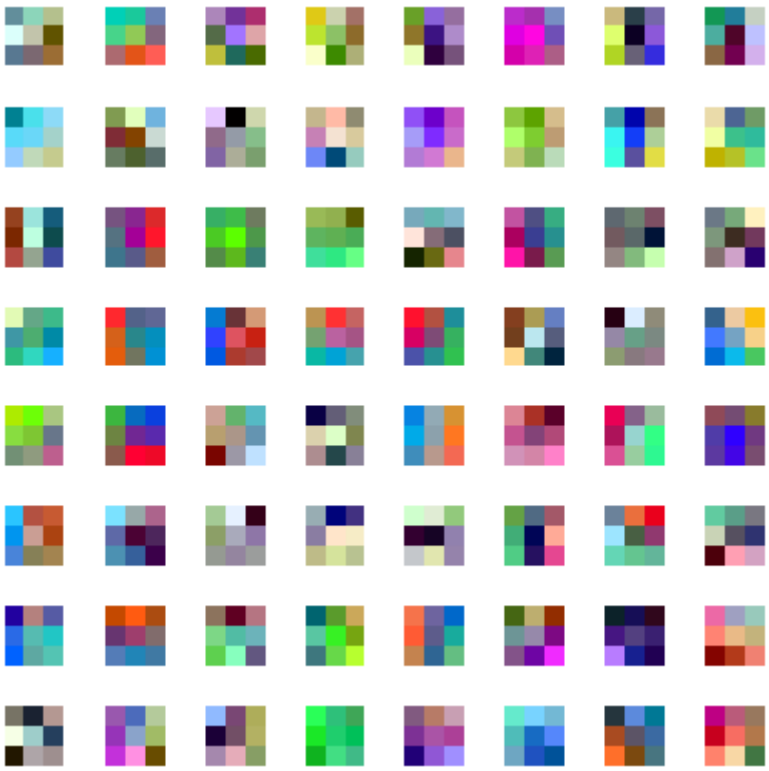
The 64 filters learned in 1st layer of the near scale CNN

Sample image from INRIA dataset

Person scaled, such that bounding box would be 128 x 48 pixels



The 64 filters learned in 1st layer of the medium scale CNN



The 64 filters learned in 1st layer of the far scale CNN

Network

Input Coding

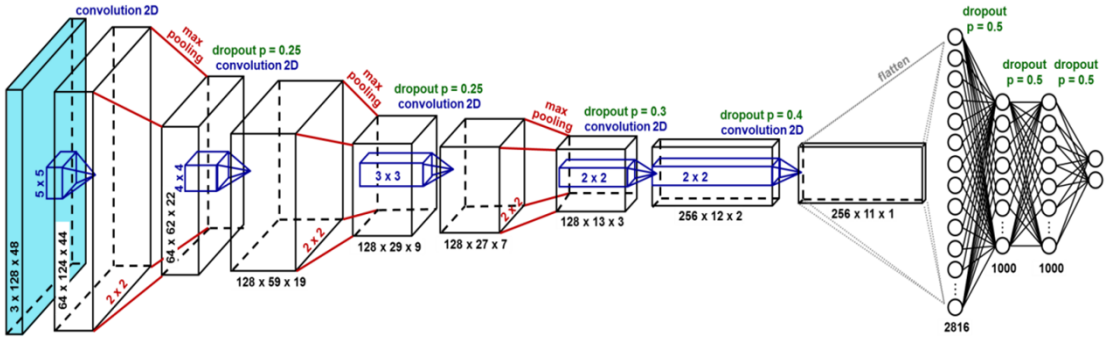
Network Topology – Training

Input

Near

128 x 48 x 3

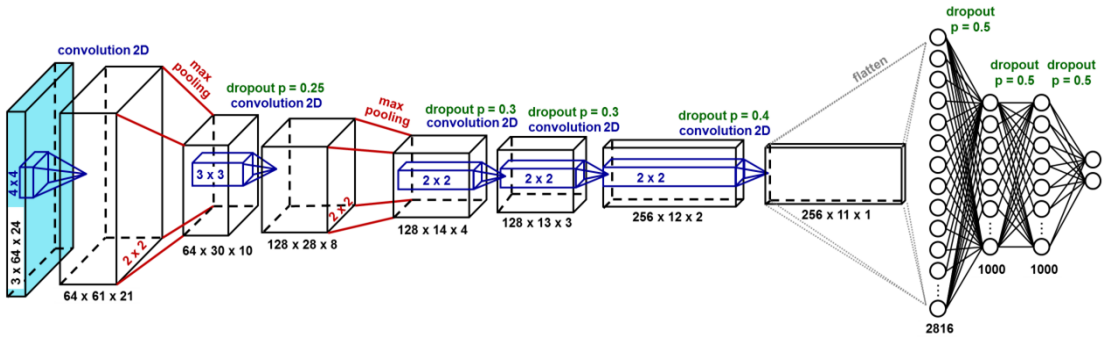
RGB [-1, 1]



Medium

64 x 24 x 3

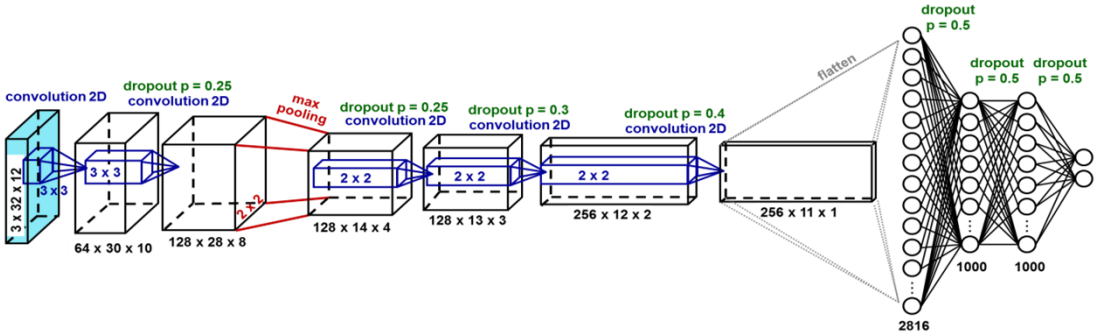
RGB [-1, 1]



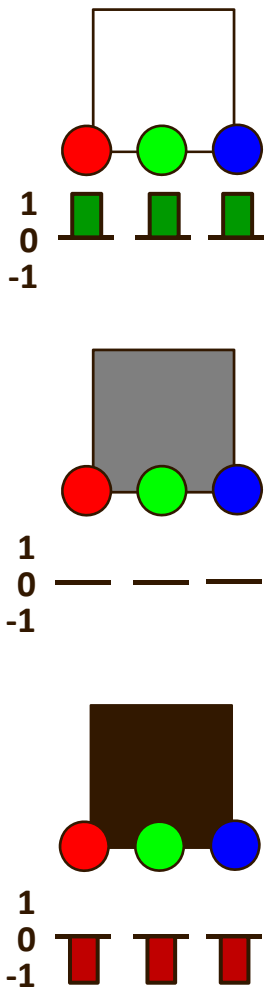
Far

32 x 12 x 3

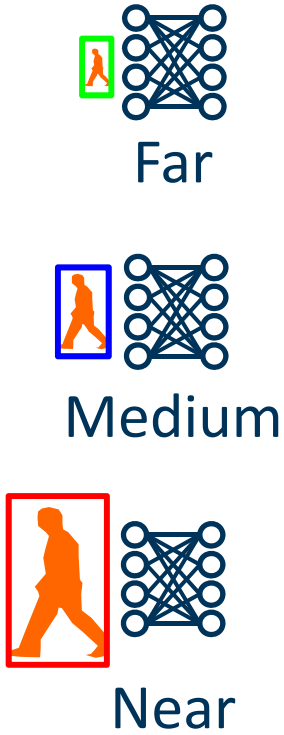
RGB [-1, 1]

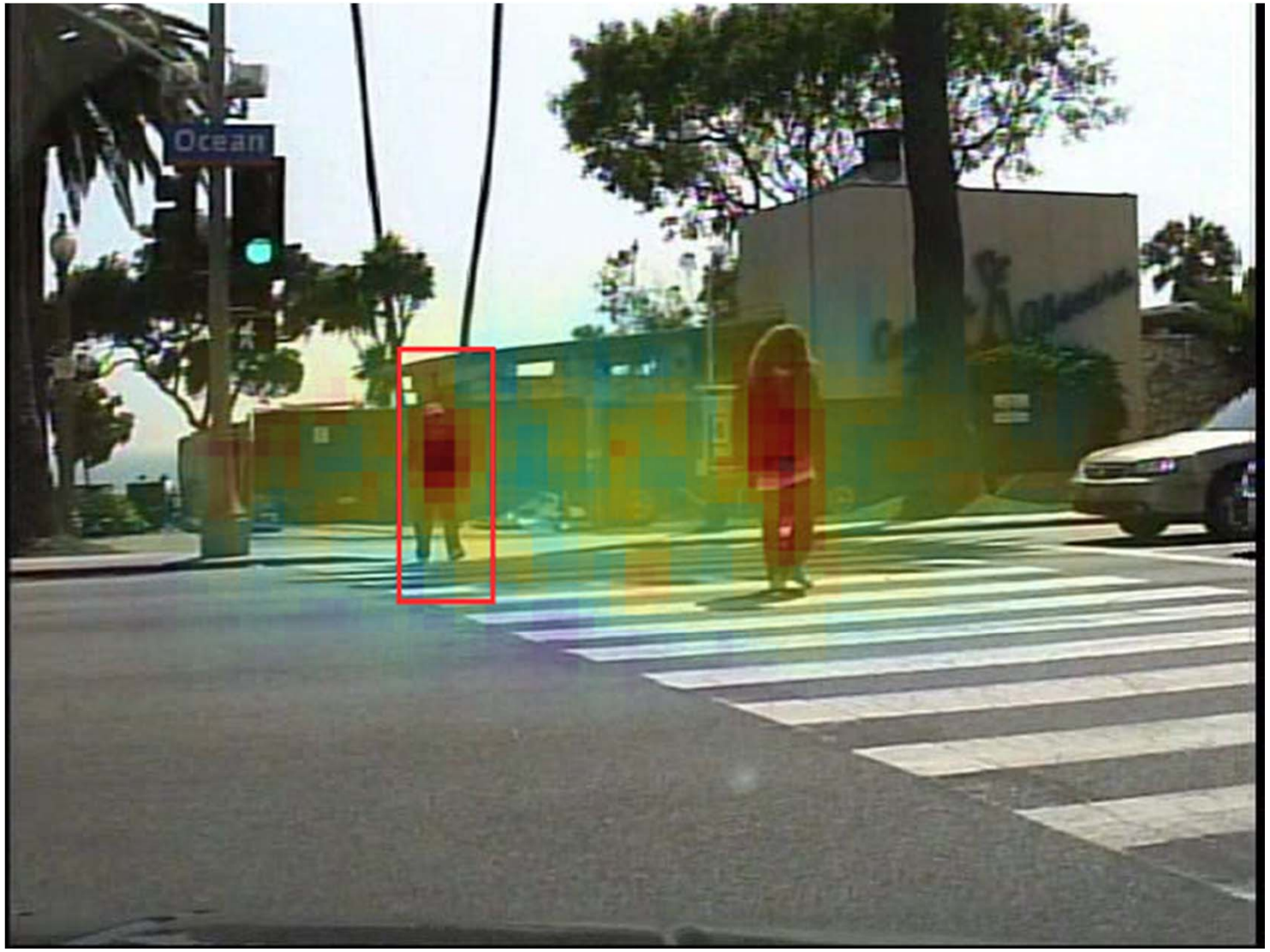


RGB coding

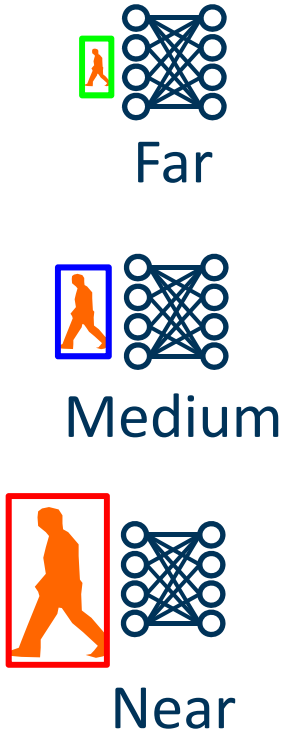


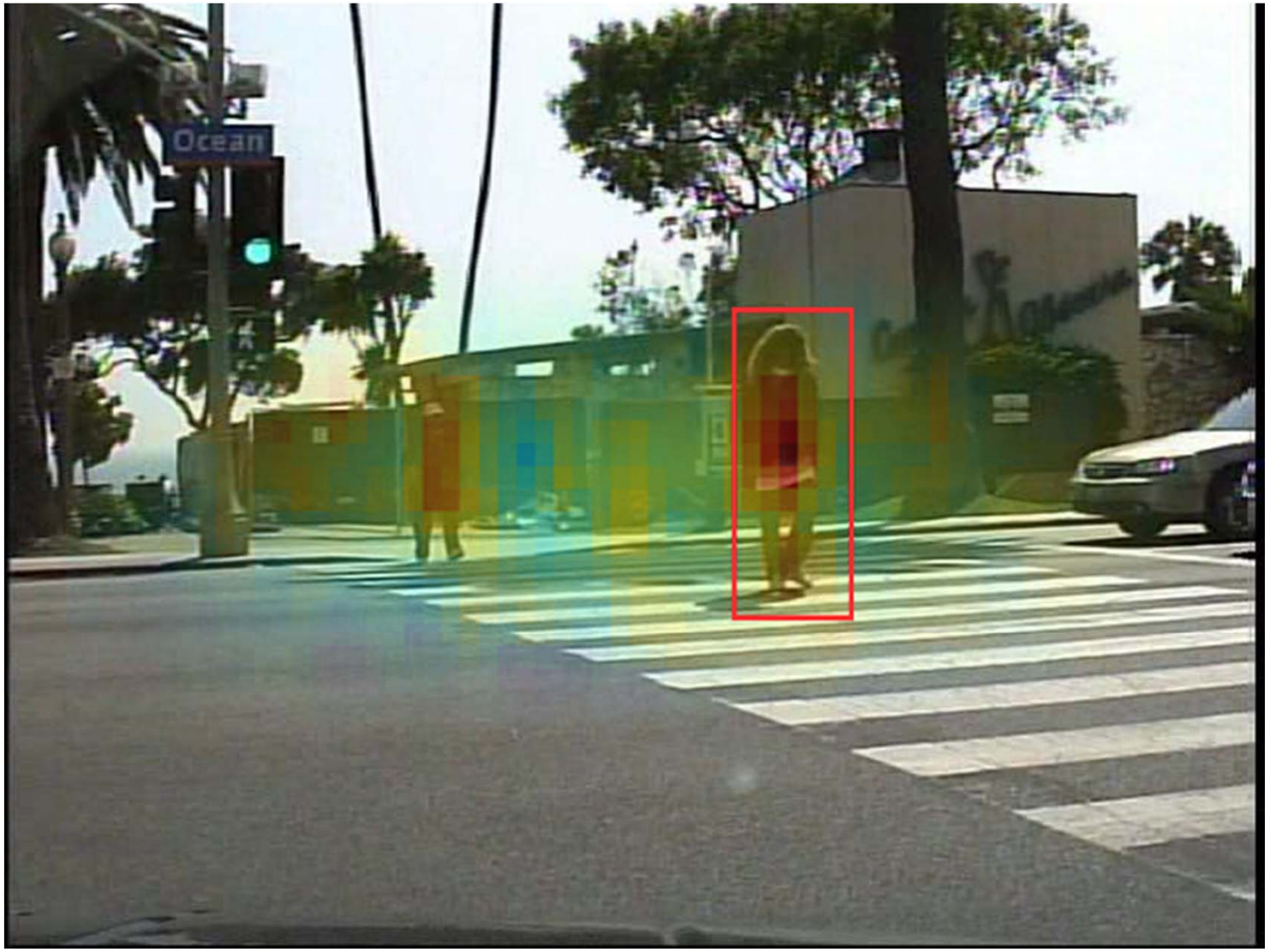
Results – Visual Examples





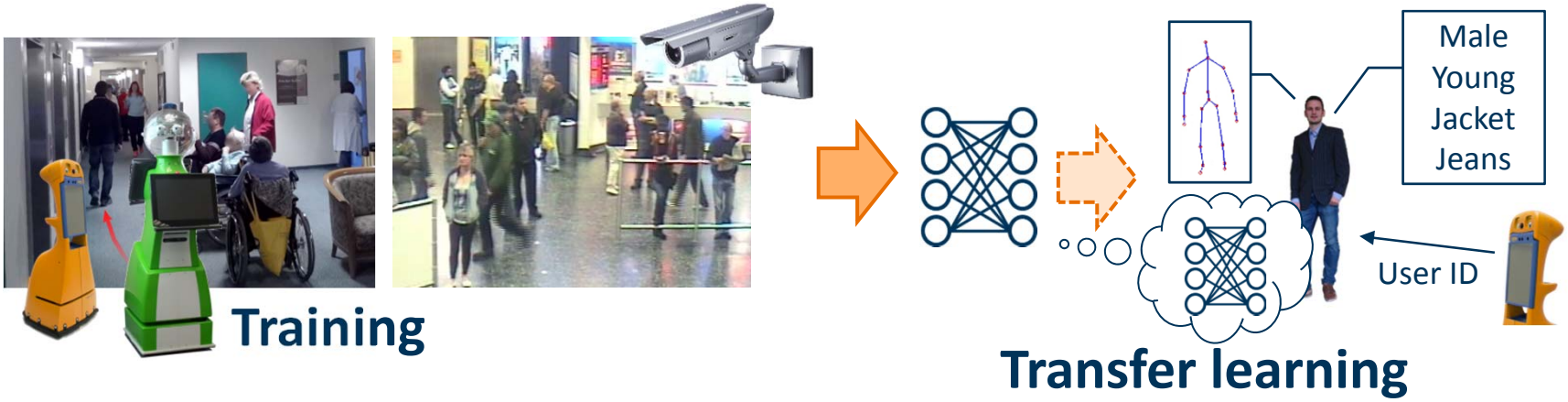
Results – Visual Examples





Future Work

- **Transfer Learning**
  - Adequate model for transfer learning
    - person re-identification
    - activity recognition
    - attribute learning
    - etc.



→ Alternative to AlexNet for person-specific data