# Learning non-maximum suppression

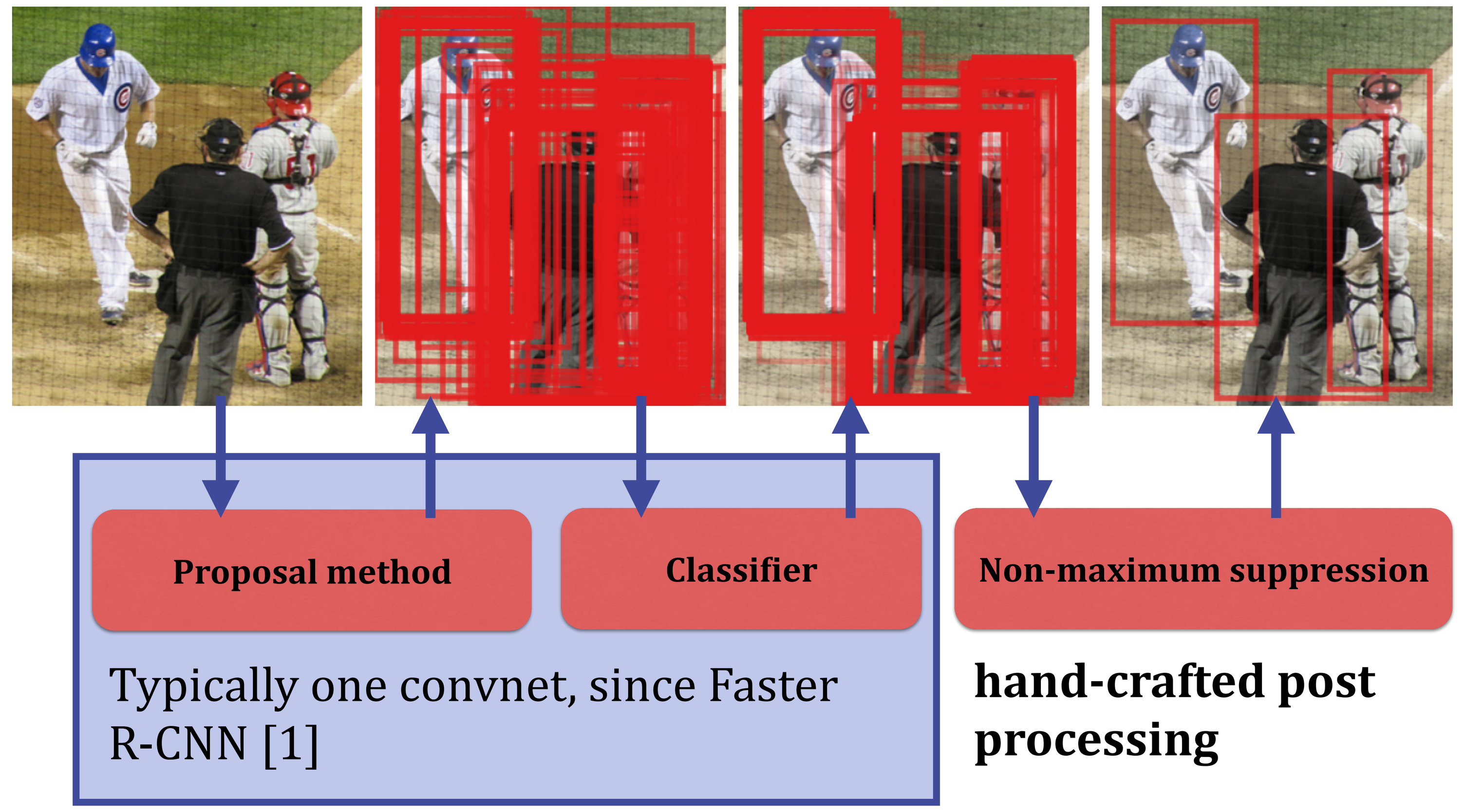
max planck institut informatik

Jan Hosang, Rodrigo Benenson, Bernt Schiele
firstname.lastname@mpi-inf.mpg.de

## Detectors still use hand-crafted NMS!

Virtually all object detectors are not trained end-to-end:

- Classifier scores all detections close to objects high
- NMS is supposed to keep exactly one detection per object




## Can we train a neural network to perform NMS?

## Approach: Neural network for NMS

Pure NMS net: For now decisions are based on detection scores and geometry

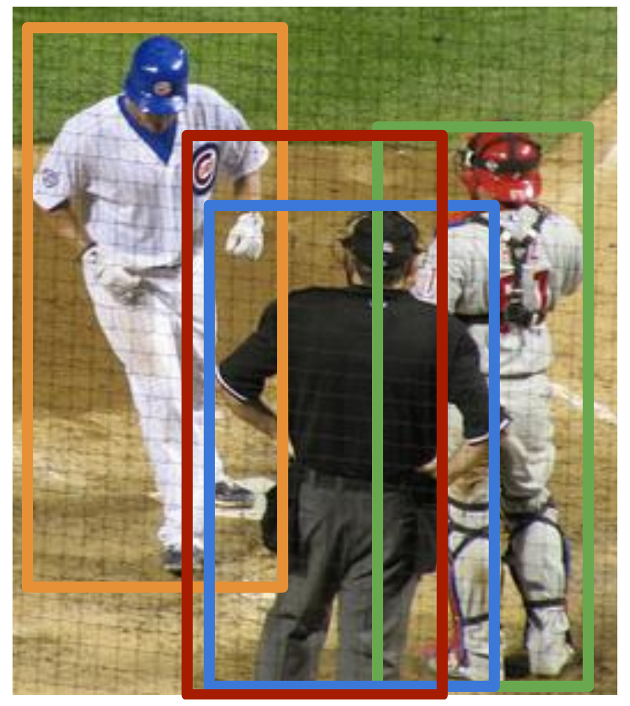
Rescoring: Update score of every detection (instead of suppressing a detection, we decrease its score)

### Key ingredients

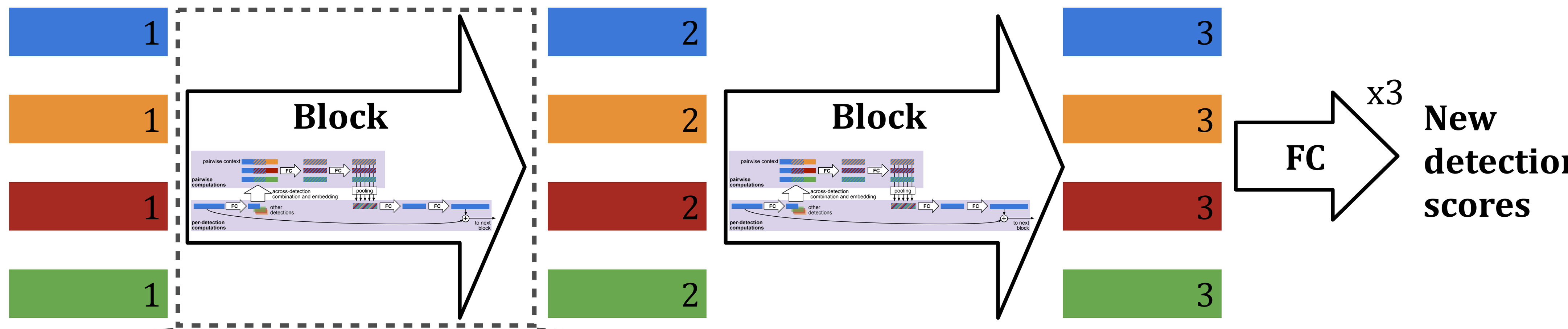
What is necessary to train a neural net to output exactly one detection per object?

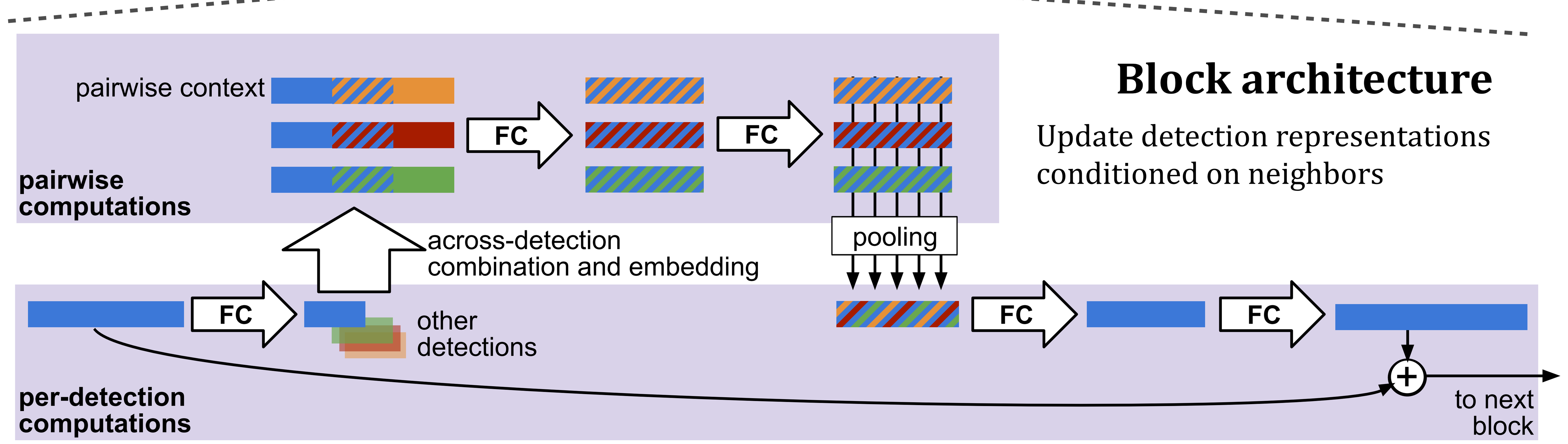
- **Matching loss:** penalize double detections
- **Joint processing** of neighbors: whether a detection is the "best one" depends on other detections close-by

## Gnet jointly rescores all detections by letting them "talk"



1. Latent detection representation (start from zeros)
2. Detection representations are updated by "blocks" that look at neighboring detections (
3. Predict new detection scores from detection representations



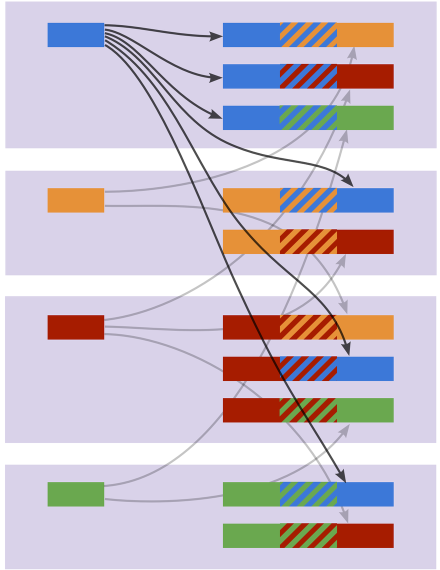

### Block architecture

Update detection representations conditioned on neighbors

## Across-detection embedding

For each detection, build pairs with all nearby detections. Concatenate

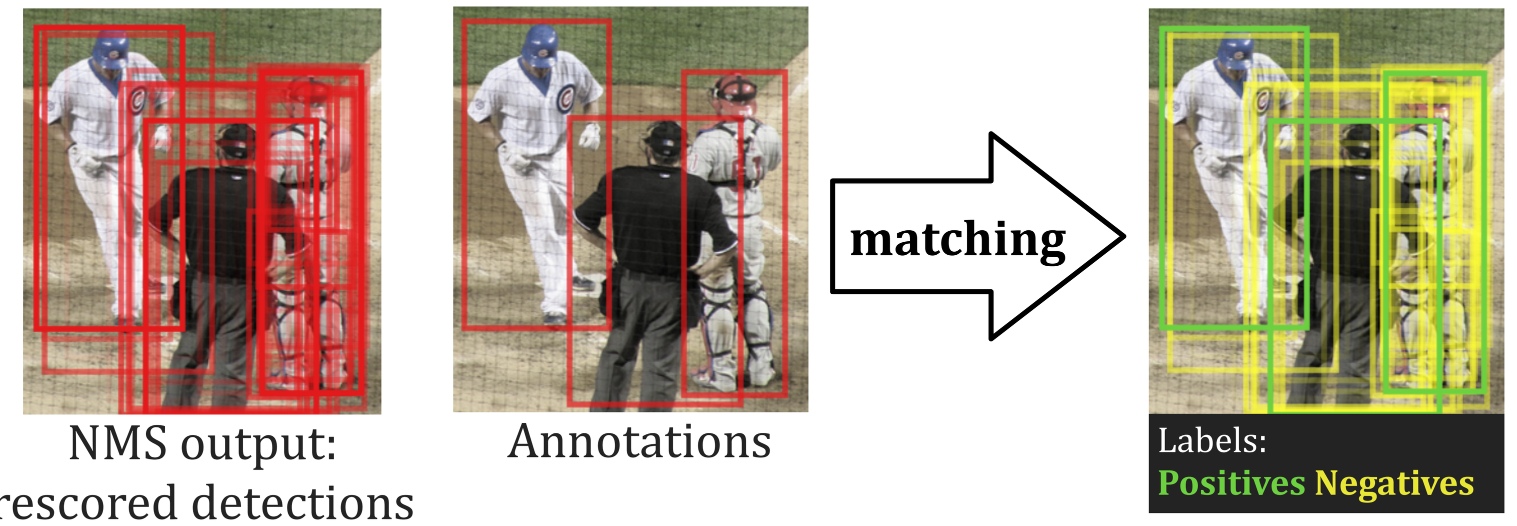
- latent detection representations
- embedding of detection scores, IoU, other geometric information



## Matching loss

Run matching to determine labels for detections to allow only one detection per object [3] (same as in evaluation): matched detections are true positives, unmatched detections are negatives

- Matching depends on scores, so labels are determined after rescoring
- Labels used in standard cross entropy loss
- **Multiclass** matching also yields binary labels




NMS output: rescored detections

Annotations

### References

[1] S. Ren, K. He, R. Girshick, J. Sun. **Faster R-CNN.** NIPS 2015

[2] J. Hosang, R. Benenson, B. Schiele. **A convnet for non-maximum suppression.** GCPR 2016

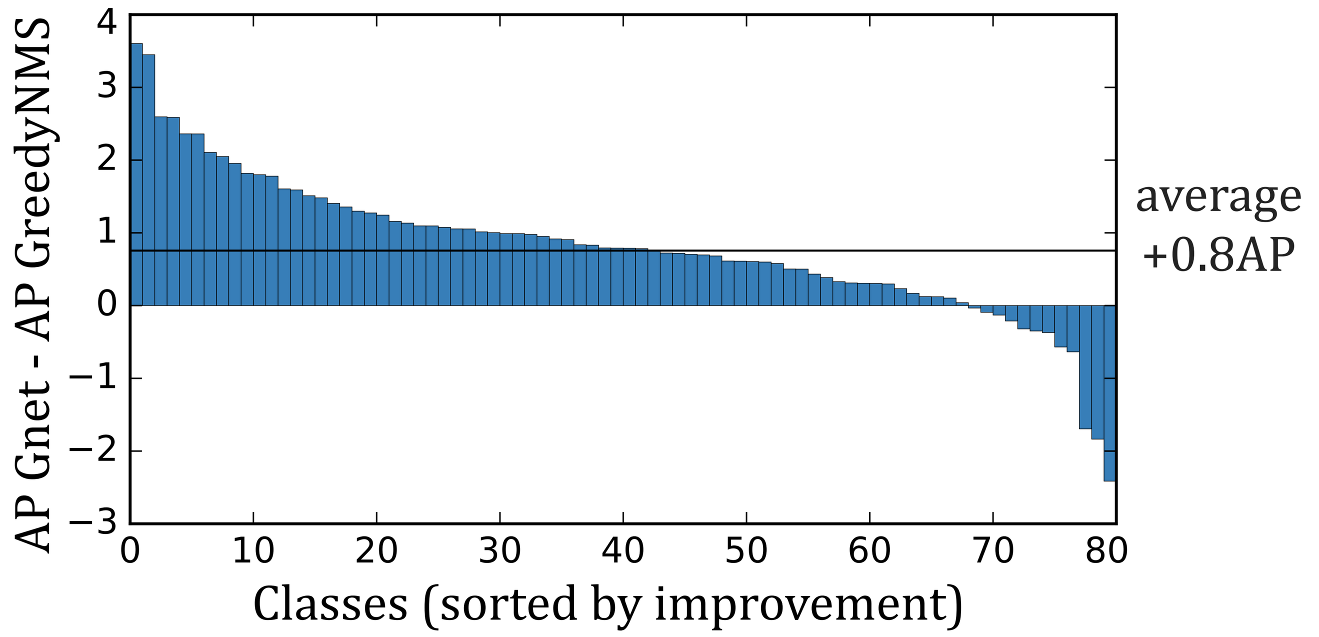
[3] R. Stewart, M. Andriluka, A. Ng. **End-to-end people detection in crowded scenes.** CVPR 2016

## Take away messages

- Neural nets can learn to perform NMS
- Learning NMS in a neural net requires a matching loss and joint rescoring
- Multiple blocks help performance

➡ *This provides the opportunity of true end-to-end learning for object detectors*

## Experiments: COCO multiclass




## Experiments: COCO persons

| | Method | All $AP_{0.5}$ | All $AP_{0.5}^{0.95}$ | Occlusion [0, 0.5) $AP_{0.5}$ | Occlusion [0, 0.5) $AP_{0.5}^{0.95}$ | Occlusion [0.5, 1] $AP_{0.5}$ | Occlusion [0.5, 1] $AP_{0.5}^{0.95}$ |
|---|---|---|---|---|---|---|---|
| val | GreedyNMS>0.5 | 65.6 | 35.6 | 65.2 | 35.2 | 35.3 | 12.1 |
| val | Gnet, 8 blocks | **67.3** | **36.9** | **66.9** | **36.7** | **36.7** | **13.1** |
| test | GreedyNMS>0.5 | 65.0 | 35.5 | 61.8 | 33.8 | 30.3 | 11.0 |
| test | Gnet, 8 blocks | **66.6** | **36.7** | **66.8** | **36.1** | **33.9** | **12.4** |