

Fast and accurate object detection in high resolution 4K and 8K video using GPUs

Vit Ruzicka, Franz Franchetti

Motivation

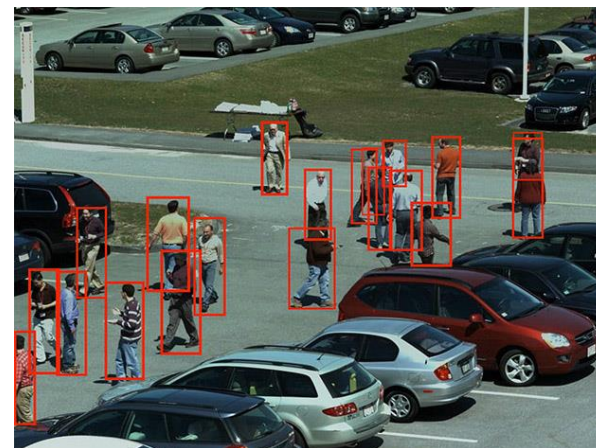
- Object detection in high resolution video



4K: 3840x2160 px

Problem

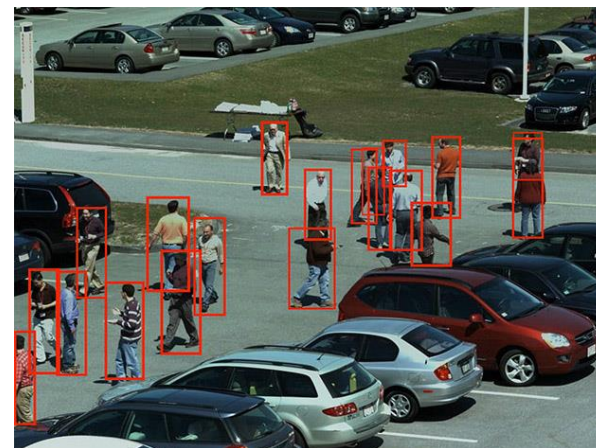
- Object detection
 - Faster RCNN
 - YOLO
 - SSD



Problem

■ Object detection

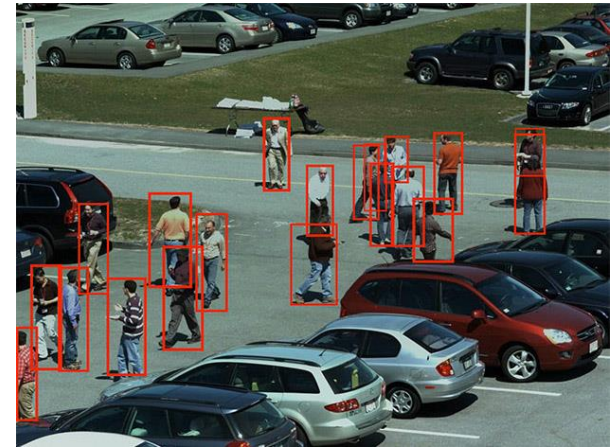
- Faster RCNN *max 1000x600 px*
- YOLO *max 608x608 px*
- SSD *max 512x512 px*



Problem

■ Object detection

- Faster RCNN *max 1000x600 px*
- YOLO *max 608x608 px*
- SSD *max 512x512 px*



■ Object detection in high resolution



4K

3840x2160 px

608x608 px

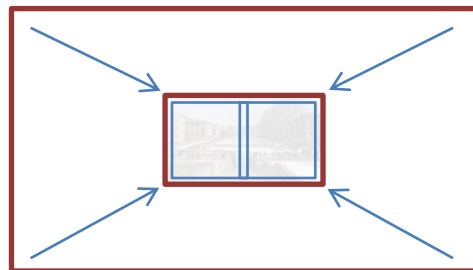


Approaches



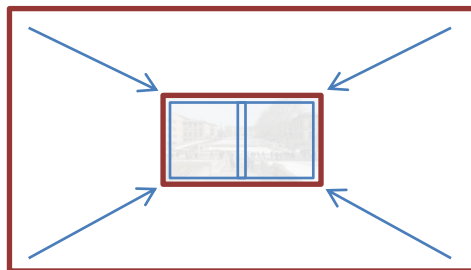
Approaches

Downscale

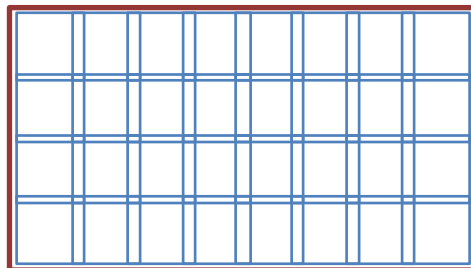


Approaches

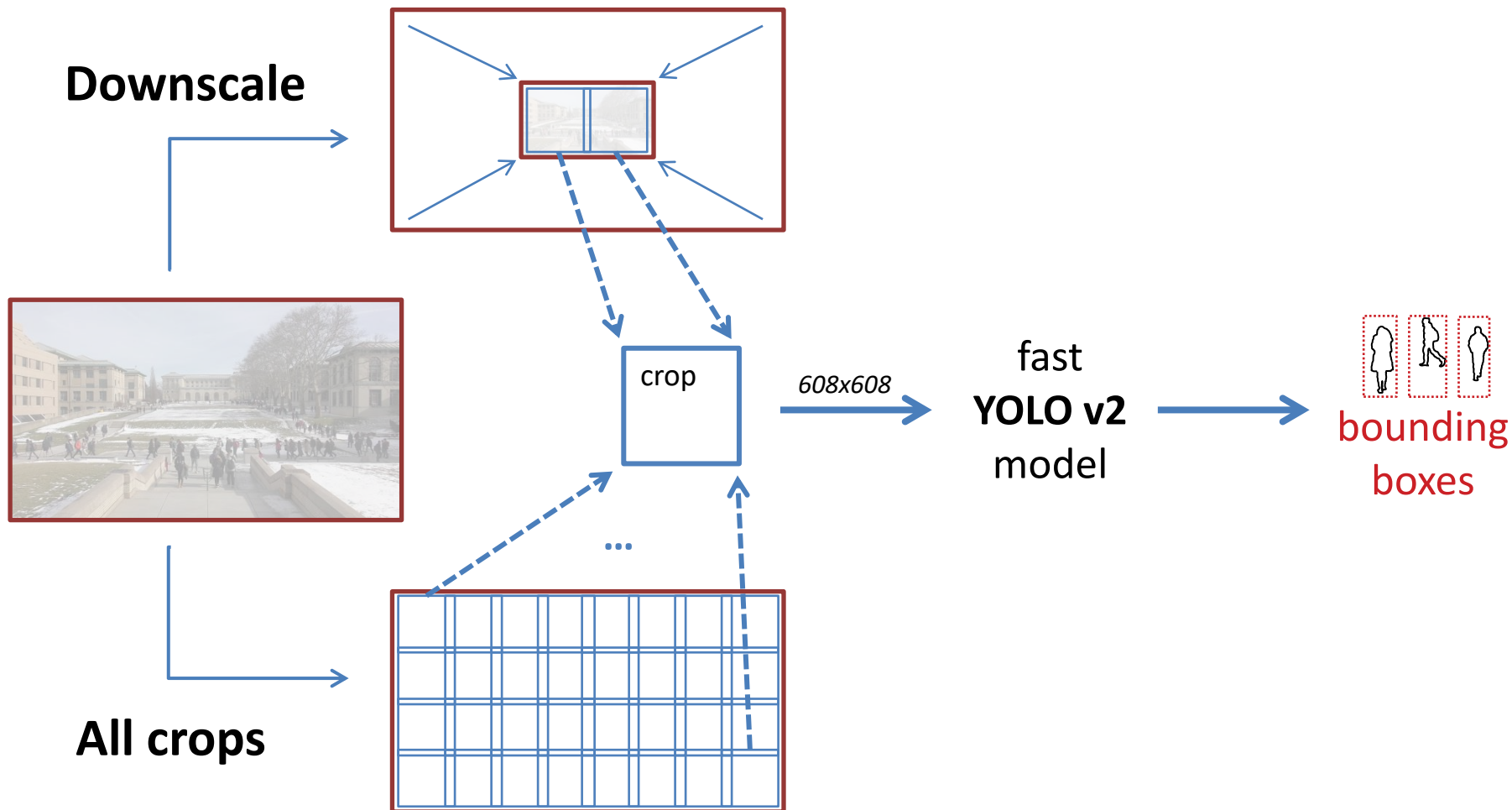
Downscale



All crops

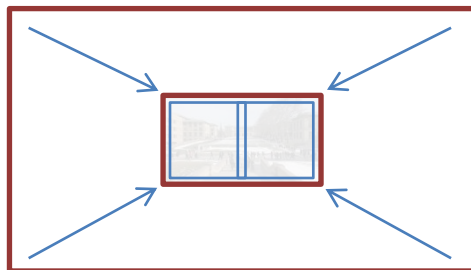


Approaches

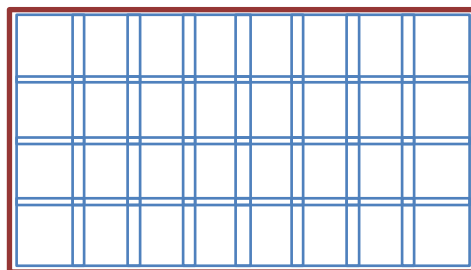


Approaches

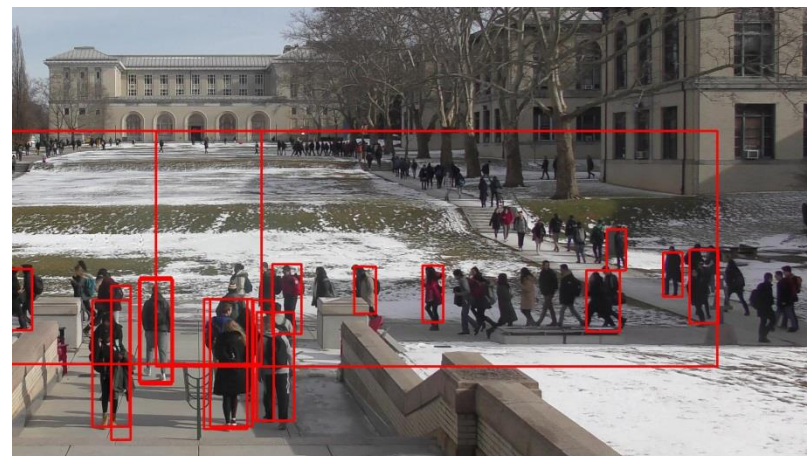
Downscale



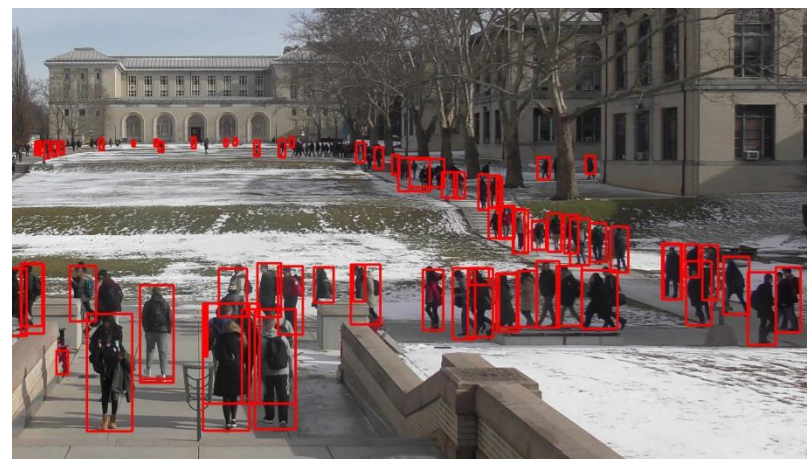
All crops



Showing a detail:

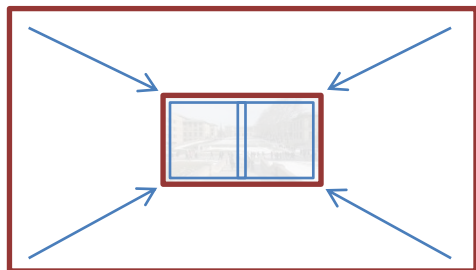


Low accuracy but **fast**



High accuracy but **slow**

Goal

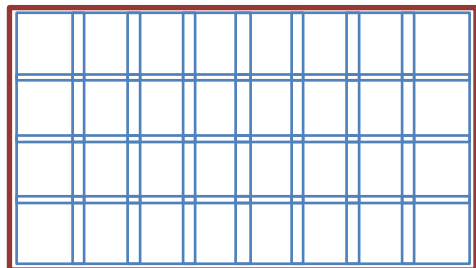


Downscale baseline

- Low accuracy
- **Fast**



Balance accuracy and speed

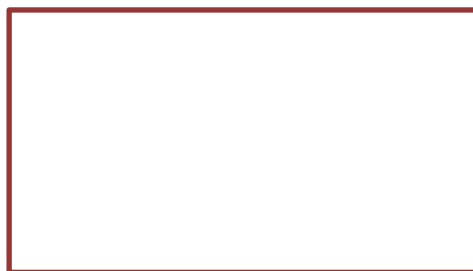


All crops baseline

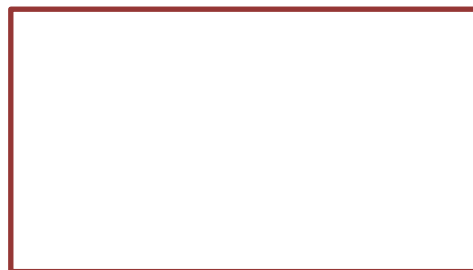
- **High accuracy**
- Slow

Proposed Solution

Attention step

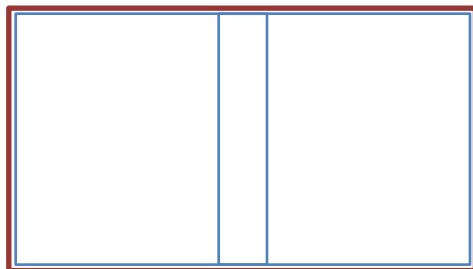


Final evaluation step



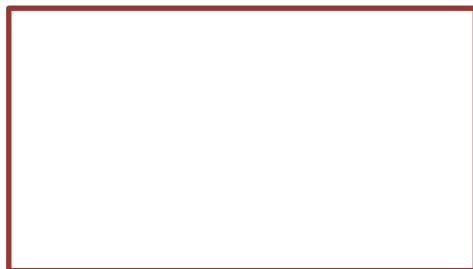
Proposed Solution

Attention step



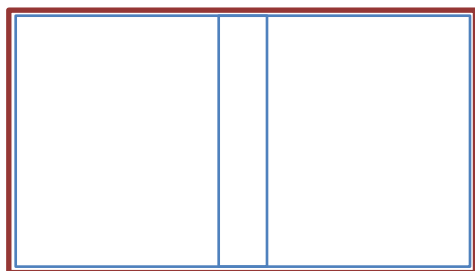
- **Fast**
- Low resolution

Final evaluation step



Proposed Solution

Attention step



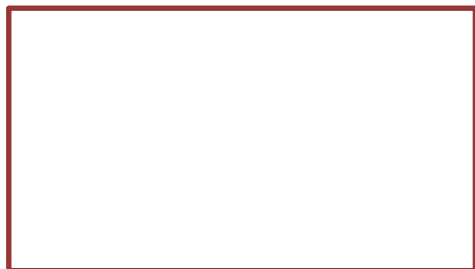
- **Fast**
- Low resolution



First rough
bounding
boxes

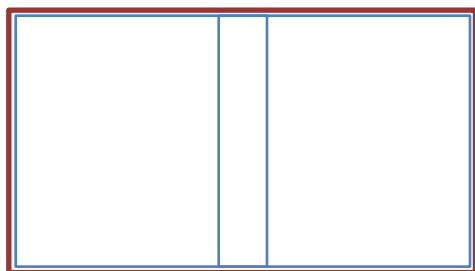


Final evaluation step



Proposed Solution

Attention step



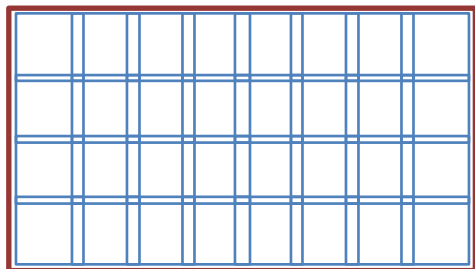
- **Fast**
- Low resolution



First rough
bounding
boxes



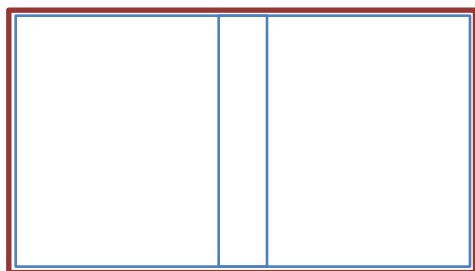
Final evaluation step



- **High resolution**
- Slow

Proposed Solution

Attention step



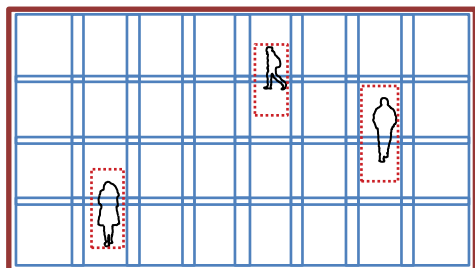
- **Fast**
- Low resolution

YOLO v2

First rough
bounding
boxes



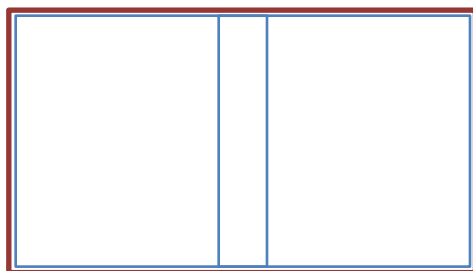
Final evaluation step



- **High resolution**
- Slow

Proposed Solution

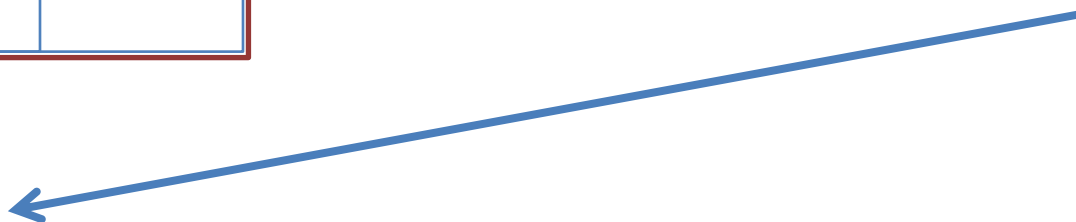
Attention step



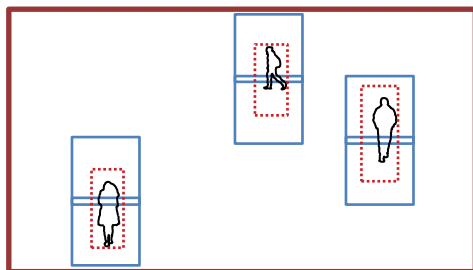
- **Fast**
- Low resolution

YOLO v2 →

First rough
bounding
boxes



Final evaluation step



only active crops

- **High resolution**
- **Slow**

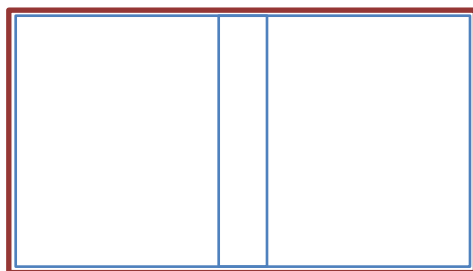
YOLO v2 →

Final
bounding
boxes

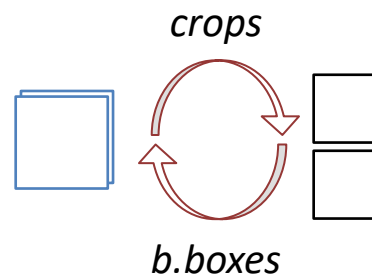


Parallel Evaluation

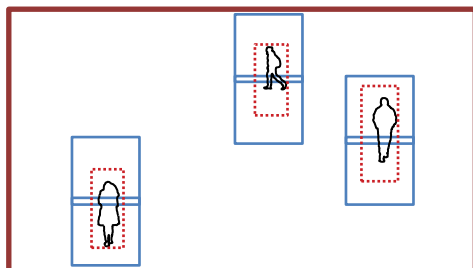
frame



**Attention
step**



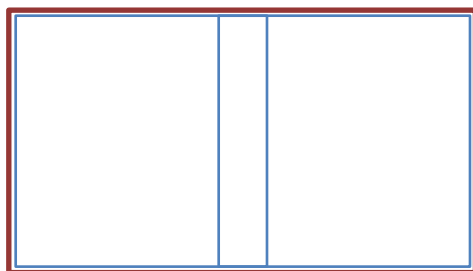
attention
evaluation
server(s)



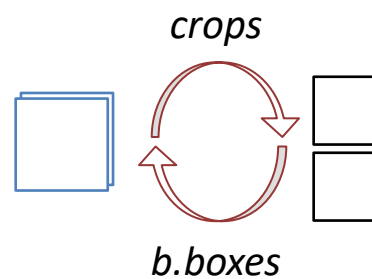
**Final
evaluation**

Parallel Evaluation

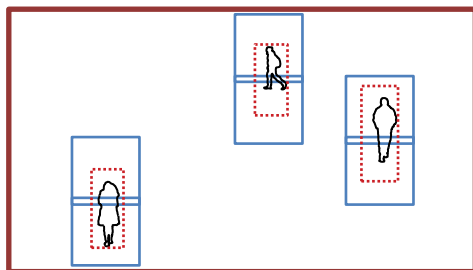
frame



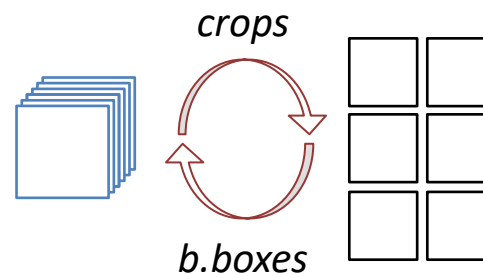
**Attention
step**



attention
evaluation
server(s)

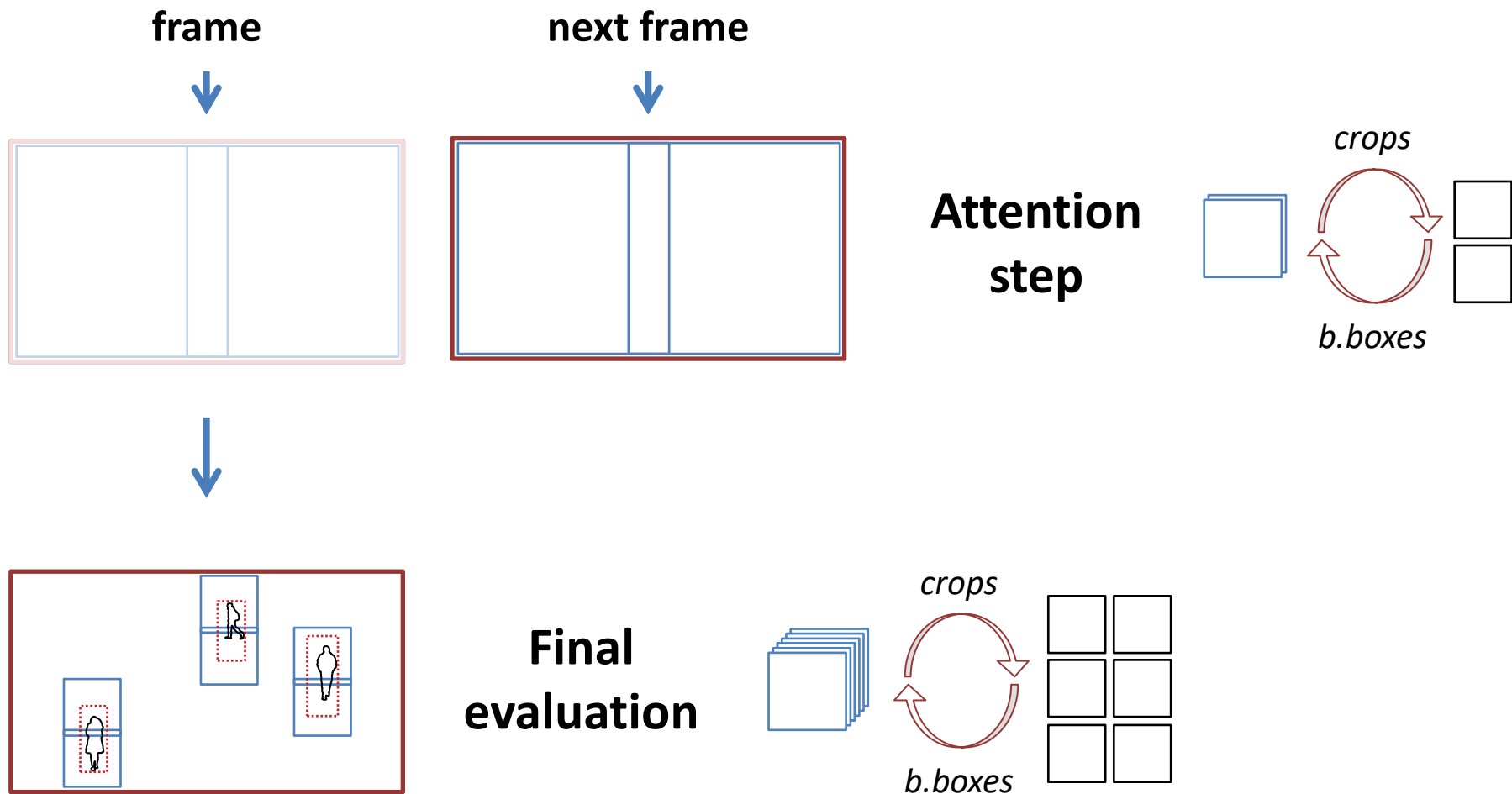


**Final
evaluation**



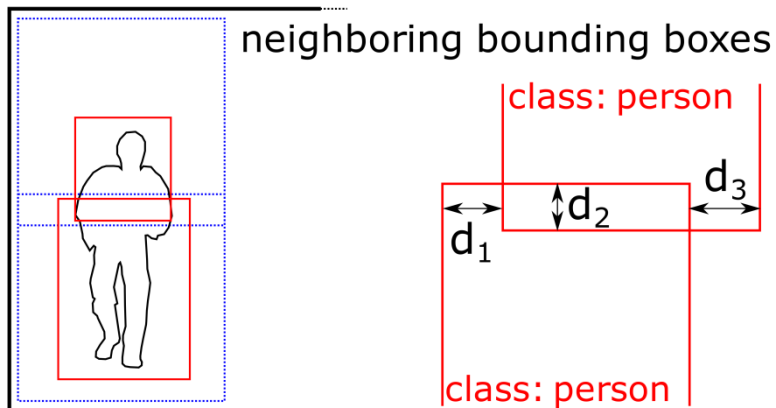
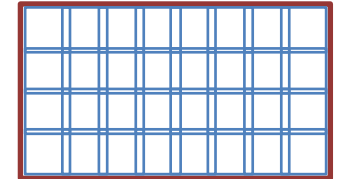
final
evaluation
server(s)

Pre-computing Attention



Post Processing

- Final bounding boxes can be cut in half by a boundary of the grid:



- **Merge** nearby proposals
- Average bounding box of class (“*person*”, ...) as guidance for thresholds

Experiments

- Accuracy

- Speed

Experiments

- Accuracy (*PASCAL VOC average precision*)



PEViD dataset *“easy”*



our dataset *“hard”*

- Speed

Experiments

- Accuracy (*PASCAL VOC average precision*)



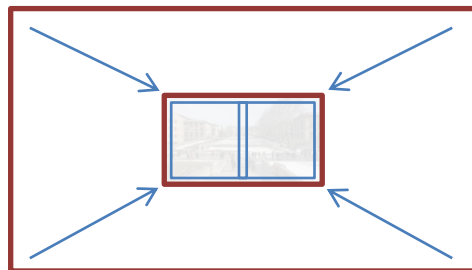
PEViD dataset “easy”



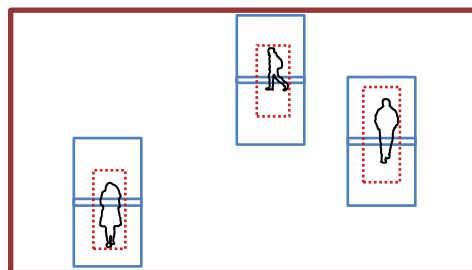
our dataset “hard”

- Speed (*time, ms*)
 - Distributed workers on server, including client \leftrightarrow server transfer
CPU: Intel Xeon E5-2683, GPUs: Tesla P100 Pascal

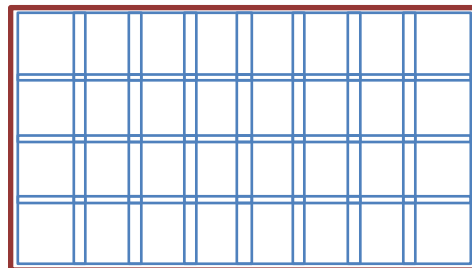
Experiments - methods



Downscale baseline



Our method



All crops baseline

Results

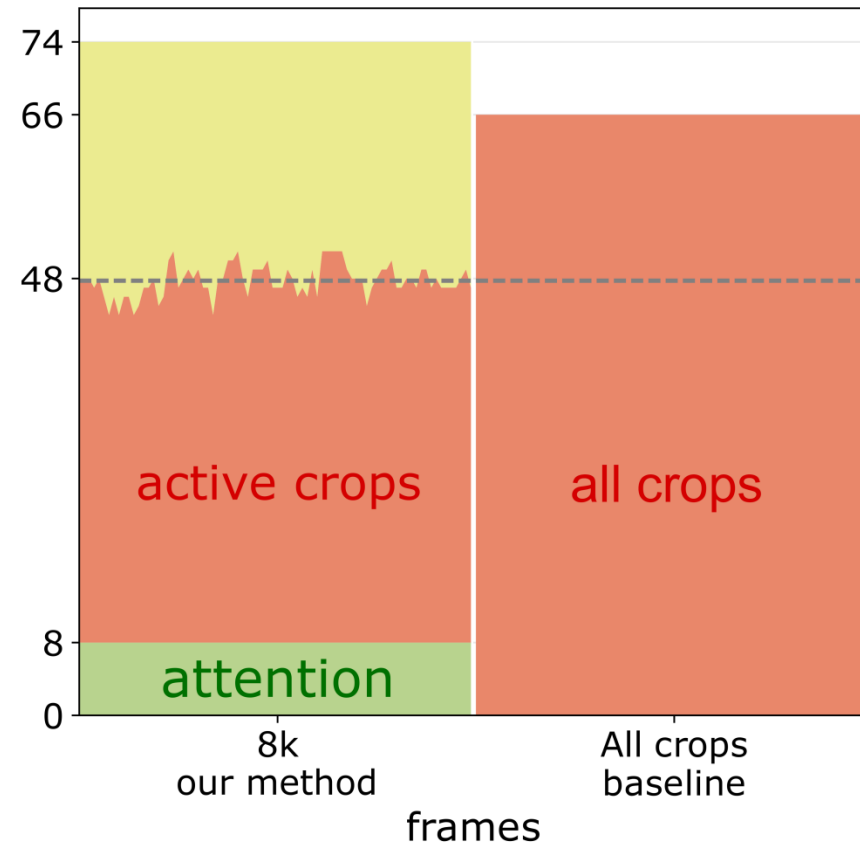


all crops



our method

■ Number of evaluated crops



Results

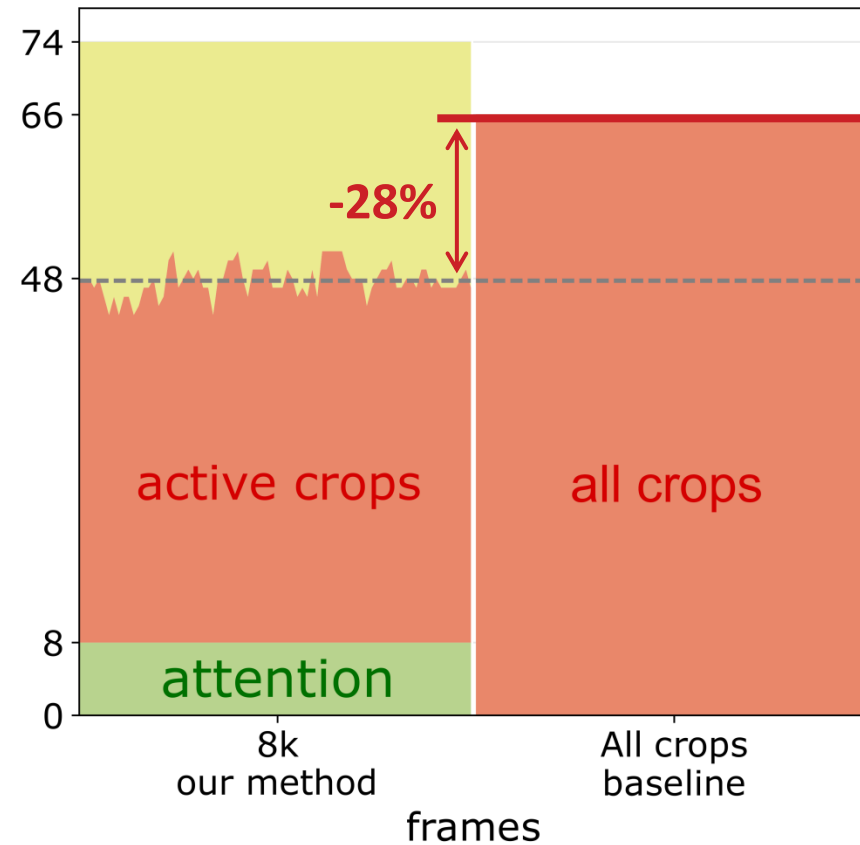


all crops

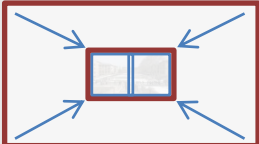
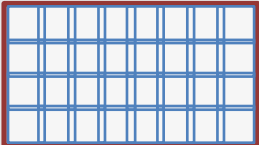


our method

Number of evaluated crops

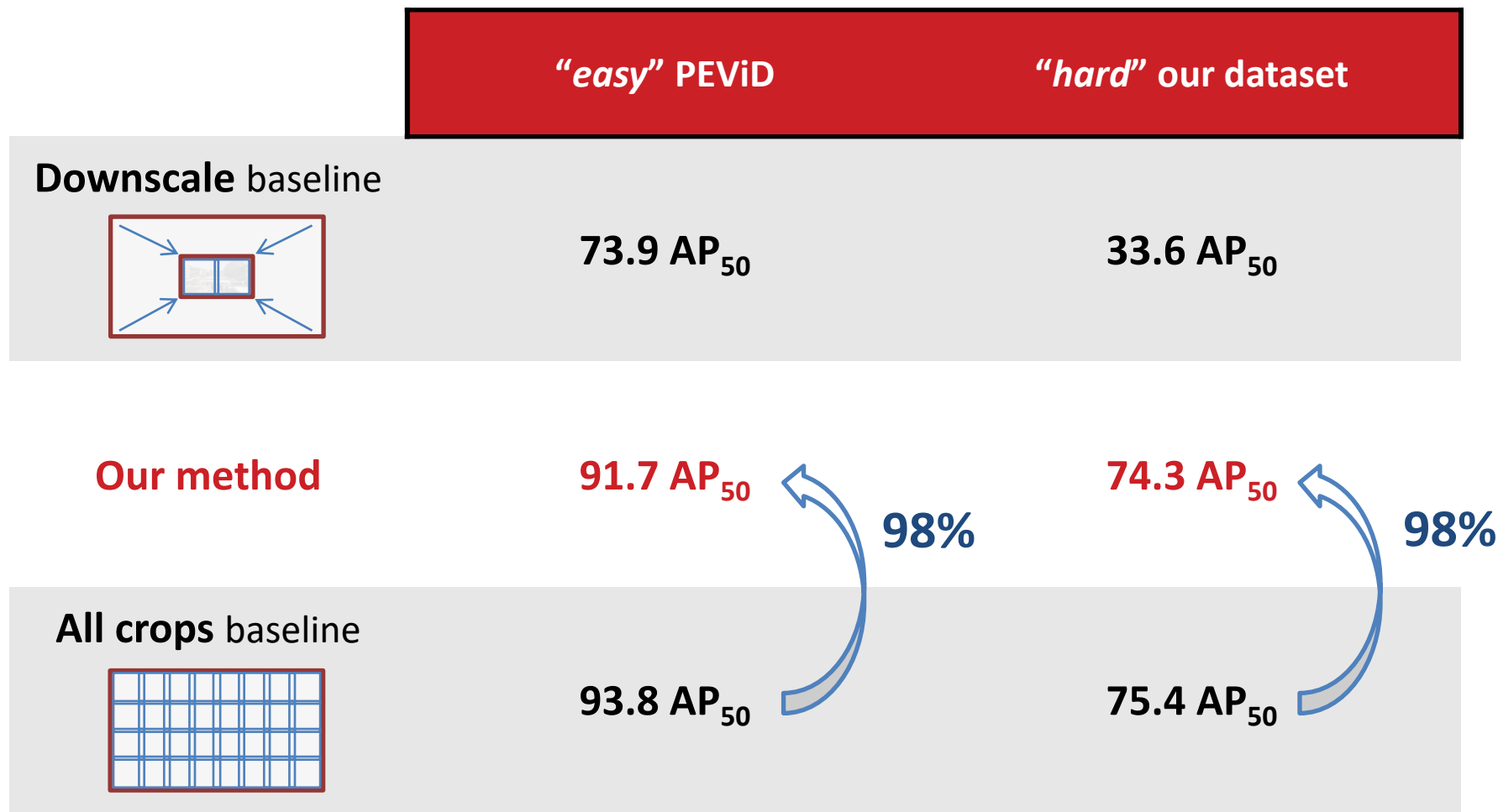


Results - Accuracy

	<i>“easy”</i> PEViD	<i>“hard”</i> our dataset
Downscale baseline 	73.9 AP₅₀	33.6 AP₅₀
Our method	91.7 AP₅₀	74.3 AP₅₀
All crops baseline 	93.8 AP₅₀	75.4 AP₅₀

AP = Average Precision between predicted and ground truth *b.boxes*, higher is better

Results - Accuracy



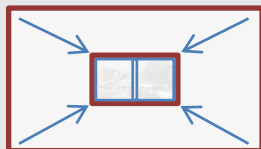
AP = Average Precision between predicted and ground truth *b.boxes*, higher is better

Results - Speed

“easy” PEViD

“hard” our dataset

Downscale baseline



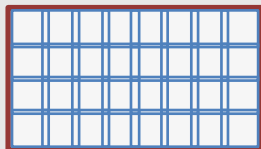
Reported 50 ms, 20 fps for YOLO v2

Our method

201 ms, 5 fps

247 ms, 4 fps

All crops baseline



249 ms, 4 fps

314 ms, 3.2 fps

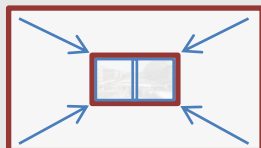
Time in **ms**, lower is better

Results - Speed

“easy” PEViD

“hard” our dataset

Downscale baseline



Reported 50 ms, 20 fps for YOLO v2

low
accuracy

Our method

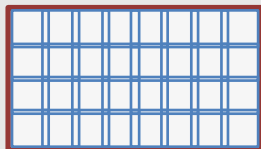
201 ms, 5 fps

-19%

247 ms, 4 fps

-21%

All crops baseline



249 ms, 4 fps

314 ms, 3.2 fps

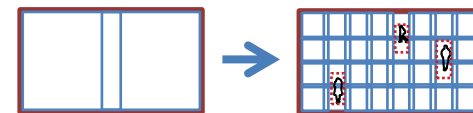
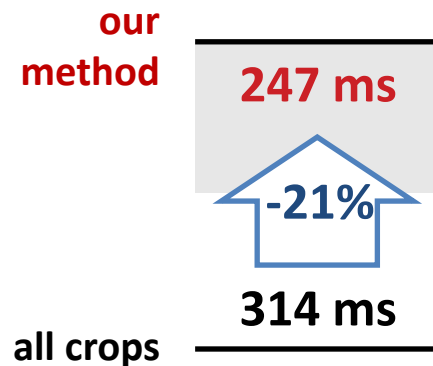
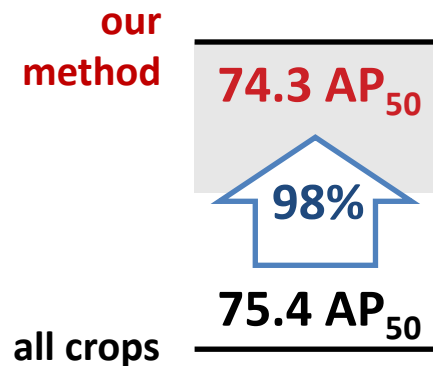
Time in ms, lower is better

Results - Video

- Video presentation, also available at:
<https://youtu.be/07wCxSItnAk>

Summary

- Our method maintains high **accuracy**
- Increases **performance** on tested datasets
- **General method** allowing for custom implementations of attention and final evaluation steps



Thank you for your attention, Questions?

Fast and accurate object detection in high
resolution 4K and 8K video using GPUs

[Vit Ruzicka](#), Franz Franchetti