PERFORMANCE OF RECENT TINY/SMALL YOLO VERSIONS IN THE CONTEXT OF TOP-VIEW FISHEYE IMAGES

Benoit Faure, Nathan Odic, Olfa Haggui, and Baptiste Magnier

IMT Mines Alès
École Mines-Télécom

LIRMM

SUMMARY

CONTEXT AND MOTIVATION

YOLO

Detection algorithm

You Only Look Once

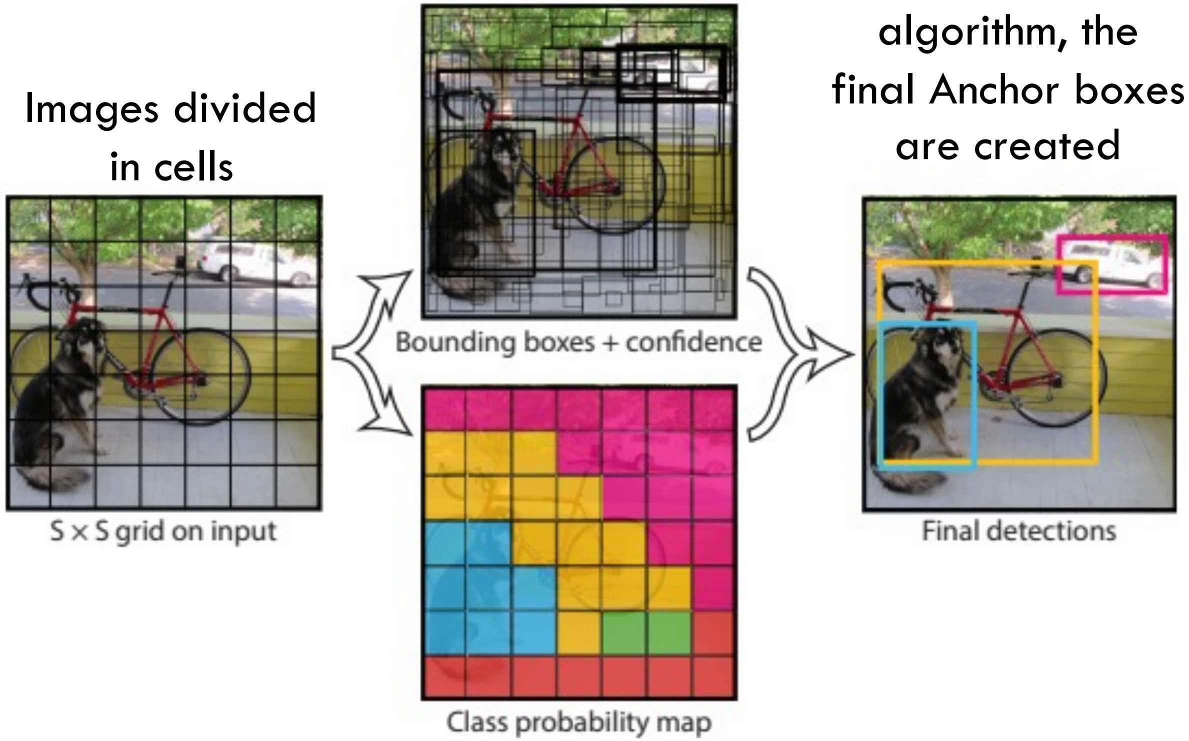
Feature detection + classification in one forward pass

Good real time application

Images divided in cells

Cells create the anchor boxes

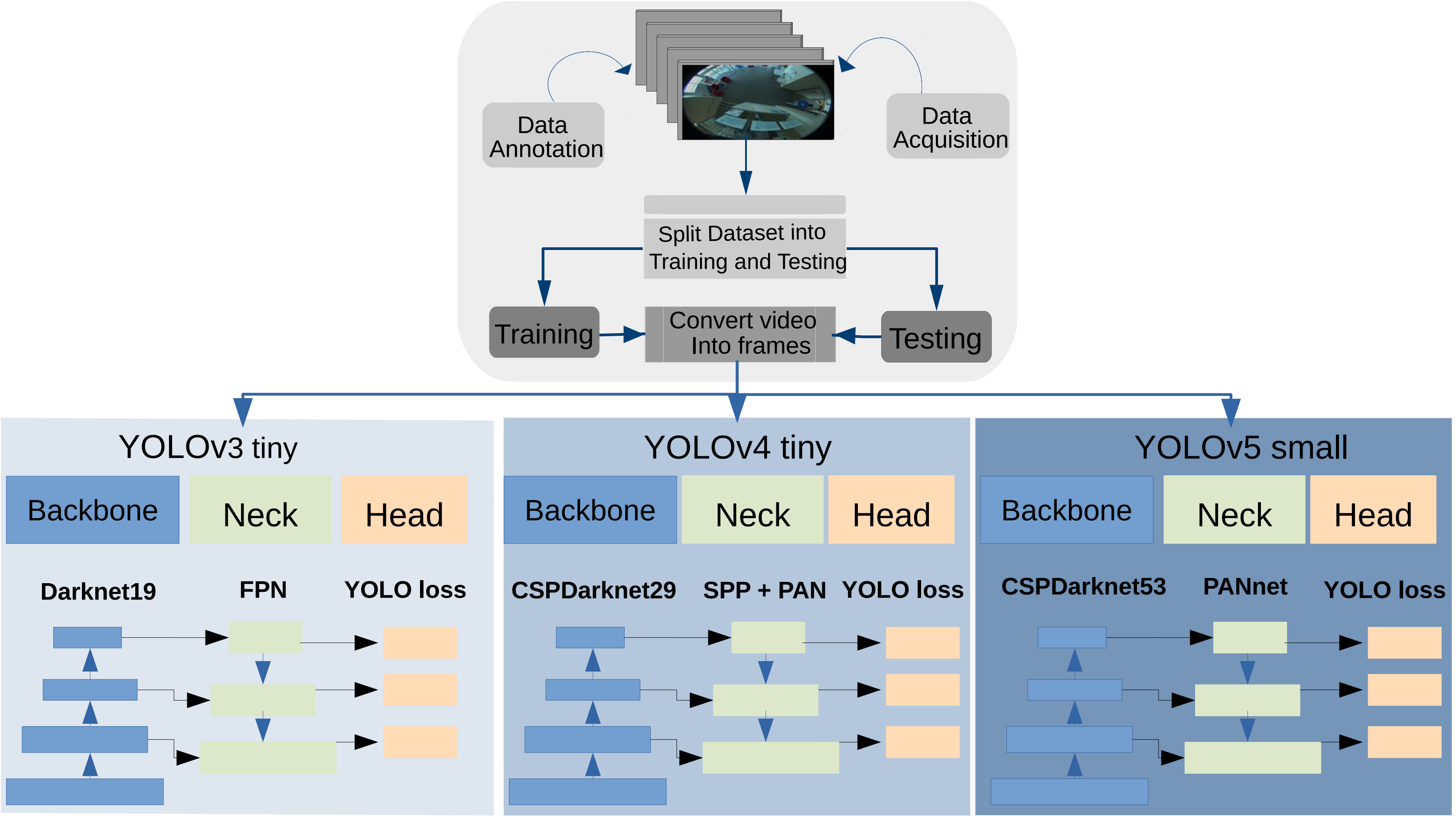
Using the non max suppression algorithm, the final Anchor boxes are created



and locally predict class probability (human, dog…)

Redmon, Joseph, et al. "You Only Look Once: Unified, Real-Time Object Detection." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, https://doi.org/10.1109/cvpr.2016.91.

DIFFERENT SMALL/TINY YOLO VERSIONS

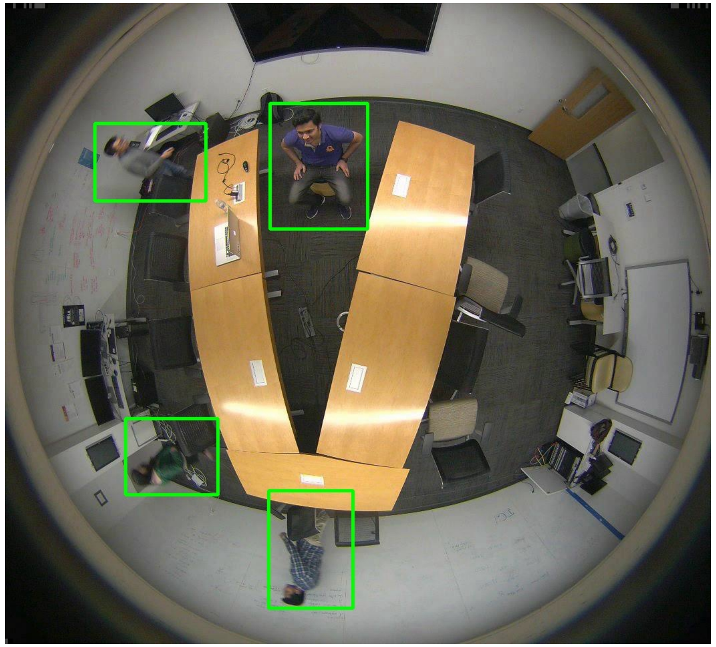


"Performance of Recent tiny/small YOLO Versions in the Context of Top-view Fisheye Images", Benoit FAURE, Nathan ODIC, Olfa HAGGUI and Baptiste MAGNIER – Figure 1

MIRROR DATASET

Use of the Mirror Worlds dataset to pre-train yolo to process fisheye images :

- Good lighting conditions
- Various environments
- Large amount of people
- Bad image sizing
- Errors in annotations



https://www2.icat.vt.edu/mirrorworlds/challenge/index.html

19 videos |821 training images | 204 testing images | Static camera

NEW DATASET

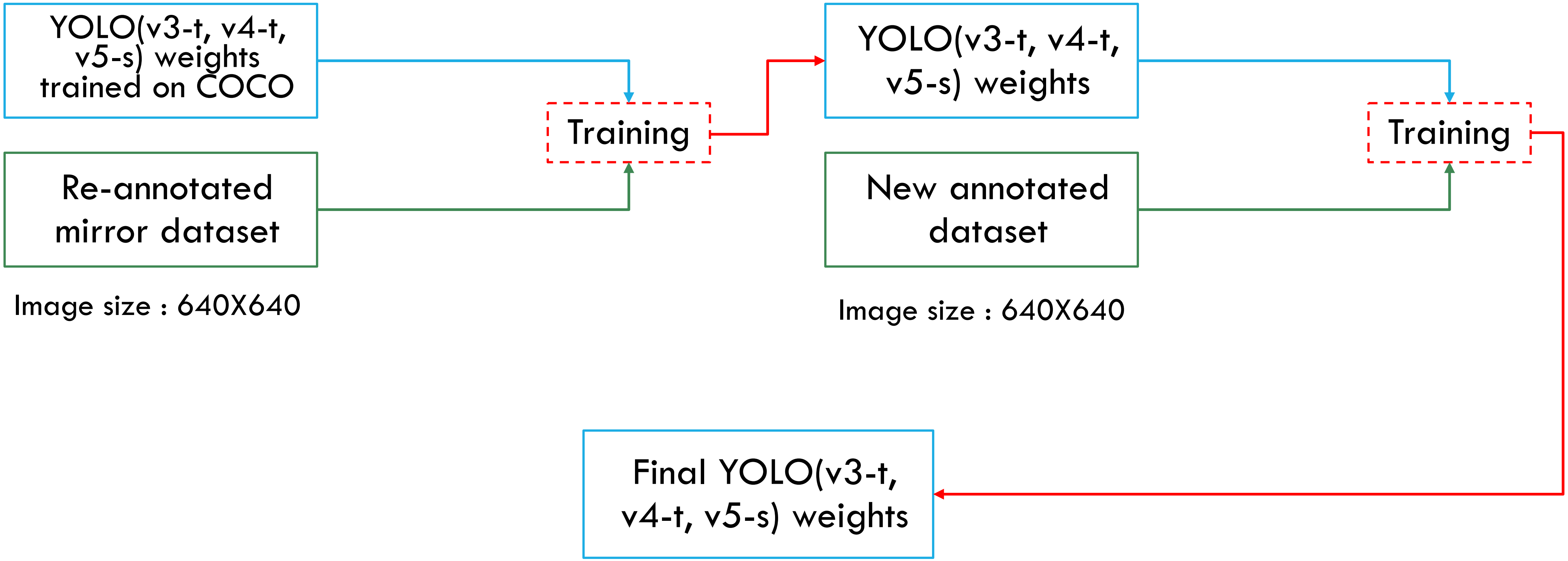
Creation of a dataset adapted to our fisheye camera for better results :

- Numerous cases and scenarios
- Various orientations and occlusions
- Different external conditions
- Number of people that varies
- Static and moving camera

https://github.com/BenoitFaureIMT/CERIS_FishEye

31 videos | 1492 training images | 377+406 testing images | Moving camera

TRAINING THE MODELS



METRICS

$$precision = \frac{total\ amount\ of\ true\ positives}{number\ of\ true\ positives\ +\ number\ of\ false\ positives}$$

How accurate the model is when declaring a detection is positive

$$recall = \frac{total\ amount\ of\ true\ positives}{total\ amount\ of\ (true\ positives + false\ negatives)}$$

It represents the percentage of objects which were not missed

$$IoU = \frac{Intersection\ area\ of\ both\ boxes}{Union\ area\ of\ both\ boxes}$$

Represents how much one box covers another

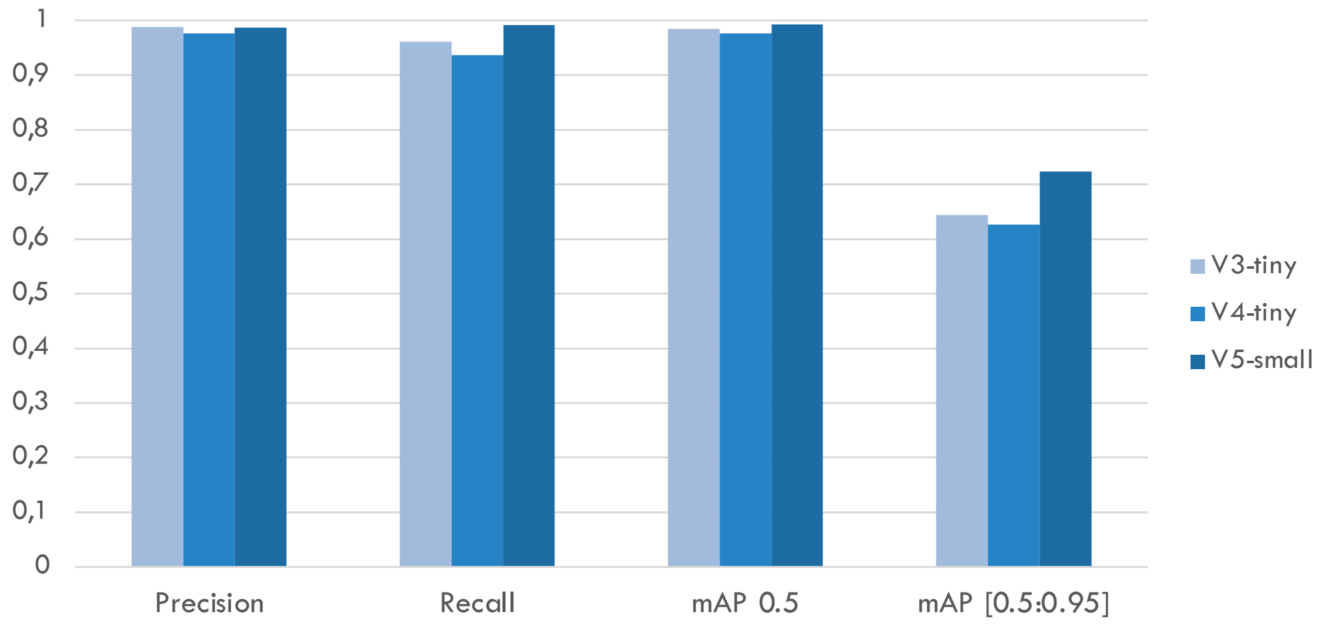
$$mAP_\alpha = \frac{Number\ of\ detections\ where\ the\ coressponding\ IoU \geq \alpha}{Number\ of\ detections\ from\ the\ neural\ network}$$

The percentage of detections which had an IoU with the ground truth greater than α

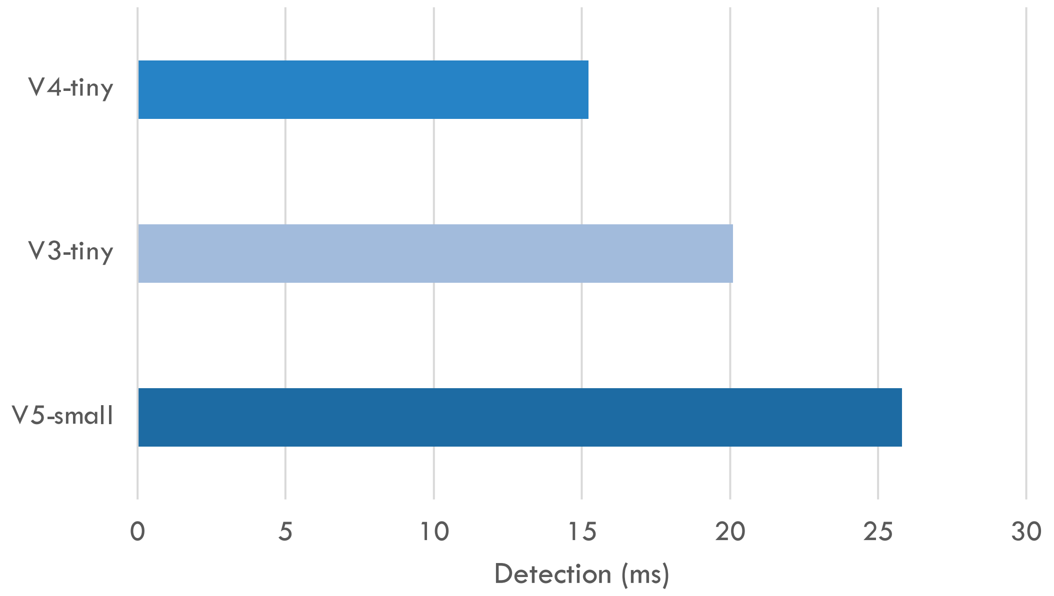
PERFORMANCE IN A FAMILIAR CONTEXT

Testing was done with images pulled from the dataset we created to train the networks.

Number of images	Mirror	Our dataset
Training	821	1492
Testing	204	377

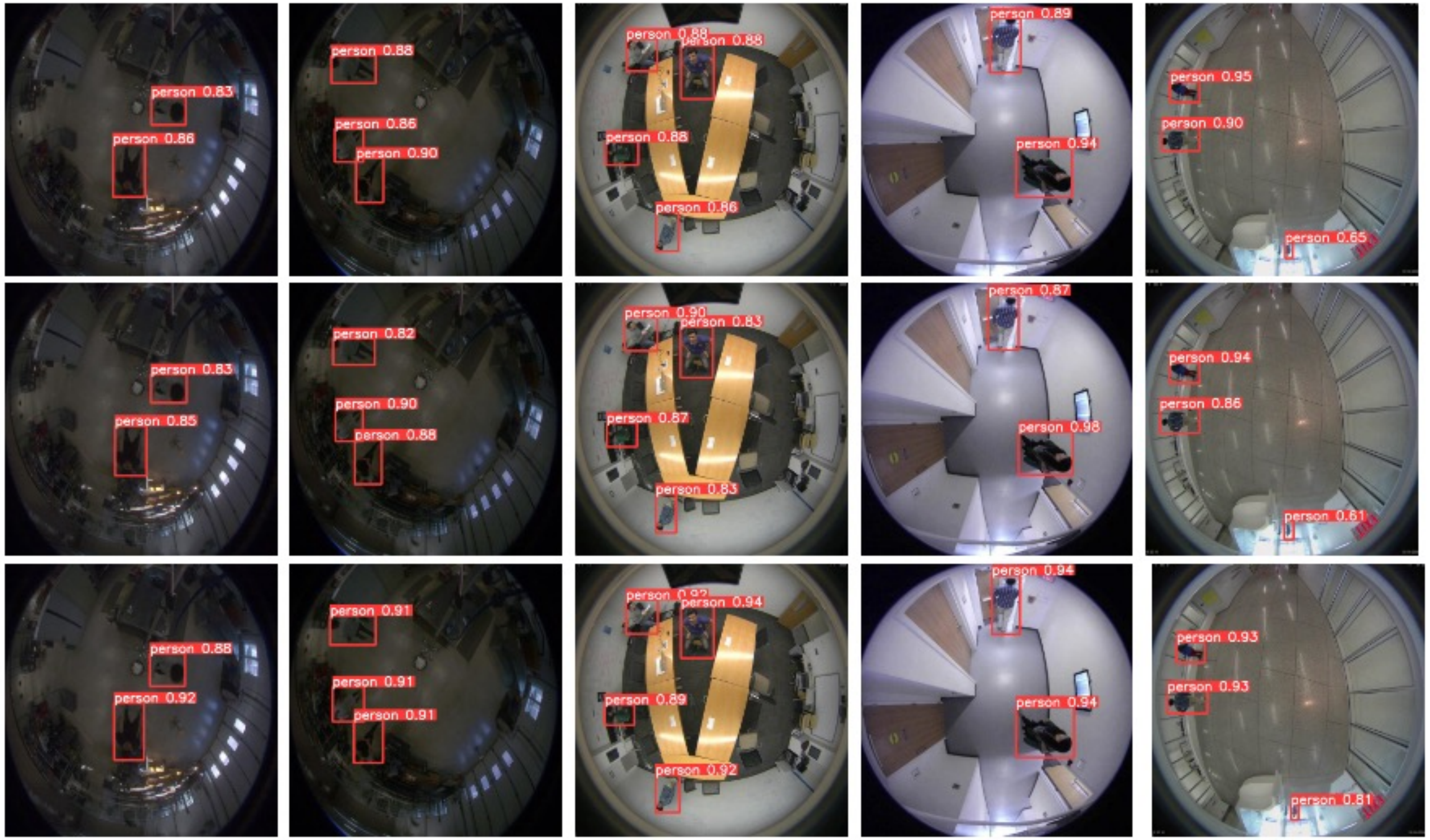


Detection metrics for each version (higher is better)



Detection speeds of each version (lower is better)

PERFORMANCE IN A FAMILIAR CONTEXT



YOLOv3-t

YOLOv4-t

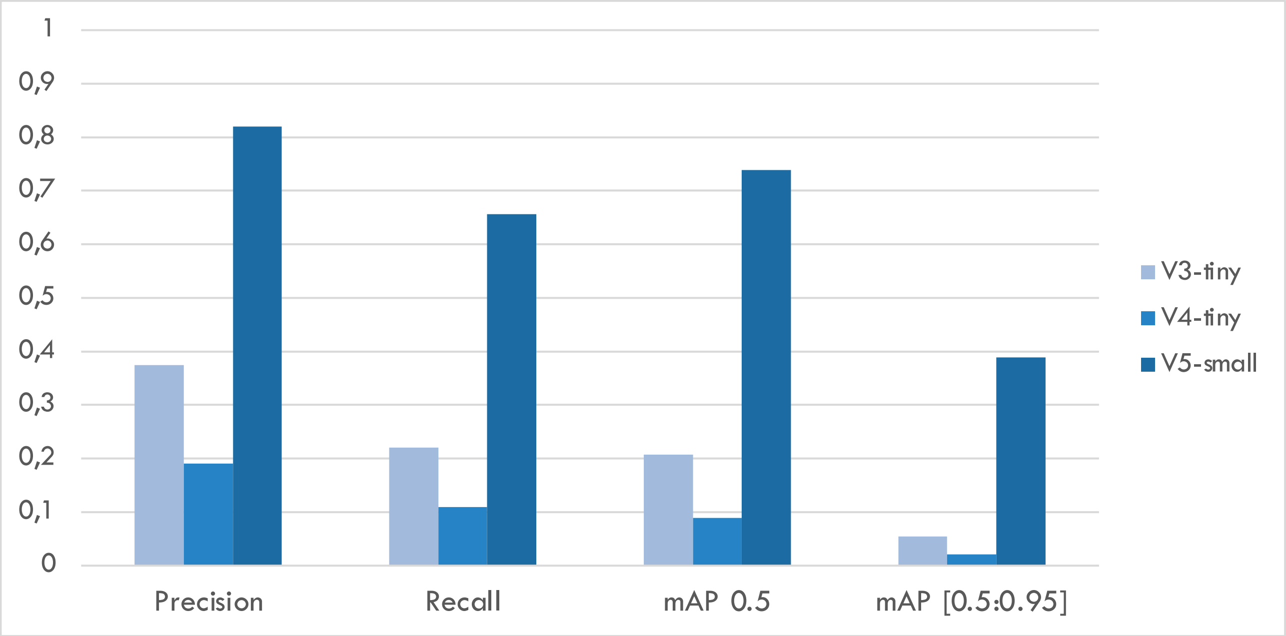
YOLOv5-s

Frame (a) Frame (b) Frame (c) Frame (d) Frame (e)

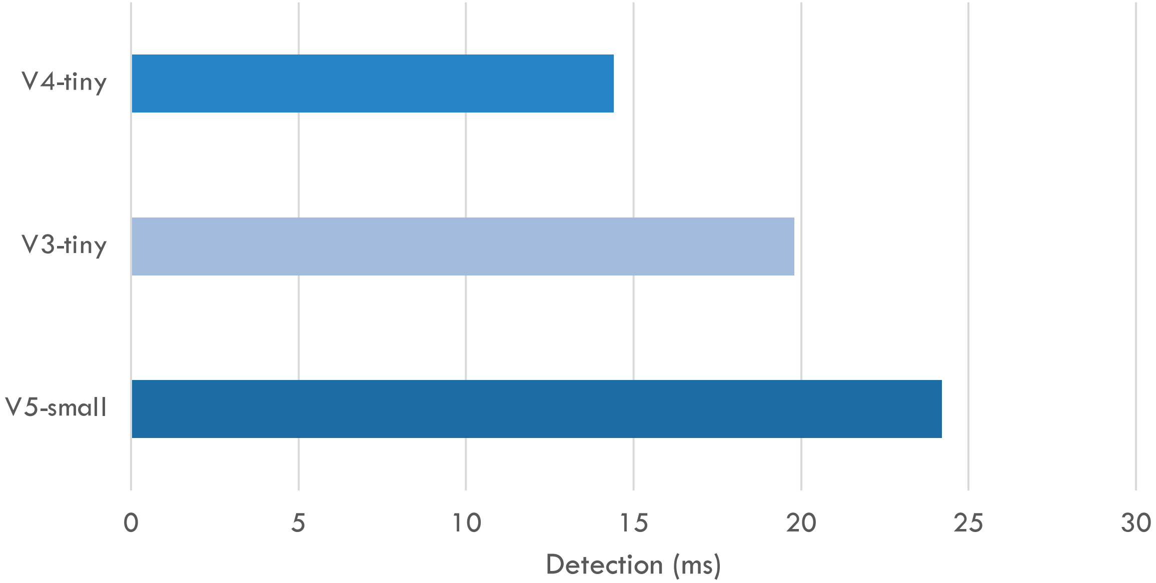
PERFORMANCE IN AN UNFAMILIAR CONTEXT

Second part of our new dataset (406 testing images)

Background and lighting vary greatly from the training images.



Detection metrics for each version (higher is better)



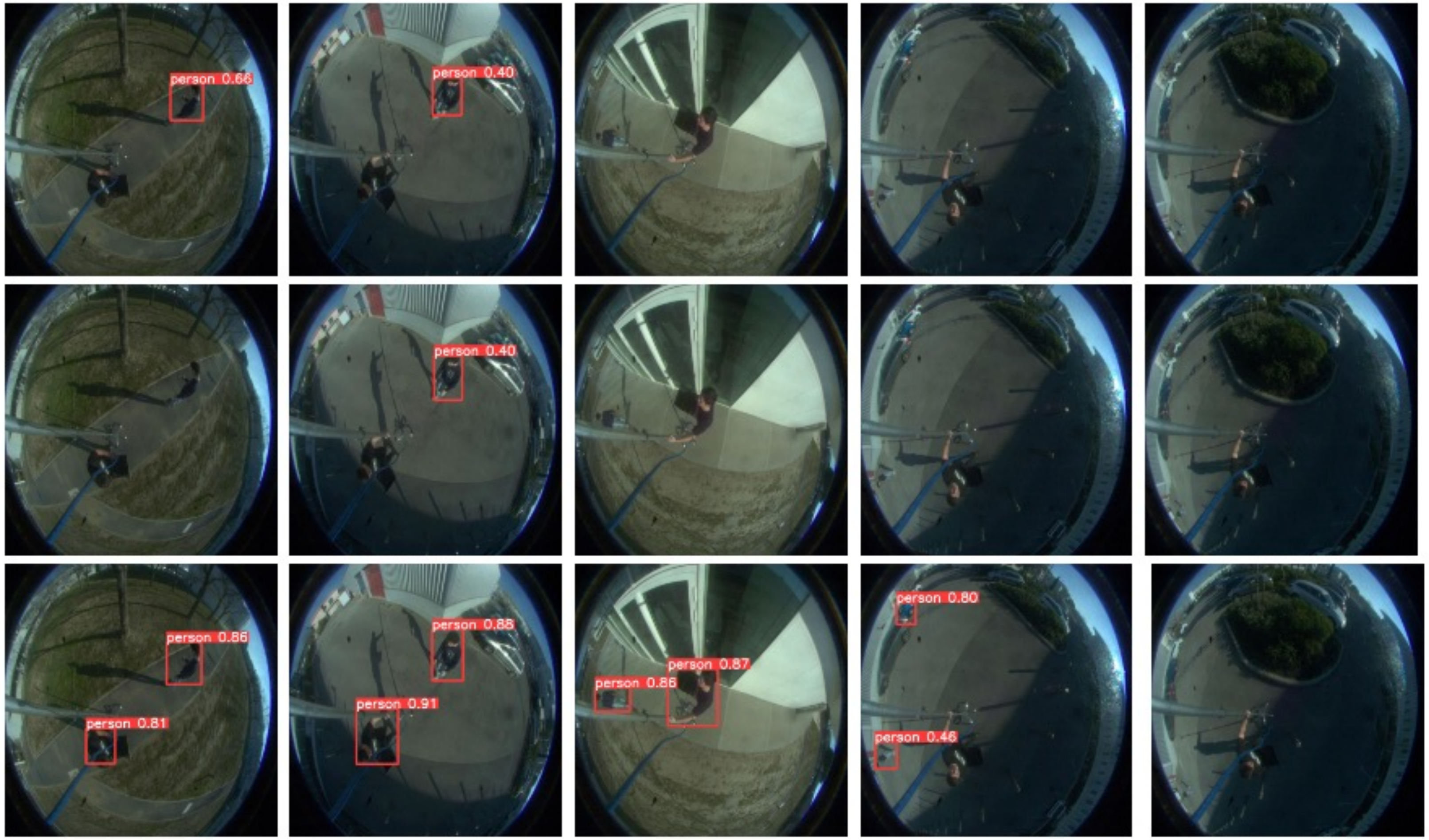
Detection speeds of each version (lower is better)

PERFORMANCE IN AN UNFAMILIAR CONTEXT

YOLOv3-t

YOLOv4-t

YOLOv5-s



Frame (a) Frame (b) Frame (c) Frame (d) Frame (e)

CONCLUSION

Detection speeds : **YOLOv4-t** > YOLOv3-t > YOLOv5-s

Detection quality : YOLOv4-t < YOLOv3-t << **YOLOv5-s**

Possible reason : YOLOv5-s network size was bigger than YOLOv3-t and YOLOv4-t. That is why we have longer detection speeds but could also explain the large increase in detection quality.

<u>Possible improvements</u> :

- Compare larger implementations of each YOLO versions
- Expand the fisheye datasets to improve training quality
- Tracking