http://aimir.isti.cnr.it

Learning Pedestrian Detection from Virtual Worlds

Giuseppe Amato, Luca Ciampi, Fabrizio Falchi, Claudio Gennaro, Nicola Messina

luca.ciampi@isti.cnr.it nicola.messina@isti.cnr.it

GitHub Pages

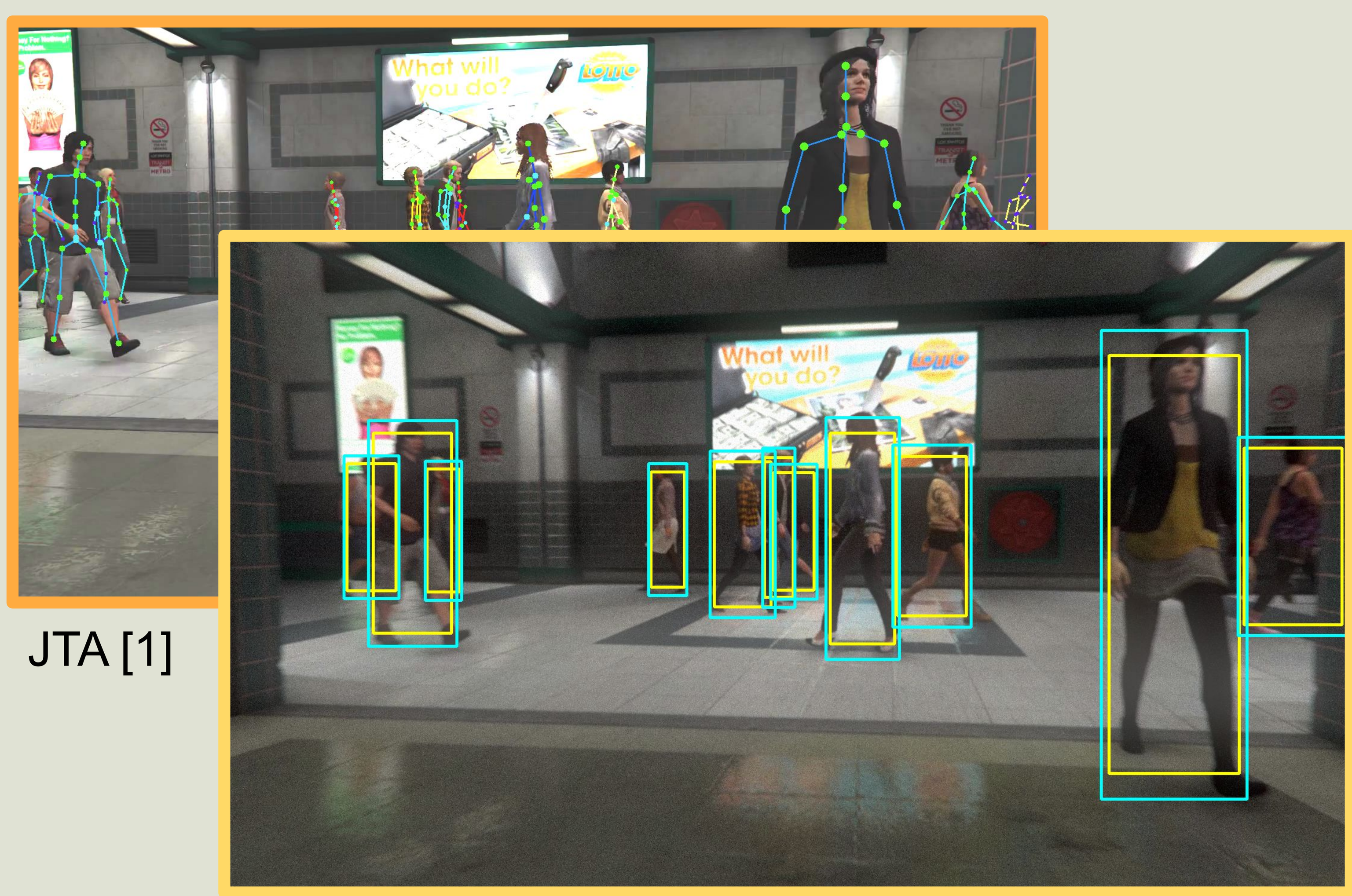
Pedestrian Detectors

- need for huge amount of data
- datasets are usually human-annotated
  - huge manual effort

Virtual Worlds

- images and labels automatically computer-generated
- images should match as much as possible real scenarios
  - **generalization** to multiple real scenarios

ViPeD - Virtual Pedestrian Dataset



JTA [1]

ViPeD

[1] Fabbri, Matteo, et al. "Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World". 2018. http://imagelab.ing.unimore.it/jta

- **data augmentation** to match real-world images
  - bloom effect, radial blur, noise
- precise **bounding box estimation** from keypoints

$$h_m^i = h_s^i + \frac{\alpha}{z^i}$$

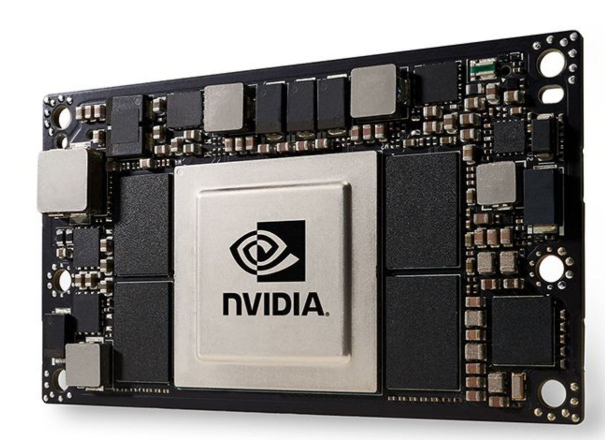
h_m^i and h_s^i are the heights of the *i*-th person and its skeleton respectively

α depends on the camera settings

z^i is the distance of the *i*-th person from camera

Model

- YOLOv3 trained on COCO
  - Low memory consumption
  - Real time on embedded devices
    3-4 FPS on NVidia Jetson TX2

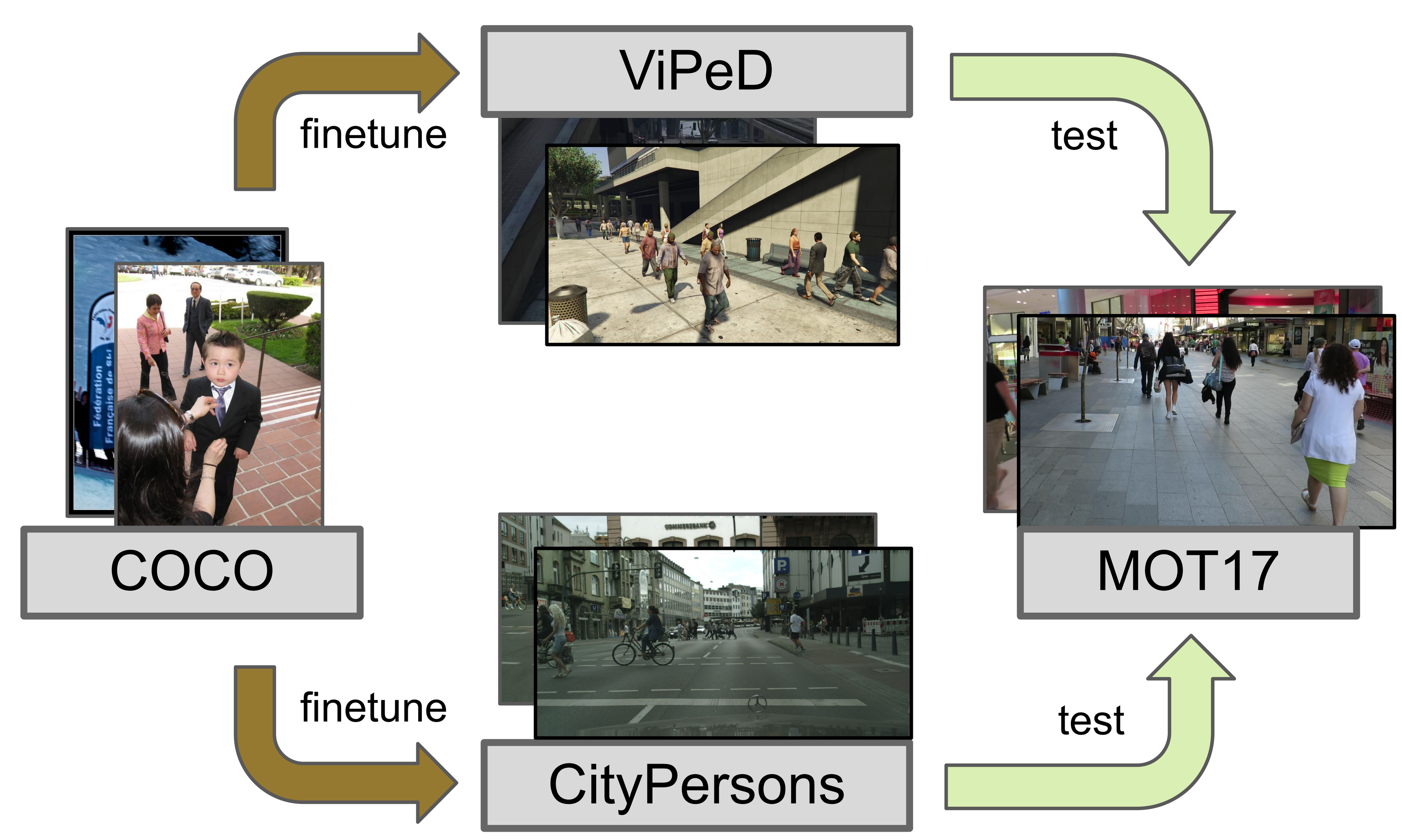


NVidia Jetson TX2

Training and Evaluation

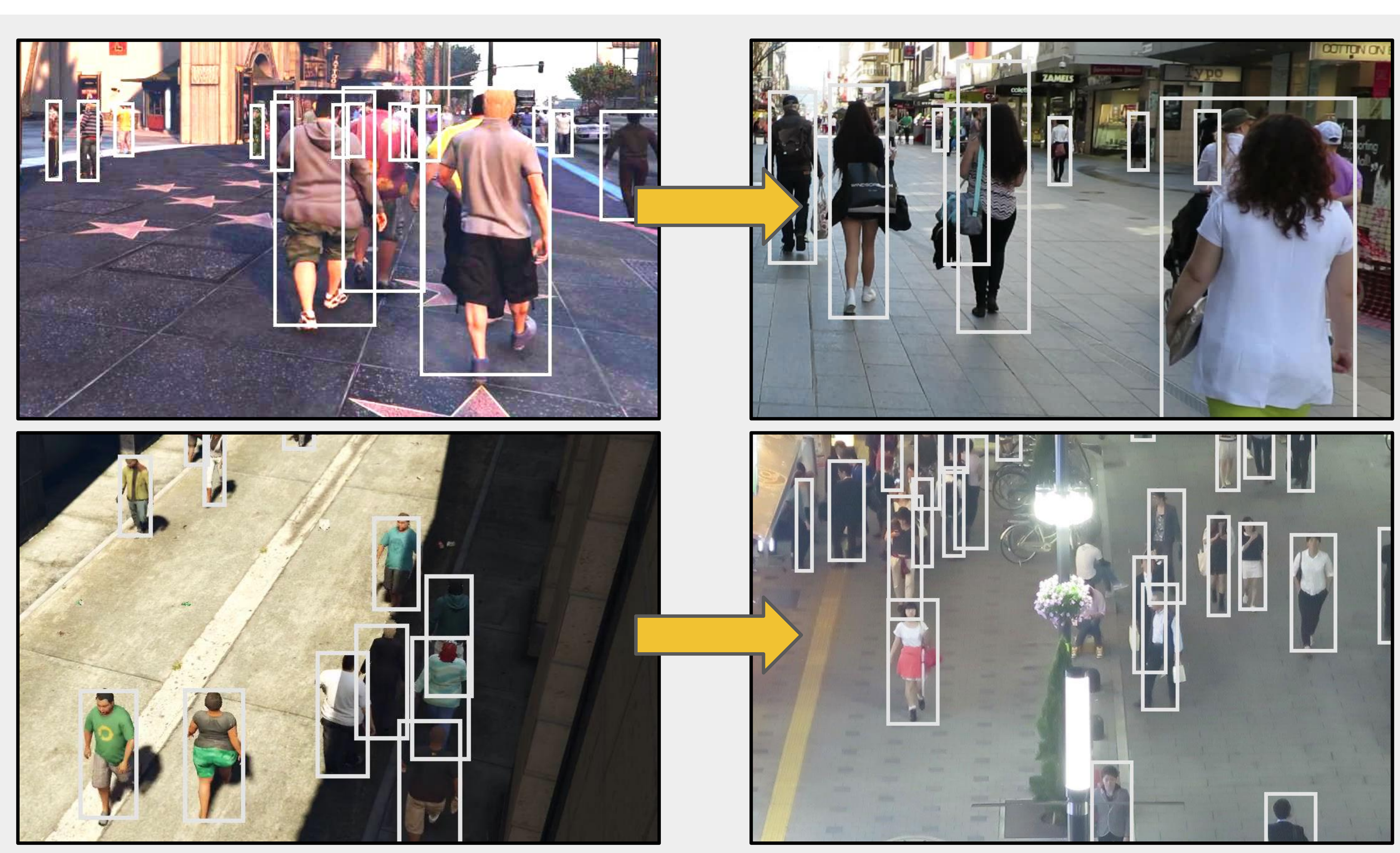
- **Finetune** on ViPeD (virtual), CityPersons (real-world)
- **Test** on MOT17 (real-world)

Baseline: YOLOv3 trained on COCO, tested on MOT17



Results with YOLOv3 on MOT17

Training dataset	MOT AP	COCO AP	Precision	Recall
COCO (baseline)	0.69	0.41	87.4	72.4
CityPersons	0.58	0.37	68.6	60.5
ViPeD - No augm.	0.63	0.40	**91.1**	69.2
ViPeD - Augm.	**0.71**	**0.48**	89.3	**73.9**



Detections on JTA images

Detections on MOT17 images