E^{2}VAD: An Energy-Efficient Video Action Detector

1st Place Winner's Solution to ICCV-LPCV UAV Track

THE UNIVERSITY OF TEXAS AT AUSTIN

TEXAS A&M UNIVERSITY

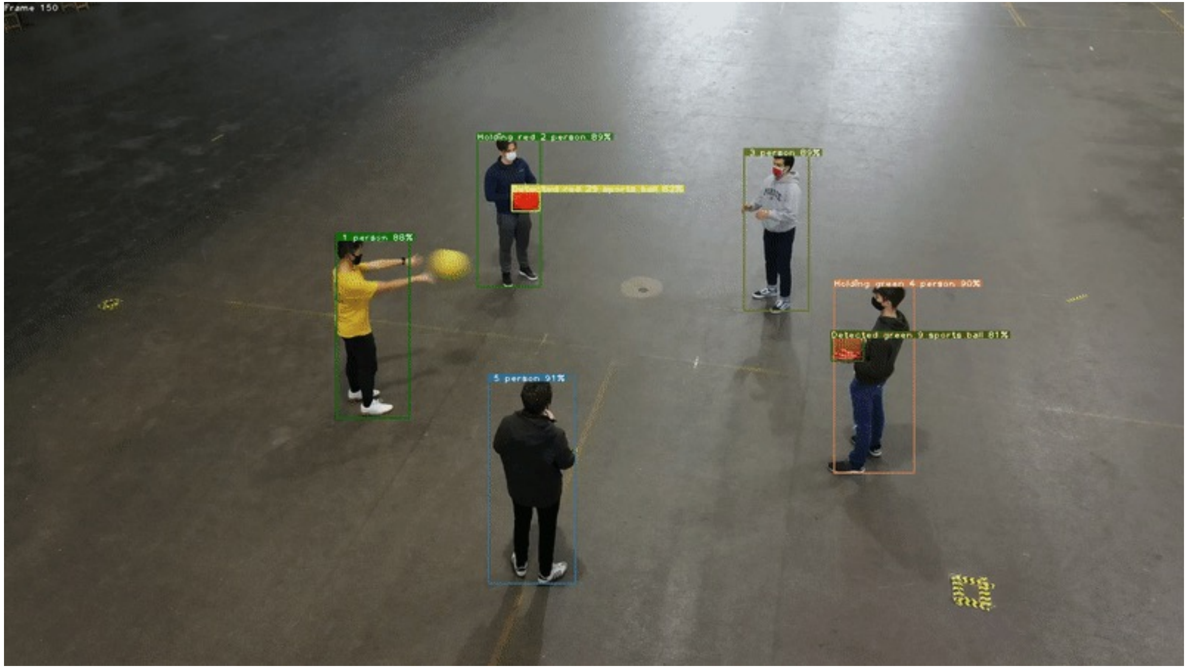
WORMPEX AI RESEARCH LLC

Outline

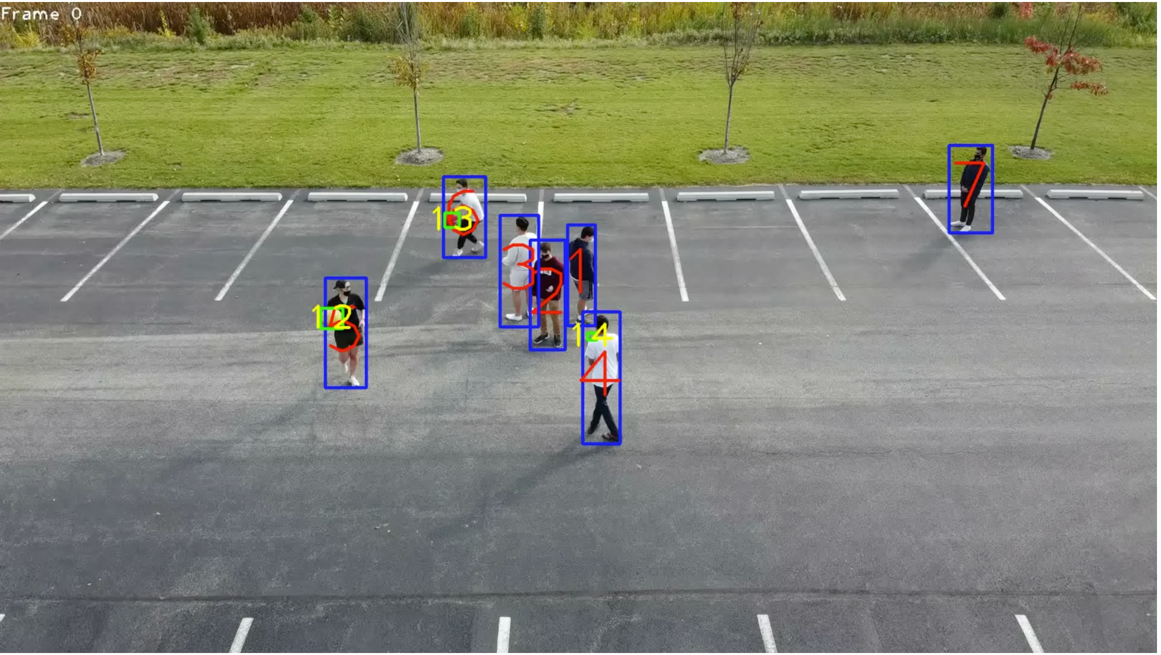
- Curating Our Ball-Person Dataset
- Two Basic Visions Tasks: Detection & Re-Identification
- Core Component: Deep Association
- Detection: Improving Efficiency & Robustness
- ReID: Improving Efficiency & Robustness
- Video Action Detection: Improving Efficiency & Robustness
- Cache-Friendly Pipeline
- Dynamic Inference

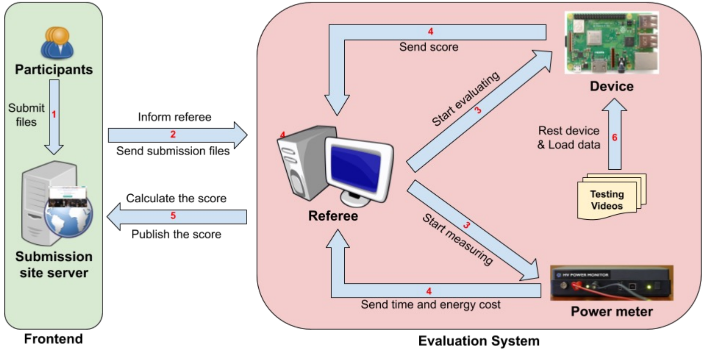
Task Overview: **LPCV Online Track - UAV Video**

- **Competition:** Track multiple moving objects in video captured by an unmanned aerial vehicle (UAV).
- **Hardware:** Raspberry Pi 3B+.
- **Software:** Standard system image + PyTorch, Built from master.


Frame 150
Holding red 2 person 89%
Detected red 29 sports ball 82%
3 person 89%
1 person 88%
Holding green 4 person 90%
Detected green 9 sports ball 81%
5 person 91%

Our Solution-Demo





Team	Score	Rank
VITA	8.473	1
美团 Meituan	7.117	2
ByteDance 字节跳动	6.962	3
ZHEJIANG UNIVERSITY THE UNIVERSITY OF SYDNEY	5.895	4

Unique Challenges

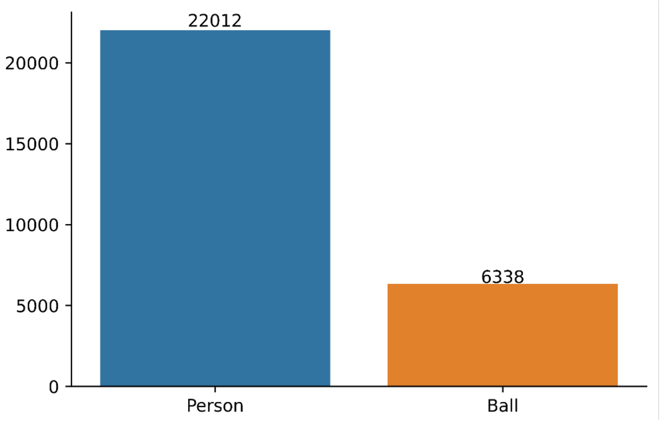
- Lack of Training Data
  - Unlike the ubiquitous “person” object found in benchmarks for detection, segmentation, pose estimation, or tracking tasks, the “ball” object could only be found in COCO’s sports ball category, with large semantic domain gap.
- Robustness
  - The irregular moving pattern of the actors and the drone has made the tracking extremely difficult. The resulting occlusion and varying view angle has brought enormous detection and association errors.
- Efficiency
  - Detecting target persons/balls, extracting their ReID features, and localizing key action spatiotemporally are computationally intensive. Considering the limited computation power and memory capacity on Pi 3B+, running these modules on Pi 3B+ in real-time would be difficult.

COCO Dataset

Expectations

- Person and ball should coexist
- No other categories of objects

Samples of COCO Ball+Person subset

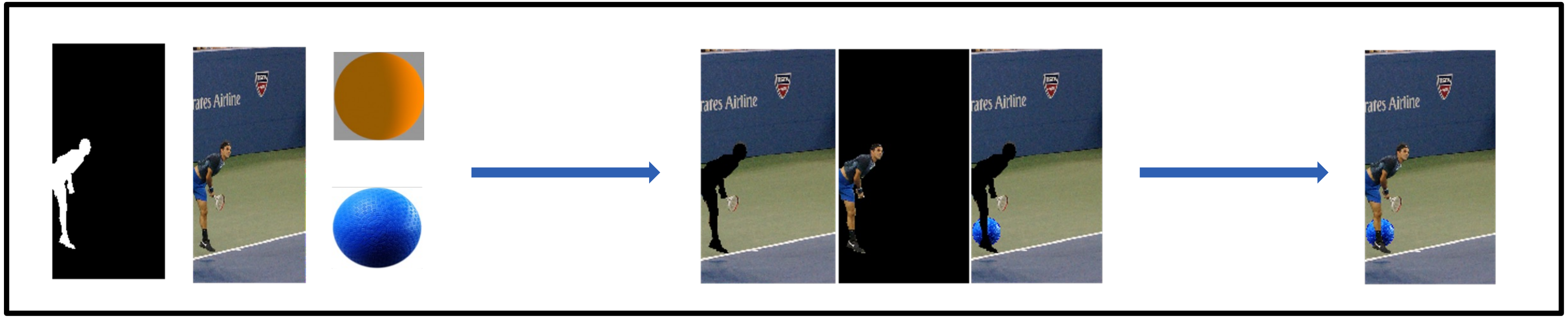


The label distribution of COCO Ball+Person subset

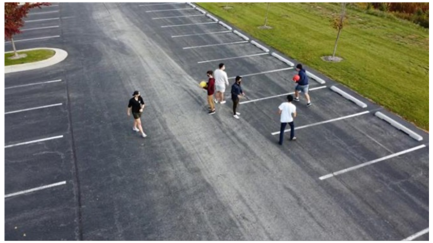
Problems

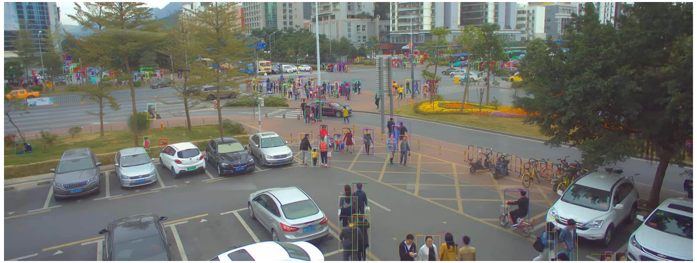
- Imbalanced classes (person:ball=3:1)
- Few occluded samples
- Large domain gap (especially ball)

Attempt #1: Augment COCO by Occlusion-Aware Copy-Paste



Attempt #2: Including Pedestrian-Related Dataset





Attempt #3: Harmonization-Aware Image Composition

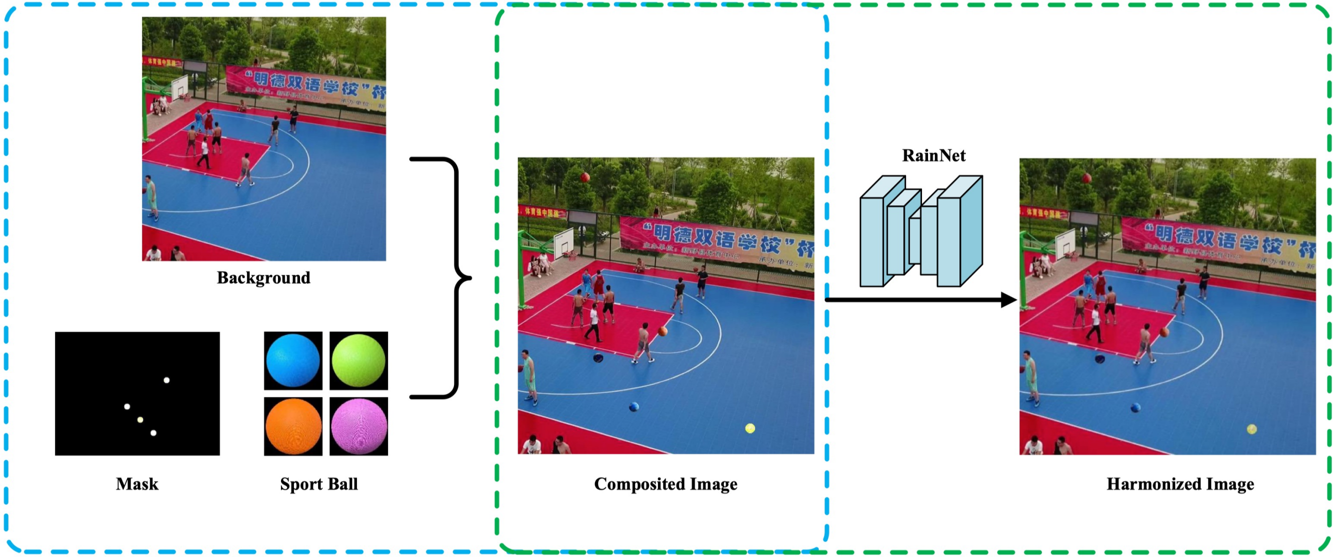
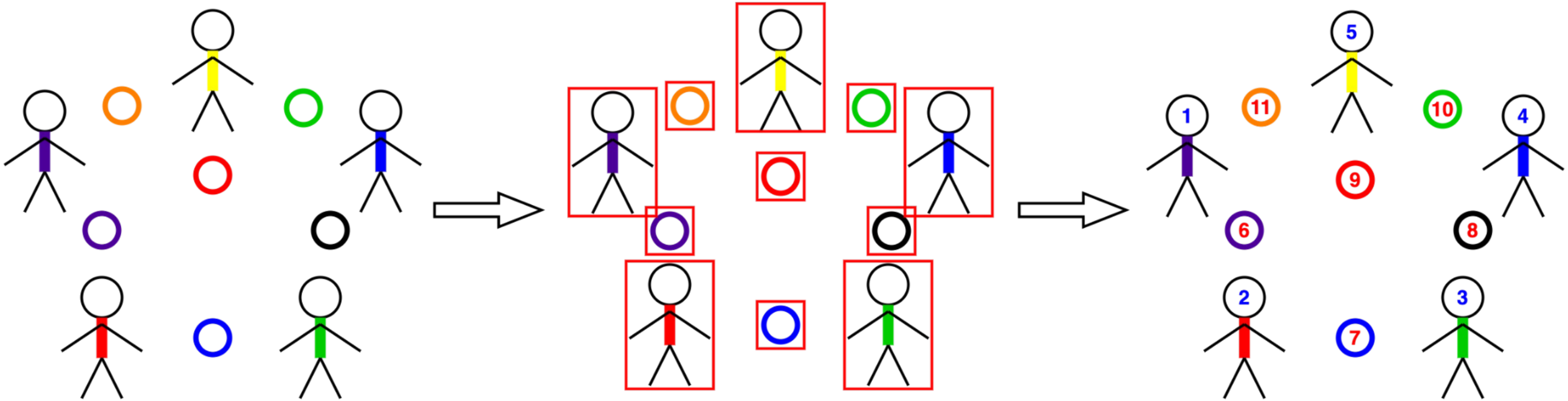


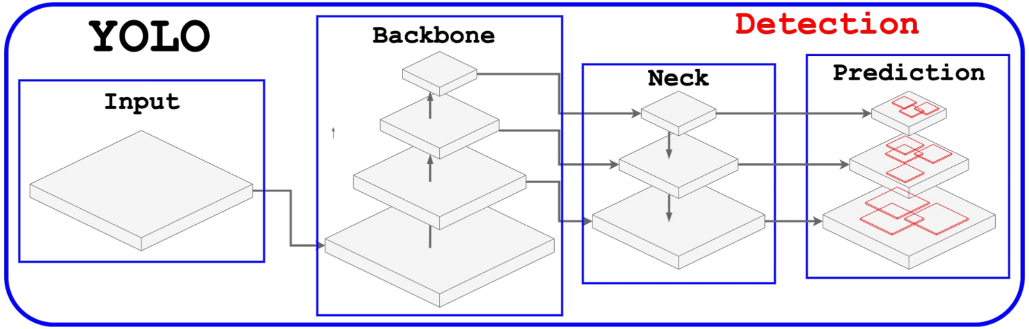
Image Composition

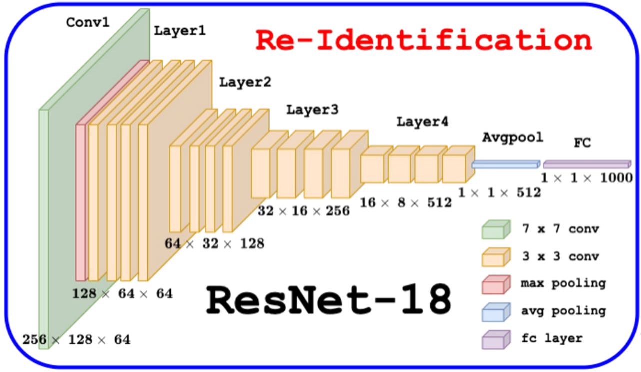
Image Harmonization

- Balls are composited into background images guided by randomly-placed masks.
- The composited images are harmonized by RainNet with region-aware instance normalization.

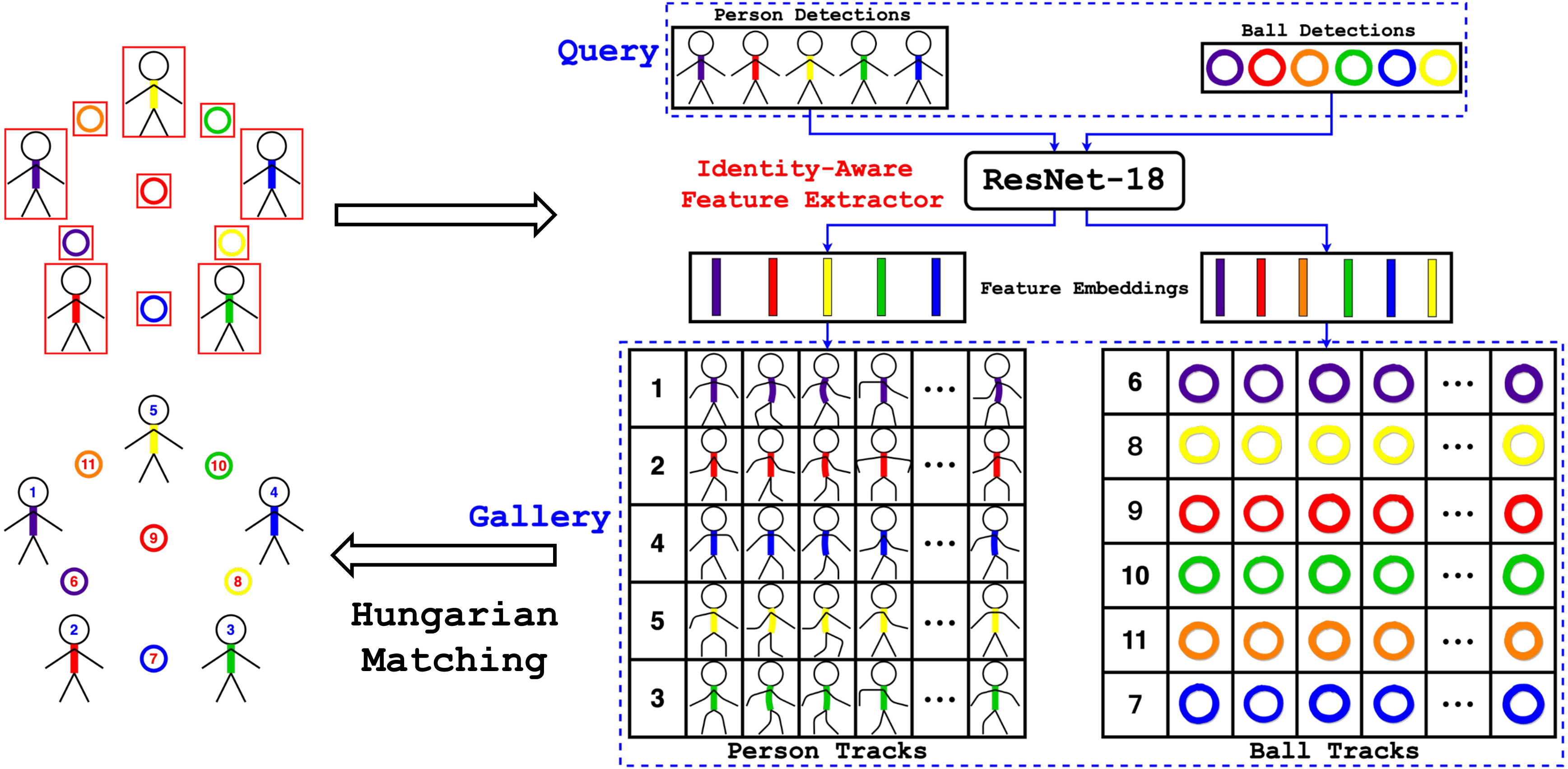
Two Basic Vision Tasks: Detection & Re-Identification







The Core Component: Deep Association



Query

Person Detections

Ball Detections

Identity-Aware Feature Extractor

ResNet-18

Feature Embeddings

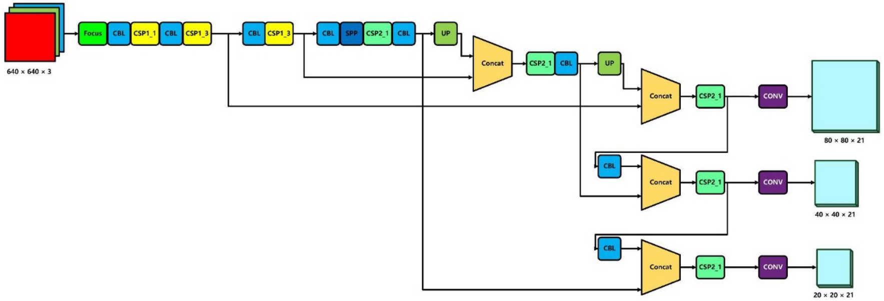
Gallery

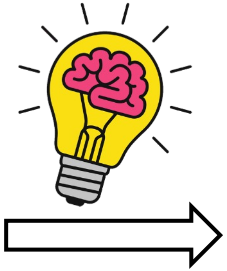
Person Tracks

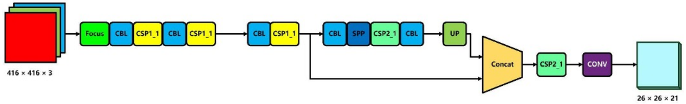
Ball Tracks

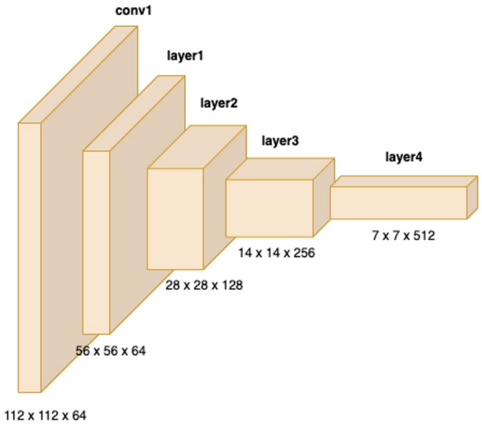
Hungarian Matching

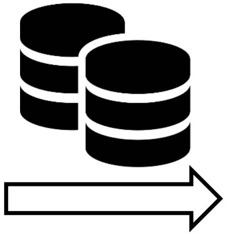
Two Ways of Neural Network Pruning: Knowledge & Data Driven

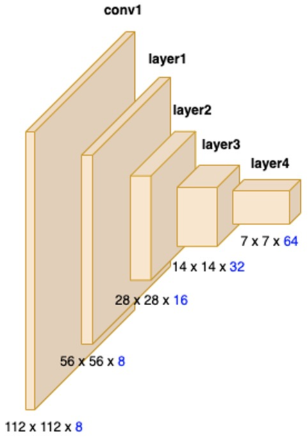




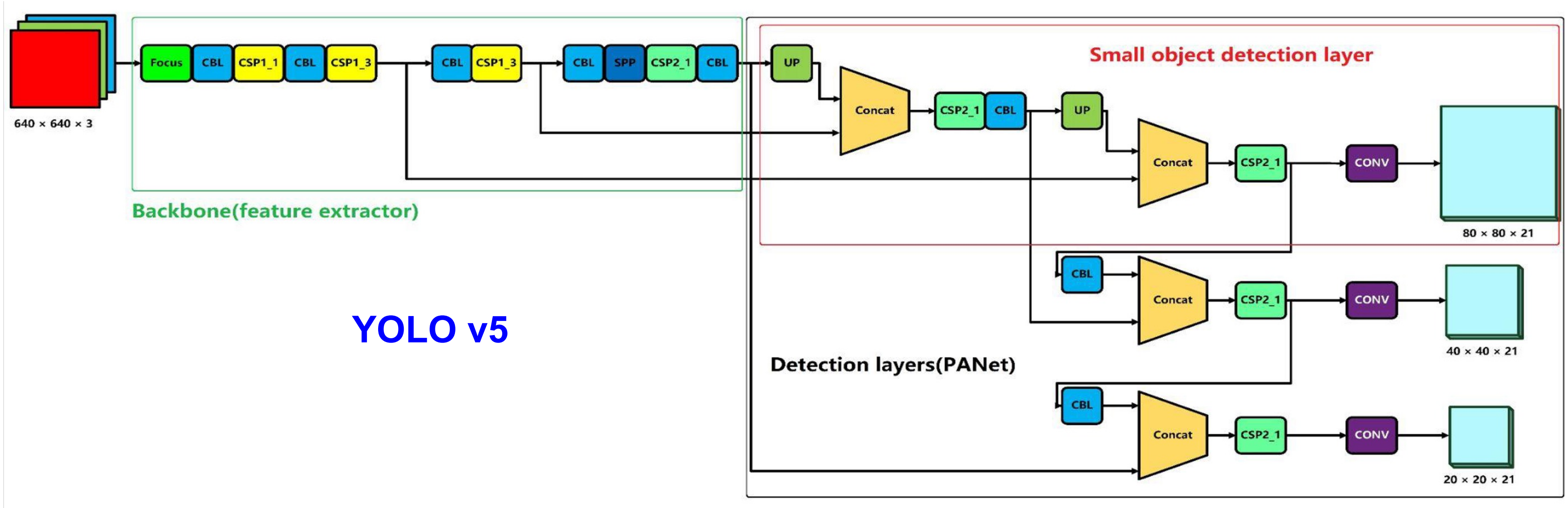




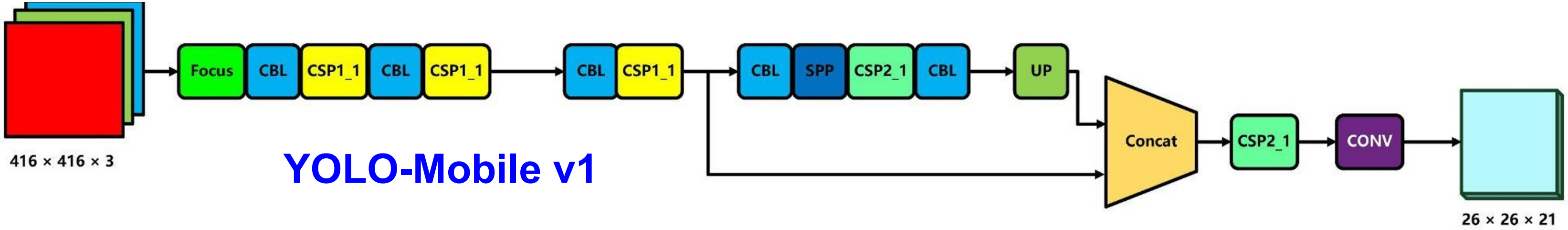




Proposed YOLO-Mobile v1

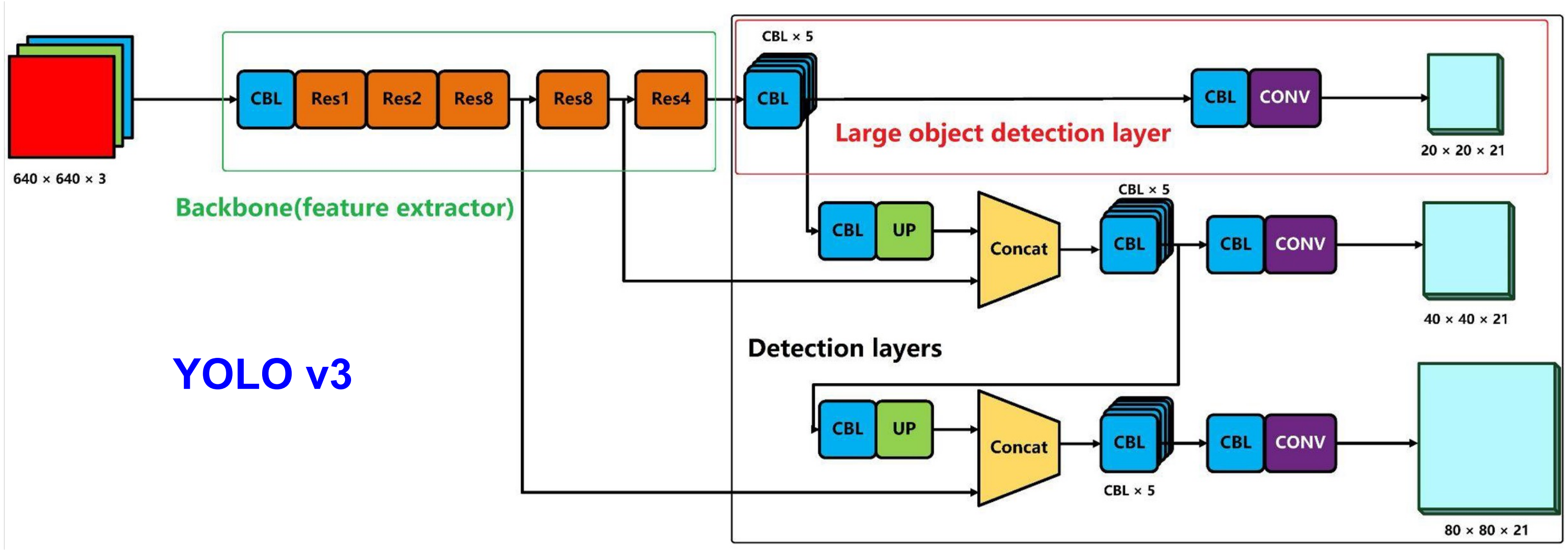


YOLO v5

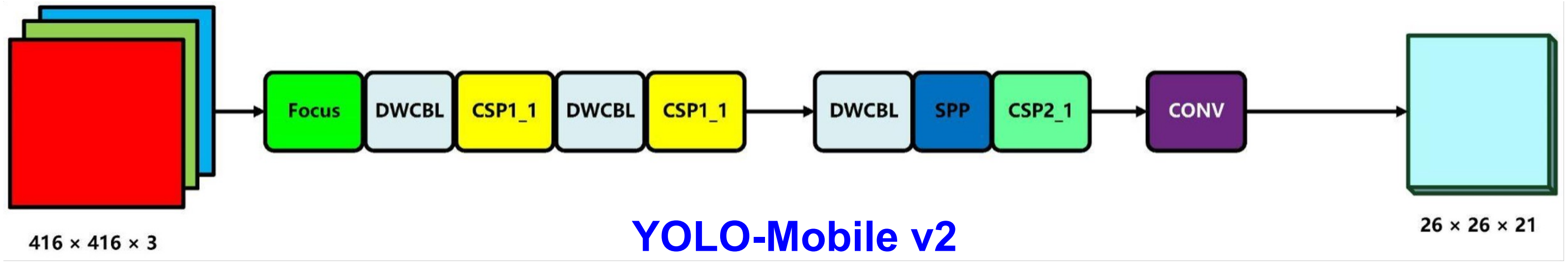


YOLO-Mobile v1

Proposed YOLO-Mobile v2

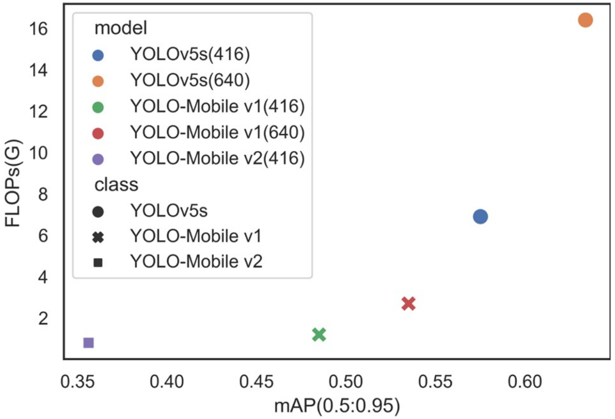


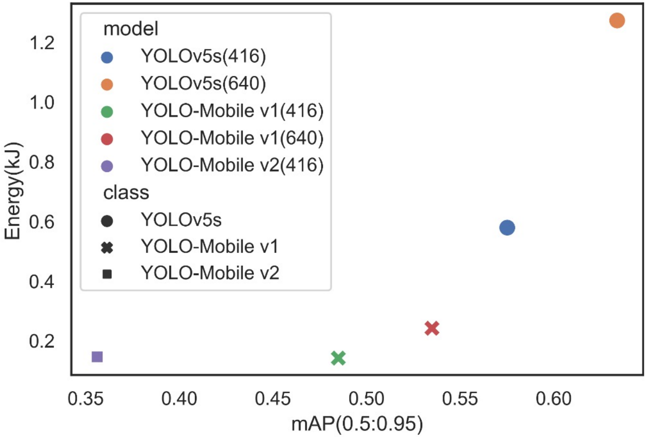
YOLO v3

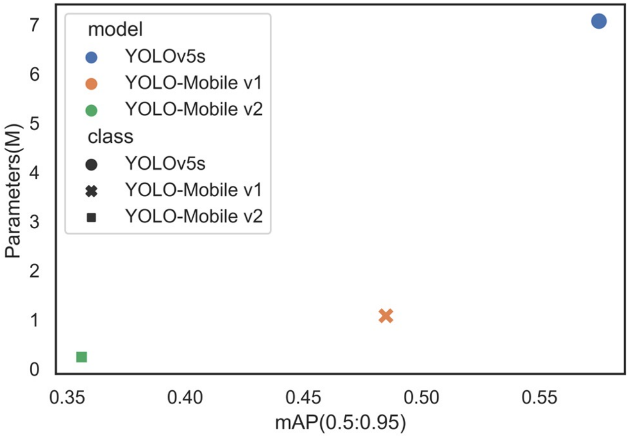


YOLO-Mobile v2

FLOPS/Energy/Parameters vs mAP





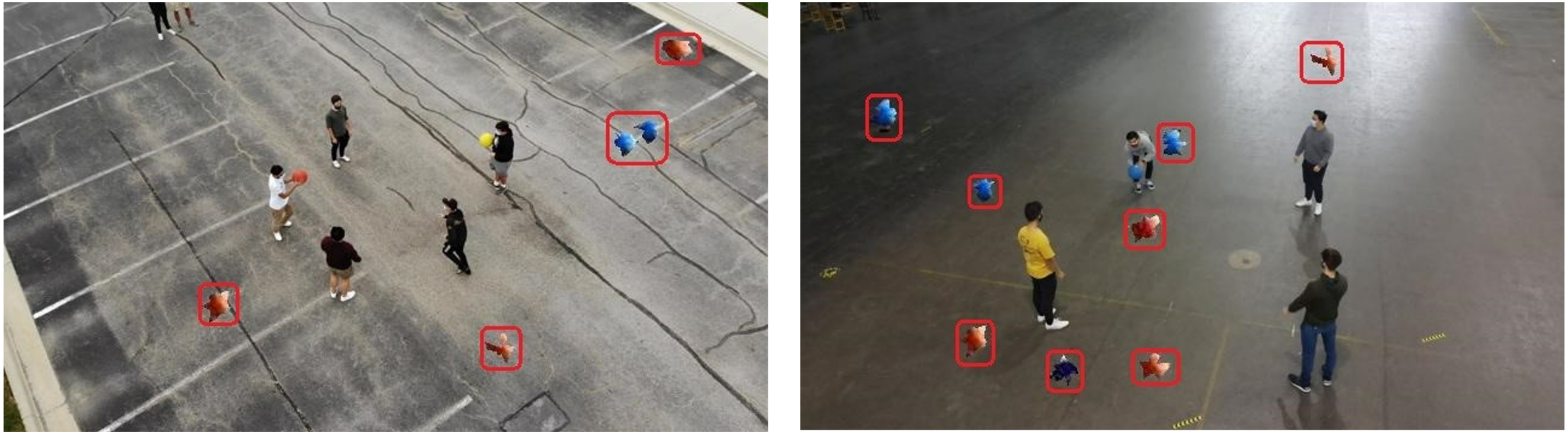


- YOLO-Mobile v1(416) reduces **16×** FLOPs at the cost of **24%** mAP loss compared to YOLO v5-small(640)

- YOLO-Mobile v1(416) reduces **10×** energy compared to YOLO v5-small(640)
- Depthwise separable convolution is **NOT efficient** energy on Raspberry Pi

- YOLO-Mobile v2 reduces **29×** parameters compared to YOLO v5-small

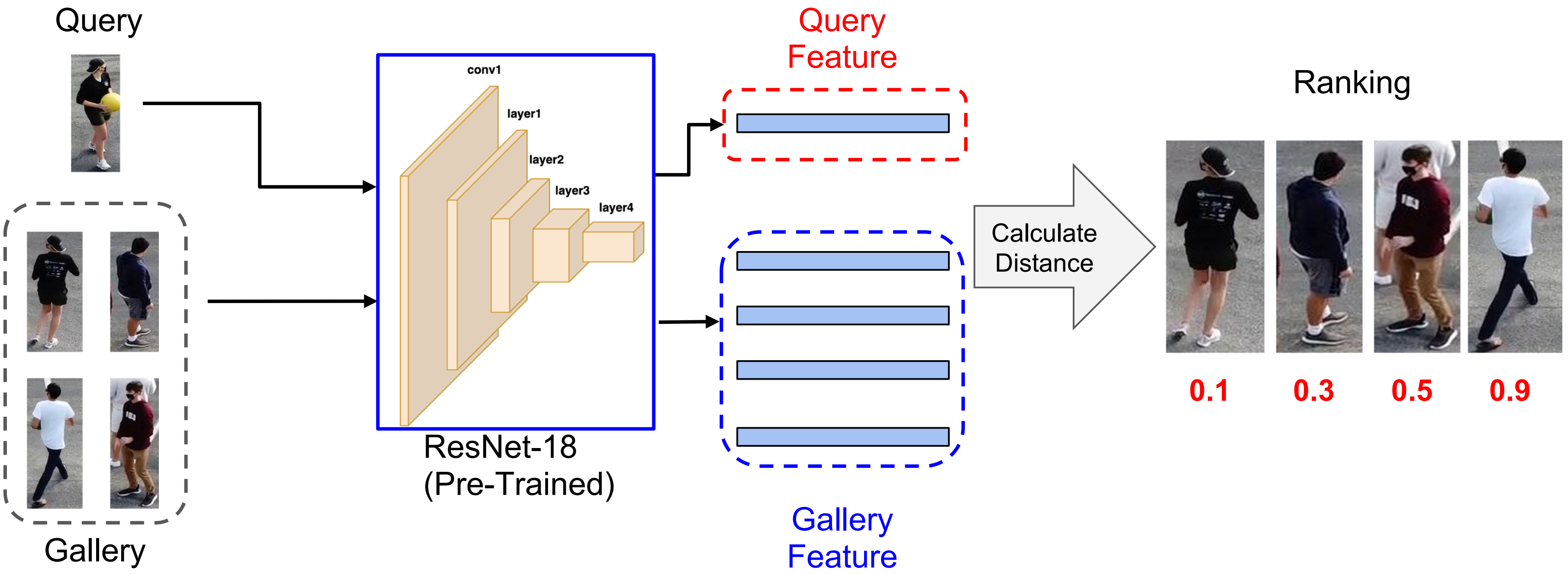
Detection Robustness Challenge: Texture-Shape Bias



Texture-Shape Debiased Training: adding negative samples for robust ball detection

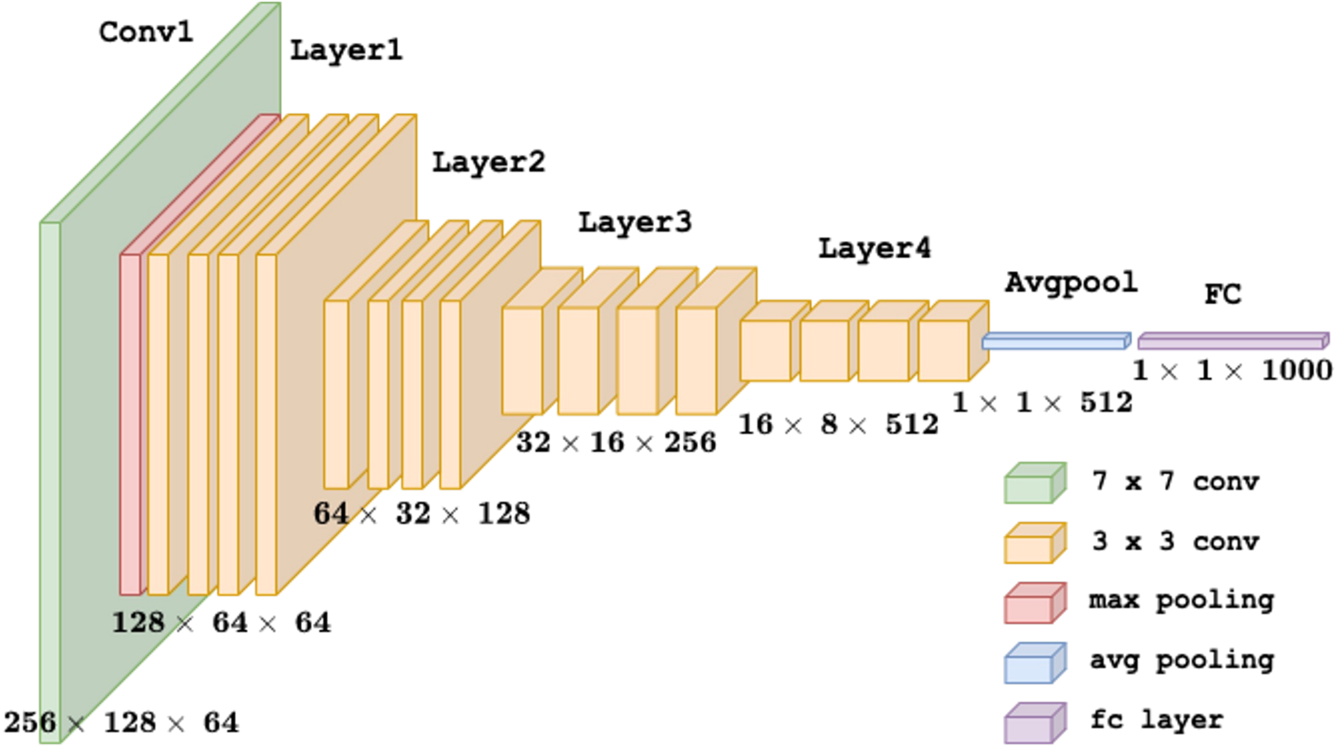
- Model trained with existing dataset is strongly biased towards texture for ball detection.
- Add random-shaped patches with homogeneous texture, which serve as negative samples for ball detection.

Re-Identification (ReID)



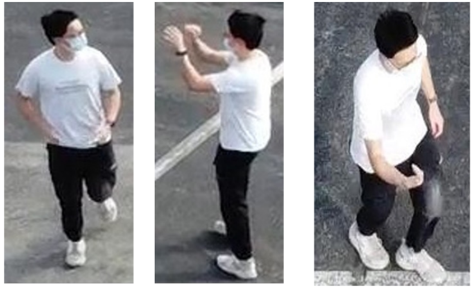
Samples | **Feature Extraction** | **Nearest Neighbor Matching**

ResNet-18 for Re-ID

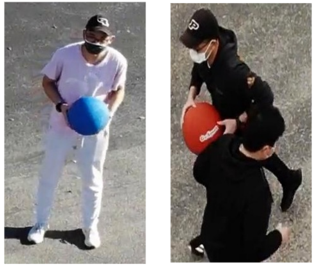


Re-ID Challenges

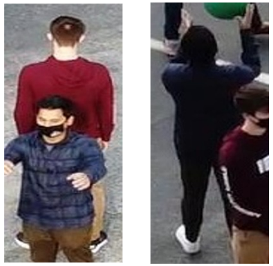
- Different camera views
- Occlusion
  - Ball-person occlusion
  - Person-Person occlusion



(a) Appearance under different camera views



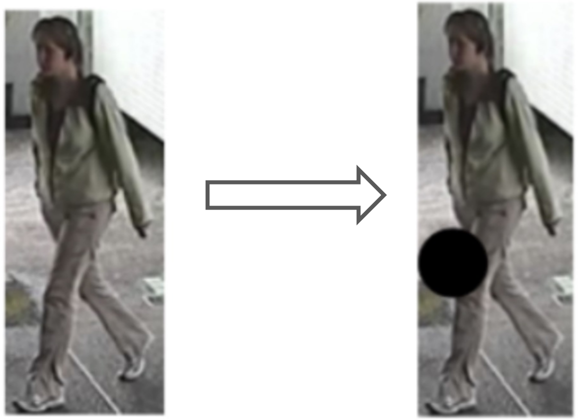
(b) Ball-Person occlusion



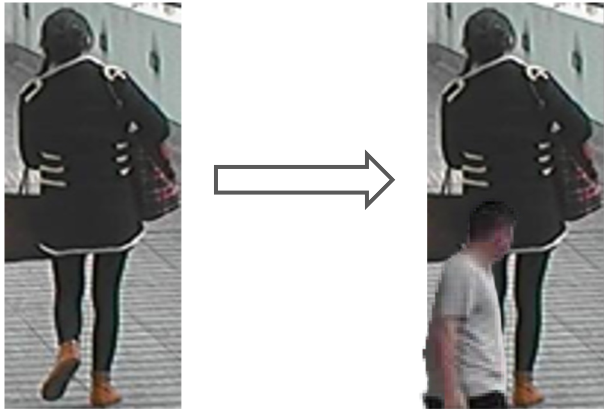
(c) Person-Person occlusion

Re-ID Robustness Challenge I: Occlusion

- Solution: Occlusion-aware data augmentation



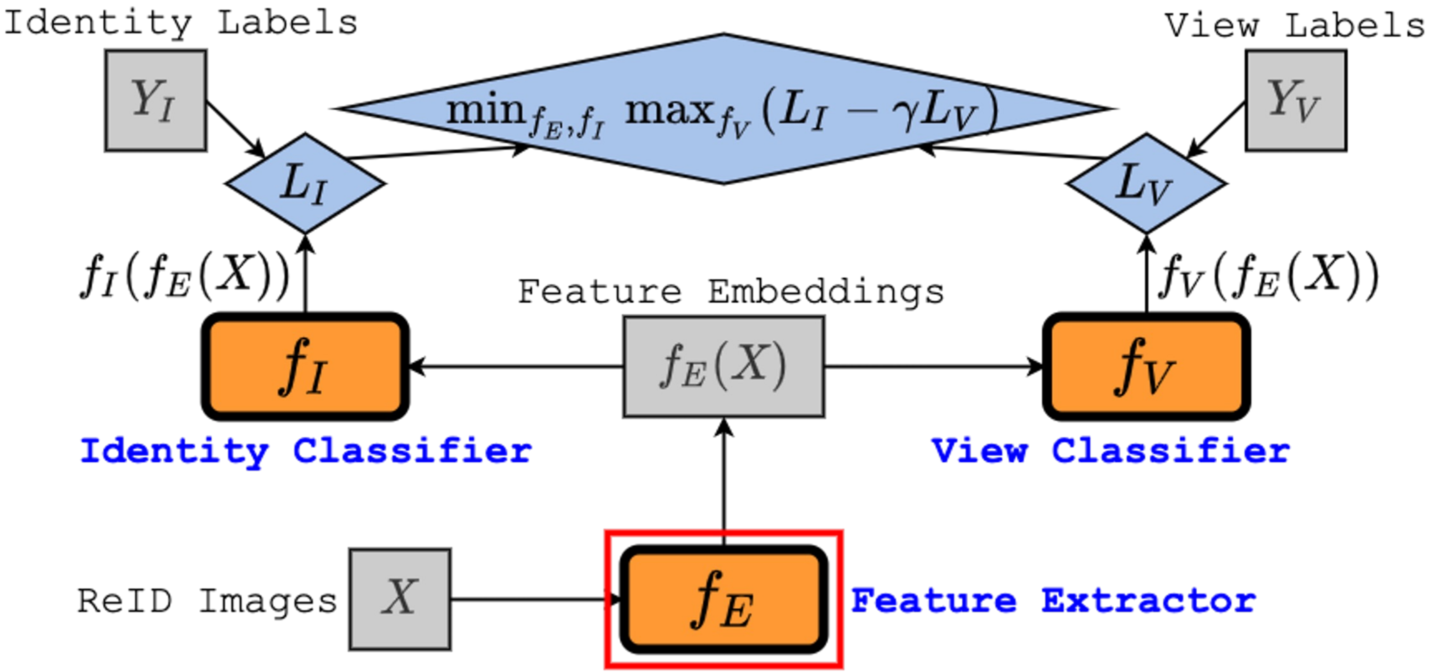
(a) Occlusion-aware data augmentation for ball.



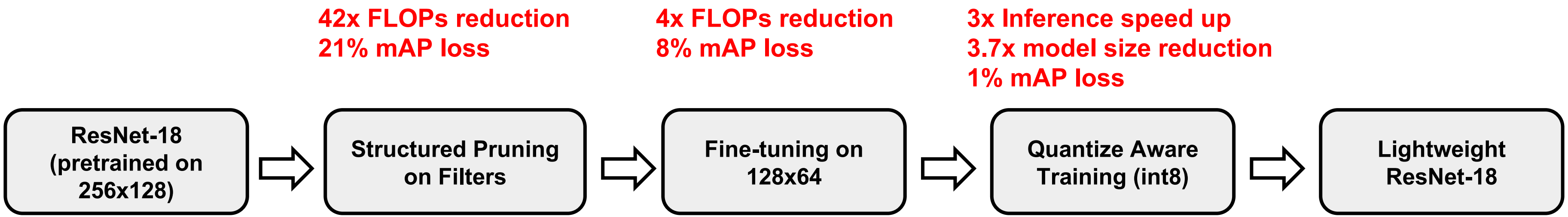
(b) Occlusion-aware data augmentation for person.

Re-ID Robustness Challenge II: Various Camera View

- Solution: Domain-Invariant Feature Learning: a Minimax Game



Re-ID Model Compression & Acceleration

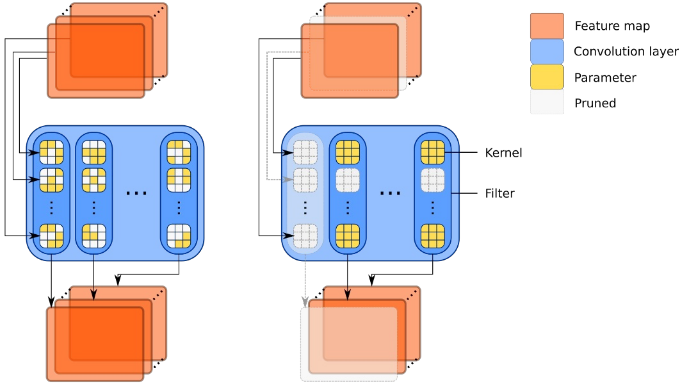


42x FLOPs reduction
21% mAP loss

4x FLOPs reduction
8% mAP loss

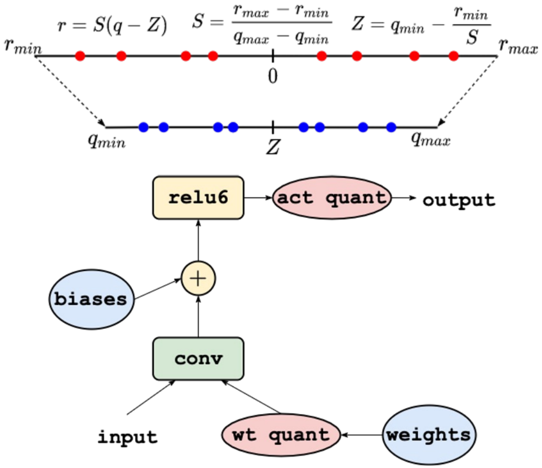
3x Inference speed up
3.7x model size reduction
1% mAP loss

ResNet-18 (pretrained on 256x128)
Structured Pruning on Filters
Fine-tuning on 128x64
Quantize Aware Training (int8)
Lightweight ResNet-18



"Unstructured" : weight pruning

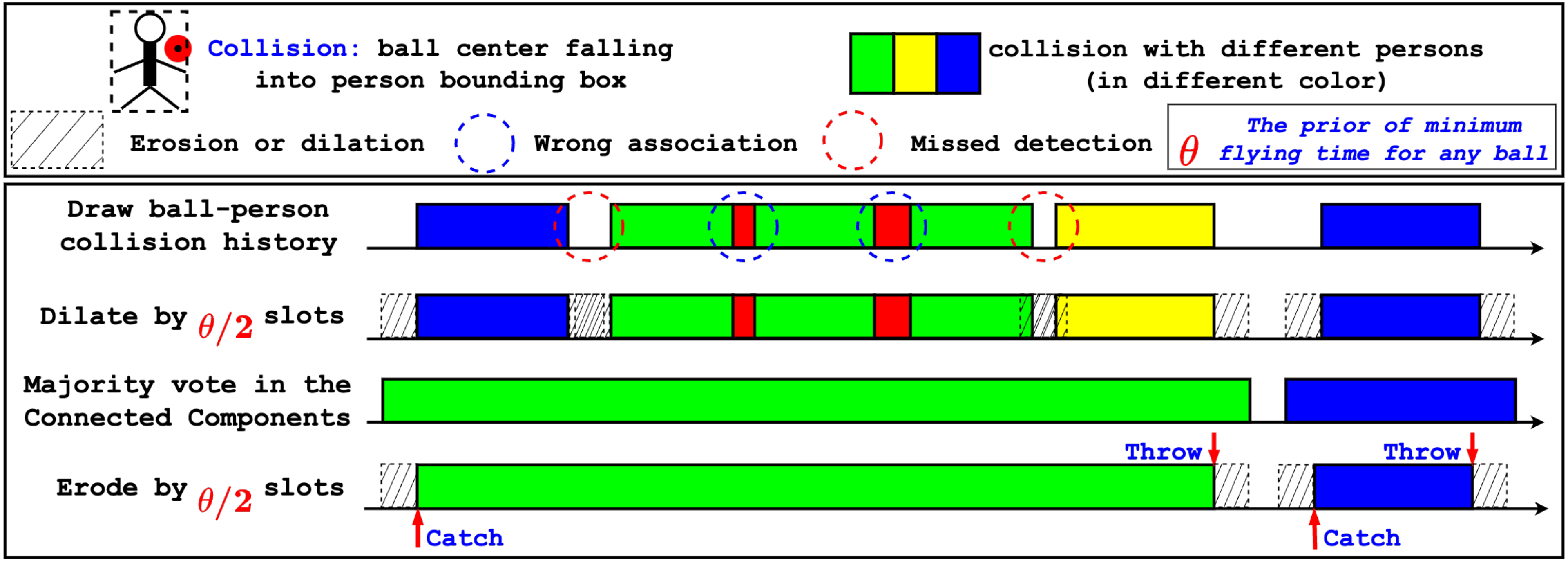
"Structured" : filter pruning



Action Detection Robustness Challenges

- Detection errors
  - Missed detections (FN): ball-person occlusions and person-person occlusions
  - False detections (FP): patches with homogeneous color are mistaken for balls
- Association errors
  - ReID errors: ball/person ReID errors occur mostly in a short time span
- 3D reasoning is expensive and difficult
  - Depth information is ill-posed in monocular camera

Robust Video Action Detection: A Heuristic Approach



Collision: ball center falling into person bounding box

collision with different persons (in different color)

Erosion or dilation

Wrong association

Missed detection

θ *The prior of minimum flying time for any ball*

Draw ball-person collision history

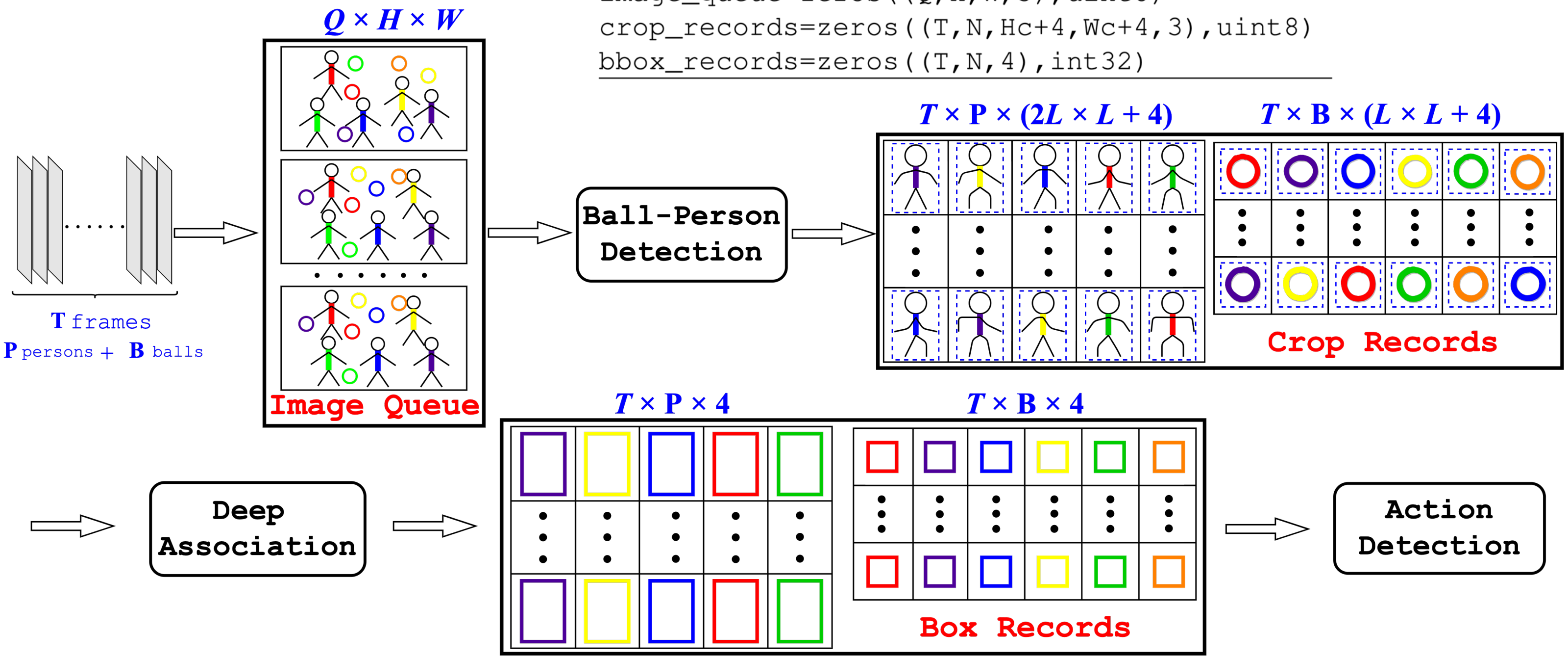
Dilate by $\theta/2$ slots

Majority vote in the Connected Components

Erode by $\theta/2$ slots

Cache-Friendly Pipeline: Overview

```
image_queue=zeros((Q,H,W,3),uint8)
crop_records=zeros((T,N,Hc+4,Wc+4,3),uint8)
bbox_records=zeros((T,N,4),int32)
```



$Q \times H \times W$

Image Queue

T frames

P persons + B balls

Ball-Person Detection

$T \times \mathrm{P} \times (2L \times L + 4)$

$T \times \mathrm{B} \times (L \times L + 4)$

Crop Records

Deep Association

$T \times \mathrm{P} \times 4$

$T \times \mathrm{B} \times 4$

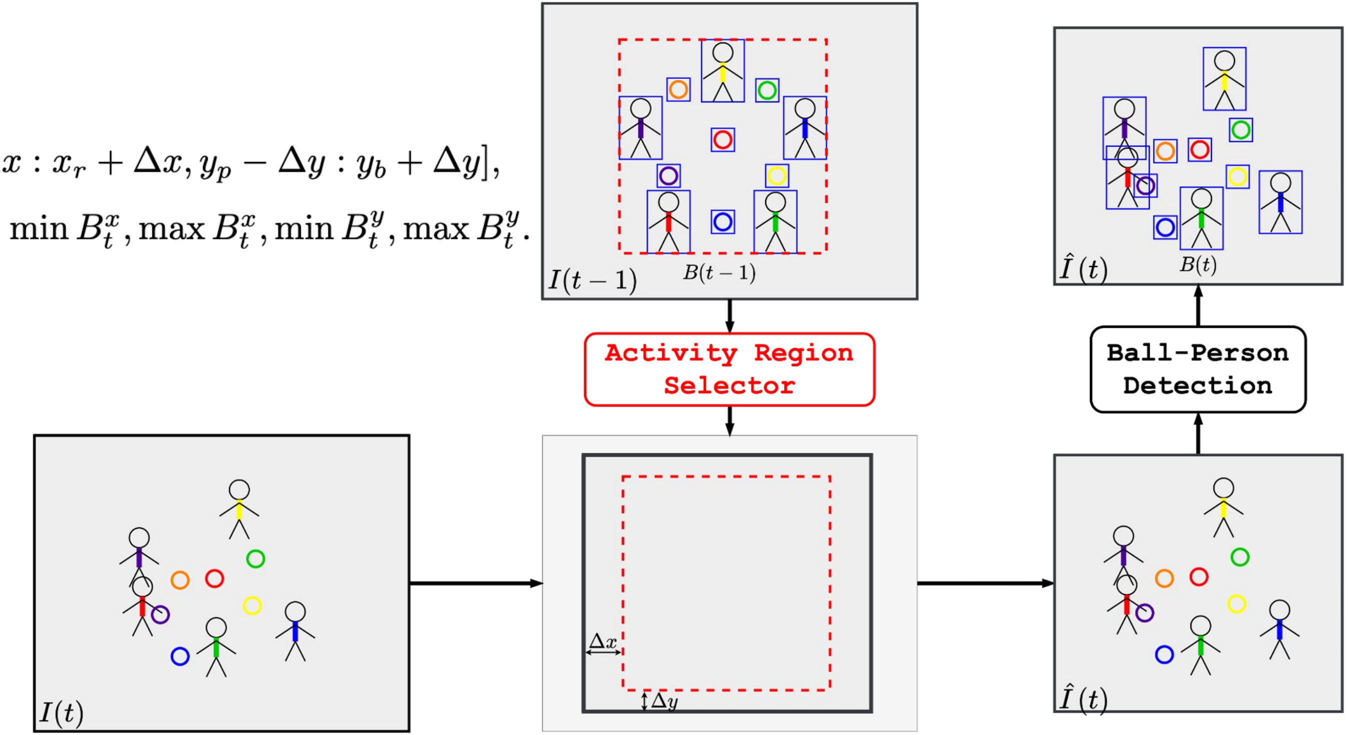
Box Records

Action Detection

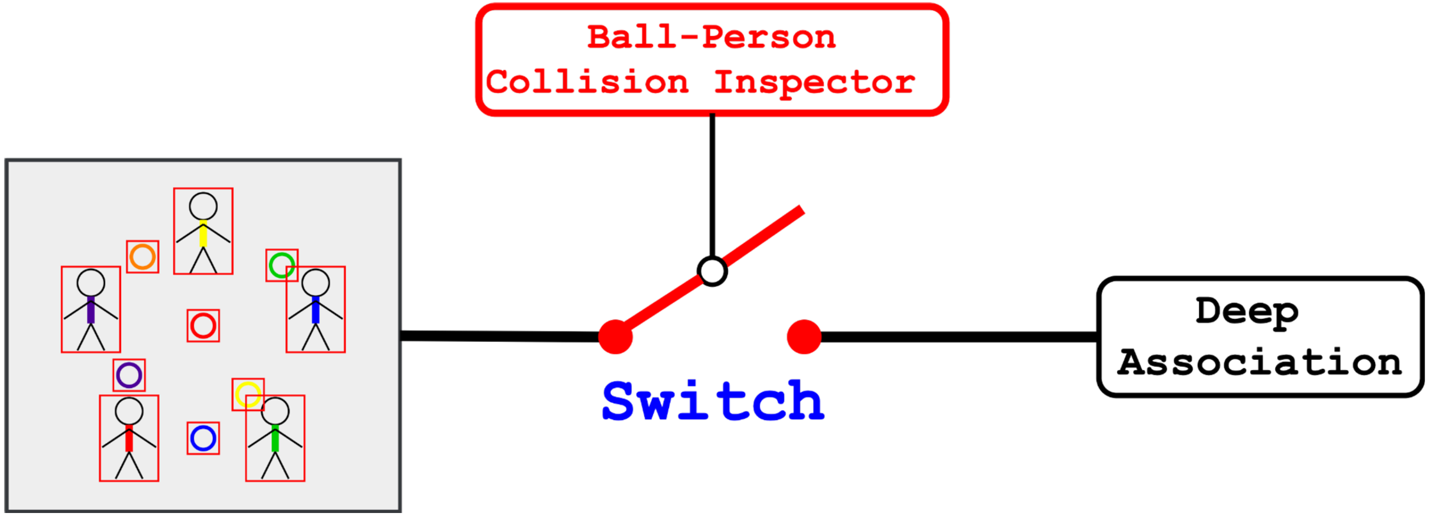
Dynamic Inference I: Activity Region Cropping (ARC)

$$\hat{I}_{t+1} = I_{t+1}[x_l - \Delta x : x_r + \Delta x, y_p - \Delta y : y_b + \Delta y],$$

$$\text{with } x_l, x_r, y_p, y_b \;=\; \min B_t^x, \max B_t^x, \min B_t^y, \max B_t^y.$$



Dynamic Inference II: Collision Inspection (CI)



Github Link of Our Solution

https://github.com/VITA-Group/21LPCV-UAV-Solution

References

[1] Bochkovskiy, Wang and Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection, Arxiv 2020.

[2] Redmon and Farhadi, YOLOv3: An Incremental Improvement, Arxiv, 2018.

[3] Redmon and Farhadi, YOLO9000: Better, Faster, Stronger, CVPR 2017.

[4] Redmon et al., You Only Look Once: Unified, Real-Time Object Detection, CVPR 2016.

[5] Chen et al., You Only Look One-level Feature, Arxiv 2021.

[6] Ge et al., YOLOX: Exceeding YOLO Series in 2021, Arxiv 2021.

[7] Ye et al., Deep Learning for Person Re-identification: A Survey and Outlook, TPAMI 2021.

[8] Deng et al., Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey, Proceedings of the IEEE 2020.

[9] Liang et al., Pruning and Quantization for Deep Neural Network Acceleration: A Survey, Arxiv 2021.

[10] Cheng et al., A Survey of Model Compression and Acceleration for Deep Neural Networks, IEEE Signal Processing Magazine 2020.

[11] Han et al., Dynamic Neural Networks: A Survey, Arxiv 2021.

[12] Bewley et al., Simple Online and Realtime Tracking, ICIP 2016

[13] Wojke, Bewley and Paulus, Simple Online and Realtime Tracking with a Deep Association Metric, ICIP 2017.

References cont.

[14] Li et al., Crop-Transform-Paste: Self-Supervised Learning for Visual Tracking, Arxiv, 2021.

[15] Naveed, Survey: Image Mixing and Deleting for Data Augmentation, Arxiv, 2021.

[16] Ghiasi et al., Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation, CVPR, 2021.

[17] Niu et al., Making Images Real Again: A Comprehensive Survey on Deep Image Composition.

[18] Geirhos et al., ImageNet-trained CNNs Are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness, ICLR, 2019.

[19] Li et al., Shape-Texture Debiased Neural Network Training, ICLR, 2021.

[20] Madan et al., Small In-Distribution Changes in 3D Perspective and Lighting Fool Both CNNs and Transformers, Arxiv, 2021.

[21] Neural Network Distiller, https://github.com/IntelLabs/distiller, Intel AI Lab.

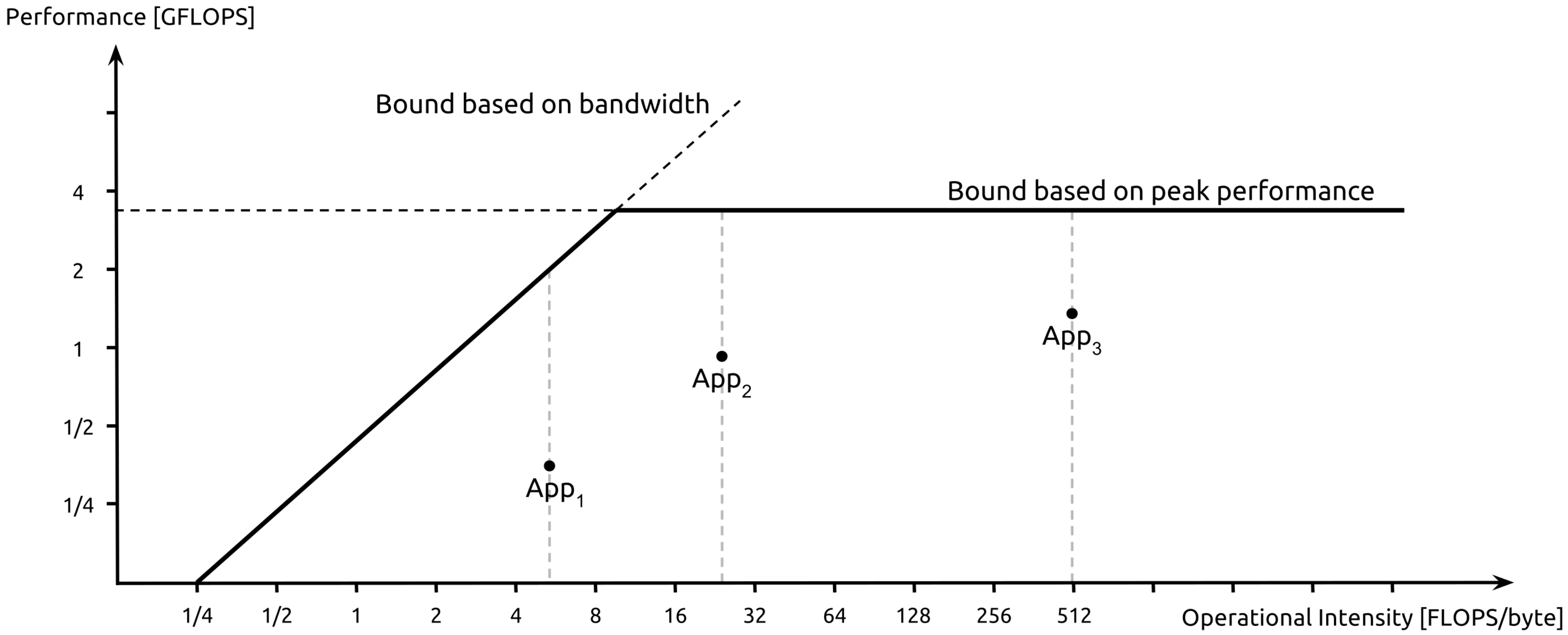
[22] Neural Network Intelligence, https://github.com/microsoft/nni, Microsoft Research.

[23] D2Go, https://github.com/facebookresearch/d2go, FaceBook Research.

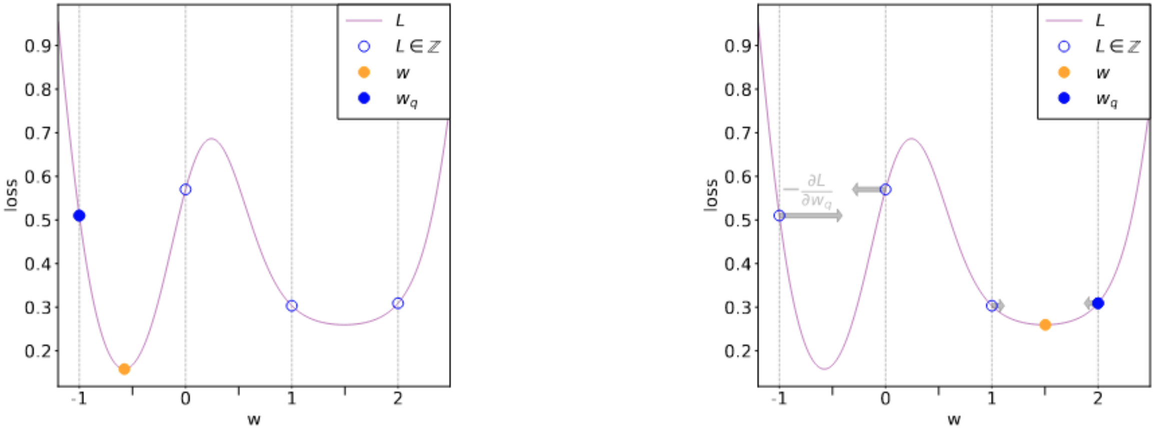
[24] YOLOv5, https://github.com/ultralytics/yolov5, Ultralytics.

[25] Pytorch ReID, https://github.com/layumi/Person_reID_baseline_pytorch, Zhedong Zheng.

Roofline Model



QAT vs. PTQ



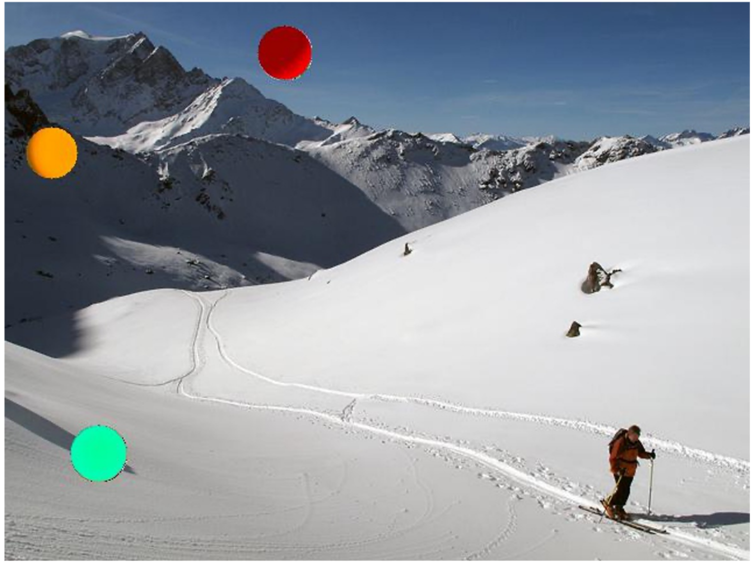
(a) Post training quantization

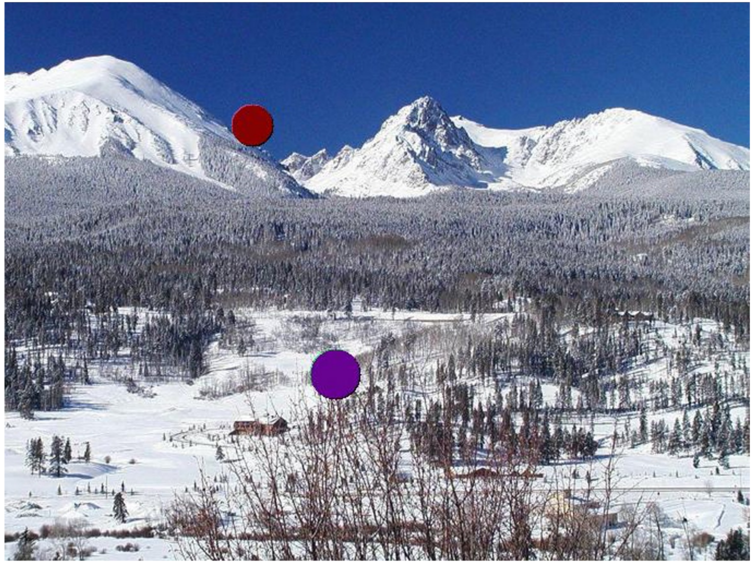
(b) After quantization aware fine-tuning

Figure 6: Example 1D loss function. The model, w, is scale quantized with scale factor 1. a) PTQ: model converges to a narrow minimum. b) QAT: model finds a wide minimum with lower loss quantization points.

Legacy

COCO-Aug v1

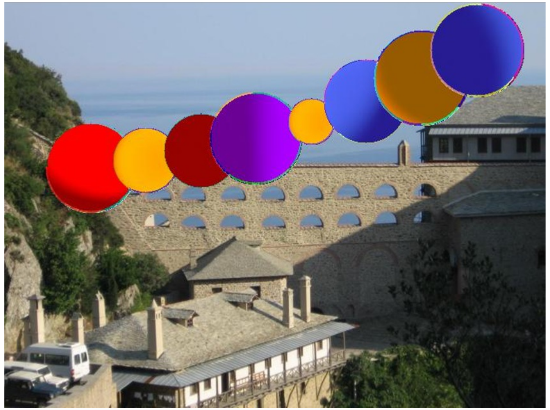


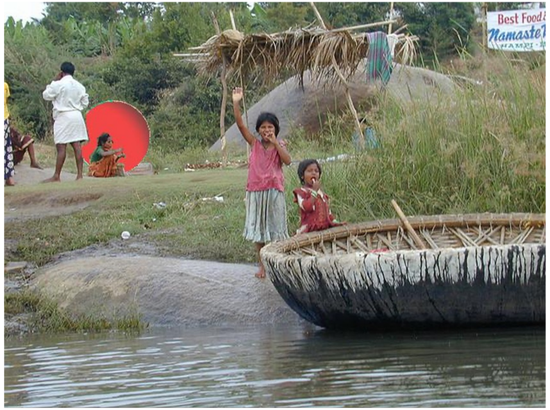


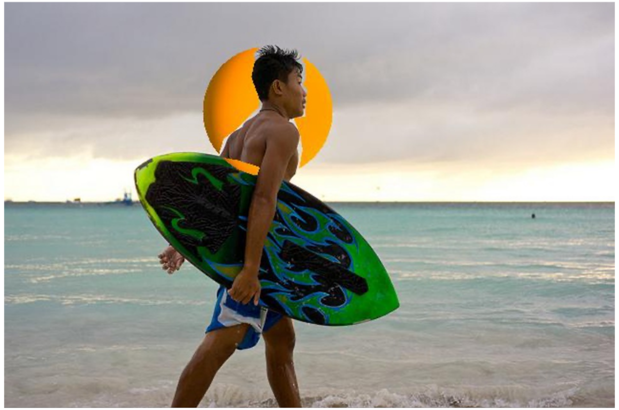
- Exist person
- Medium size balls

- No person
- Small size balls

COCO-Aug v2





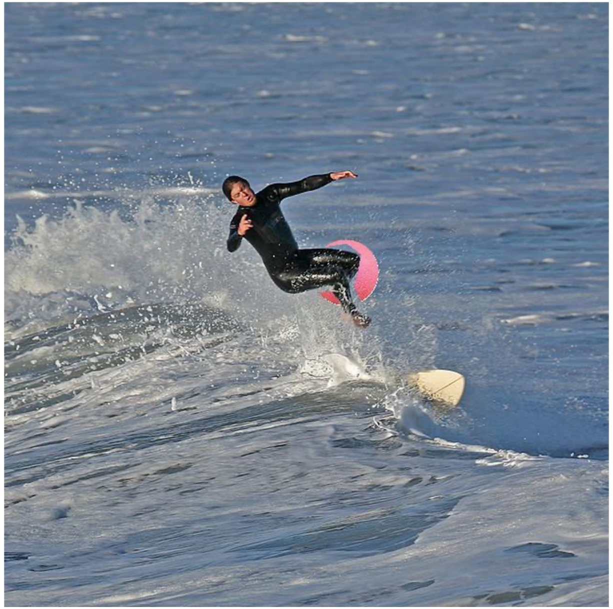


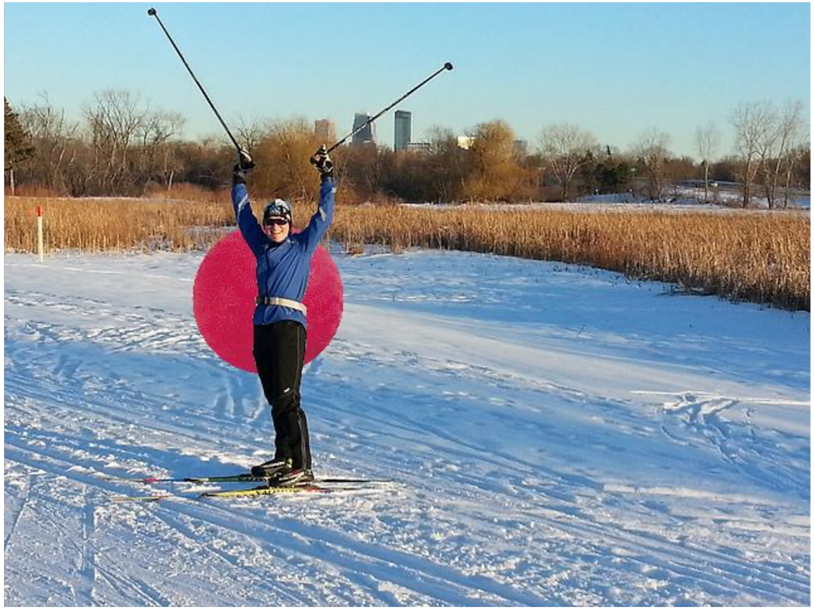
- No person
- Different size balls
- Occluded by ball

- Multiple person
- One size ball
- Occluded by person

- One person
- One size ball
- Occluded by person

COCO-Aug v3



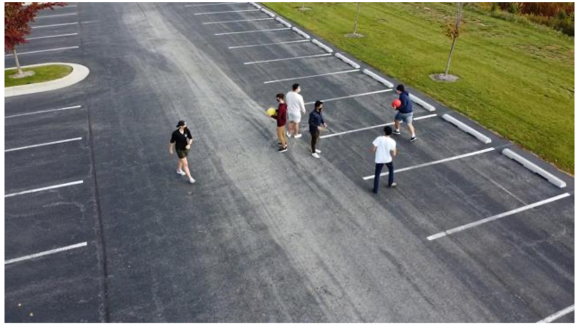


- Small size ball
- Large size ball

Existing Datasets

Dataset	Number	Characteristics
COCO Ball+Person Subset	4256	1:1 ball person ratio Domain gap large (especially ball) Few occluded samples
Organizer Data	277	Close to testing case Only one scenario

COCO Ball Person Data


Sample Videos

Organizer Data

COCO Dataset

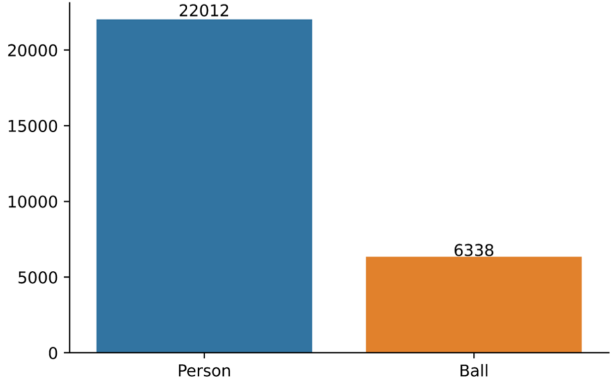
Conditions

- Person, ball coexist
- No additional objects
- The size of person is medium

COCO samples

Problems

- Imbalanced label(person:ball=3:1)
- Few occluded samples
- Large domain gap



Label distribution for COCO Ball+Person Subset

Image Composition

Occluded images

- Person image & mask from coco, real ball & 3D modeling ball
- Radius ~= person mask height/6

Non-occluded images

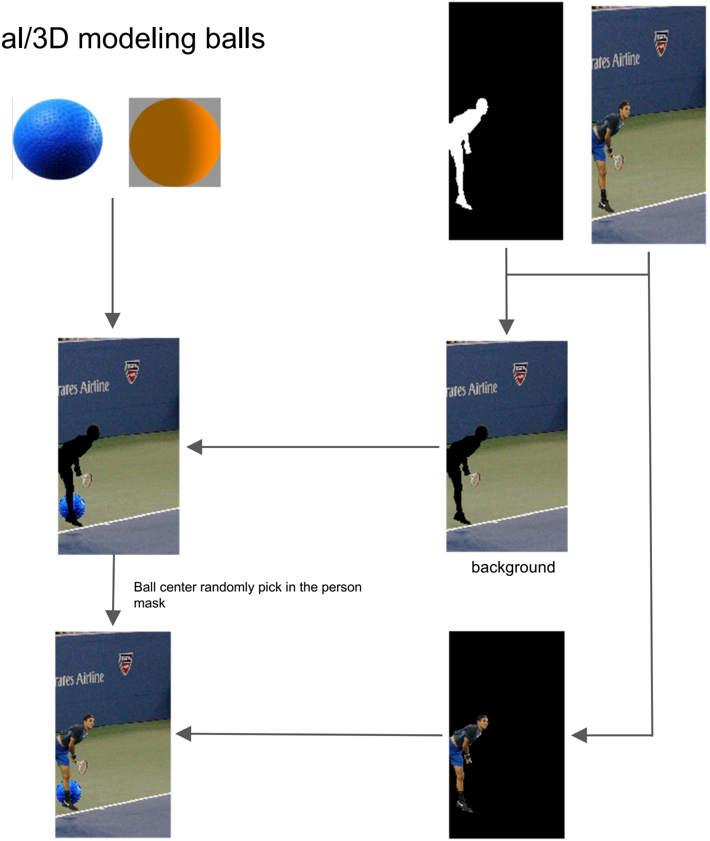
- Stack ball over images without overlapping person bbox

Non-occluded images

But domain gap is still large!
Need image harmanization!

Person images & mask

Real/3D modeling balls



background

Ball center randomly pick in the person mask

Occluded image

Person

Extended COCO Dataset

Dataset	Image number	Characteristic
Extension_v1	3304	● No occlusion ● Different size of ball ● Ball is synthesized with 3d texture ● There is no person in some images
Extension_v2	10026	● Different size of ball ● Ball is synthesized with 3d texture ● Ball is randomly occluded by person and ball ● There is no person in some images
Extension_v3	2923	● Different size of ball ● Ball is cropped from real images ● Ball is randomly occluded by person

Synthetic datasets