Computer Vision

Instance/panoptic segmentation and landmark localization

Tae-Hyun Oh (오태현)
전자전기공학과
POSTECH

Slide by Dongmin Choi (최동민)
TAs: { Dongmin Choi , Jongha Kim, Juyong Lee, Sungbin Kim} (in alphabetic order)

boostcamp AI Tech

1. Instance segmentation

1.1 What is instance segmentation?

1.2 Instance segmenters

2. Panoptic segmentation

2.1 What is panoptic segmentation?

2.2 UPSNet & VPSNet

3. Landmark localization

3.1 What is landmark localization?

3.2 Coordinate regression vs. heatmap classification

3.3 Hourglass network

3.4 Extensions

4. Detecting objects as keypoints

4.1 CornerNet & CenterNet

1.

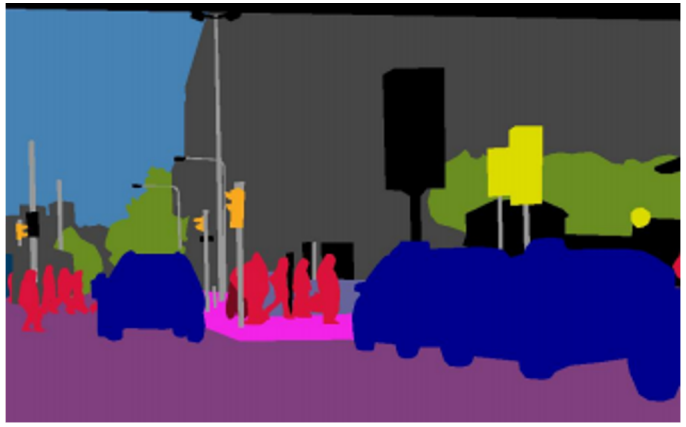
Instance segmentation

1.1 What is instance segmentation?

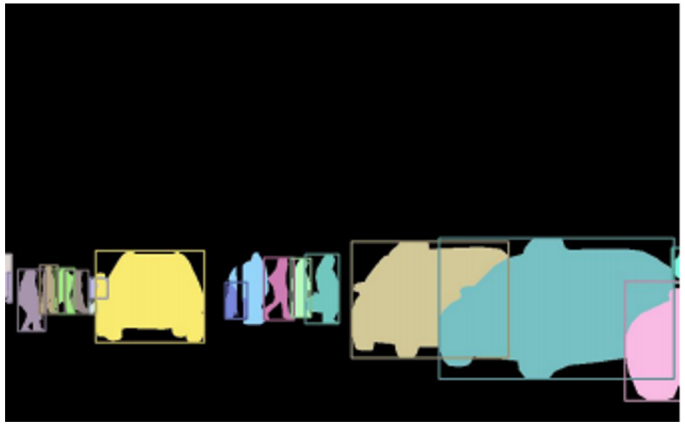
Fundamental image recognition tasks

[Kirillov et al., CVPR 2019]

Image



Semantic segmentation



Instance segmentation

Panoptic segmentation

1.1 What is instance segmentation?

Input image | Semantic segmentation | Instance segmentation

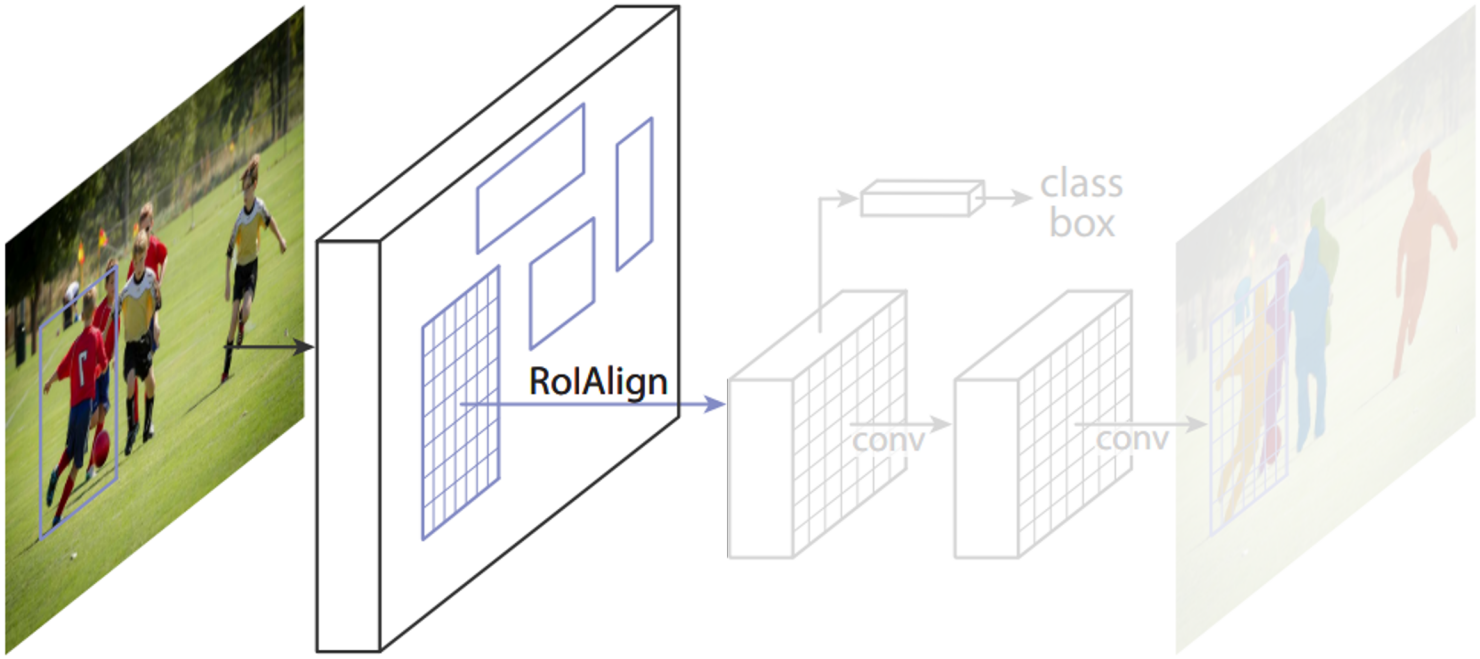
Instance segmentation = Semantic segmentation
+ distinguishing instances

[https://towardsdatascience.com/review-deepmask-instance-segmentation-30327a072339]

1.2 Instance segmenters

Mask R-CNN

[He et al., ICCV 2017]

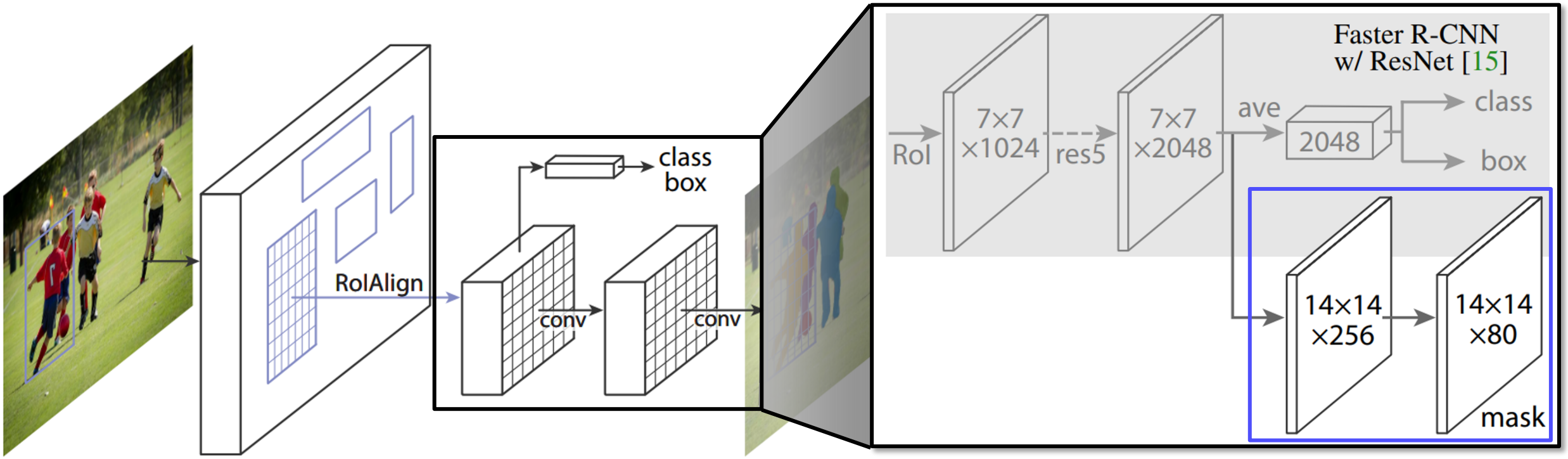


RoI extraction through RoIAlign,
an improved version of RoI Pooling

1.2 Instance segmenters

Mask R-CNN

[He et al., ICCV 2017]

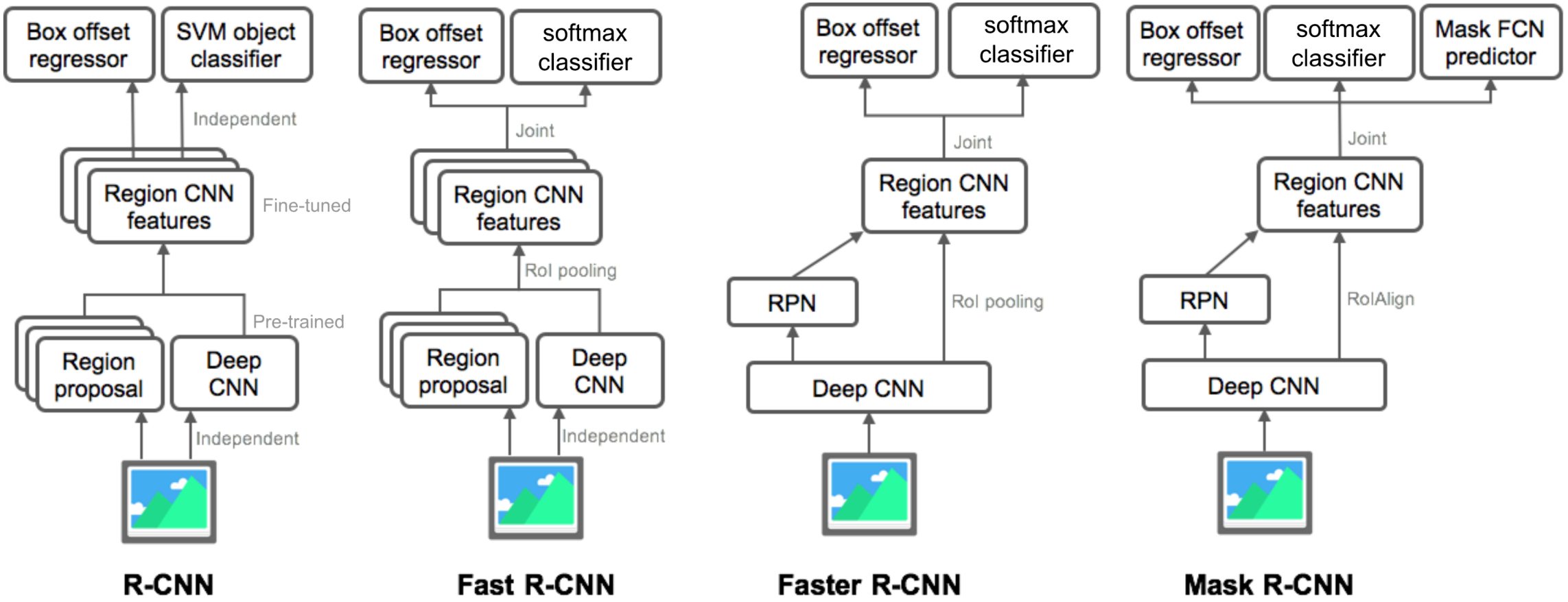


Binary classification for each class
(BG vs FG)

Mask R-CNN = Faster R-CNN + Mask branch

1.2 Instance segmenters

Summary of the R-CNN family

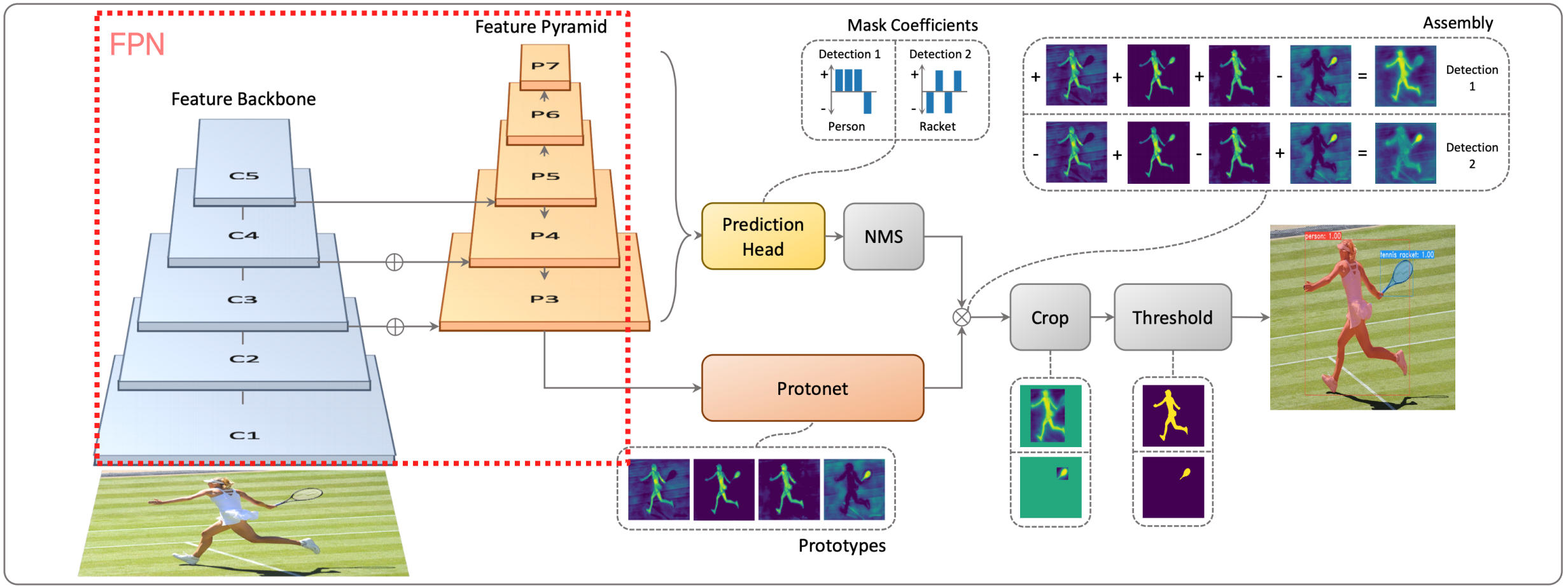


[https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html]

1.2 Instance segmenters

YOLACT (You Only Look At CoefficienTs)

[Bolya et al., ICCV 2019]

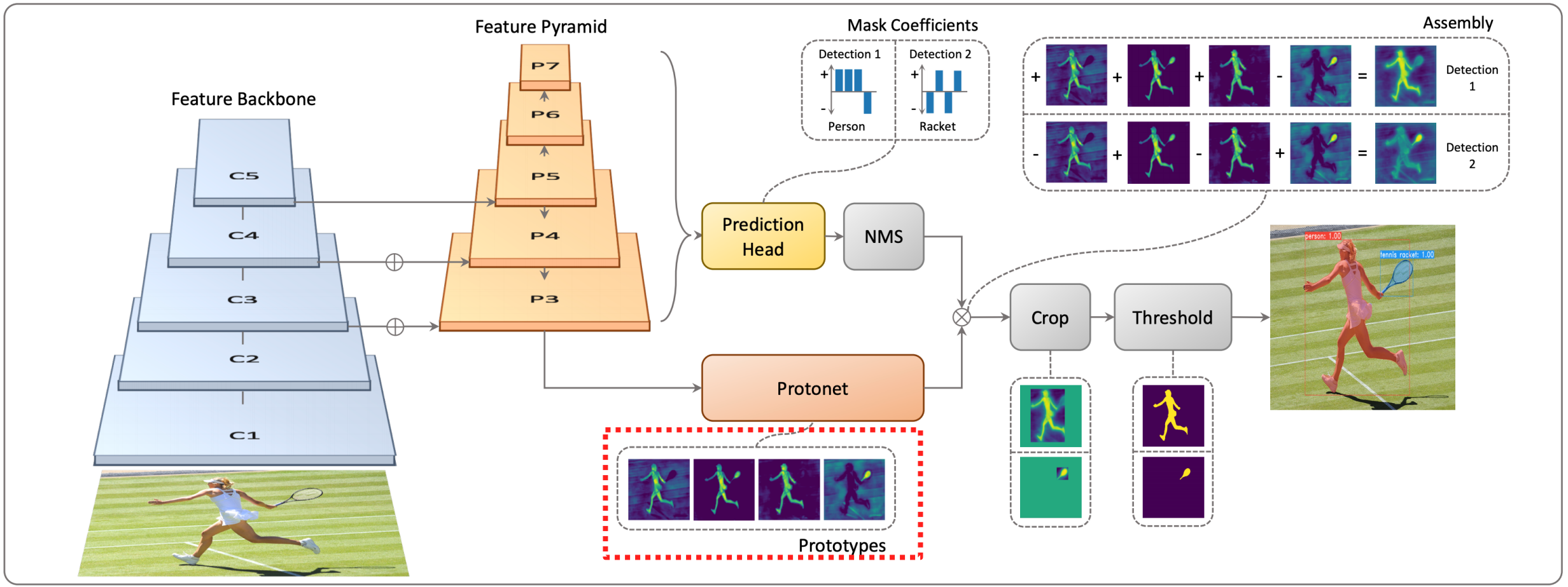


Real-time instance segmentation network

1.2 Instance segmenters

YOLACT (You Only Look At CoefficienTs)

[Bolya et al., ICCV 2019]

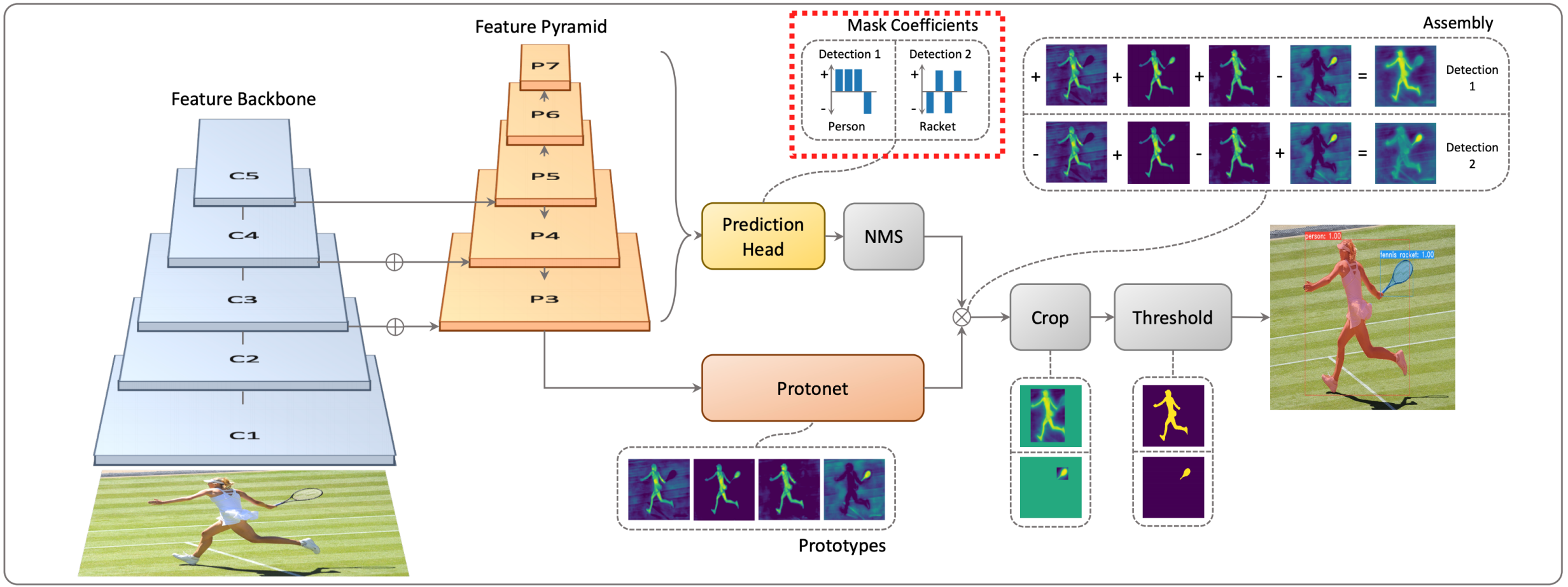


Real-time instance segmentation network

1.2 Instance segmenters

YOLACT (You Only Look At CoefficienTs)

[Bolya et al., ICCV 2019]



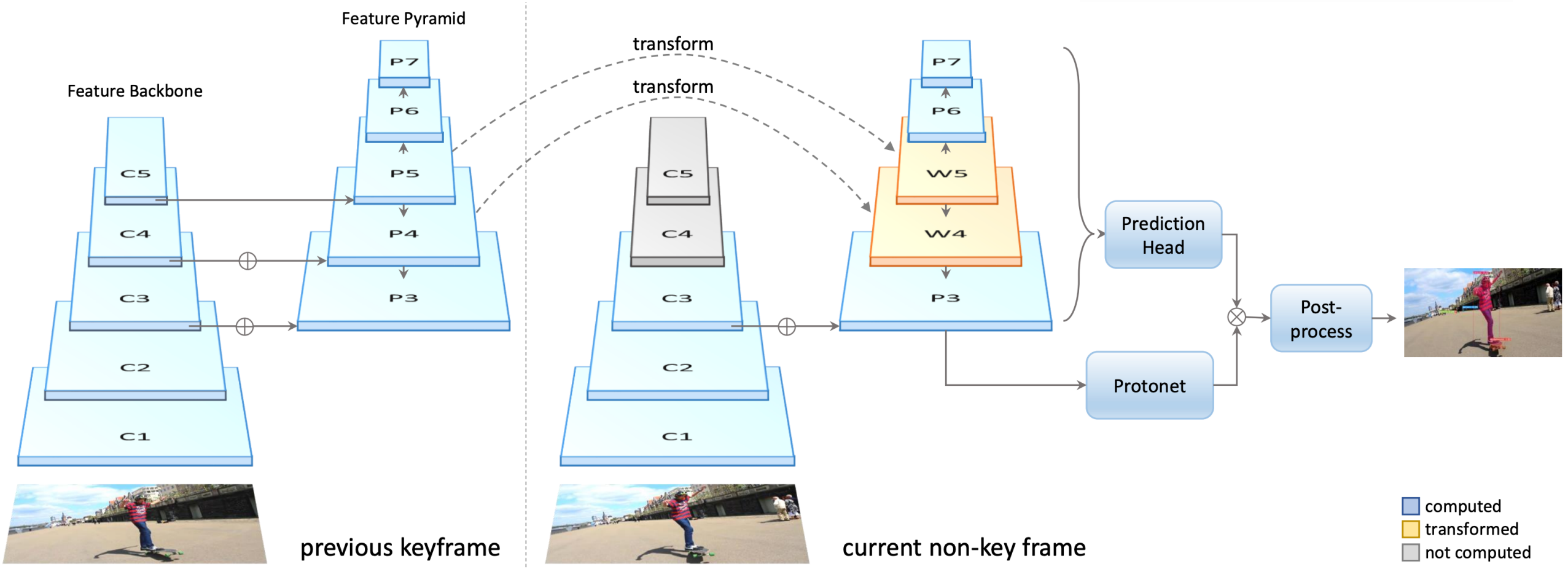
Real-time instance segmentation network

1.2 Instance segmenters

YolactEdge

[Liu et al., arXiv 2020]

Extend YOLACT to video



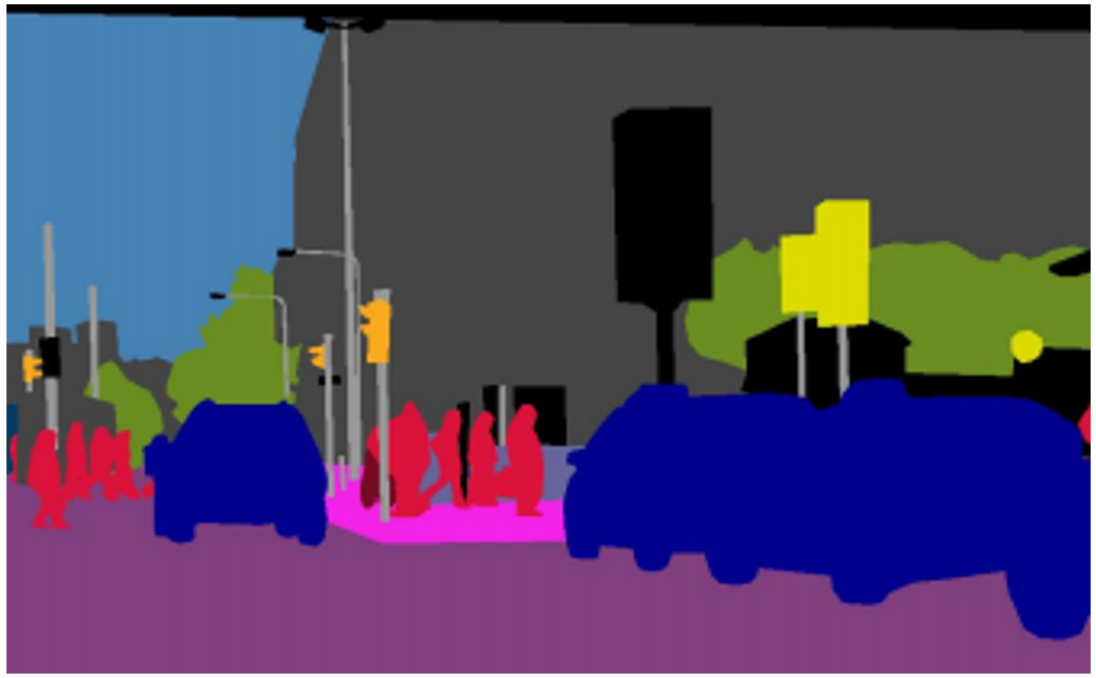
2.

Panoptic segmentation

2.1 What is panoptic segmentation?

Semantic segmentation vs. panoptic segmentation

[Kirillov et al., CVPR 2019]



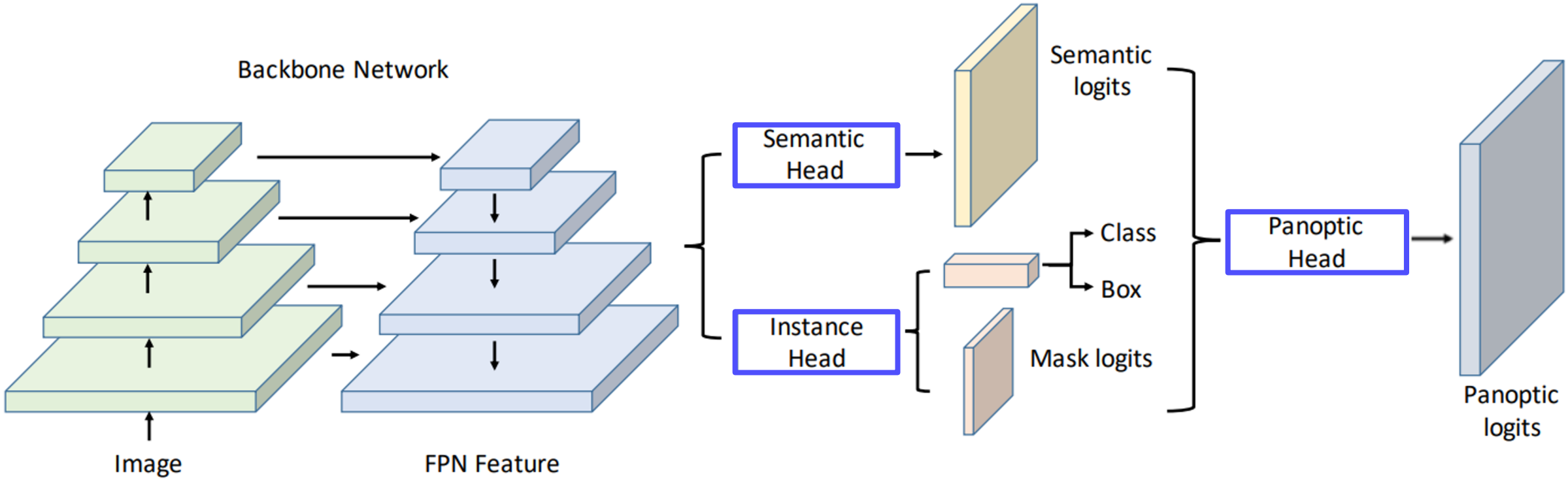
Semantic segmentation
(Stuff + Things)

Panoptic segmentation
(Stuff + Instances of Things)

2.2 UPSNet

UPSNet

[Xiong et al., CVPR 2019]

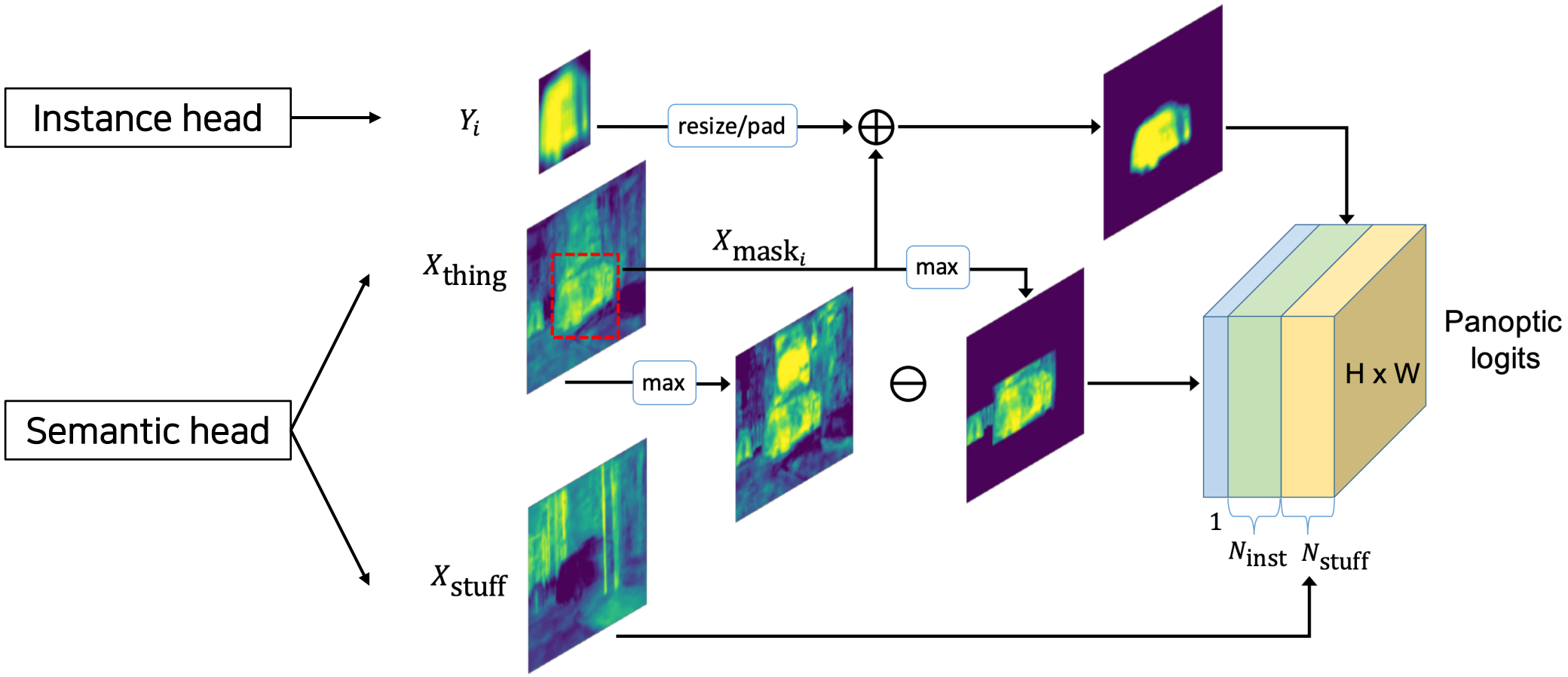


Semantic & Instance head → Panoptic head → Panoptic logits

2.2 UPSNet

UPSNet

[Xiong et al., CVPR 2019]



Architecture of the panoptic segmentation head

2.2 VPSNet

VPSNet (for video)

[Kim et al., CVPR 2020]

1. Align reference features onto the target feature map (Fusion at pixel level)

Reference frame

$\tau \in \{t-5 : t+5\}$

Target frame

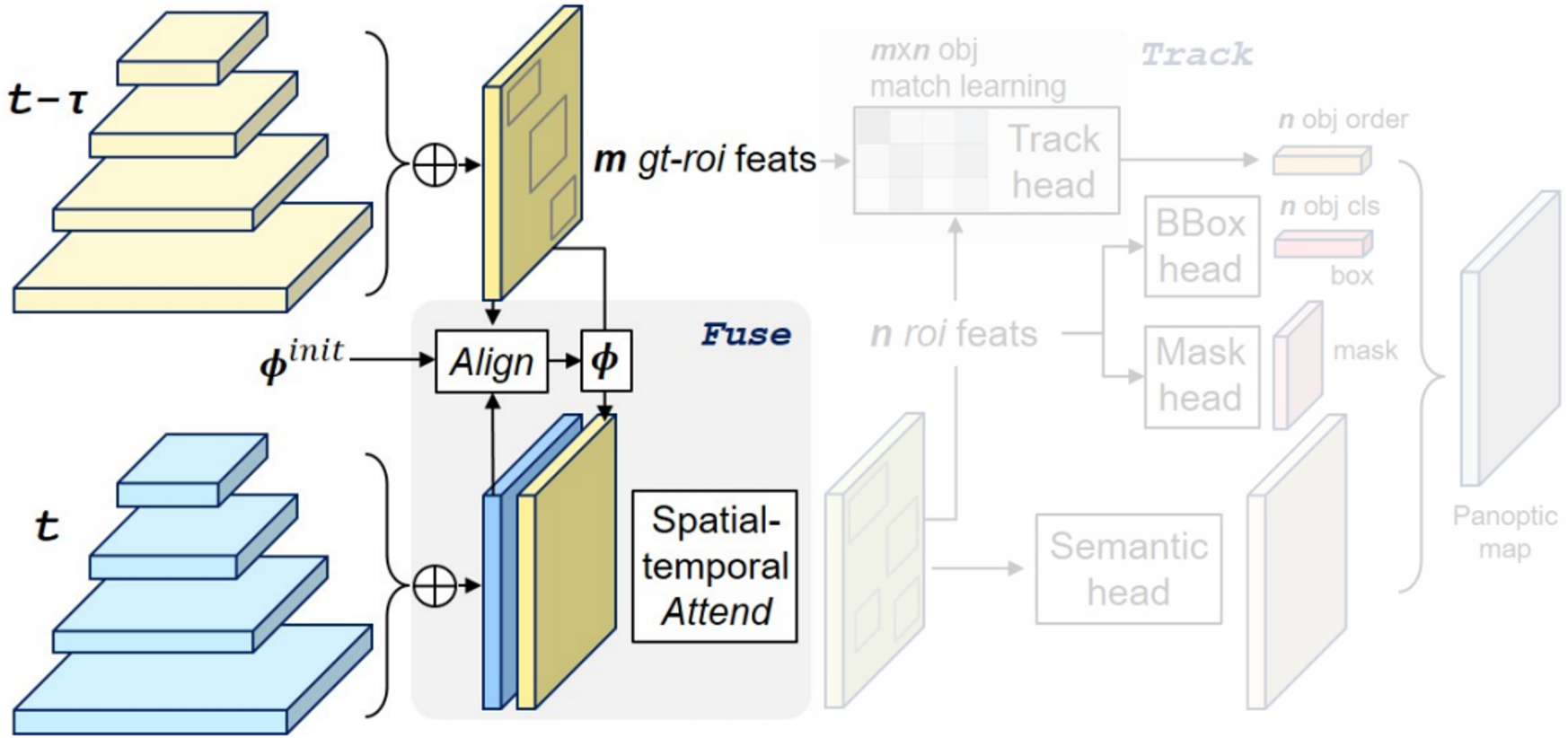


Figure 3: **Overall architecture of our VPSNet.**

[https://www.youtube.com/watch?v=AW6ZMy8TWdI]

2.2 VPSNet

VPSNet (for video)

[Kim et al., CVPR 2020]

2. Track module associates different object instances (Track at instance level)

Reference frame

$\tau \in \{t-5 : t+5\}$

Target frame

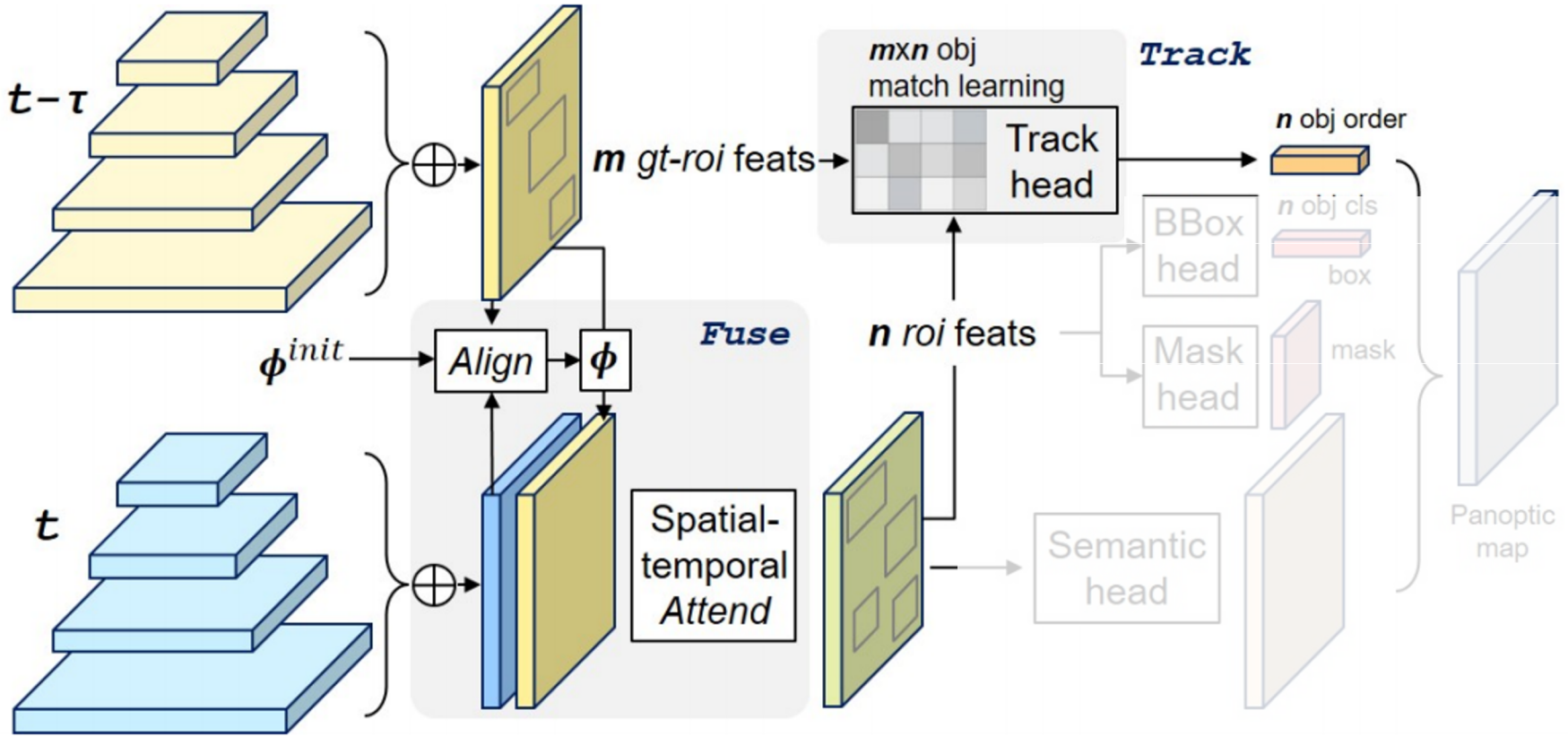


Figure 3: **Overall architecture of our VPSNet.**

[https://www.youtube.com/watch?v=AW6ZMy8TWdI]

2.2 VPSNet

VPSNet (for video)

[Kim et al., CVPR 2020]

3. Fused-and-tracked modules are trained to synergize each other

Reference frame

$\tau \in \{t-5 : t+5\}$

Target frame

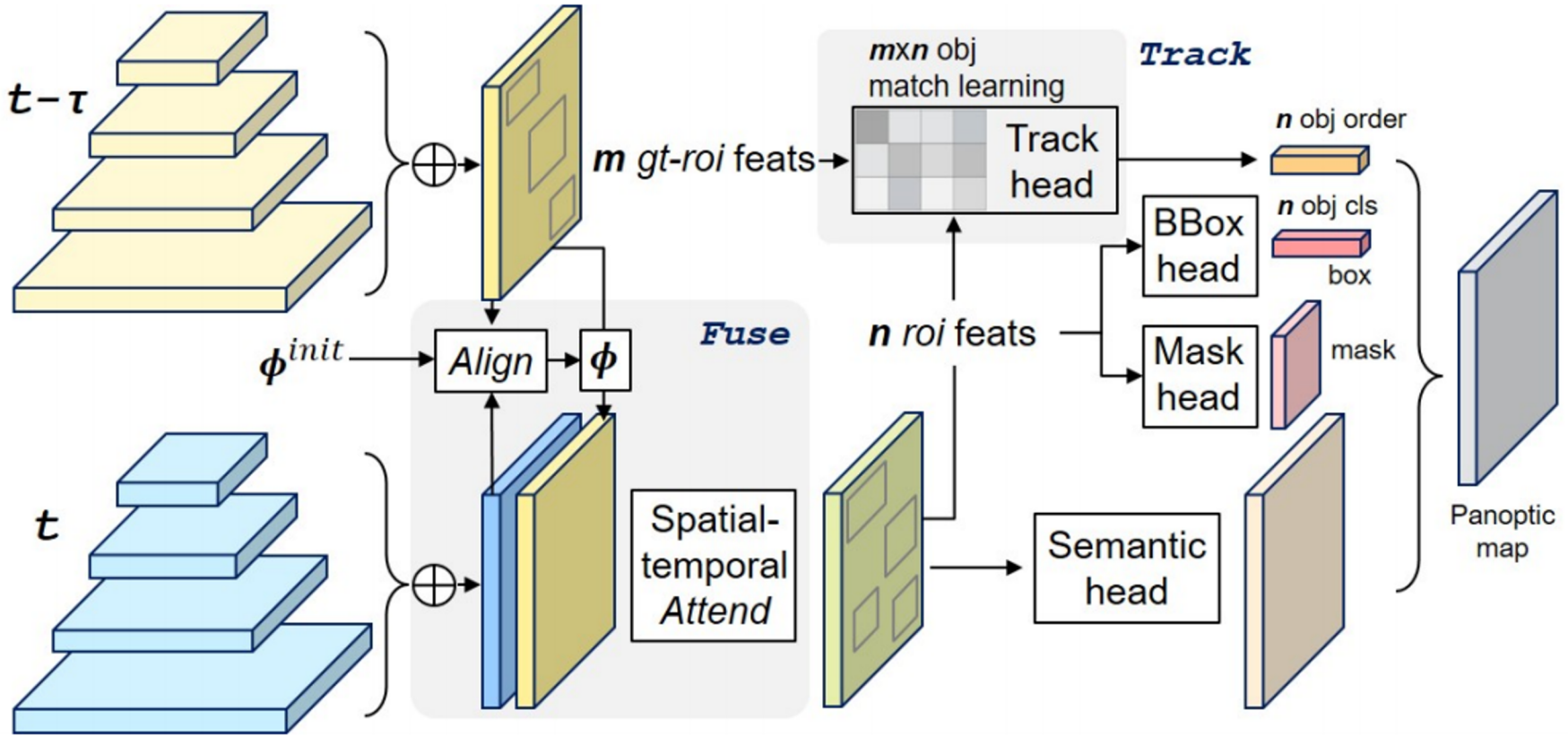


Figure 3: **Overall architecture of our VPSNet.**

[https://www.youtube.com/watch?v=AW6ZMy8TWdI]

2.2 VPSNet

VPSNet (for video)

[Kim et al., CVPR 2020]

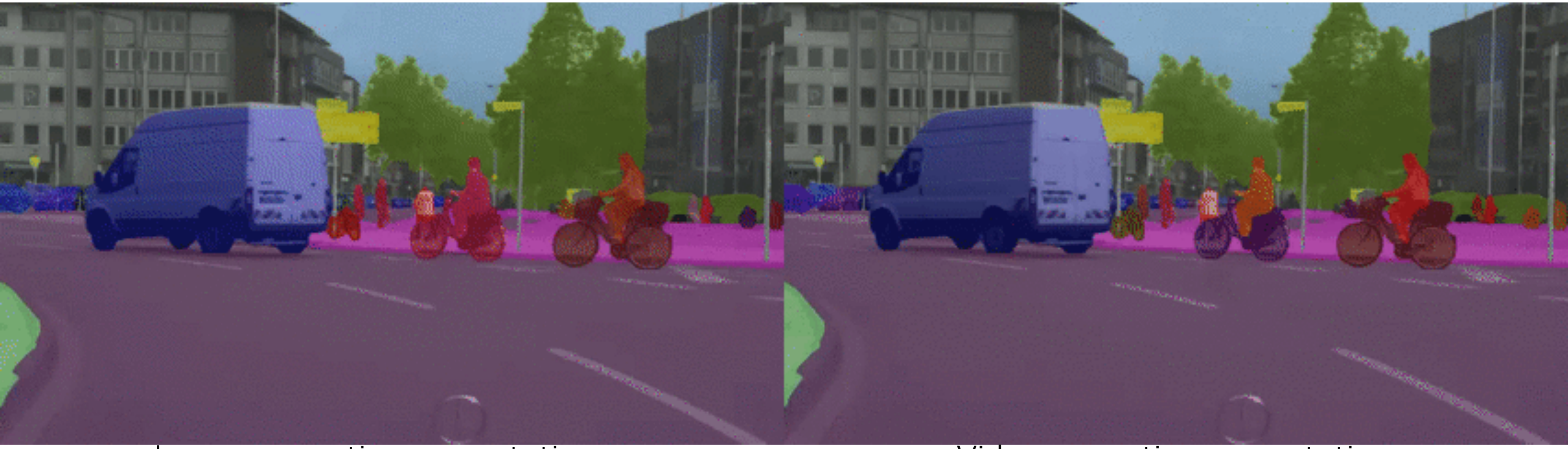


Image panoptic segmentation

×

Video panoptic segmentation

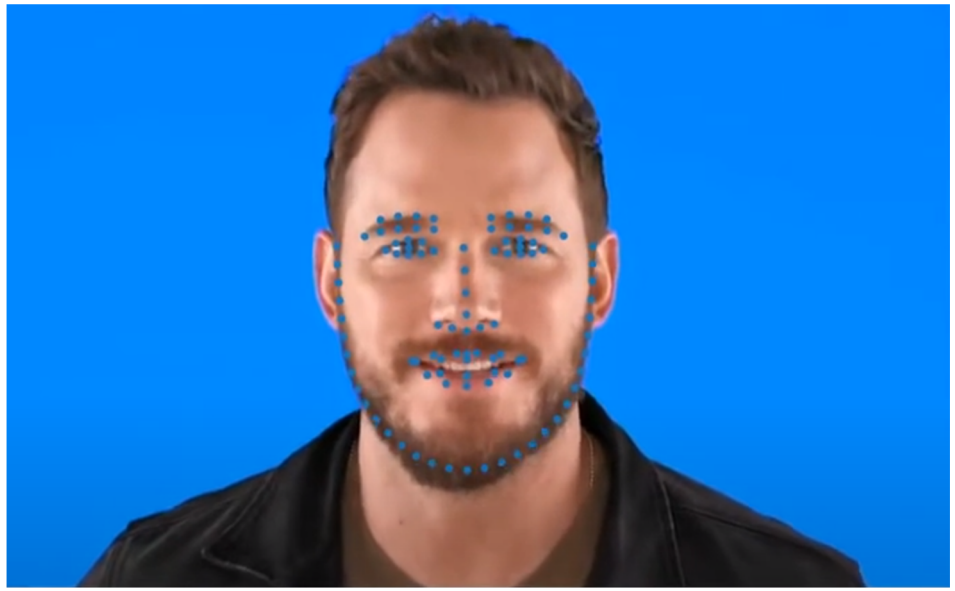
[https://www.youtube.com/watch?v=AW6ZMy8TWdI]

3.

Landmark localization

3.1 What is landmark localization?

[Cao et al., TPAMI 2019]



Facial landmark localization

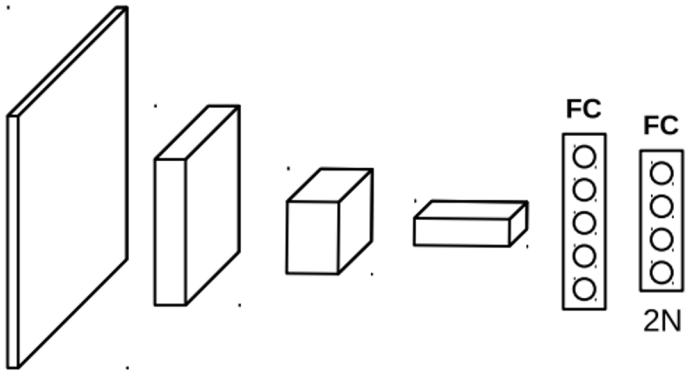
Human pose estimation

Landmark localization (= keypoint estimation): Predicting the coordinates of keypoints.

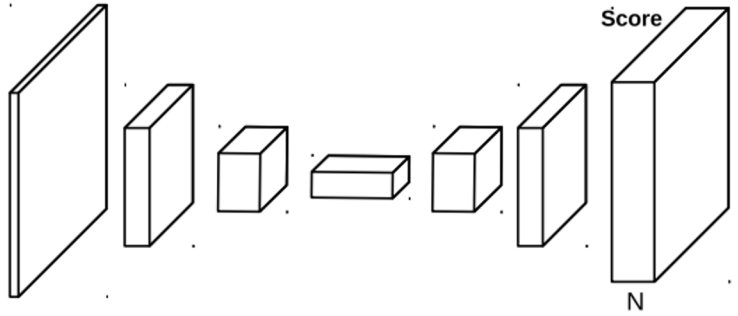
[https://wywu.github.io/projects/LAB/LAB.html]

3.2 Coordinate regression vs. heatmap classification

[Jin et al., arXiv 2020]
[Wang et al., ICCV 2019]

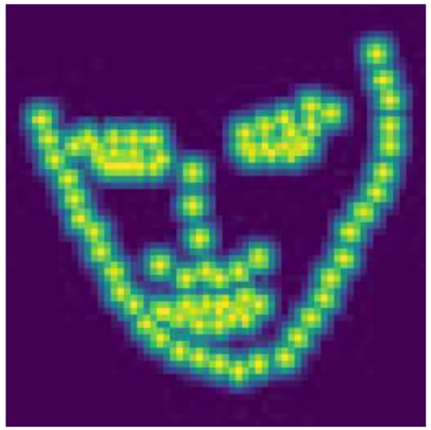


Coordinate regression



Heatmap classification

Image



Ground Truth

(a) Coordinate regression: usually inaccurate and biased
(b) Heatmap classification: better performance but high computational cost

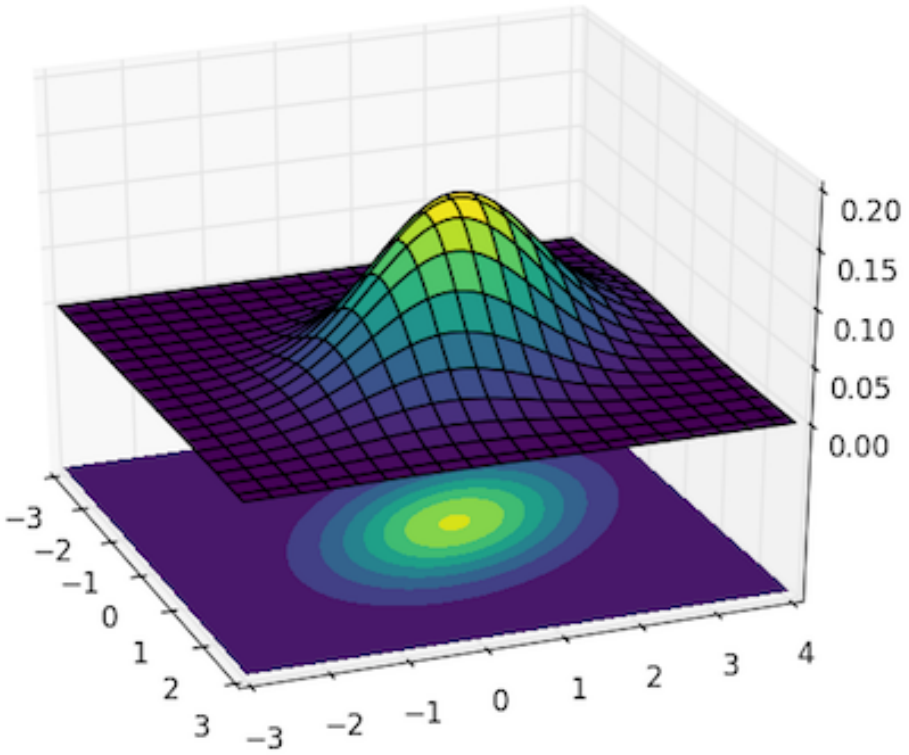
3.2 Coordinate regression vs heatmap classification

Landmark location to Gaussian heatmap

* (x_c, y_c) = center location

$$G_\sigma(x, y) = \exp\left(-\frac{(x - x_c)^2 + (y - y_c)^2}{2\sigma^2}\right)$$

```
# Generate gaussian
size = 6 * sigma + 1
x = np.arange(0, size, 1, float)
y = x[:, np.newaxis]
x0 = y0 = size // 2
# The gaussian is not normalized, we want the center value to equal 1
if type == 'Gaussian':
    g = np.exp(- ((x - x0) ** 2 + (y - y0) ** 2) / (2 * sigma ** 2))
elif type == 'Cauchy':
    g = sigma / (((x - x0) ** 2 + (y - y0) ** 2 + sigma ** 2) ** 1.5)
```



[https://scipython.com/blog/visualizing-the-bivariate-gaussian-distribution/]

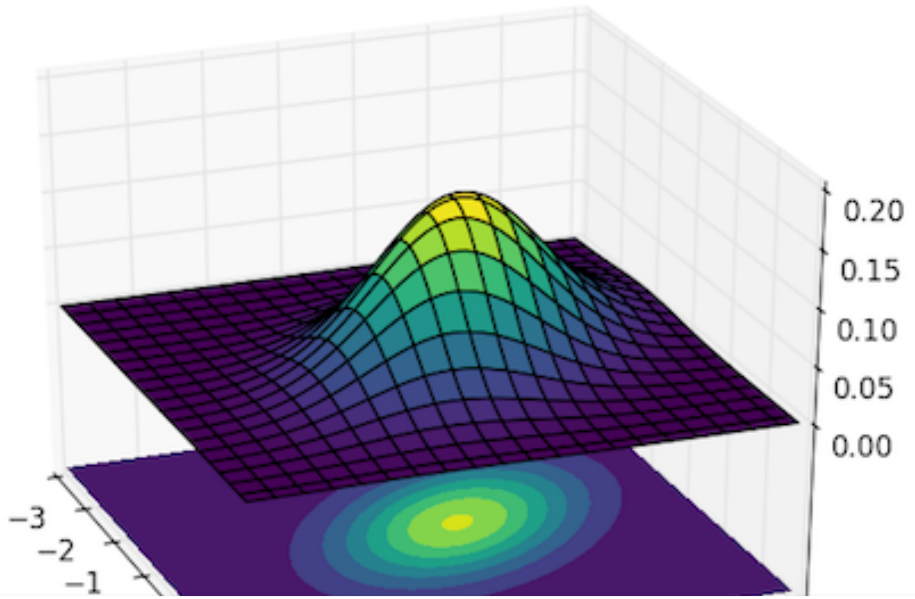
3.2 Coordinate regression vs heatmap classification

Landmark location to Gaussian heatmap

* (x_c, y_c) = center location

$$G_\sigma(x, y) = \exp\left(-\frac{(x - x_c)^2 + (y - y_c)^2}{2\sigma^2}\right)$$

```
# Generate gaussian
size = 6 * sigma + 1
x = np.arange(0, size, 1, float)
y = x[:, np.newaxis]
x0 = y0 = size // 2
# The gaussian is not normalized, we want the center value to equal 1
if type == 'Gaussian':
    g = np.exp(- ((x - x0) ** 2 + (y - y0) ** 2) / (2 * sigma
elif type == 'Cauchy':
    g = sigma / (((x - x0) ** 2 + (y - y0) ** 2 + sigma ** 2)
```



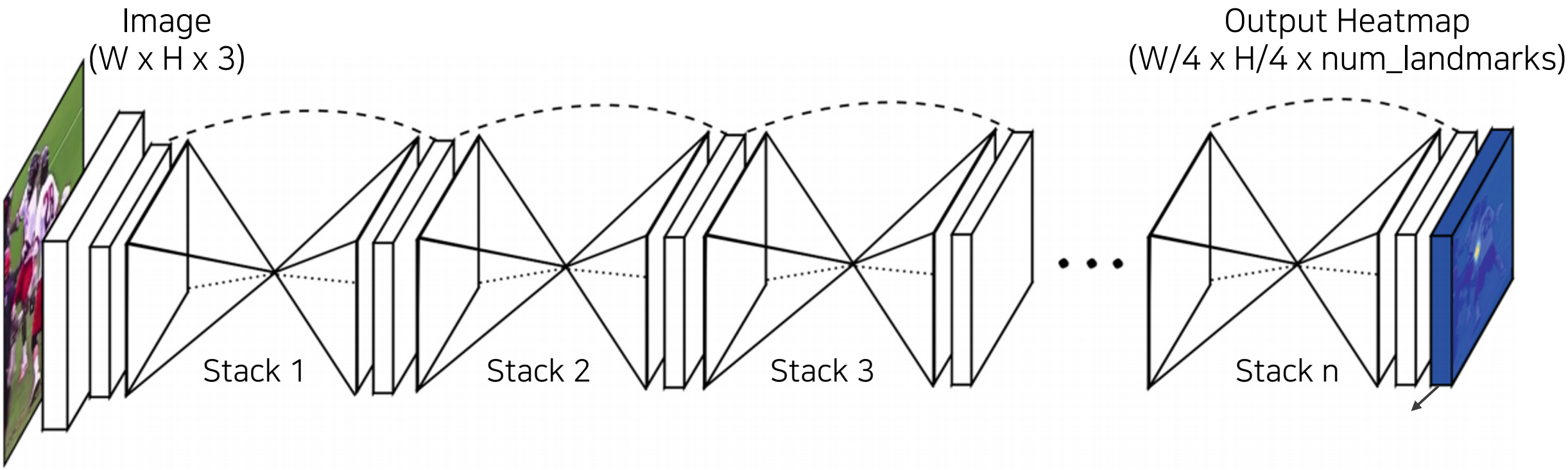
Then, how would you convert
a Gaussian heatmap to a landmark location?

⇒ Take-home assignment

[https://scipython.com/blog/visualizing-the-bivariate-gaussian-distribution/]

3.3 Hourglass network

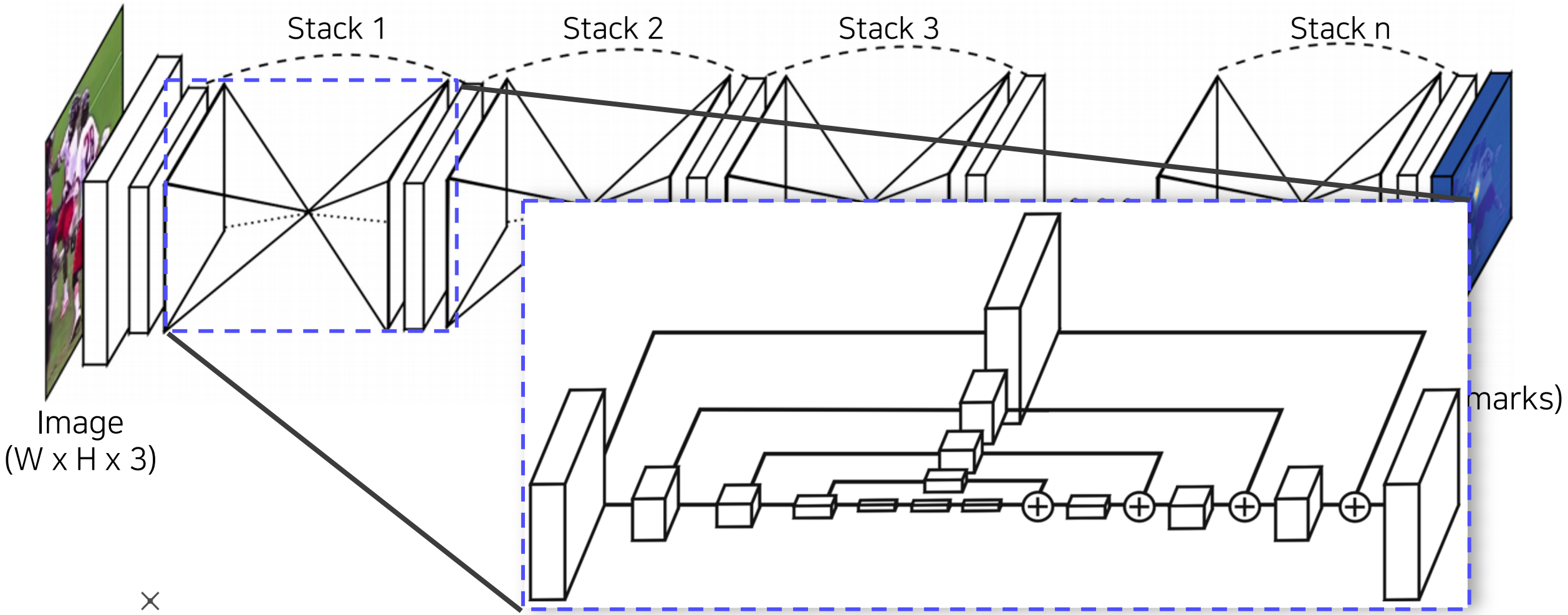
[Newell et al., ECCV 2016]



Stacked hourglass modules allow for repeated bottom-up and top-down inference that refines the output of the previous hourglass module

3.3 Hourglass network

[Newell et al., ECCV 2016]



3.4 Extensions

DensePose

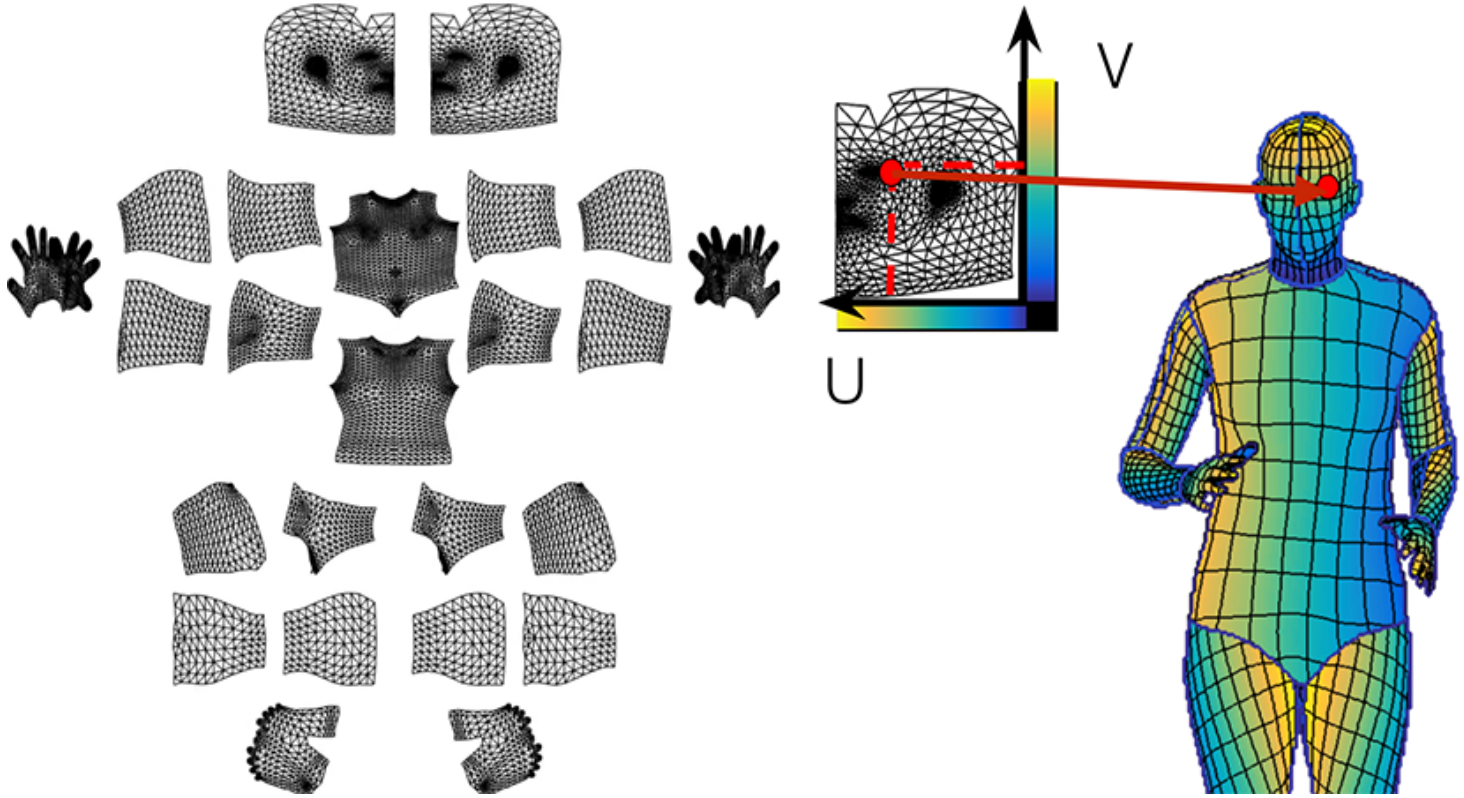
[Guler et al., CVPR 2018]

All pixels → 3D surface of the human body

3.4 Extensions

DensePose

[Guler et al., CVPR 2018]

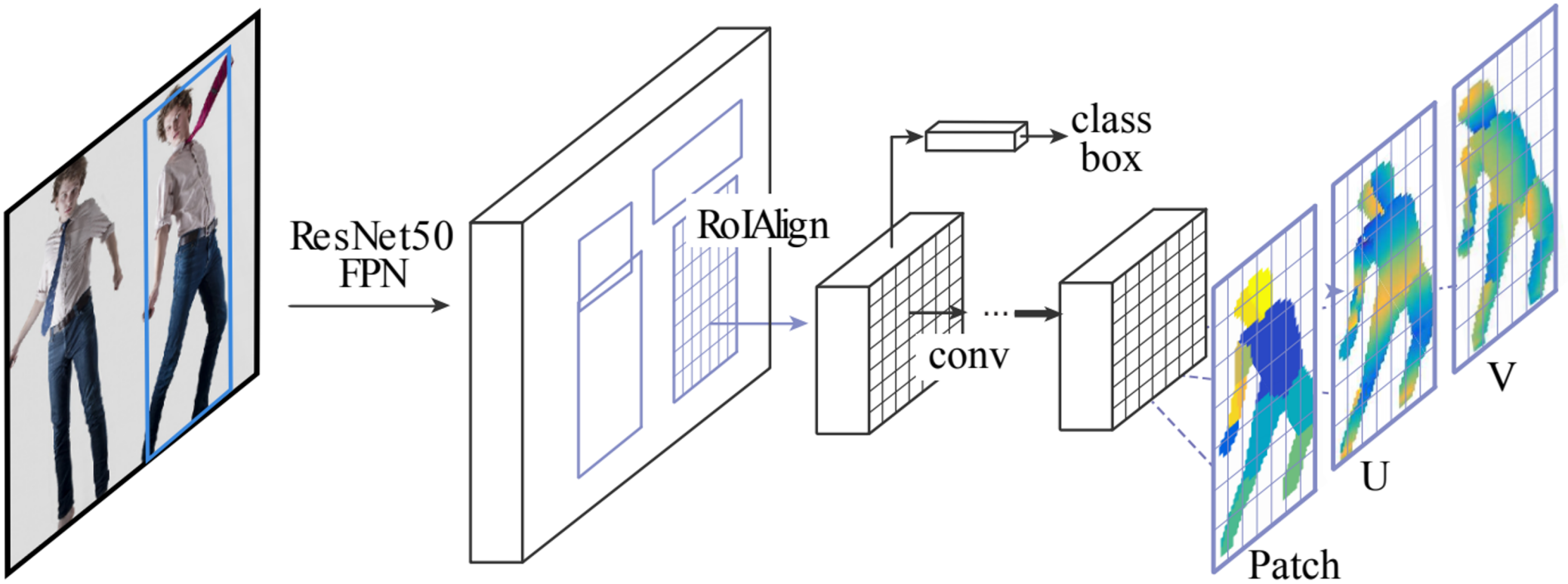


UV map is a flattened representation of 3D geometry
Also, UV map is invariant to motion (i.e., canonical coordinate)

3.4 Extensions

DensePose

[Guler et al., CVPR 2018]

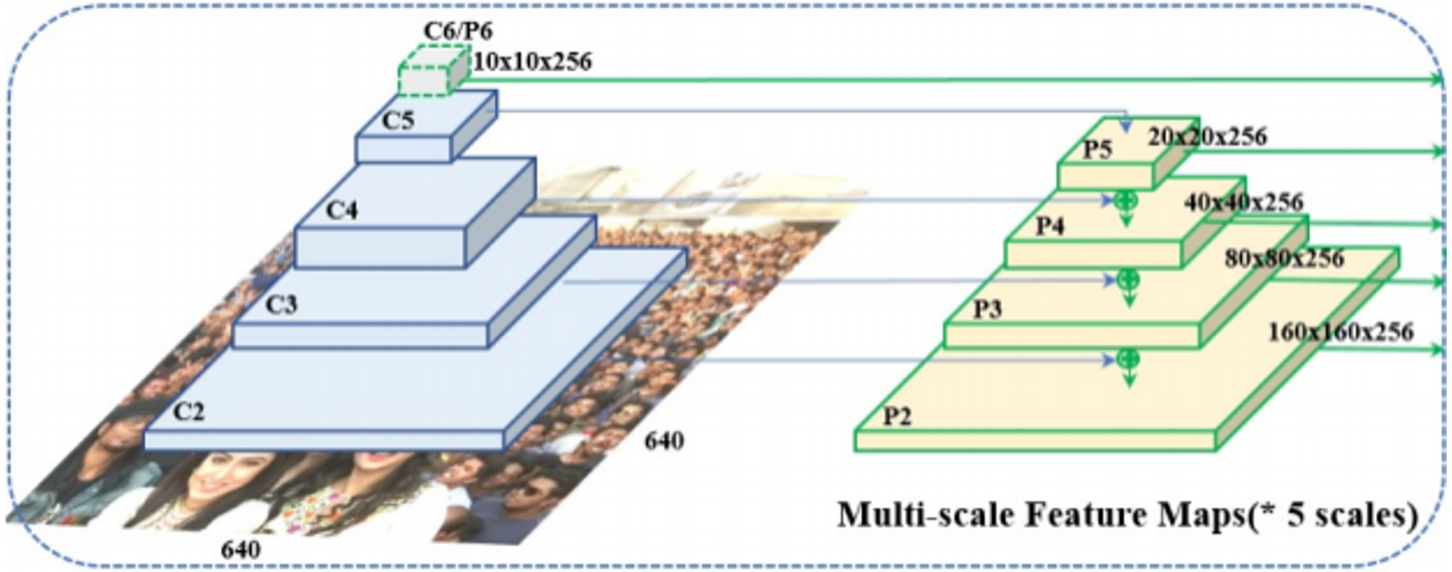


DensePose R-CNN = Faster R-CNN + 3D surface regression branch

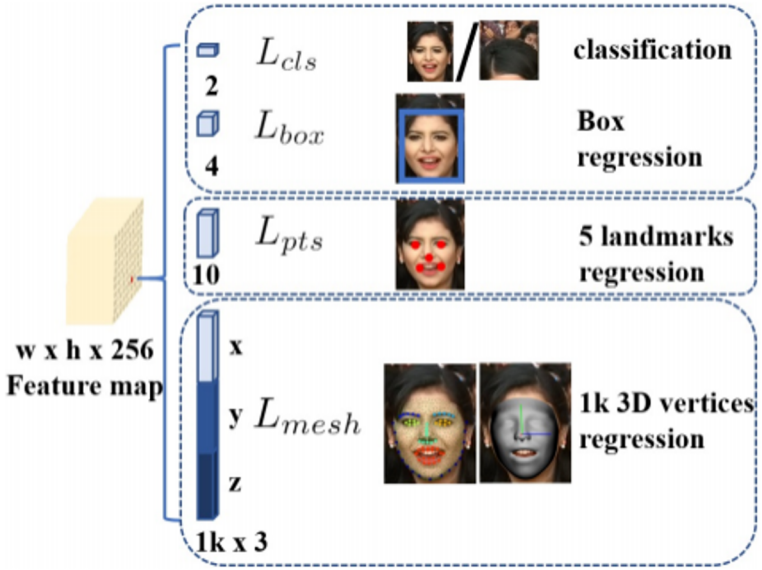
3.4 Extensions

RetinaFace

[Deng et al., CVPR 2020]



FPN



Multi-task branches

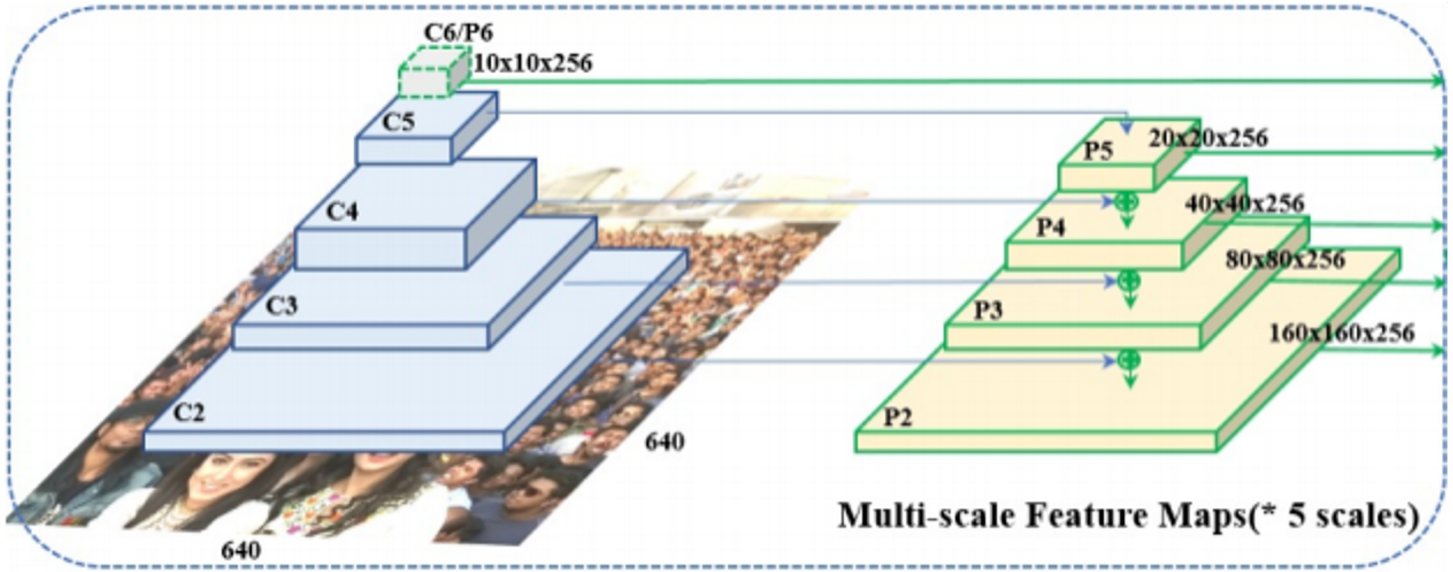
RetinaFace = FPN + Multi-task branches

(classification, bounding box, 5 point regression, mesh regression)

3.4 Extensions

Extension pattern

[Deng et al., CVPR 2020]



FPN

FPN + Target-task branches

4.

Detecting objects as keypoints

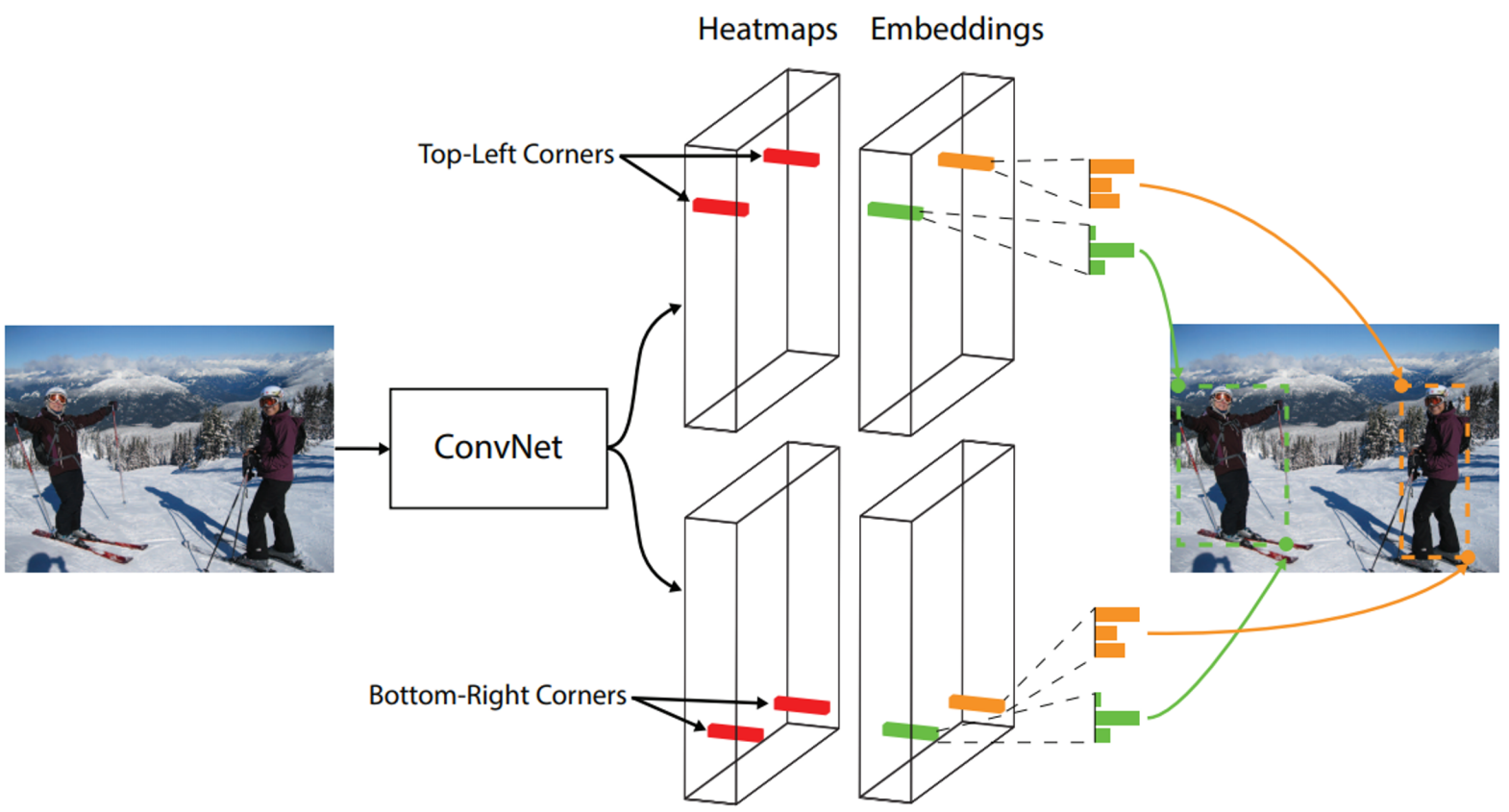
4.1 CornerNet & CenterNet

Detecting objects as keypoints

CornerNet

[Law et al., ECCV 2018]

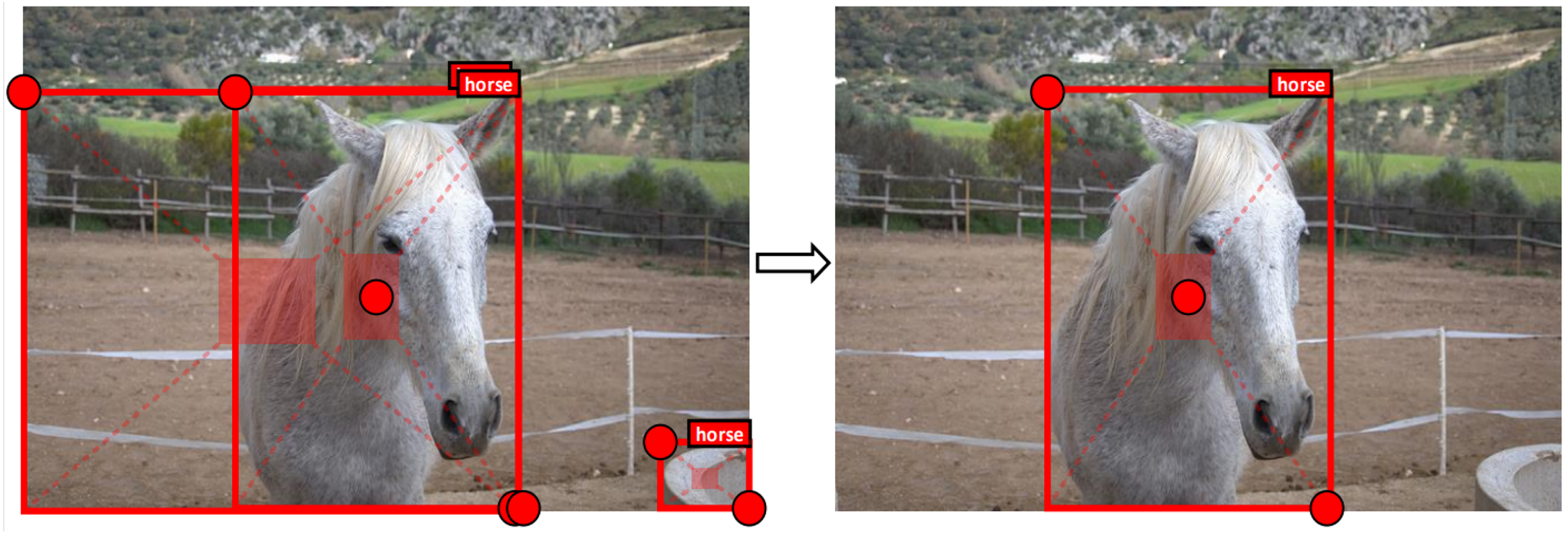
Bounding box = {Top-left, Bottom-right} corners



CenterNet (1)

[Duan et al., ICCV 2019]

Bounding box = {Top-left, Bottom-right, Center} points



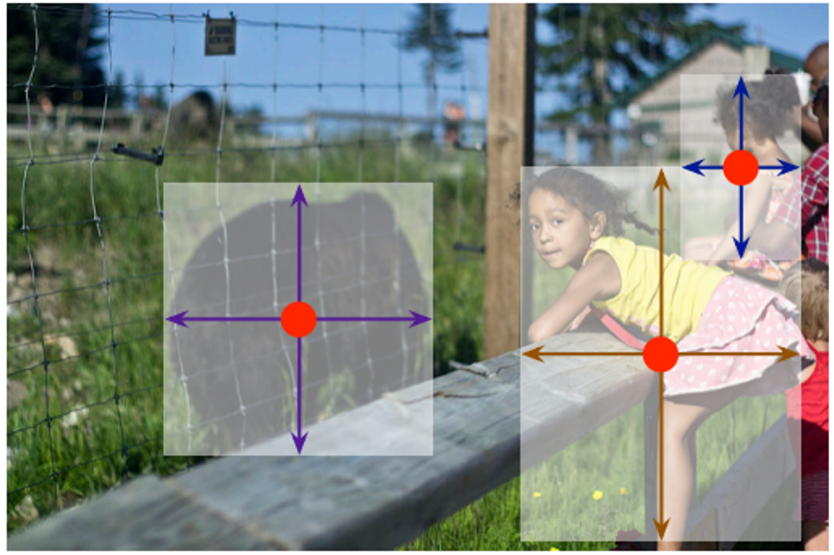
Center helps final decision!

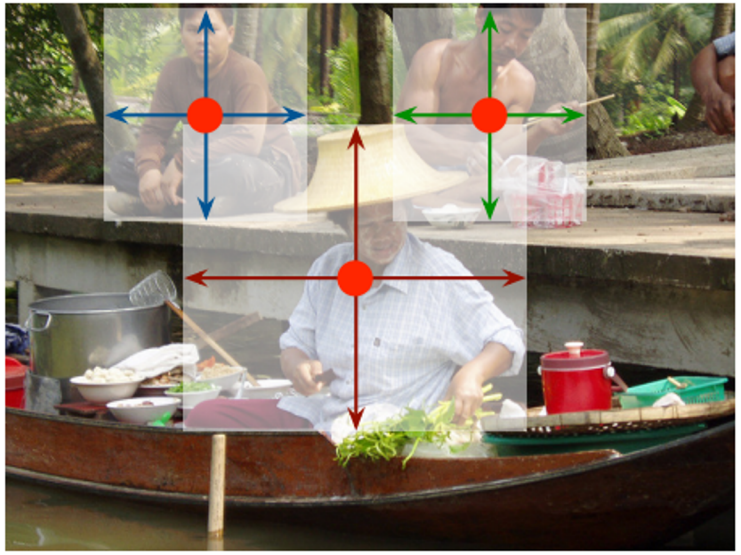
4.1 CornerNet & CenterNet

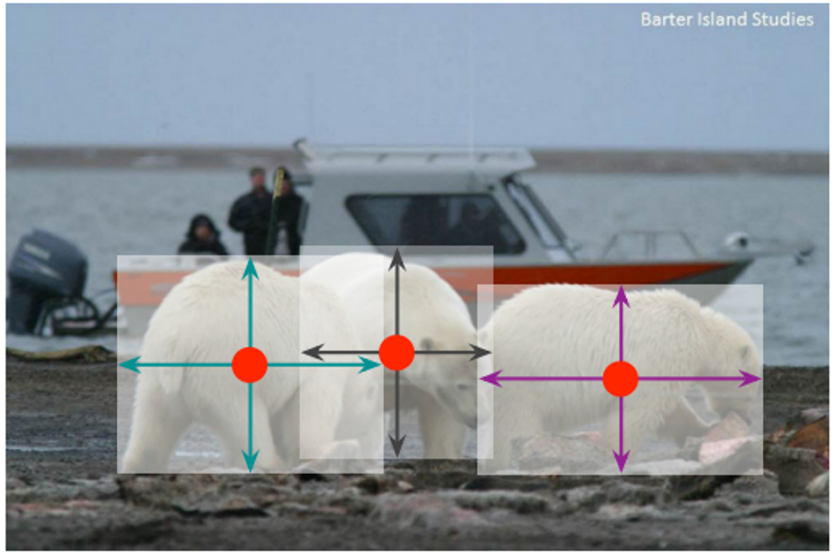
CenterNet (2)

[Zhou et al., arXiv 2019]

Bounding box = {Width, Height, Center} points

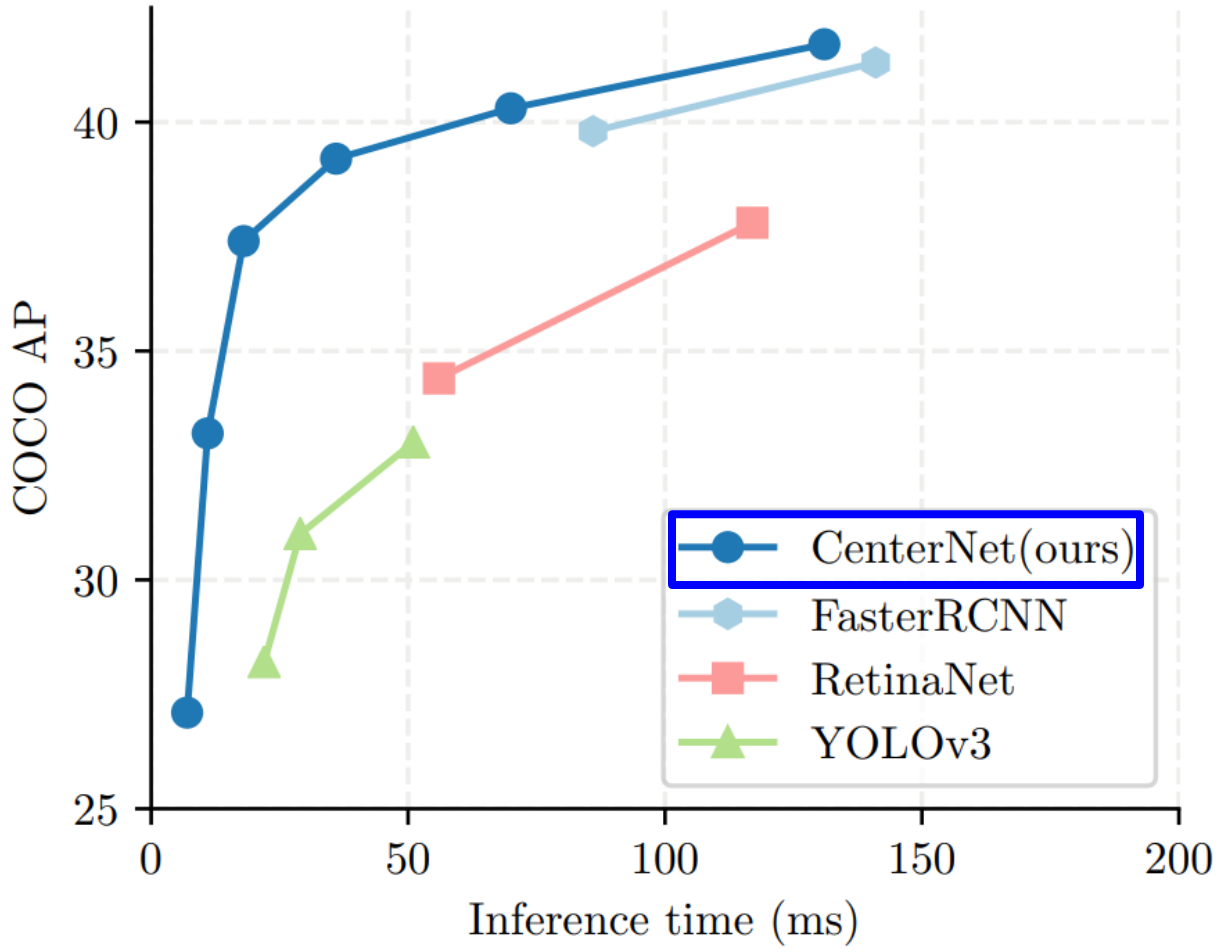






CenterNet (2)

[Zhou et al., arXiv 2019]



Reference

1. Instance segmentation

- Kirillov et al., Panoptic segmentation, CVPR 2019
- He et al., Mask R-CNN, ICCV 2017
- Bolya et al., YOLACT Real-time Instance Segmentation, ICCV 2019
- Liu et al., YolactEdge: Real-time Instance Segmentation on the Edge (Jetson AGX Xavier: 30 FPS, RTX 2080 Ti: 170 FPS), arXiv 2020

2. Panoptic segmentation

- Xiong et al., UPSNet: A Unified Panoptic Segmentation Network, CVPR 2019
- Kim et al., Video Panoptic Segmentation, CVPR 2020

3. Landmark localization

- Cao et al., OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, IEEE TPAMI 2019
- Jin et al., Pixel-in-Pixel Net: Towards Efficient Facial Landmark Detection in the Wild, arXiv 2020
- Wang et al., Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression, ICCV 2019
- Newell et al., Stacked Hourglass Networks for Human Pose Estimation, ECCV 2016
- Guler et al., DensePose: Dense Human Pose Estimation in the Wild, CVPR 2018

Reference

4. Detecting objects as keypoints

- Law et al., CornerNet: Detecting Objects as Paired Keypoints, ECCV 2018
- Duan et al., CenterNet: Keypoint Triplets for Object Detection, ICCV 2019
- Zhou et al., Objects as Points, arXiv 2019

End of Document
Thank You.