Video Instance Segmentation (VIS)

ICCV 2019

Lingjie Yang, Yuchen Fan, Ning Xu

Presented by Amit, Michael, and Jun

VIS takes inspiration from the image domain

Motivation

- Detects boundaries of objects
- Classifies objects
- Demarcates separate instances of each class
- Each pixel is assigned a specific object instance (or to the background)

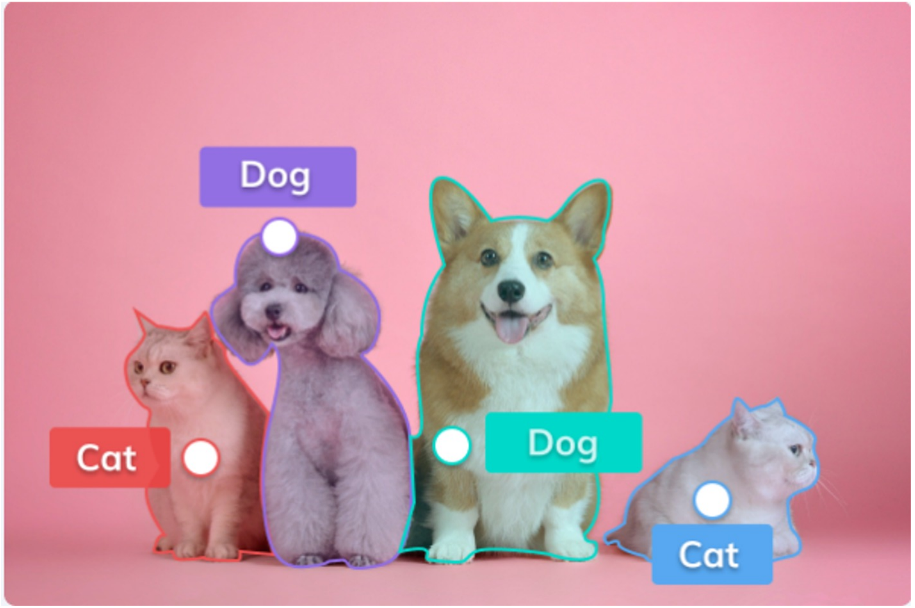


Image Instance Segmentation

Motivation

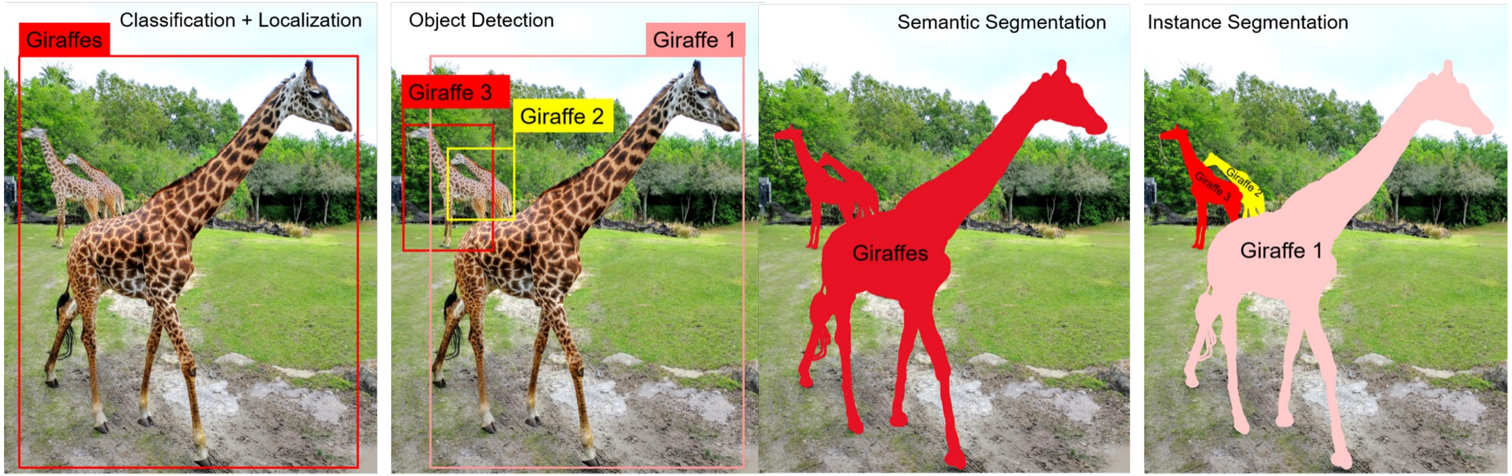
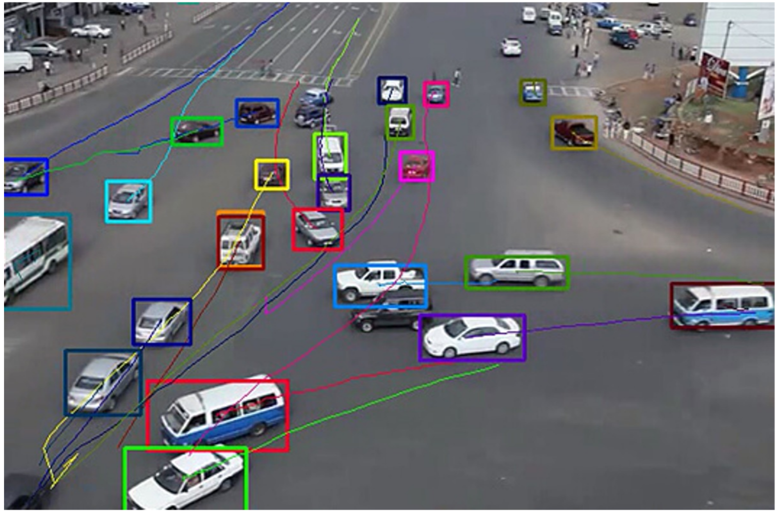


Image Instance Segmentation **combines** ideas of other image-based tasks

VIS brings this idea to the video domain

Past Work

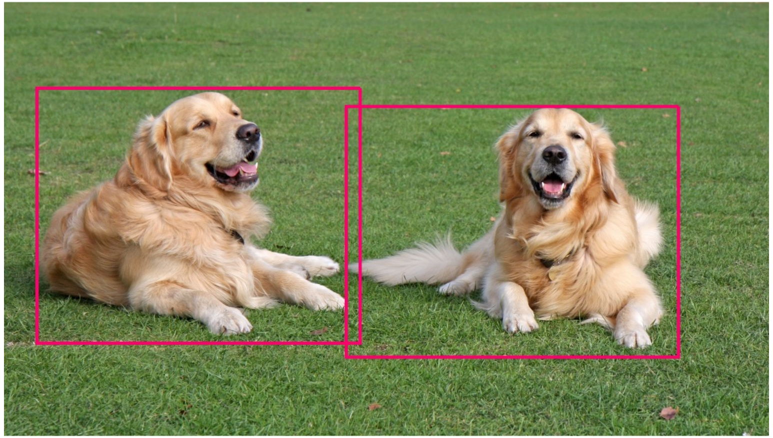
Video Object Tracking	Video Object Detection	Video Semantic Segmentation	Video Object Segmentation



Track objects in a video given their initial bounding box

Past Work

Video Object Tracking | **Video Object Detection** | Video Semantic Segmentation | Video Object Segmentation



Detect objects within a video without any initialization

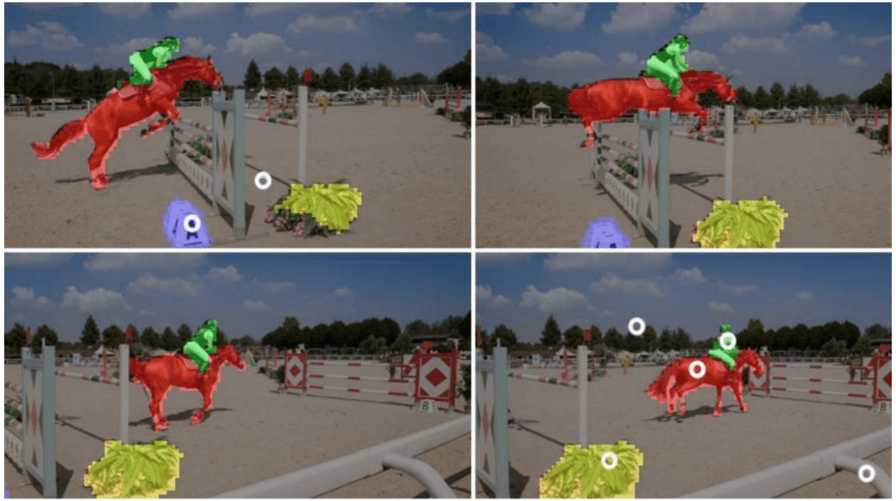
Past Work

Video Object Tracking | Video Object Detection | **Video Semantic Segmentation** | Video Object Segmentation

Image pixels are predicted as different semantic classes
to understand objects and regions in a video

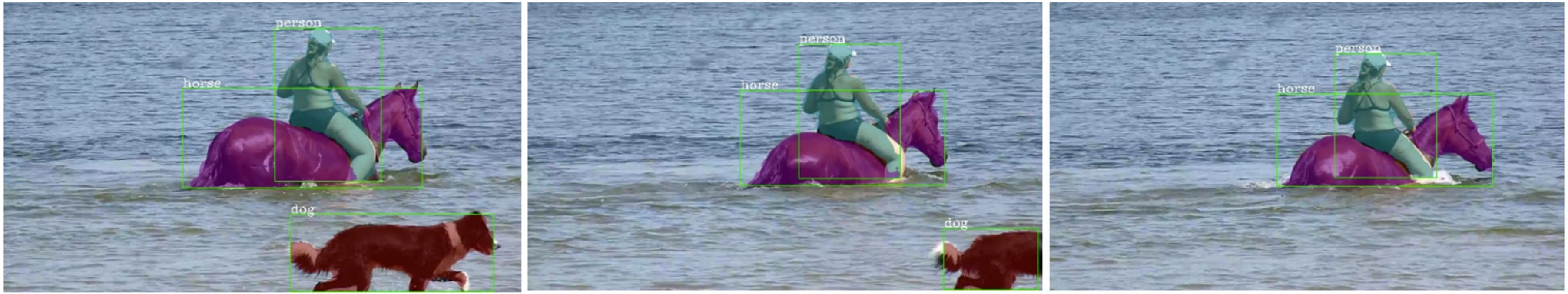
Past Work

Video Object Tracking | Video Object Detection | Video Semantic Segmentation | Video Object Segmentation



Segment the object from the background and follow changes in movement

Video Instance Segmentation



Simultaneous detection, segmentation, and tracking of object instances in videos across frames

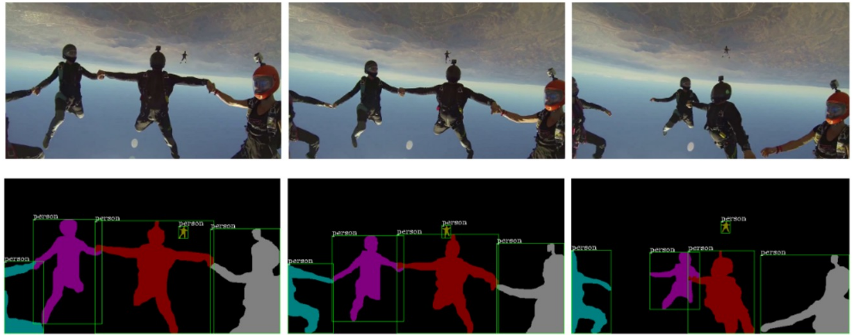
To embark on a new research field, you need

1. A newly annotated benchmark that provides temporal instance labels.
   a. No existing large-scale dataset can serve the purpose of VIS.

1. A newly designed model that can do
   a. Object detection
   b. Instance segmentation (object classification + segmentation)
   c. Instance tracking

The dataset: Youtube-VIS

Youtube-VOS → Youtube-VIS



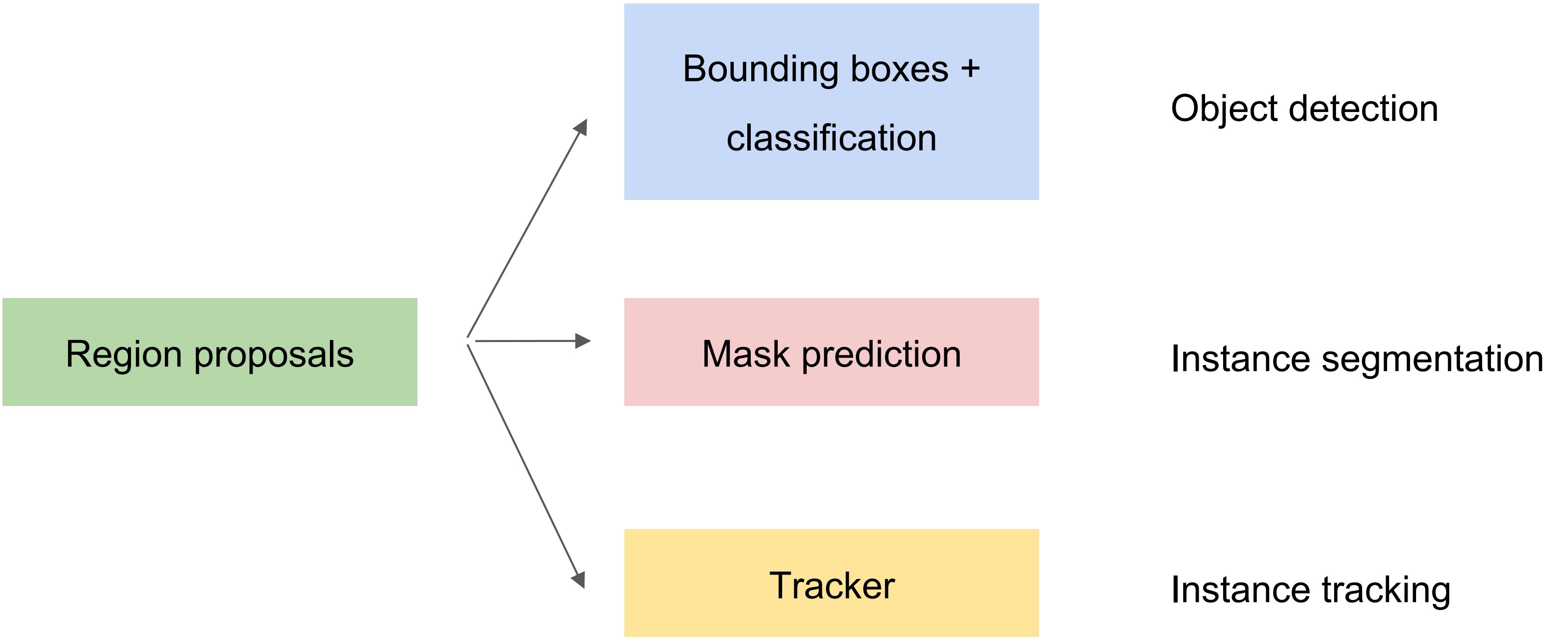
- 4,453 youtube videos
- 94 categories
- 6,048 objects
- Object masks are not exhaustive

Selected:

- ~ 2,900 videos
- 40 categories
- 4,883 unique objects
- Exhaustively annotated

The model: Mask-Track R-CNN

Model breakdown

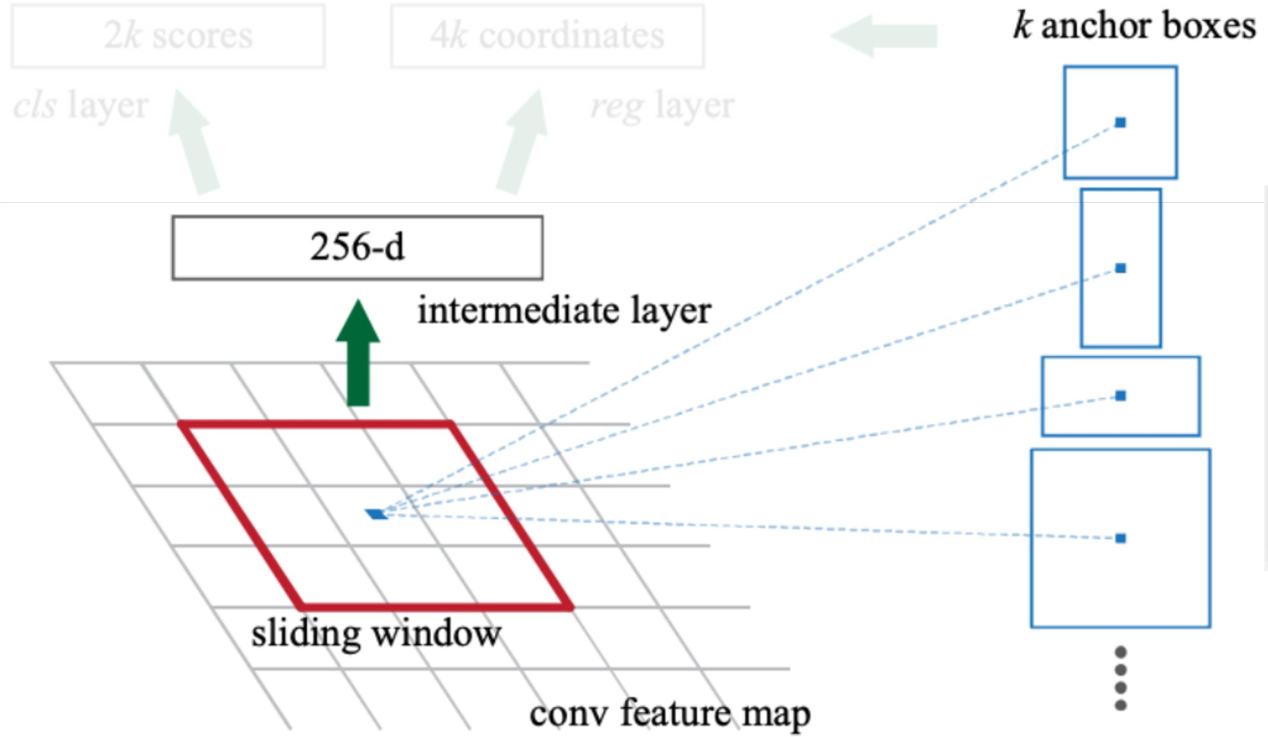


Region proposals

Bounding boxes + classification

Mask prediction

Tracker

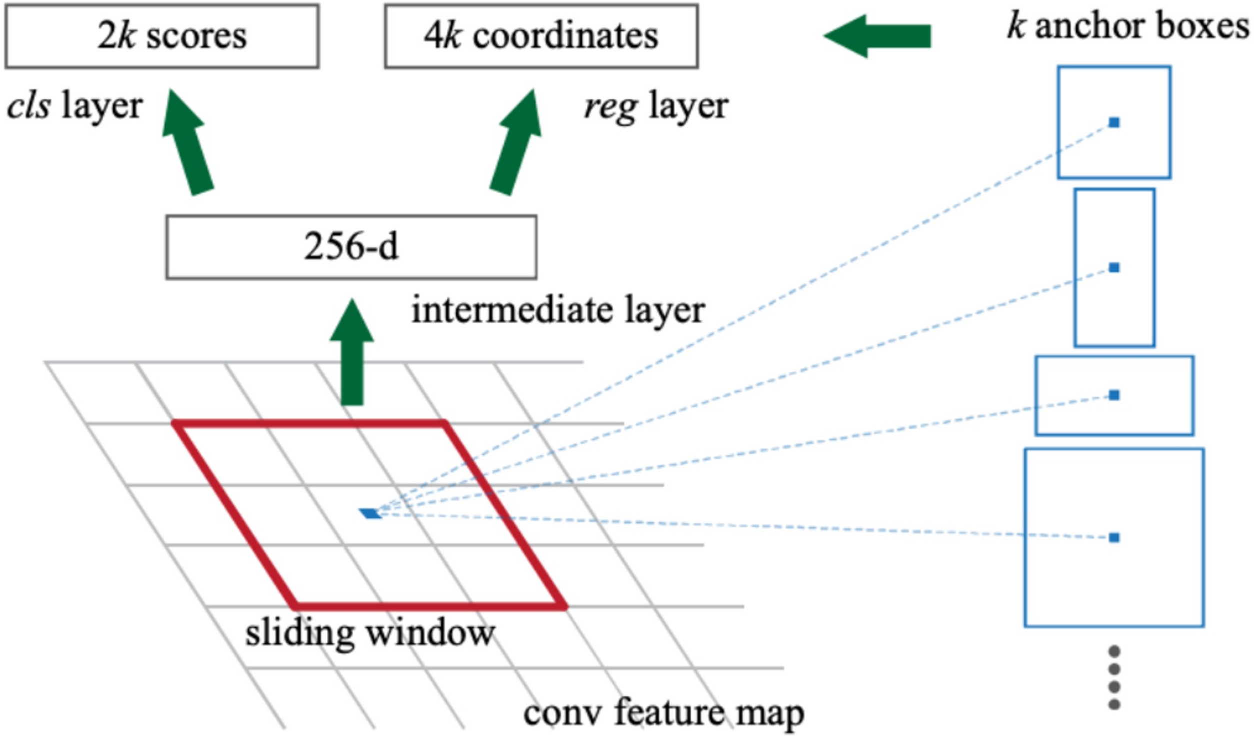


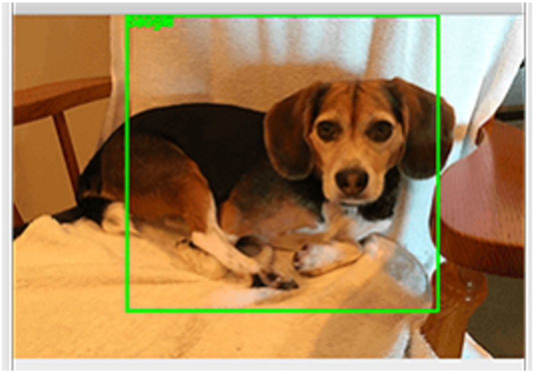
Region proposals

Bounding boxes + classification

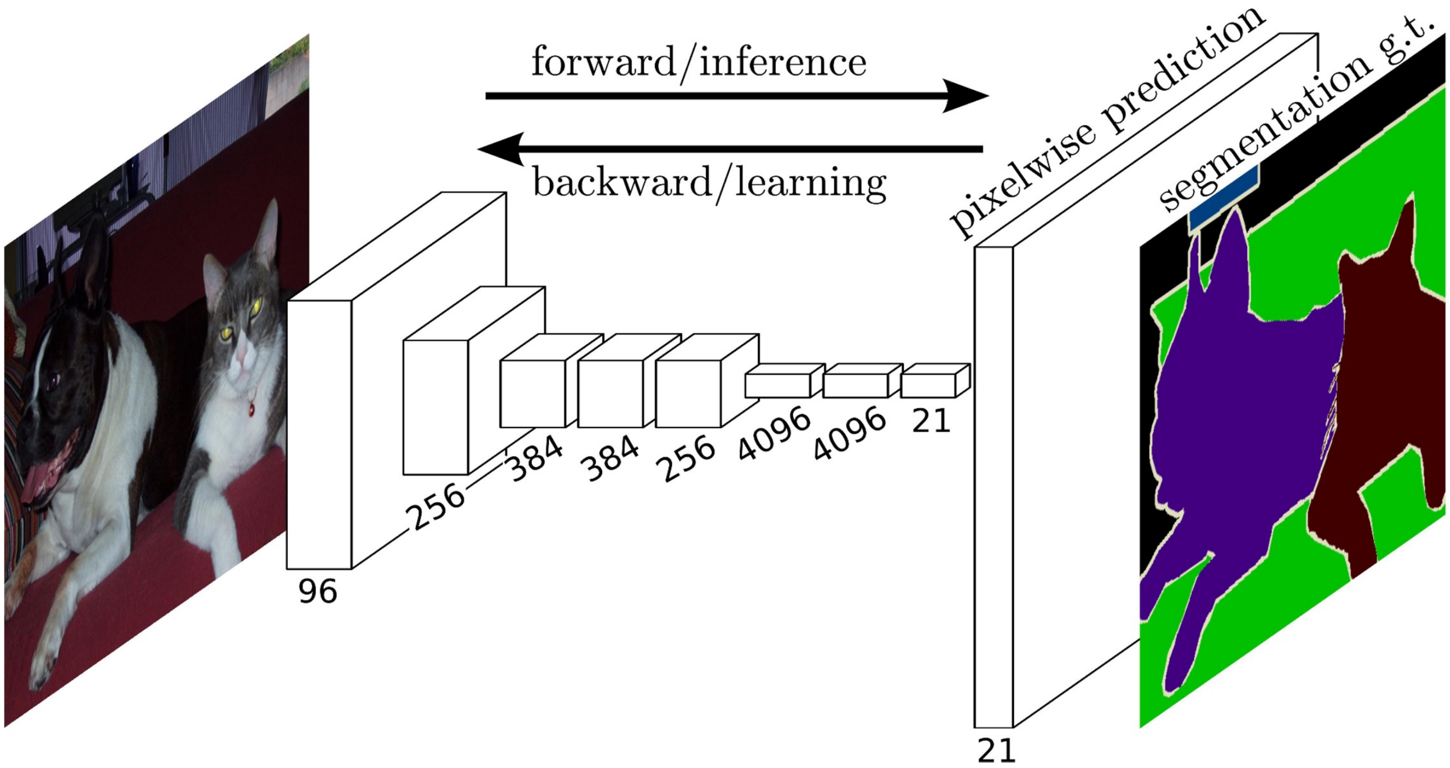
Mask prediction

Tracker





Region proposals | Bounding boxes + classification | **Mask prediction** | Tracker



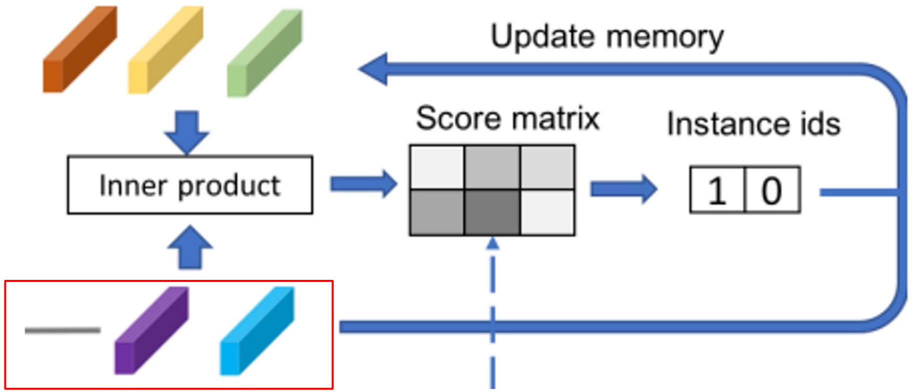
Region proposals

Bounding boxes + classification

Mask prediction

Tracker

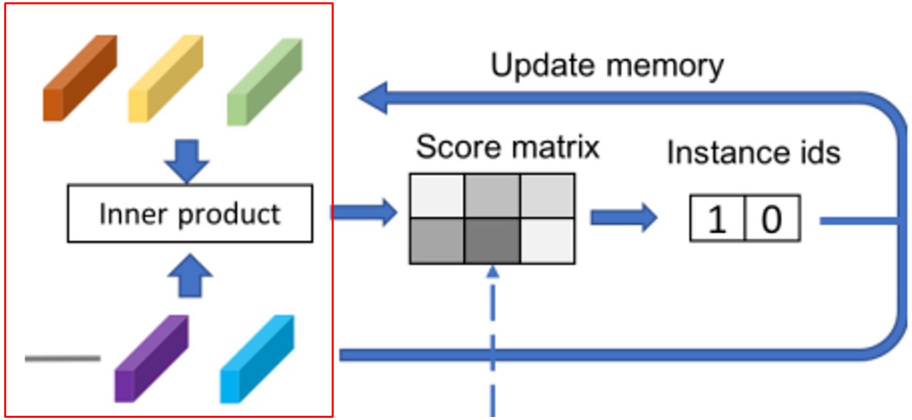
1. Extract feature vectors from the current frame



Region proposals | Bounding boxes + classification | Mask prediction | **Tracker**

2. Similarity comparison (dot product)

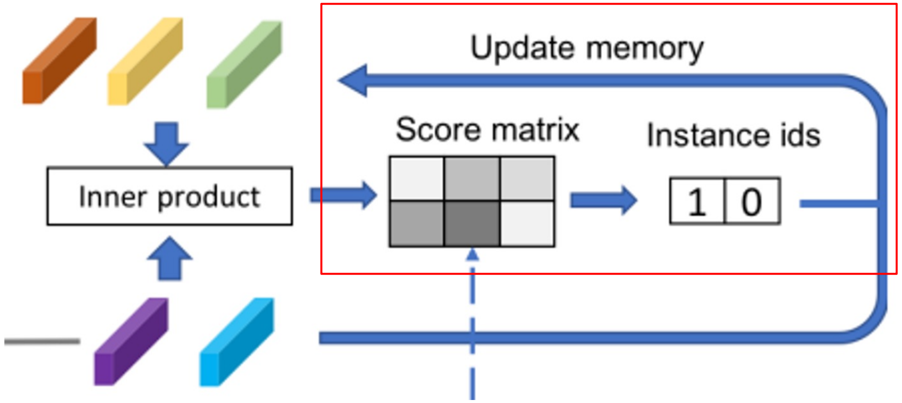
Feature vectors from previous frames

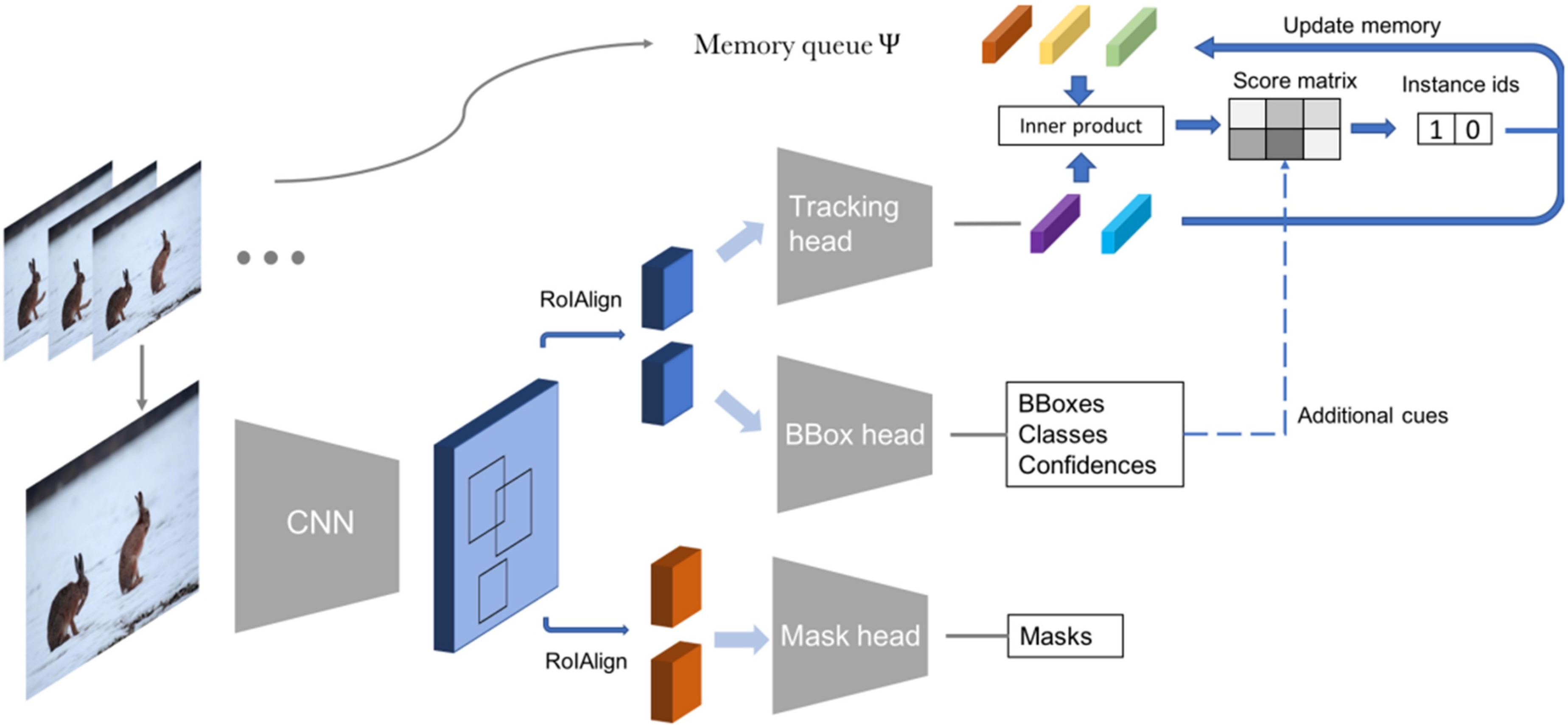


Feature vectors from the current frame

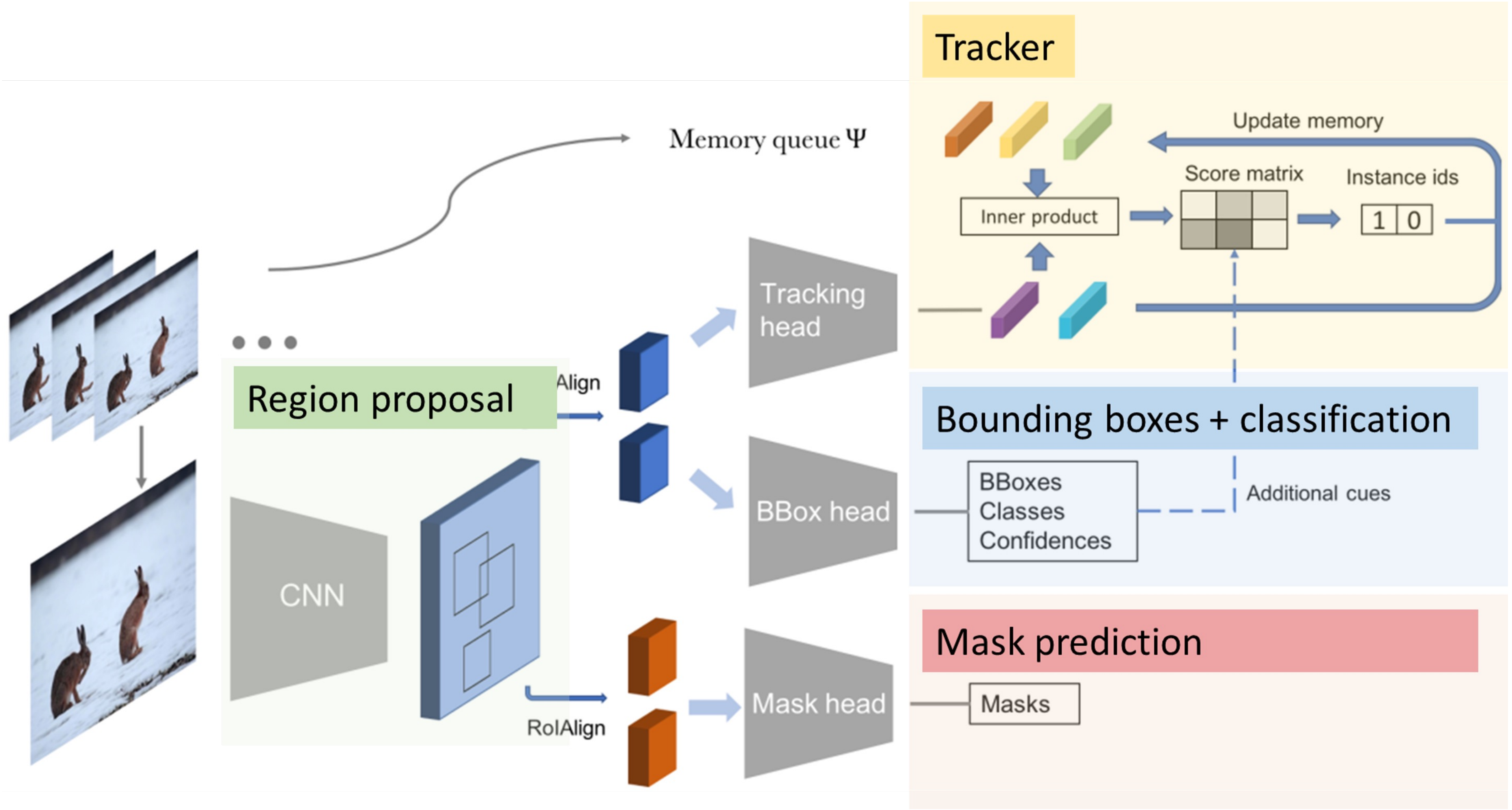
Region proposals | Bounding boxes + classification | Mask prediction | **Tracker**

3. Assign instance labels and update the memory bank





Memory queue Ψ
Update memory
Inner product
Score matrix
Instance ids
1 0
Tracking head
RoIAlign
BBox head
BBoxes
Classes
Confidences
Additional cues
CNN
Mask head
Masks
RoIAlign



Tracker
Memory queue Ψ
Update memory
Score matrix
Instance ids
Inner product
1 0
Tracking head
Region proposal
Align
BBox head
Bounding boxes + classification
BBoxes
Classes
Confidences
Additional cues
CNN
Mask prediction
Mask head
Masks
RoIAlign

Metrics

1. Average Precision (AP) : the area under the precision-recall curve
    a. Precision : TP / (TP + FP)
    b. Recall : TP / (TP + FN)
    c. Intersection-over-union (IOU)
    d. Precision-recall curve

1. Average Recall (AR)

Metrics

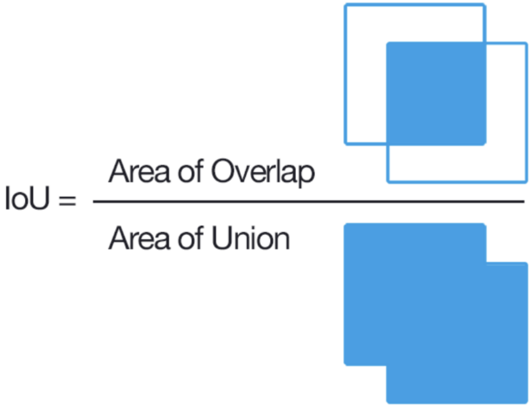
1. Average Precision (AP) : the area under the precision-recall curve
    a. Precision : TP / (TP + FP)
    b. Recall : TP / (TP + FN)
    c. Intersection-over-union (IOU)
    d. Precision-recall curve

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

1. Average Recall (AR)

Metrics

1. Average Precision (AP) : the area under the precision-recall curve
    a. Precision : TP / (TP + FP)
    b. Recall : TP / (TP + FN)
    c. Intersection-over-union (IOU)
    d. Precision-recall curve

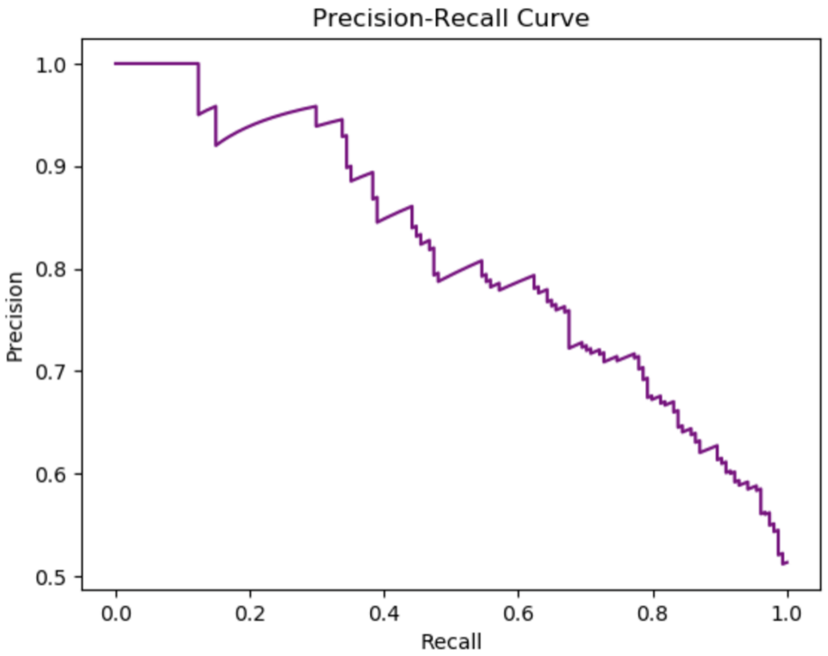


1. Average Recall (AR)

$$\mathrm{IoU}(i,j) = \frac{\Sigma_{t=1}^{T} |\mathbf{m}_t^i \cap \tilde{\mathbf{m}}_t^j|}{\Sigma_{t=1}^{T} |\mathbf{m}_t^i \cup \tilde{\mathbf{m}}_t^j|}$$

Metrics

1. Average Precision (AP) : the area under the precision-recall curve
    a. Precision : TP / (TP + FP)
    b. Recall : TP / (TP + FN)
    c. Intersection-over-union (IOU)
    d. Precision-recall curve



1. Average Recall (AR)

Metrics

1. Average Precision (AP) : the area under the precision-recall curve
    a. Precision : TP / (TP + FP)
    b. Recall : TP / (TP + FN)
    c. Intersection-over-union (IOU)
    d. Precision-recall curve

2. Average Recall (AR) : the maximum recall given some fixed number of segmented instance per video
    a. Recall : TP / (TP + FN)

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Metrics

1. Average Precision (AP) : the area under the precision-recall curve
    a. Precision : TP / (TP + FP)
    b. Recall : TP / (TP + FN)
    c. Intersection-over-union (IOU)
    d. Precision-recall curve

2. Average Recall (AR) : the maximum recall given some fixed number of segmented instance per video
    a. Recall : TP / (TP + FN)

    Evaluated per category -> averaged over the category set

    Higher is better

Result & Experiments The video-level prediction corrects these mis- takes by majority voting of all frames.

1. Video-level prediction corrects mistakes by majority voting of all frames

1. Track the object after it disappears and reoccurs

Result & Experiments

Quantitative comparison to others

Methods		validation set					test set				
		AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Mask propagation	OSMN [36]	23.4	36.5	25.7	28.9	31.1	27.3	44.4	28.0	28.8	34.0
	FEELVOS [31]	26.9	42.0	29.7	29.9	33.4	29.6	45.4	30.7	33.4	36.8
Track-by-detect	IoUTracker+	23.6	39.2	25.5	26.2	30.9	25.2	41.9	26.2	28.7	33.7
	OSMN [36]	27.5	45.1	29.1	28.6	33.1	27.3	44.4	28.0	28.8	34.0
	DeepSORT [33]	26.1	42.9	26.1	27.8	31.3	27.2	44.0	29.2	29.1	33.3
	SeqTracker	27.5	45.7	28.7	29.7	32.5	29.5	48.1	31.2	32.0	34.5
	MaskTrack R-CNN	**30.3**	**51.1**	**32.6**	**31.0**	**35.5**	**32.3**	**53.6**	**34.2**	**33.6**	**37.3**

Thank you