
Deep Learning for Video Instance Segmentation

Jianping Fan
Department of Computer Science
UNC-Charlotte

Course Website:

<http://webpages.uncc.edu/jfan/itcs5152.html>

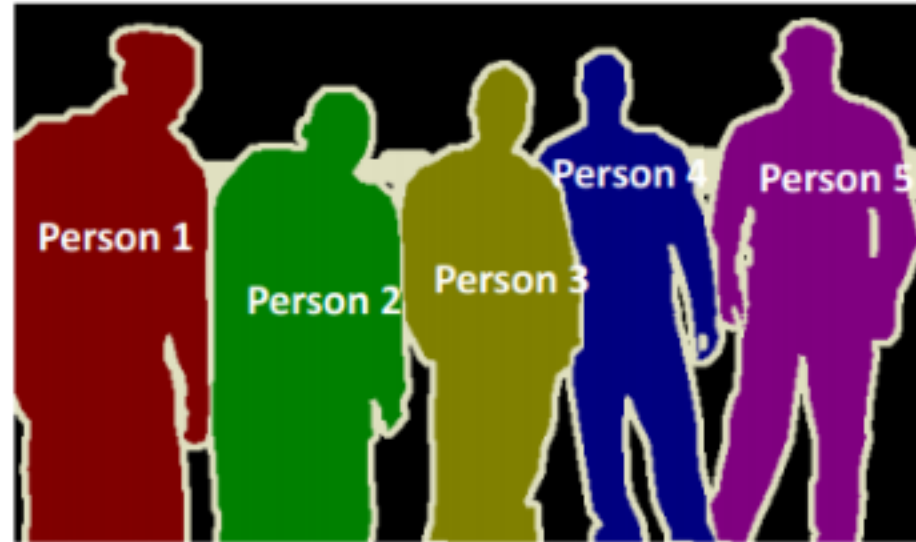
The approaches for video instance segmentation can be categorized into two groups:

- (a) frame-by-frame approaches by performing instance segmentation on every frame;
- (b) segmentation+tracking approaches by performing instance segmentation on the first frame and following by temporal tracking along frames.

Instance Segmentation vs. Semantic Segmentation



Semantic Segmentation



Instance Segmentation

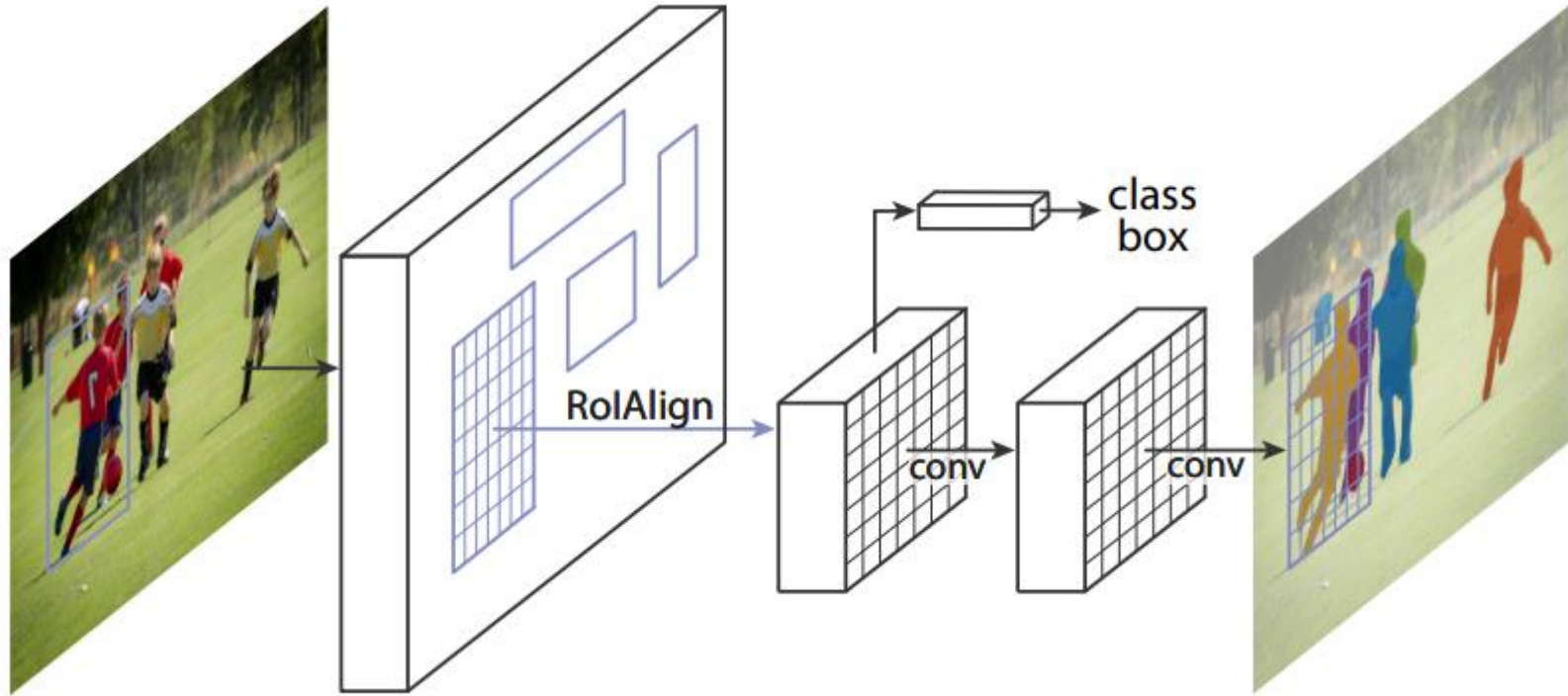
One more operation on identification should be added.

Mask RCNN, ICCV 2017

Mask R-CNN efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. Mask R-CNN extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN, running at 5 fps.

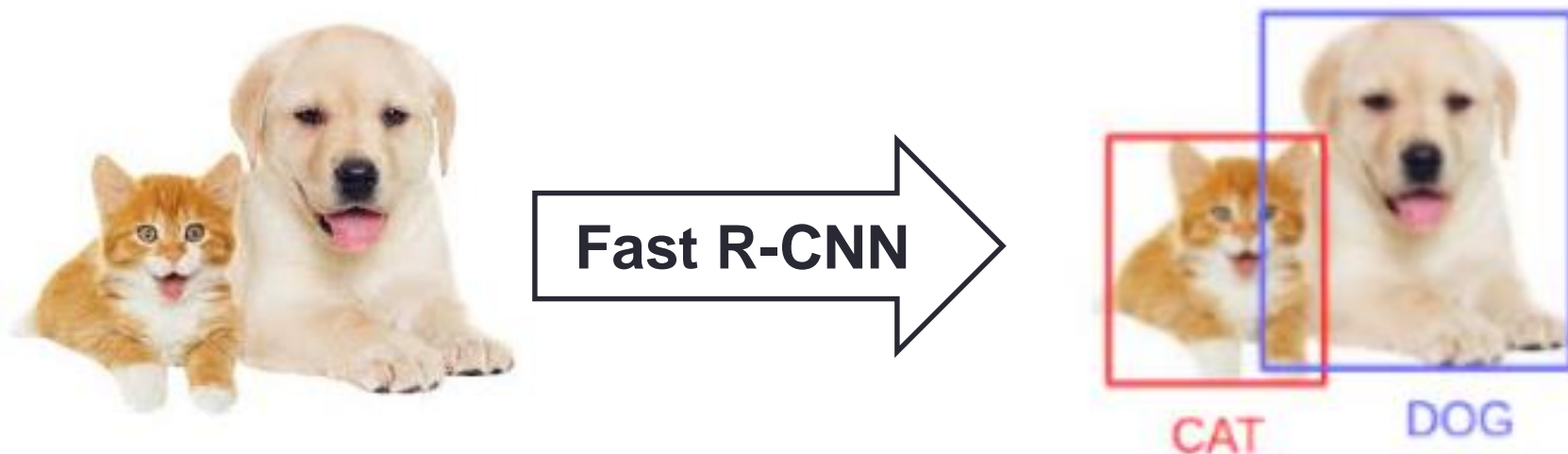
The added branch can predict segmentation masks on each Region of Interest (RoI), in parallel with the existing branch for classification and bounding box regression. The mask branch is a small FCN applied to each RoI, predicting a segmentation mask in a pixel-to-pixel manner. Mask R-CNN is simple to implement and train given the Faster R-CNN framework, which facilitates a wide range of flexible architecture designs. Additionally, the mask branch only adds a small computational overhead, enabling a fast system and rapid experimentation.

Mask RCNN, ICCV 2017



Mask RCNN, ICCV 2017

The Mask R-CNN framework is built on top of Faster R-CNN.



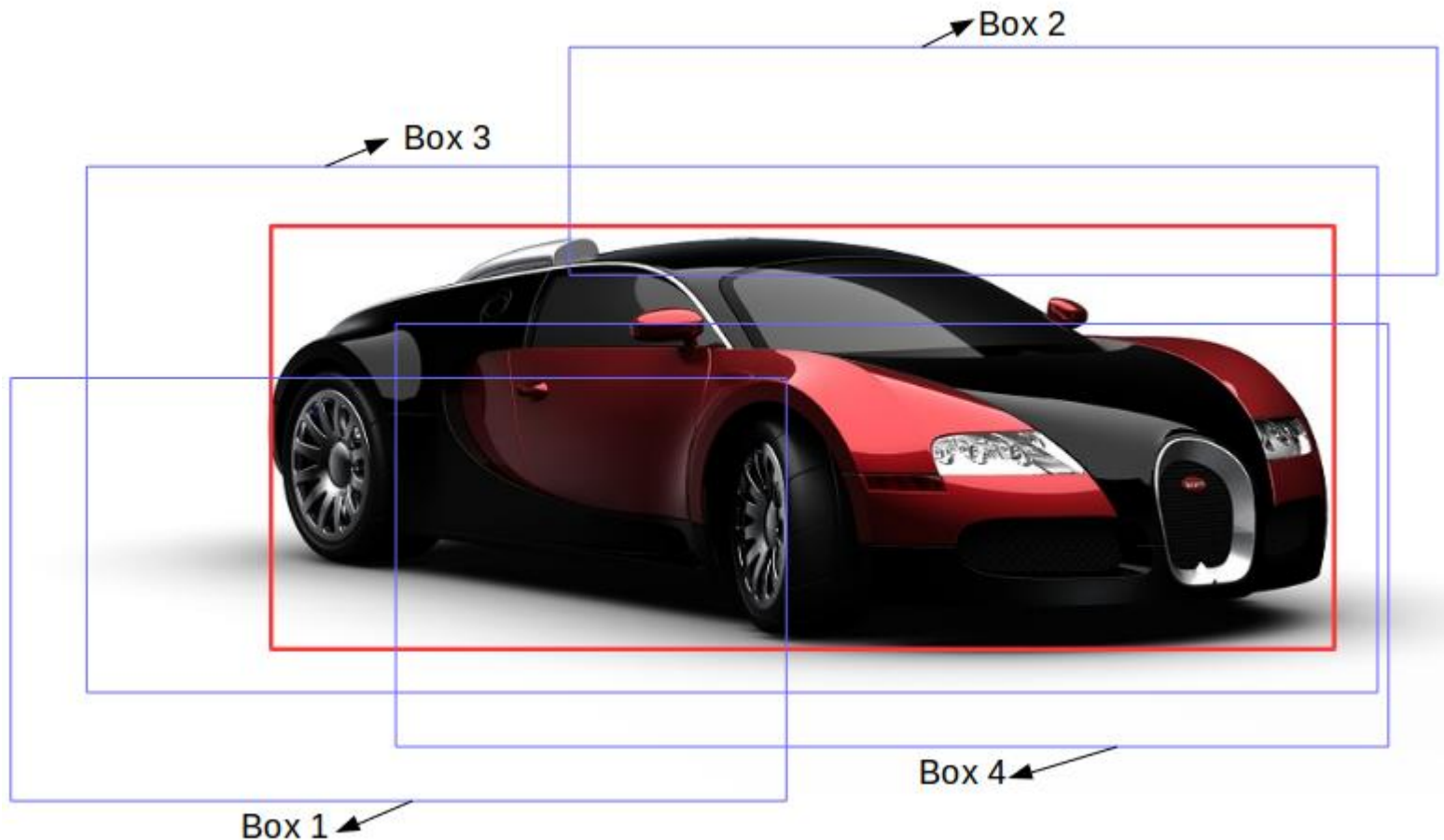
(a) Faster R-CNN first uses a ConvNet to extract feature maps from the images

(b) These feature maps are then passed through a Region Proposal Network (RPN) which returns the candidate bounding boxes

(c) An RoI pooling layer is applied on these candidate bounding boxes to bring all the candidates to the same size

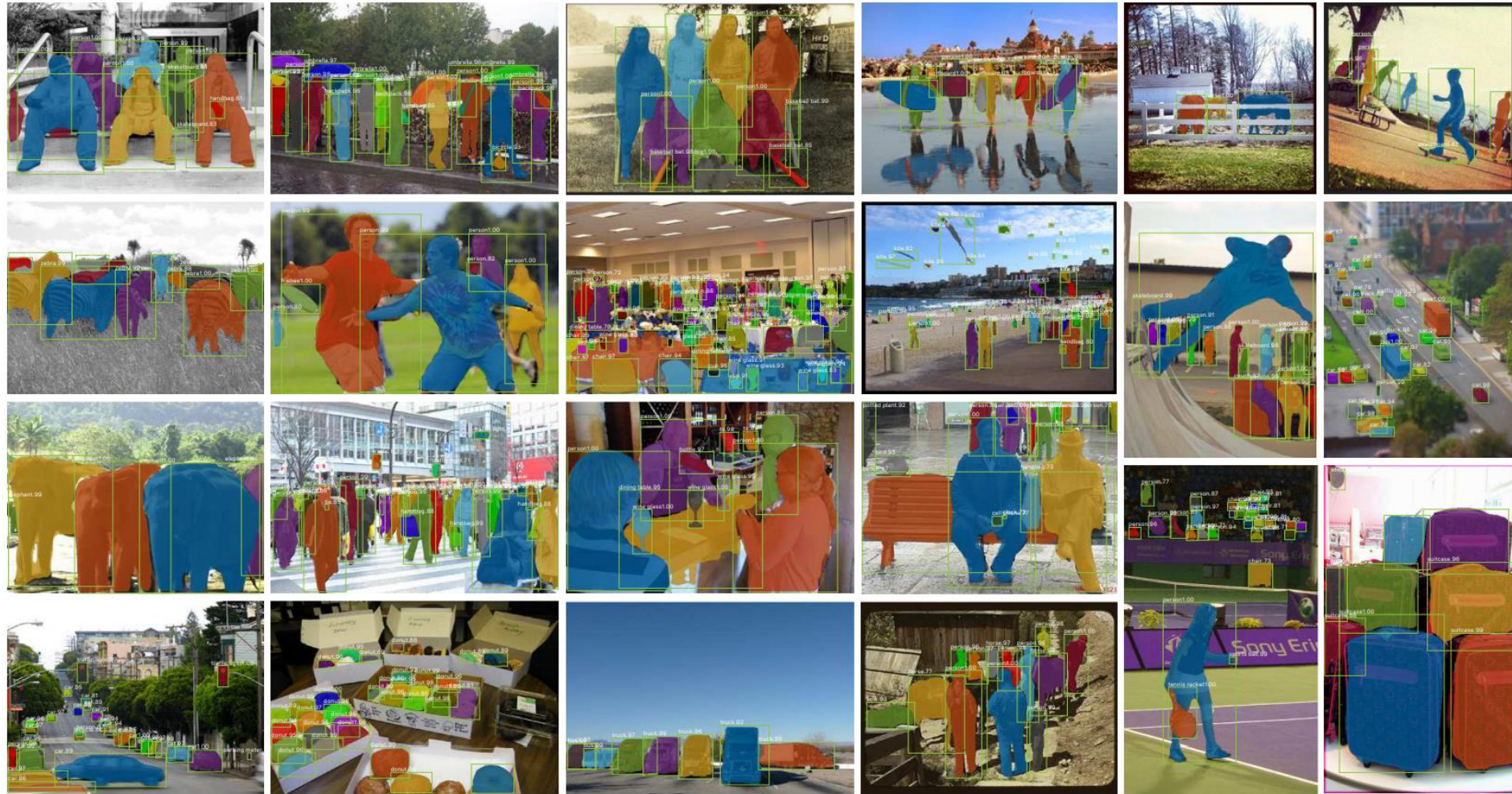
(d) Finally, the proposals are passed to a fully connected layer to classify and output the bounding boxes for objects

Mask RCNN, ICCV 2017



Only if the IoU is greater than or equal to 0.5, we consider that as a region of interest. Otherwise, we neglect that particular region. We do this for all the regions and then select only a set of regions for which the IoU is greater than 0.5.

Mask RCNN, ICCV 2017



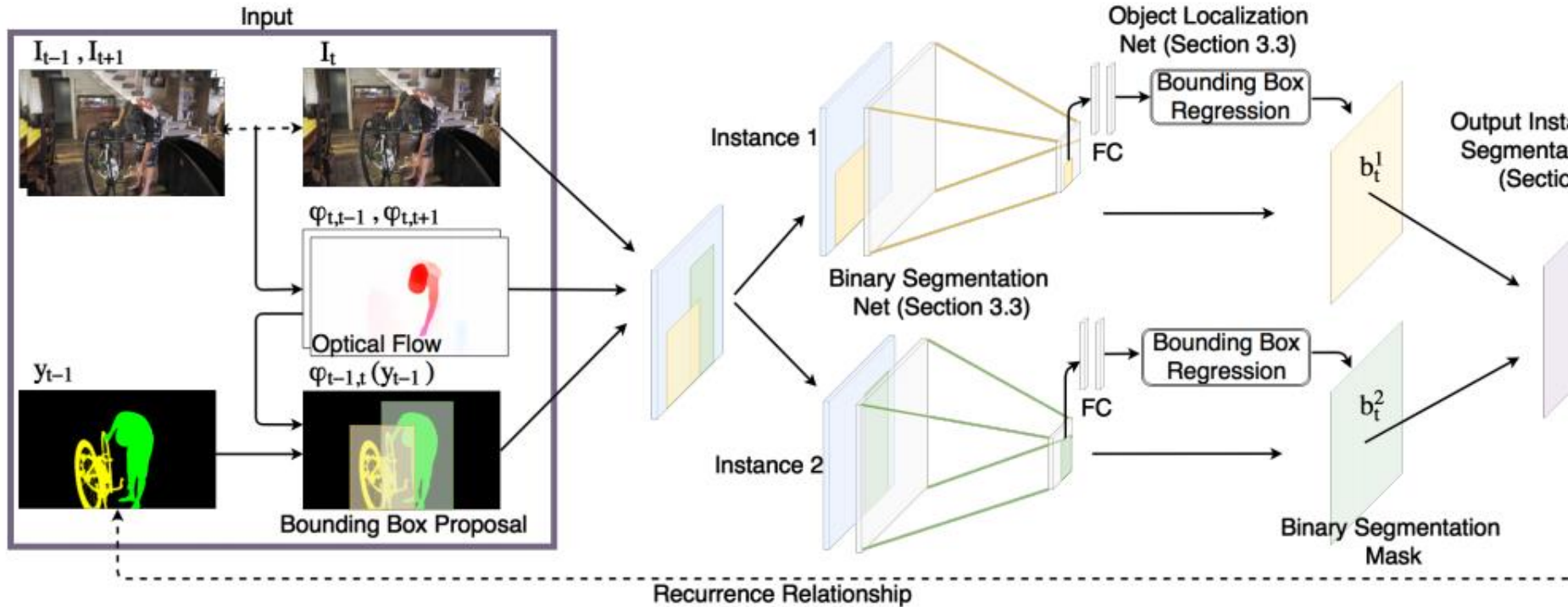
MaskRNN: Instance Level Video Object Segmentation, NIPS 2017

To capture the temporal coherence, MaskRNN uses a recurrent neural net approach which fuses in each frame the output of two deep nets for each object instance — a binary segmentation net providing a mask and a localization net providing a bounding box.

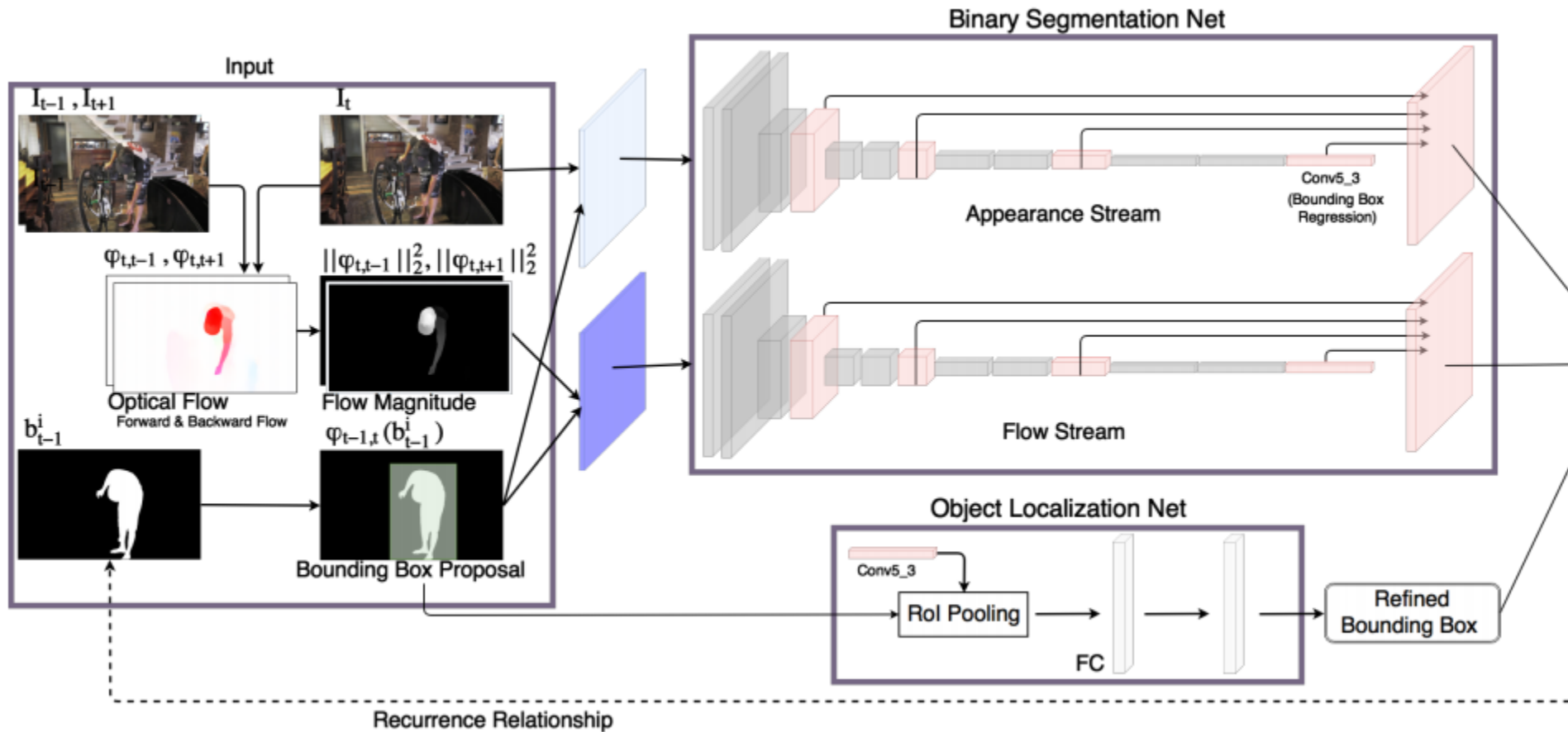
Due to the recurrent component and the localization component, our method is able to take advantage of long-term temporal structures of the video data as well as rejecting outliers.

MaskRNN performs instance level object segmentation by combining binary segmentation with effective object tracking via bounding boxes. To benefit from temporal dependencies, MaskRNN employs a recurrent neural net component to connect prediction over time in a unifying framework.

MaskRNN: Instance Level Video Object Segmentation, NIPS 2017



MaskRNN: Instance Level Video Object Segmentation, NIPS 2017

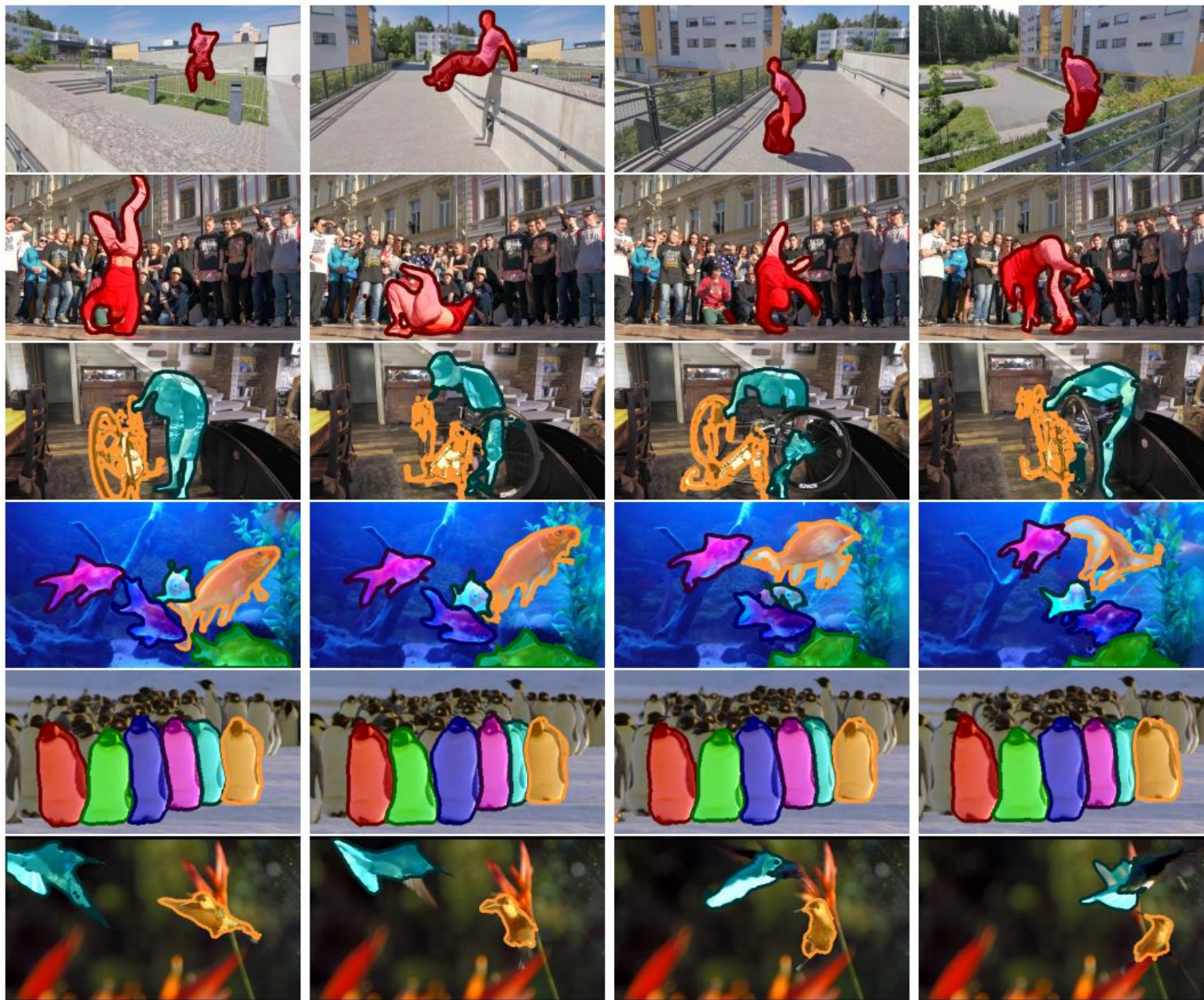


MaskRNN: Instance Level Video Object Segmentation, NIPS 2017

Binary Segmentation Net: The objective for each of the binary segmentation nets is to predict the foreground-background mask $b_t^i \in [0, 1]^{H \times W}$ for its corresponding object instance $i \in \{1, \dots, N\}$. To achieve this task, the binary segmentation net is split into two streams, *i.e.*, the appearance stream and the flow stream. The input of the appearance stream is the concatenation of the current frame I_t and the warped prediction of the previous frame y_{t-1} , denoted as $\phi_{t-1,t}(y_{t-1})$. The warping function $\phi_{t-1,t}(\cdot)$ transforms the input based on the optical flow field from frame I_{t-1} to frame I_t . The input of the flow stream is the concatenation of the magnitude of the flow field from I_t to I_{t-1} and I_t to I_{t+1} and, again, the warped prediction of the previous frame $\phi_{t-1,t}(y_{t-1})$. The architecture of both streams is identical and follows the subsequent description.

Object Localization Net: Usage of an object localization net is inspired by tracking approaches which regularize the prediction by assuming that the object is less likely to move drastically between temporally adjacent frames. The object localization network computes the location for the i -th object in the current frame via bounding box regression. First, we find the bounding box proposal on the warped mask $\phi_t(b_{t-1}^i)$. Similarly to the bounding box regression in Fast-RCNN [15], with the bounding box proposal as the region of interest, we use the conv5_3 feature in the appearance stream of the segmentation net to perform RoI-pooling, followed by two fully connected layers. Their output is used to regress the bounding box position. We refer the reader to [15] for more details on bounding box regression.

MaskRNN: Instance Level Video Object Segmentation, NIPS 2017



Fast Video Object Segmentation by Reference-Guided Mask Propagation, CVPR 2018

Video object segmentation methods typically rely on two important cues: (a) Object Appearances; (b) Object Motions.

(1) Propagation-based methods mainly leverage the temporal coherence of object motion and formulate this problem as object mask propagation (i.e. pixel-level tracking) starting from a given annotated frame. These methods rely on the spatiotemporal connections between pixels, and thus can adapt to complex deformation and movement of a target object as long as the changes in the appearance and the location are smooth. However, these methods are vulnerable to temporal discontinuities like occlusions and rapid motion, and can suffer from drifting once the propagation becomes unreliable.

Fast Video Object Segmentation by Reference-Guided Mask Propagation, CVPR 2018

(2) Detection-based methods learn the appearance of the target object from a given annotated frame, and perform a pixel-level detection of the target object at each frame. As they rarely depend on temporal consistency, they are robust to occlusion and drifting. However, as their estimation is mostly based on the object appearance in an annotated frame(s), they often fail to adapt to appearance changes and have difficulty separating object instances with similar appearances.

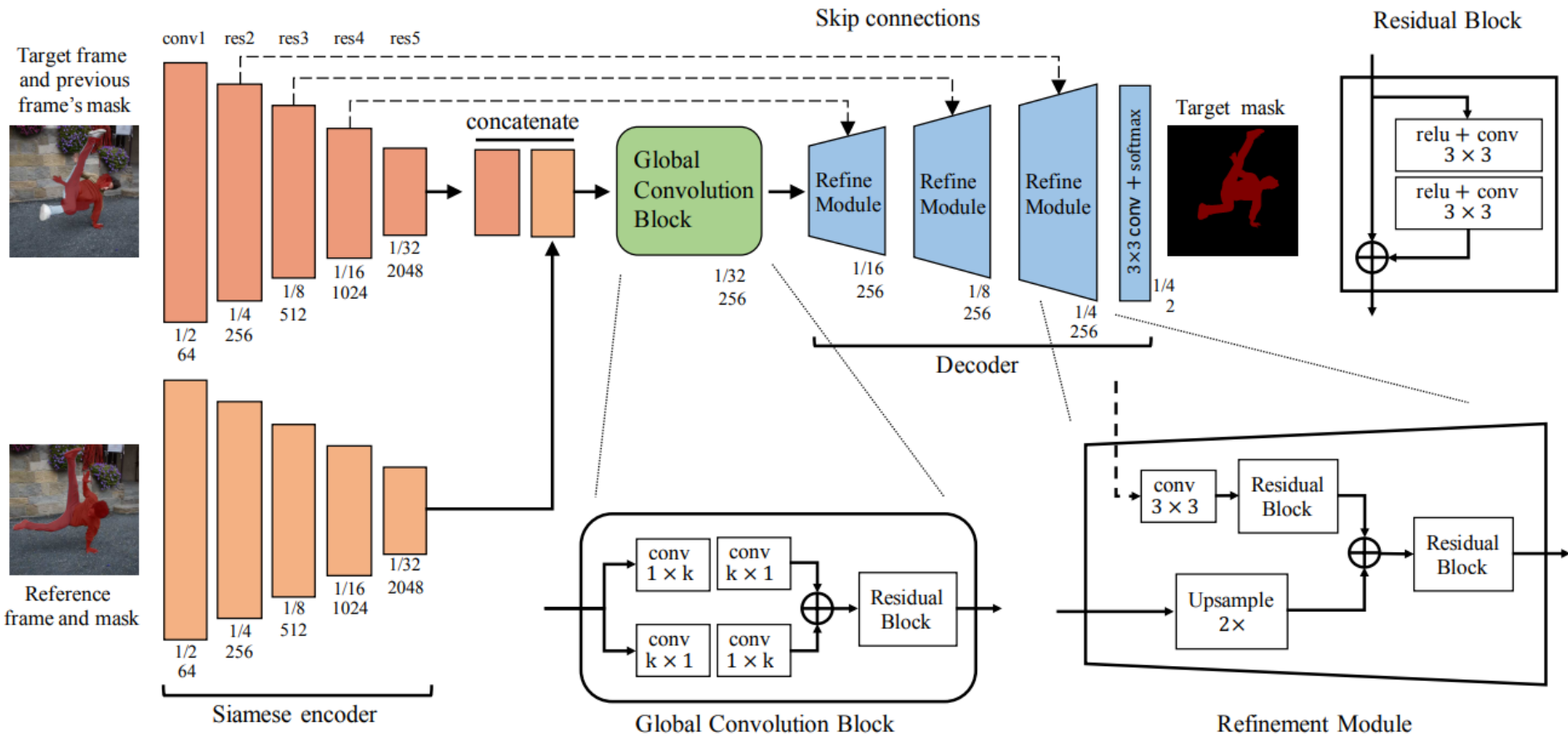
Fast Video Object Segmentation by Reference-Guided Mask Propagation, CVPR 2018

(3) Hybrid method: This paper presents a new hybrid method for semi-supervised video object segmentation, where a Siamese encoder-decoder network is constructed that simultaneously makes use of both the previous mask to be propagated to the current frame and the reference frame which specifies the target object to be detected in the current frame. The network is designed to generate a sharp object mask without time-consuming post-processing.

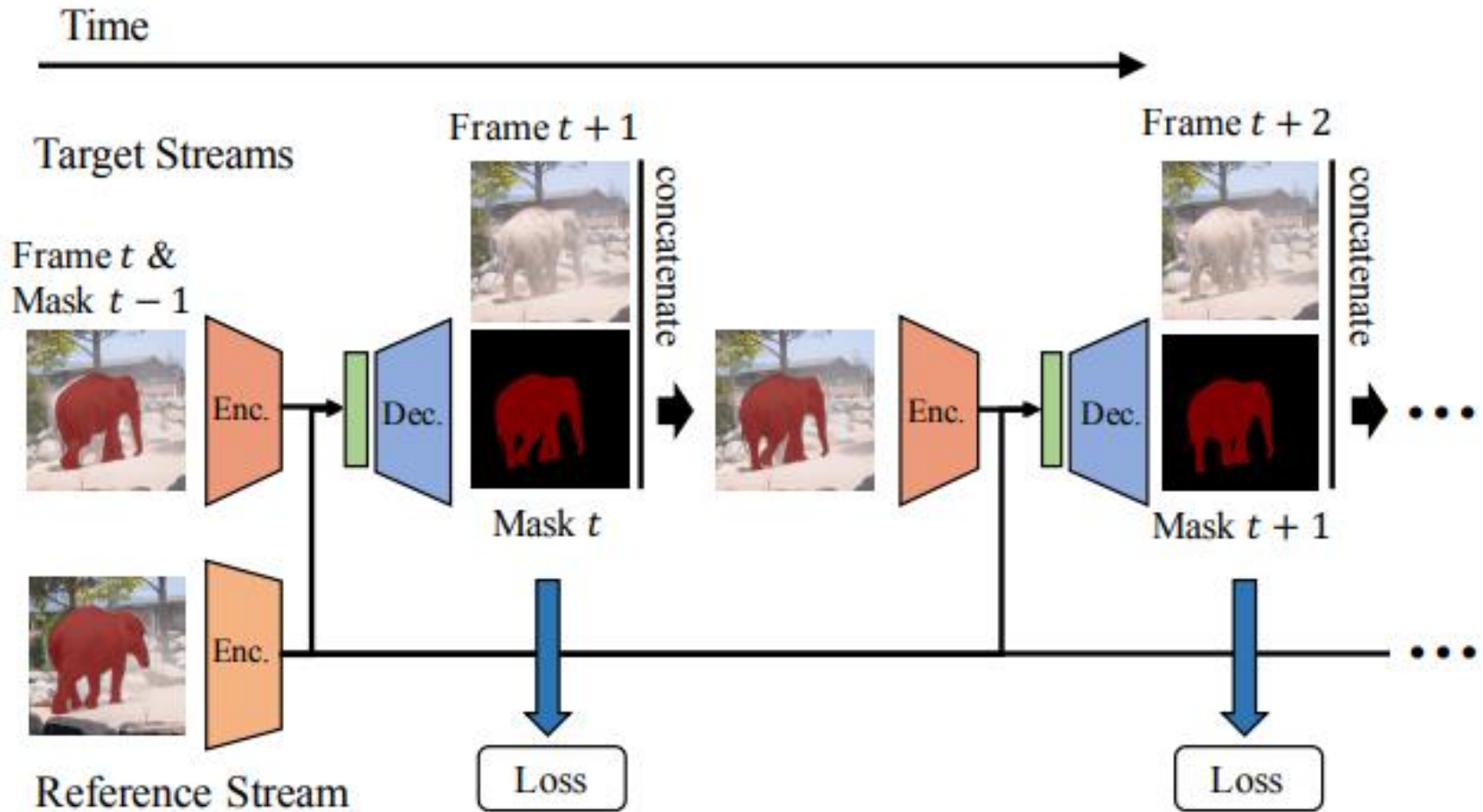
To address the lack of large segmented training video datasets, a two-stage scheme is used that pre-trains the network on synthetically generated image data and then fine-tunes it on video data.

The network architecture and training scheme have been carefully designed to take advantage of both propagation and detection cues.

Fast Video Object Segmentation by Reference-Guided Mask Propagation, CVPR 2018

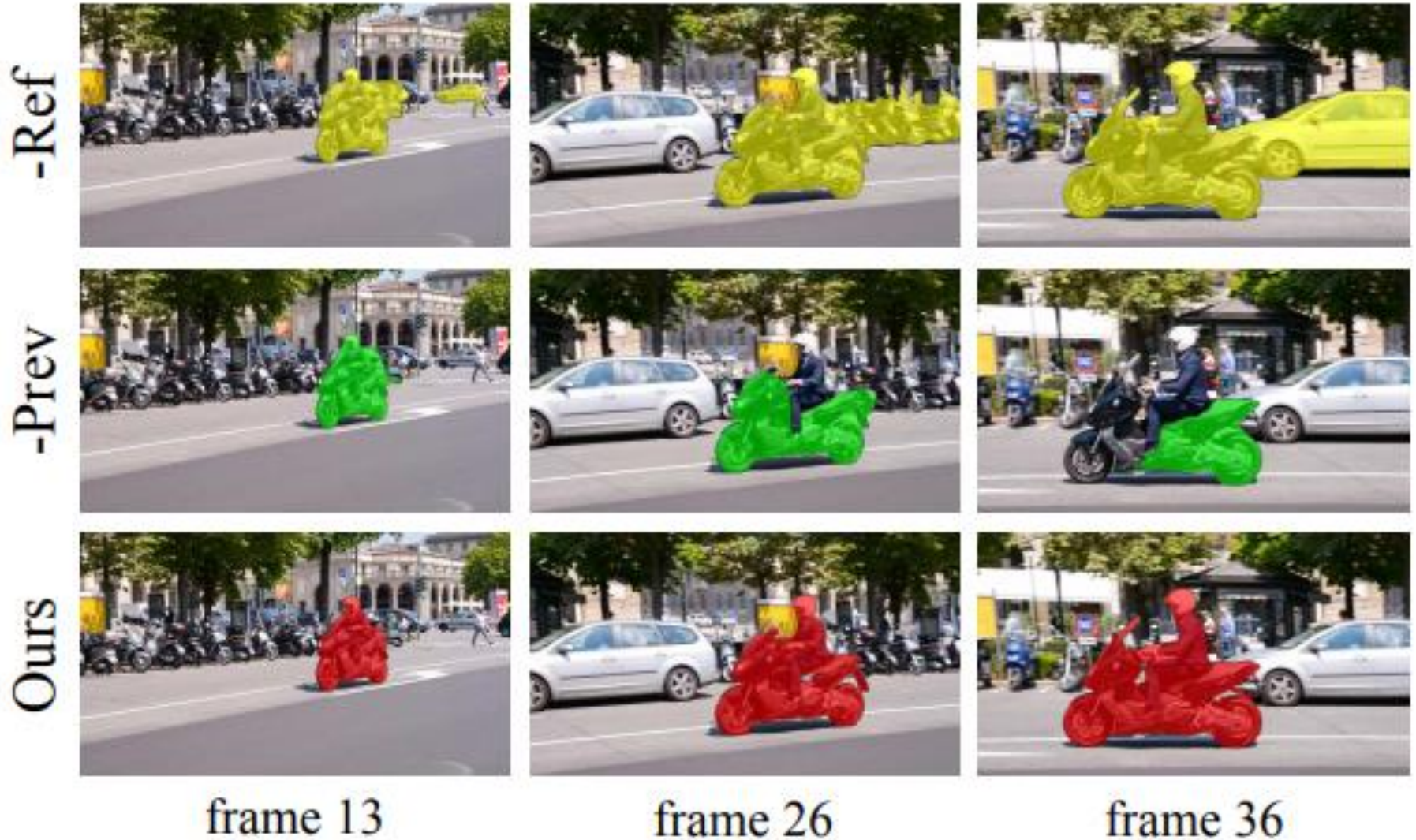


Fast Video Object Segmentation by Reference-Guided Mask Propagation, CVPR 2018



Training with recurrence

Fast Video Object Segmentation by Reference-Guided Mask Propagation, CVPR 2018

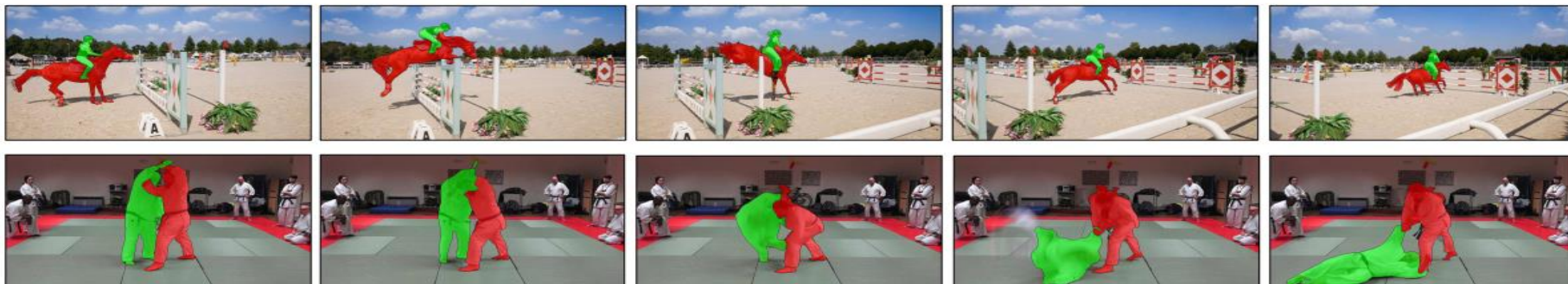


Fast Video Object Segmentation by Reference-Guided Mask Propagation, CVPR 2018

DAVIS-2016



DAVIS-2017



SegTrack v2



0%

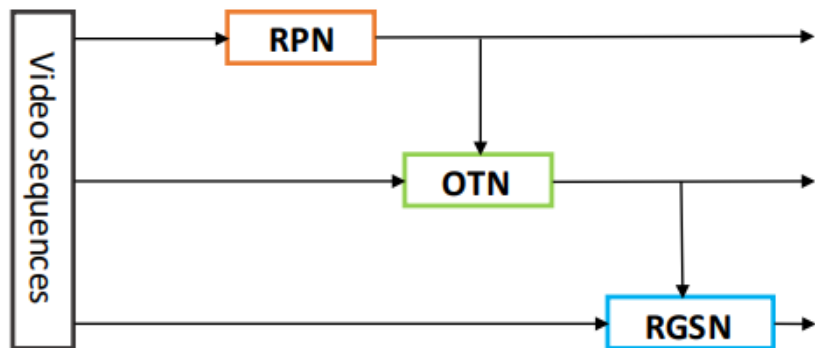
25%




50%

75%

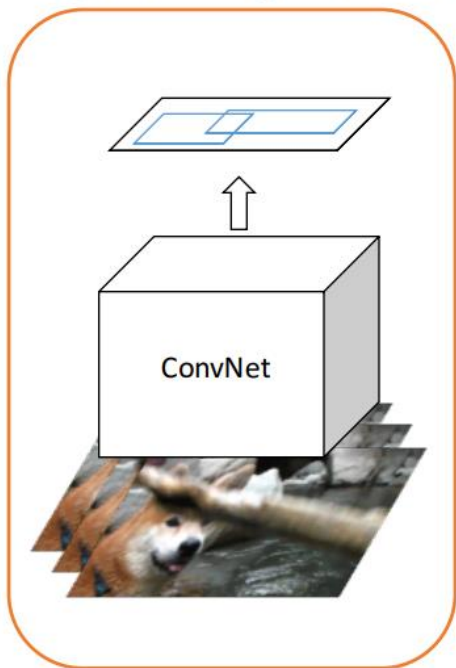
100%

Proposal Tracking and Segmentation (PTS): A cascaded network for video object segmentation, ECCV 2018

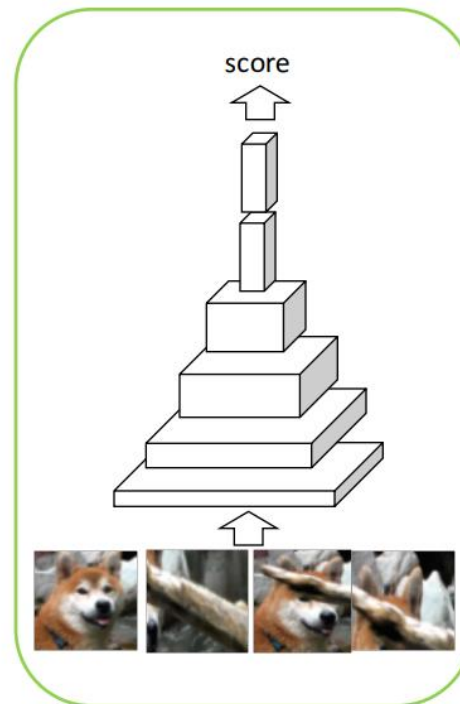


-  RPN: Region Proposal Network (2000 boxes)
-  OTN: Object Tracking Network (1 box)
-  RGSN: Reference-Guided Segmentation Network

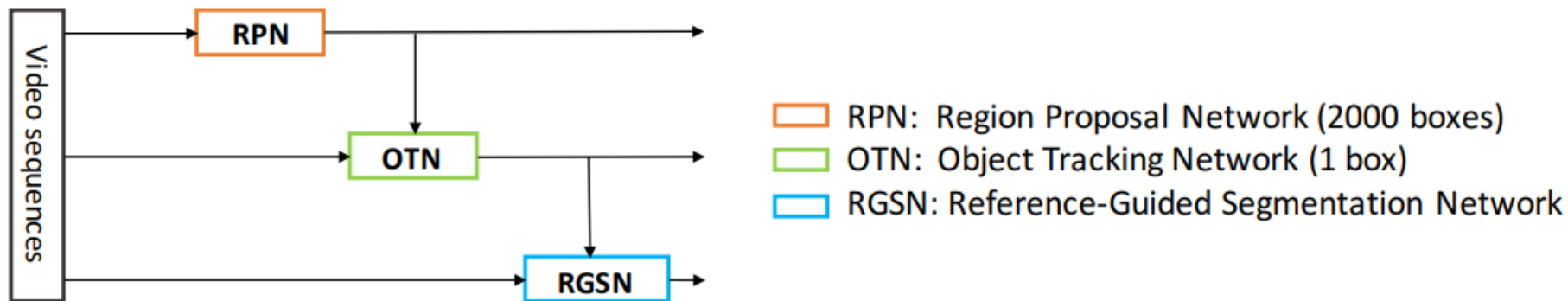
RPN: Region Proposal Network



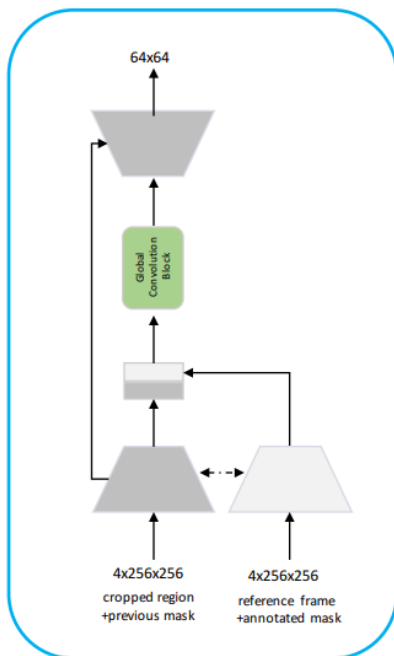
OTN: Object Tracking Network



Proposal Tracking and Segmentation (PTS): A cascaded network for video object segmentation, ECCV 2018

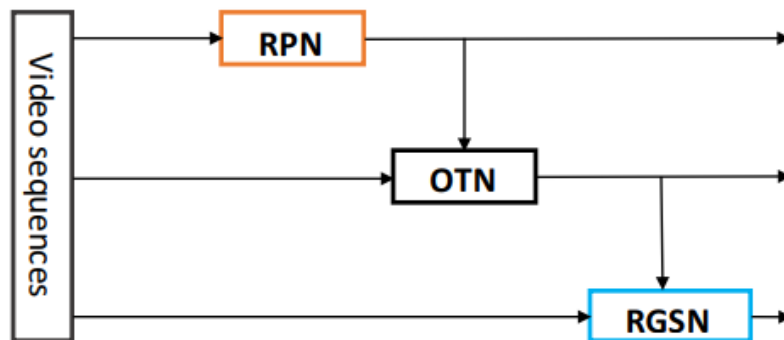


RGSN: Reference-Guided Segmentation Network



Proposal Tracking and Segmentation (PTS): A cascaded network for video object segmentation, ECCV 2018

Offline Training



RPN

RPN adapts Resnet-152 as backbone and is trained on COCO

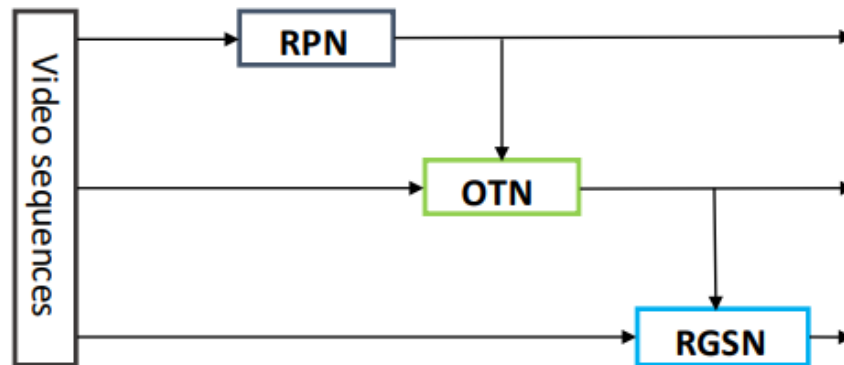
RGSN

RGSN adapts Resnet-50 as backbone and is trained on YouTube-VOS training dataset AUG:

1. Random select two frames as a current frame and a reference frame.
2. Sample bounding boxes around the ground truth box and random scale from 1.5~2.0
3. Encode the previous mask as a heatmap with a two-dimensional Gaussian distribution

Proposal Tracking and Segmentation (PTS): A cascaded network for video object segmentation, ECCV 2018

Online Training



OTN

Update model during inference

RGSN

Fine-tune with first annotated frame before inference for only one time

AUG:

1. Sample bounding boxes around the ground truth box and random scale from 1.5~2.0
2. Encode the previous mask as a heatmap with a two-dimensional Gaussian distribution

Proposal Tracking and Segmentation (PTS): A cascaded network for video object segmentation, ECCV 2018



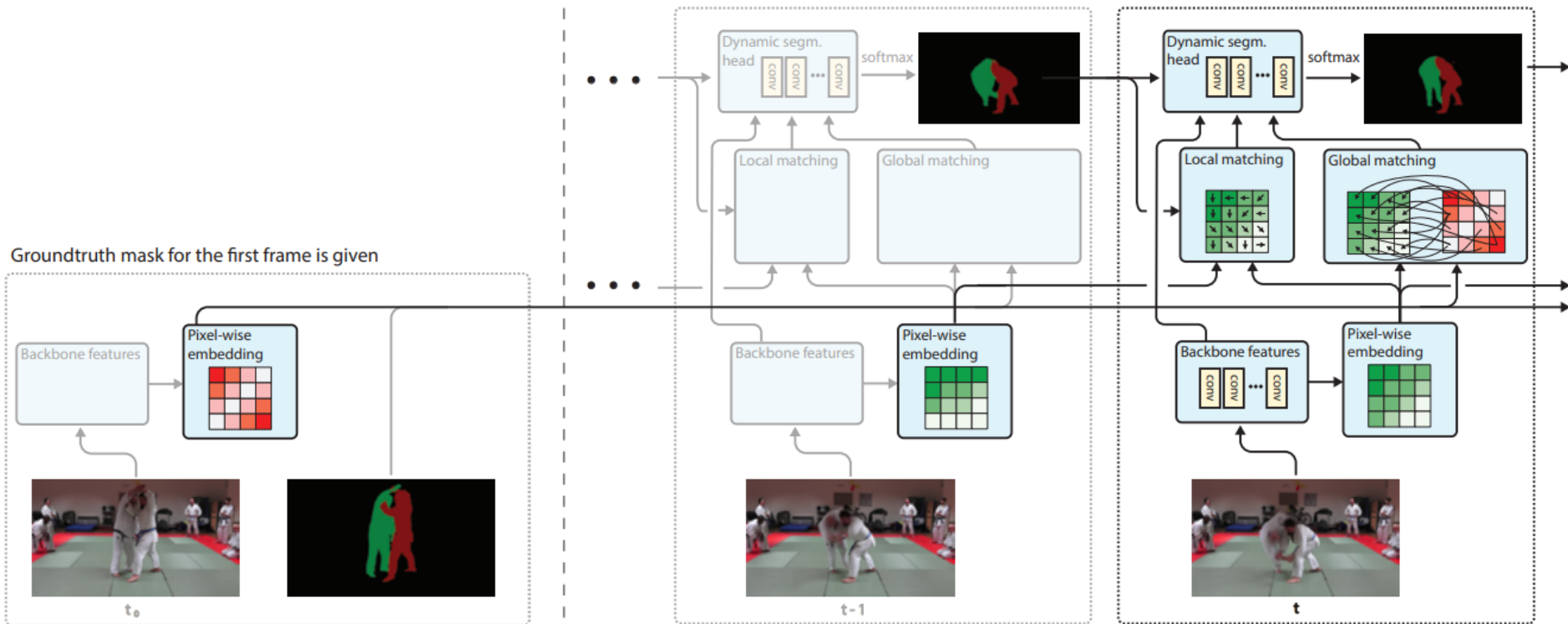
FEELVOS: Fast End-to-End Embedding Learning for Video Object Segmentation, CVPR 2019

Many of the recent successful methods for video object segmentation (VOS) are overly complicated, heavily rely on fine-tuning on the first frame, and/or are slow, and are hence of limited practical use.

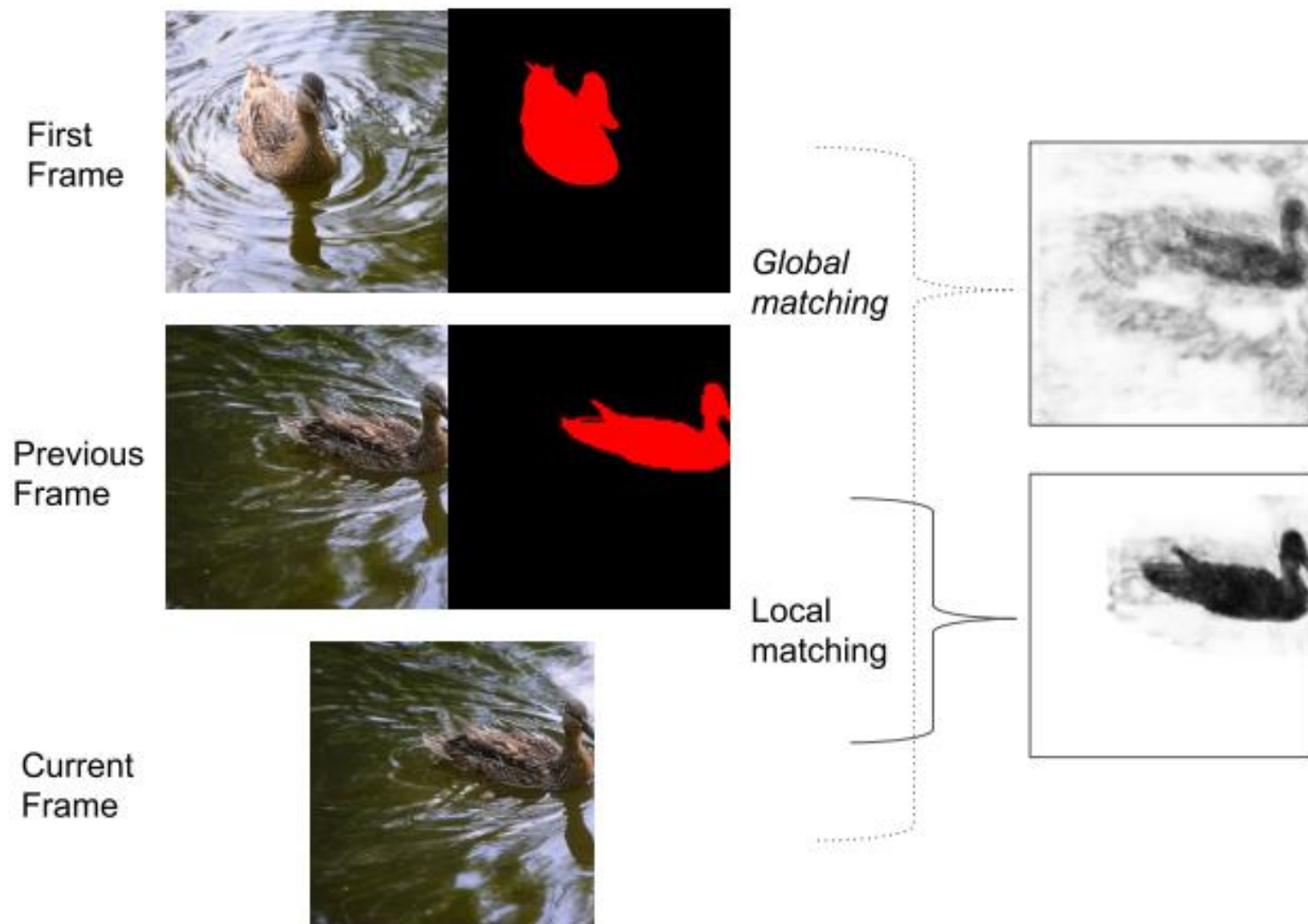
FEELVOS is a simple and fast method which does not rely on fine-tuning. In order to segment a video, for each frame FEELVOS uses a semantic pixel-wise embedding together with a global and a local matching mechanism to transfer information from the first frame and from the previous frame of the video to the current frame.

In contrast to previous work, FEELVOS is only used as an internal guidance of a convolutional network. The novel dynamic segmentation head can train the network, including the embedding, end-to-end for the multiple object segmentation task with a cross entropy loss.

FEELVOS: Fast End-to-End Embedding Learning for Video Object Segmentation, CVPR 2019

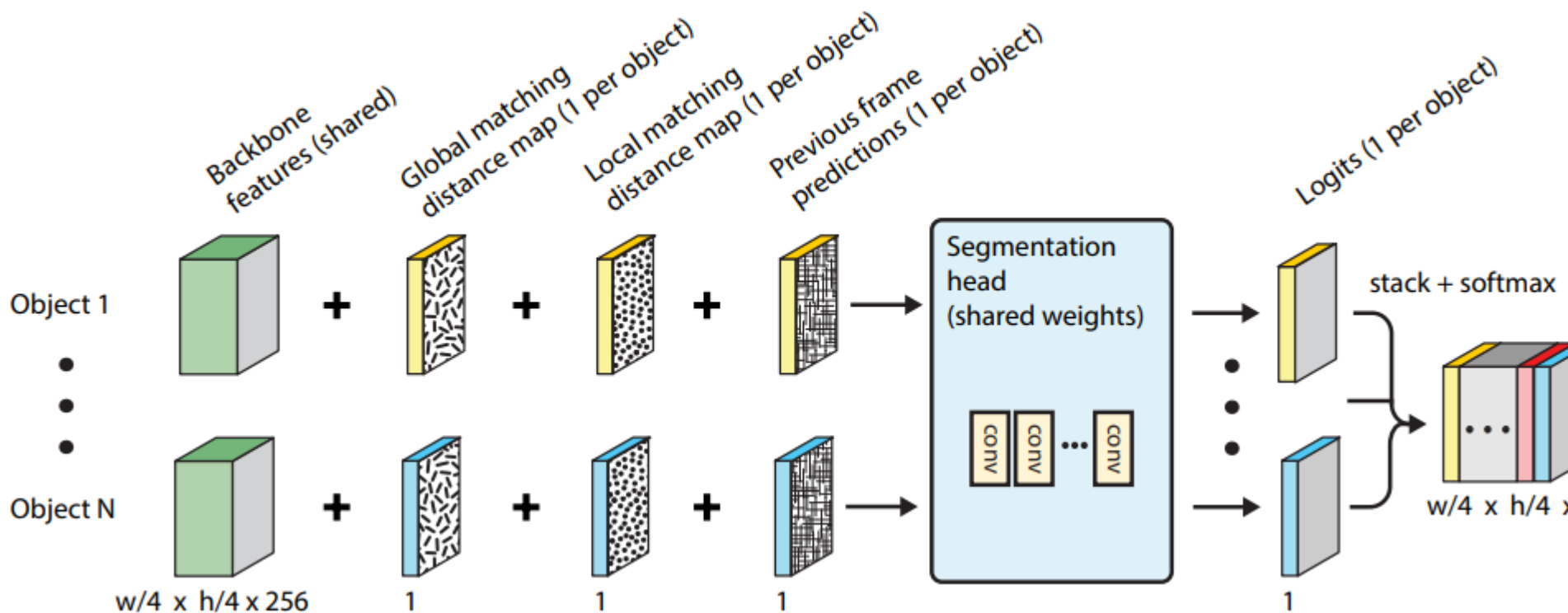


FEELVOS: Fast End-to-End Embedding Learning for Video Object Segmentation, CVPR 2019



Global & Local Matching

FEELVOS: Fast End-to-End Embedding Learning for Video Object Segmentation, CVPR 2019

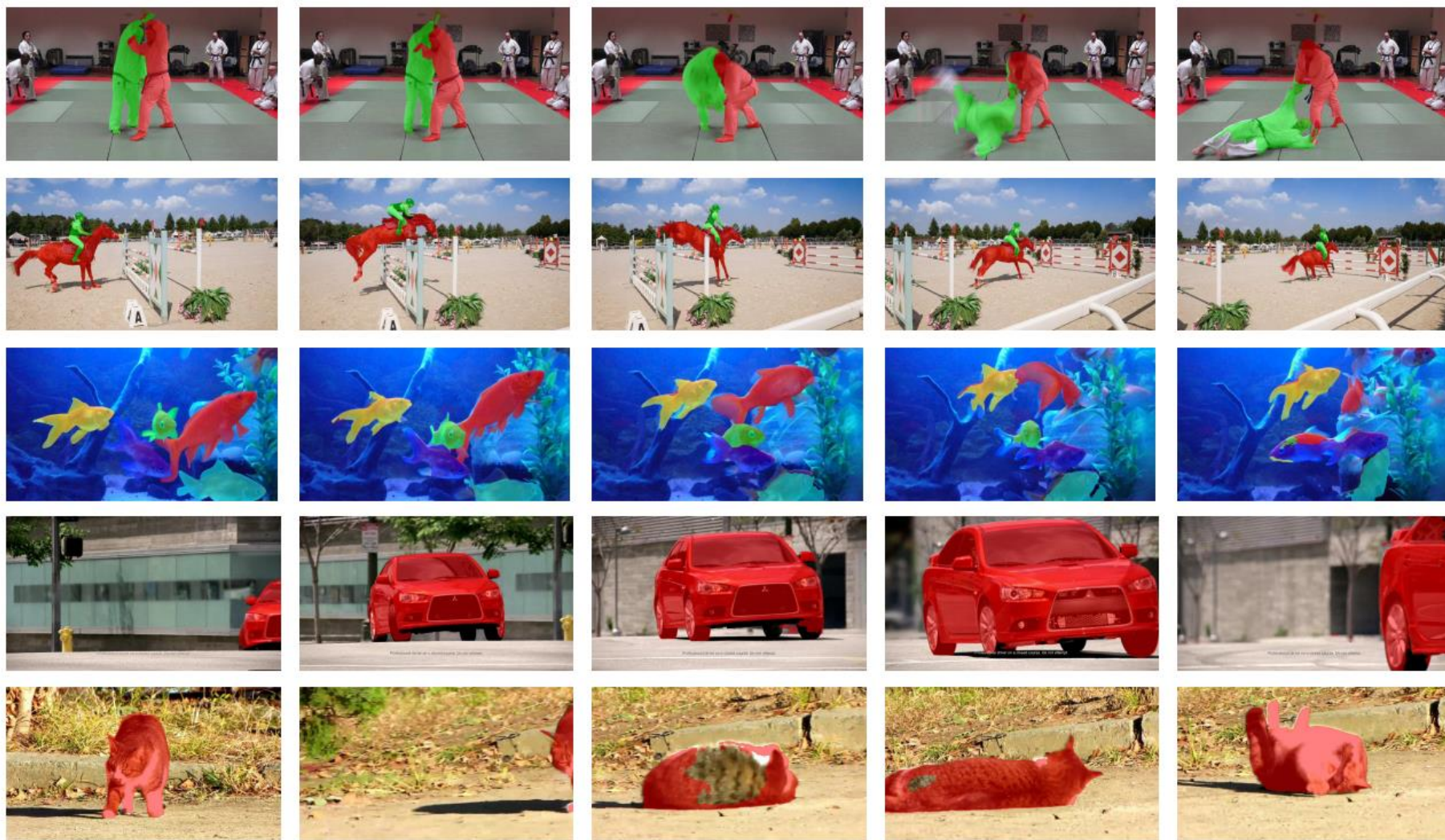


Dynamic segmentation head for systematic handling of multiple objects. The lightweight segmentation head is dynamically instantiated once for each object in the video and produces a one-dimensional feature map of logits for each object. The logits for each object are then stacked together and softmax is applied. The dynamic segmentation head can be trained with a standard cross entropy loss.

FEELVOS: Fast End-to-End Embedding Learning for Video Object Segmentation, CVPR 2019

DAVIS 2017

YouTube Objects



0%

25%

50%

75%

100%

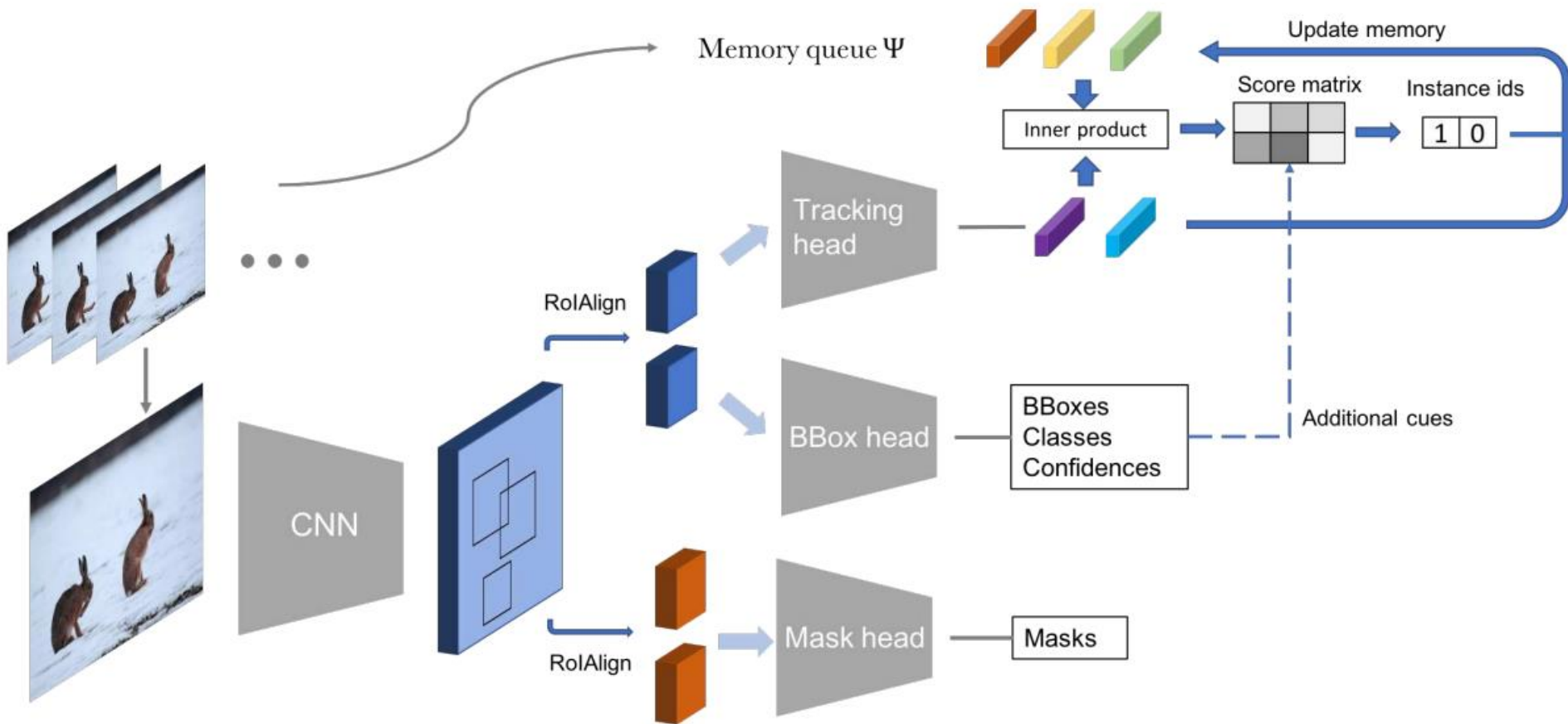
Video Instance Segmentation, ICCV 2019

MaskTrack R-CNN introduces a new tracking branch to Mask R-CNN to jointly perform the detection, segmentation and tracking tasks simultaneously.

Different from image instance segmentation, MaskTrack R-CNN aims at simultaneous detection, segmentation and tracking of object instances in videos.

Video instance segmentation is more challenging than image instance segmentation in that it not only requires instance segmentation on individual frames, but also the tracking of instances across frames. On the other hand, video content contains richer information than a single image such as motion pattern of different objects and temporal consistency, and thus provides more cues for object recognition and segmentation.

Video Instance Segmentation, ICCV 2019



Video Instance Segmentation, ICCV 2019

Mask R-CNN is a high-performing method for image instance segmentation. It consists of two stages. In the **first stage**, a RPN takes an image as input and proposes a set of candidate object bounding boxes. In the **second stage**, features are extracted by the RoIAlign operation from each candidate box and further used to perform classification, bounding box regression and binary segmentation in parallel by three dedicated branches.

Our network adopts the same two-stage procedure, with an identical **first** stage which proposes a set of object bounding boxes at each frame. In the **second** stage, in parallel to the **three** branches (i.e. classification, bounding box regression, binary segmentation), we add the **forth** branch to assign an instance label to each candidate box. Suppose there are already N instances identified by our algorithm from previous frames. Then a new candidate box can only be assigned to one of the N identities.

Video Instance Segmentation, ICCV 2019

The probability of assigning label n to a candidate box i is defined as

$$p_i(n) = \begin{cases} \frac{e^{\mathbf{f}_i^\top \mathbf{f}_n}}{1 + \sum_{j=1}^N e^{\mathbf{f}_i^\top \mathbf{f}_j}} & n \in [1, N] \\ \frac{1}{1 + \sum_{j=1}^N e^{\mathbf{f}_i^\top \mathbf{f}_j}} & n = 0 \end{cases} \quad (2)$$

where \mathbf{f}_i and $\mathbf{f}_j, j \in [1, N]$ denote the new features extracted by our tracking branch from the candidate box and the N identified instances. Our tracking branch has two fully connected layers which project the feature maps extracted by RoIAlign into new features. Since the features of previously identified instances have already been computed, we use an external memory to store them for efficiency. The cross entropy loss is used for our tracking branch, *i.e.* $L_{track} = -\sum_i \log(p_i(y_i))$ where y_i is the ground truth instance label.

Video Instance Segmentation, ICCV 2019

