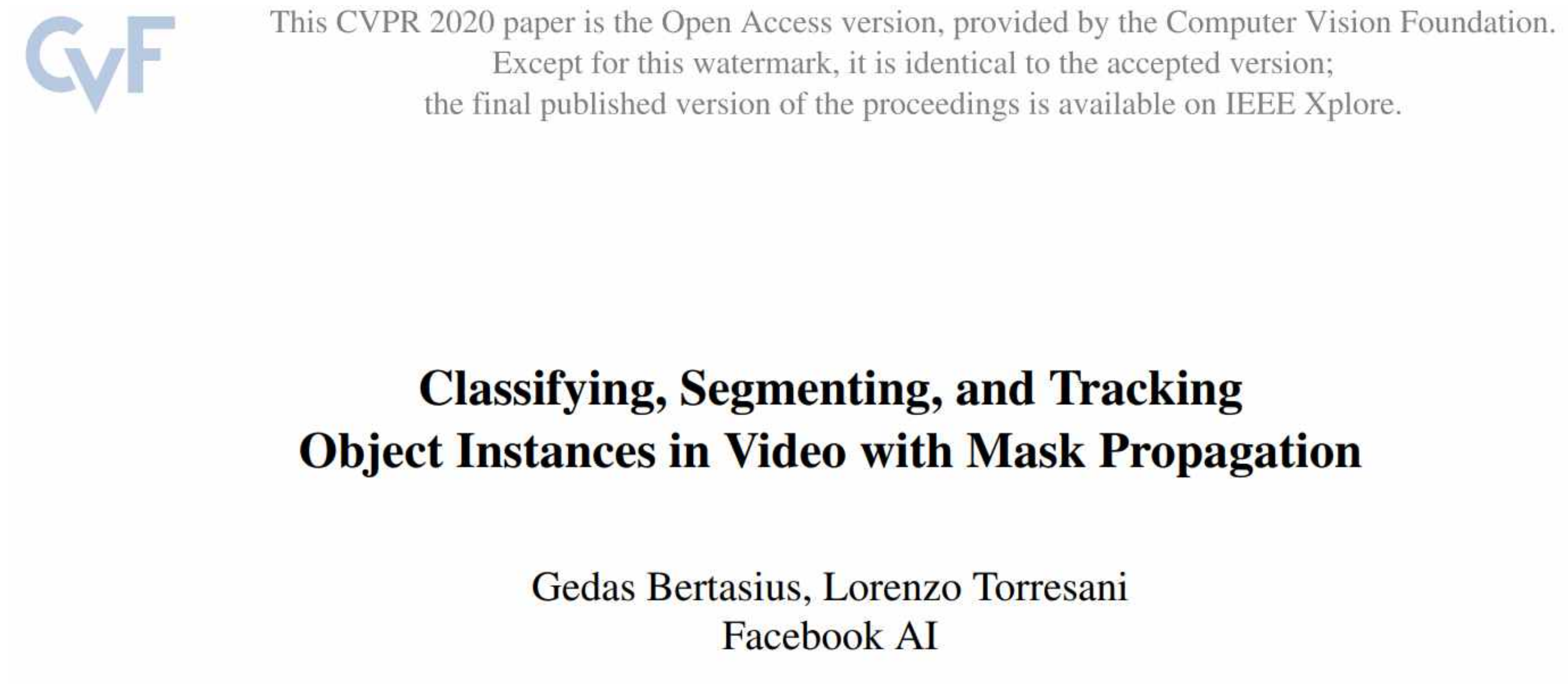
Oral paper



# Classifying, Segmenting, and Tracking Object Instances in Video with Mask Propagation

Gedas Bertasius, Lorenzo Torresani
Facebook AI

2020.07.15

Presented by Kyungmin Jo

# Introduction

- Task : Object instance segmentation in video
- This task is challenging because of very large spatial resolution, analyzing multiple video frames simultaneously
  - Reduce the spatial resolution of the input ➔ Performance ↓
  - Segmentation on individual frames and then link them temporally ➔ Suboptimal results, because two tasks are closely intertwined

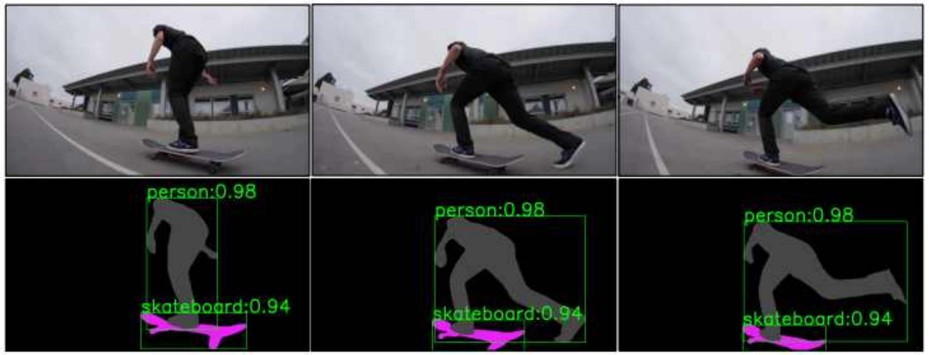
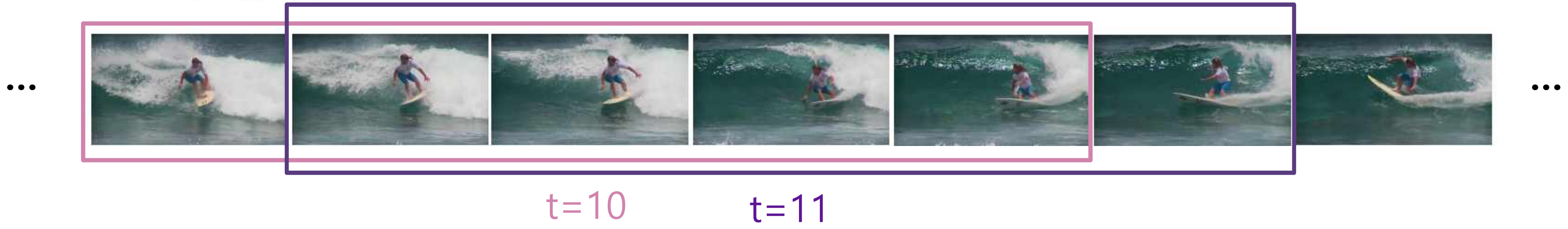


Figure 1: In this paper, we tackle the problem of video instance segmentation, which requires classifying, segmenting, and tracking object instances in a given video sequence. Our proposed Mask Propagation framework (MaskProp) provides a simple and effective way for solving this task.

# Mask propagation

- Aim : segment and temporally link all object instances that are visible for at least one frame in video
- Two steps of Mask propagation method
  1. Clip-level instance segmentation
     - 1.1 Instance feature computation
     - 1.2 Instance feature propagation
     - 1.3 Propagated instance segmentation
  2. Video-level instance segmentation

Video : $V \in \mathcal{R}^{L \times 3 \times H \times W}$



Clip : $V_{t-T:t+T} \in \mathcal{R}^{(2T+1) \times 3 \times H \times W} \quad t = 1, 2, \ldots, L$

# 1. Clip-level instance segmentation

- Based on the Mask R-CNN
- Adds a mask propagation branch
- Loss function

| Pred | GT | sIoU Numerator | sIoU Denominator |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| q | 0 | 0 | q |
| 0 | 1 | 0 | 1 |
| q | 1 | q | 1 |

$$L_t = L_t^{cls} + L_t^{box} + L_t^{mask} + L_t^{prop}$$

$$L_t^{prop} = \sum_{i}^{\tilde{N}_t} \sum_{t'=t-T}^{t+T} 1 - sIoU(M^i_{t-T:t+T}(t'), \tilde{M}^i_{t-T:t+T}(t'))$$

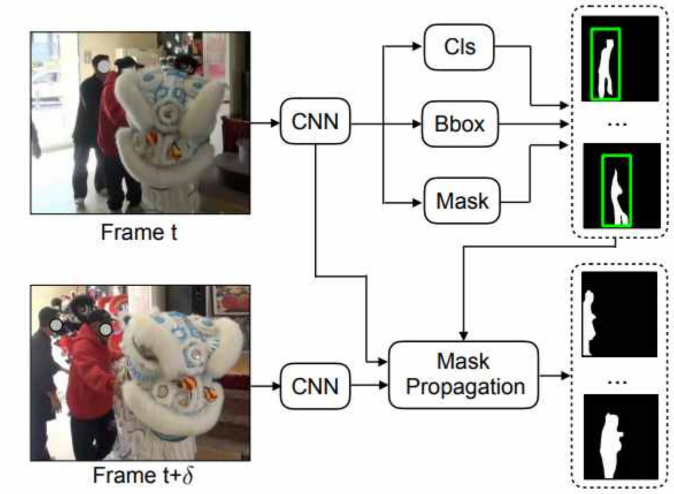
$$sIoU(A, B) = \frac{\sum_p A(p)B(p)}{\sum_p A(p) + B(p) - A(p)B(p)}$$

$M^i_{t-T:t+T}(t') \in [0, 1]$ : the segmentation at time ***t'*** for an instance ***i*** predicted from a clip centered at ***t***

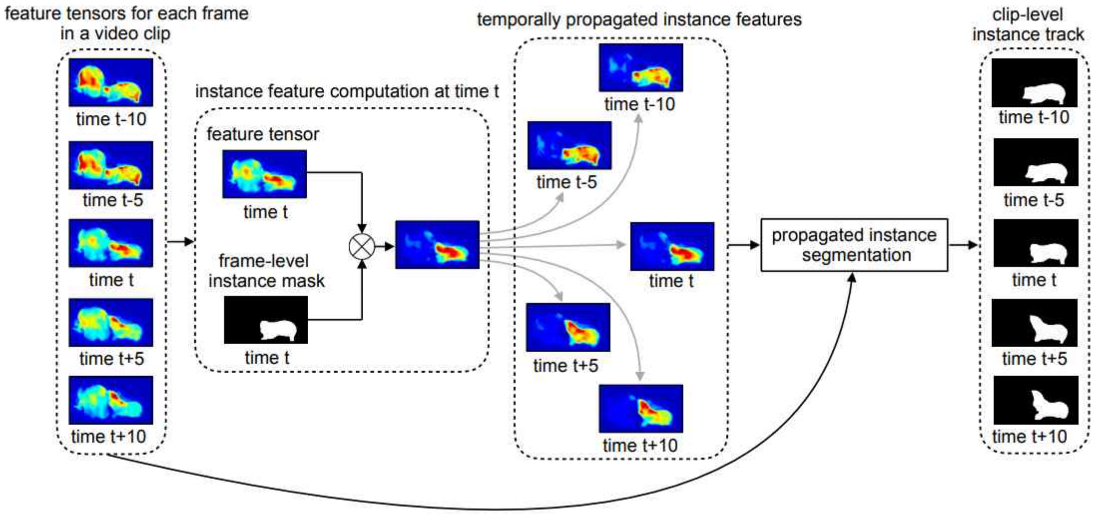
$\tilde{M}^i_{t-T:t+T}(t')$ : corresponding ground truth mask

$\tilde{N}_t$ : the number of ground truth object instances in frame ***t***

*p* : pixel location



▲ Mask R-CNN + mask propagation

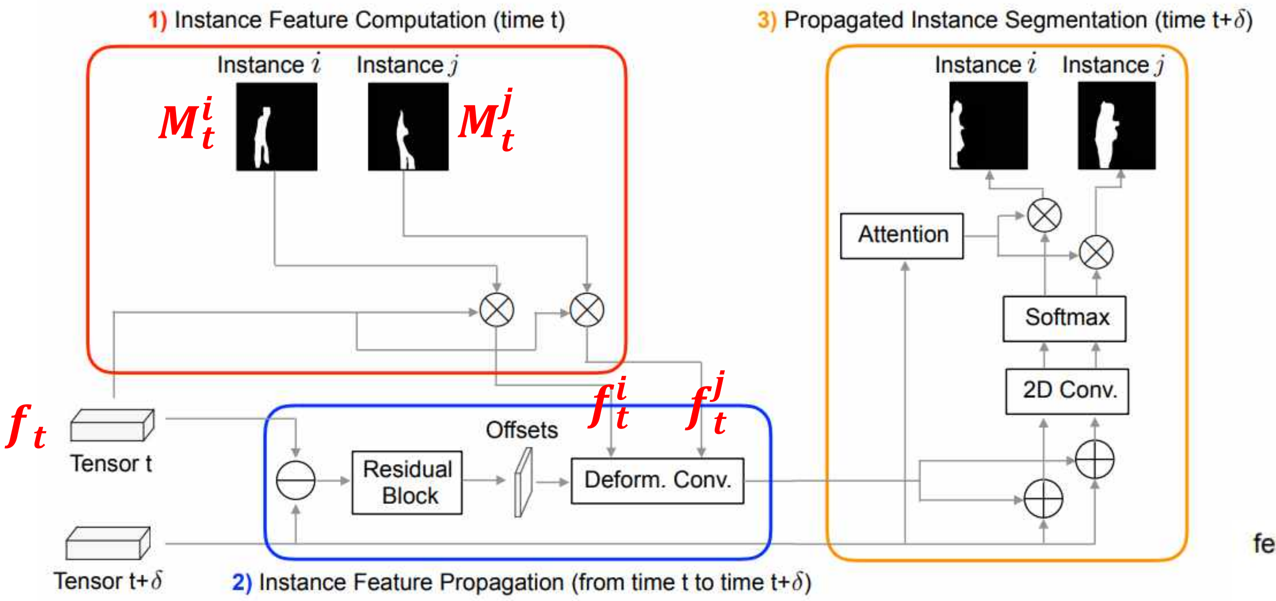

feature tensors for each frame in a video clip
time t-10
time t-5
time t
time t+5
time t+10
instance feature computation at time t
feature tensor
time t
frame-level instance mask
time t
temporally propagated instance features
time t-10
time t-5
time t
time t+5
time t+10
propagated instance segmentation
clip-level instance track
time t-10
time t-5
time t
time t+5
time t+10

▲ An illustration of Mask propagation

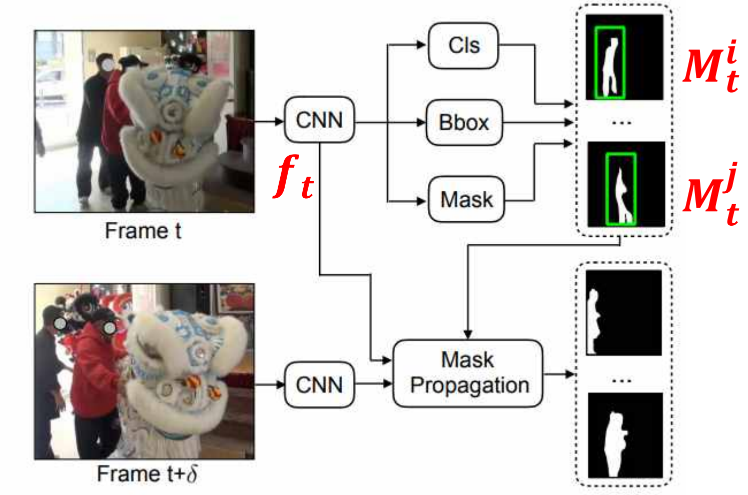
Method

# 1.1 Instance feature computation

- Yield a set of instance-specific feature $f_t^i \in \mathcal{R}^{C \times H' \times W'}$



▲Mask propagation

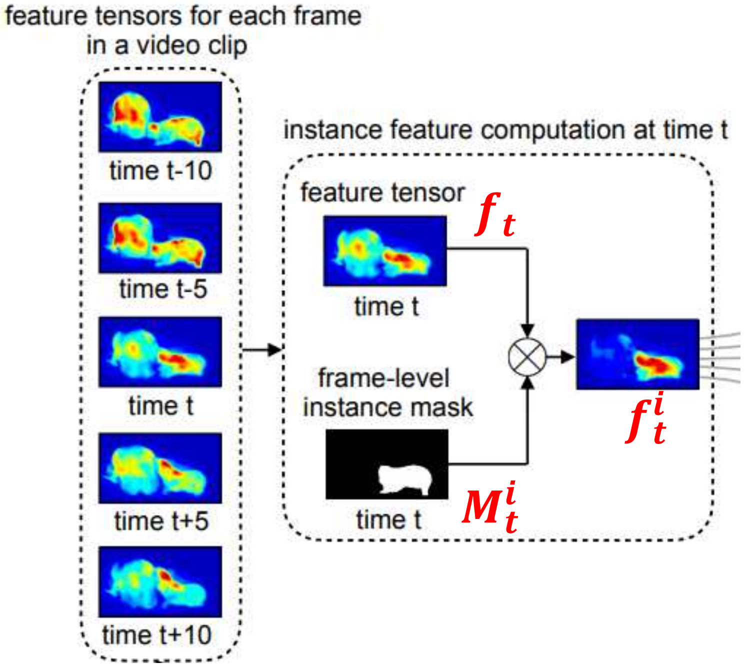


$M_t^i \in \mathcal{R}^{1 \times H' \times W'}$ : Frame-level instance masks

$f_t$ : Feature tensor from the backbone network

*i* : Object instance (*1, .., Nt* )

*Nt* : The number of object instances detected in frame *t*

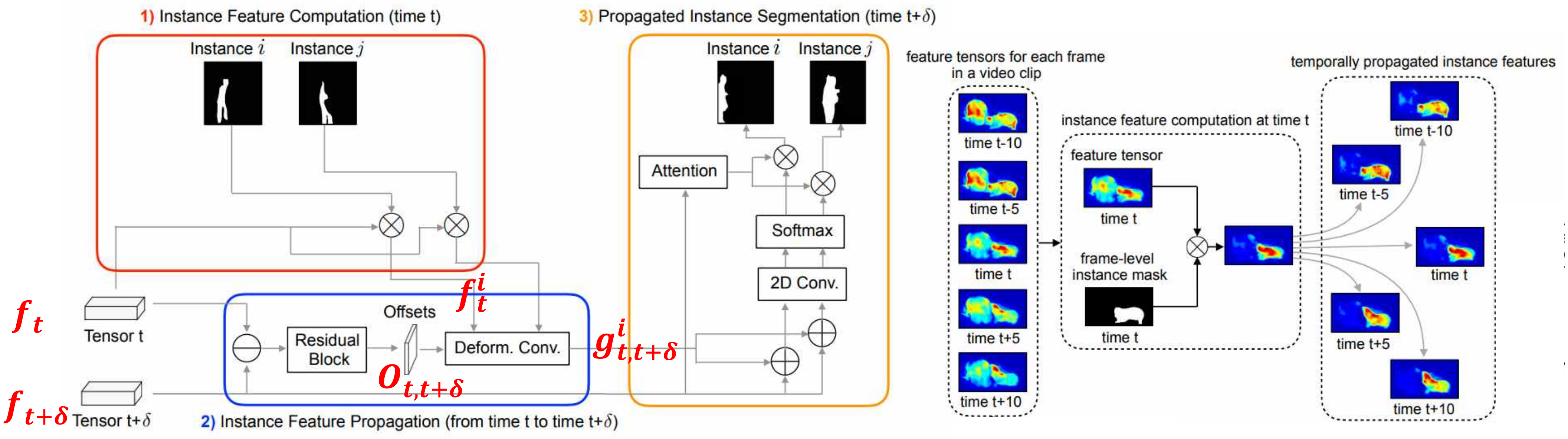


feature tensors for each frame in a video clip
time t-10
time t-5
time t
time t+5
time t+10
instance feature computation at time t
feature tensor
$f_t$
time t
frame-level instance mask
time t
$M_t^i$
$f_t^i$

Method

# 1.2 Instance feature propagation

- Generates a propagated instance feature $g^i_{t,t+\delta}$
- No explicit ground truth alignment is available between frames. The deformable convolutional kernels are supervised implicitly by optimizing $L^{prop}_t$



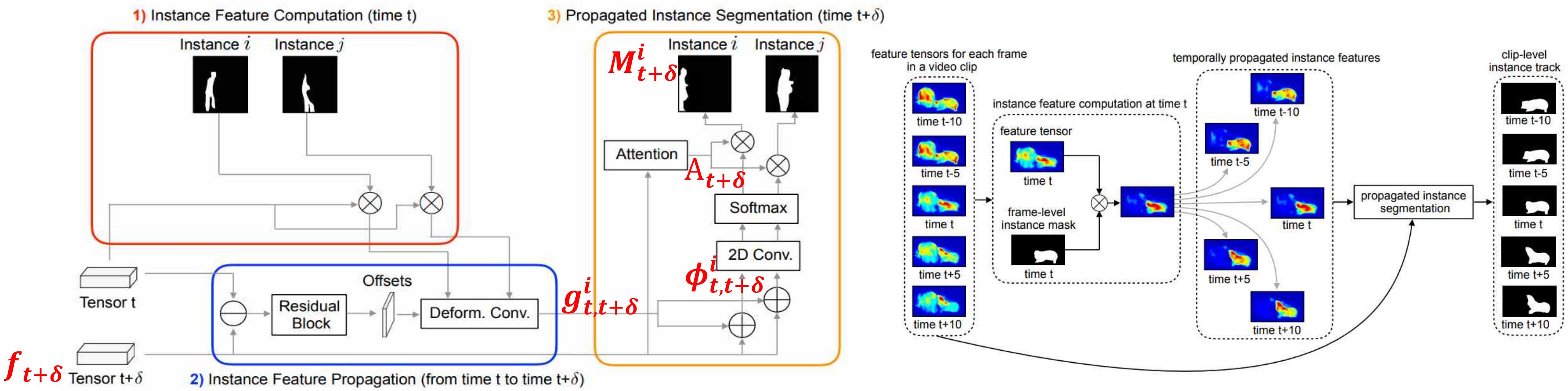
1) Instance Feature Computation (time t)
Instance $i$
Instance $j$
3) Propagated Instance Segmentation (time t+$\delta$)
Instance $i$
Instance $j$
Attention
Softmax
2D Conv.
$f_t$
Tensor t
$f_{t+\delta}$
Tensor t+$\delta$
Offsets
$f^i_t$
Residual Block
Deform. Conv.
$o_{t,t+\delta}$
$g^i_{t,t+\delta}$
2) Instance Feature Propagation (from time t to time t+$\delta$)
feature tensors for each frame in a video clip
time t-10
time t-5
time t
time t+5
time t+10
instance feature computation at time t
feature tensor
time t
frame-level instance mask
time t
temporally propagated instance features
time t-10
time t-5
time t
time t+5
time t+10

$o_{t,t+\delta} \in \mathcal{R}^{2k^2 \times H' \times W'}$ : motion offsets

※ J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. **Deformable convolutional networks**. In 2017 IEEE International Conference on Computer Vision (ICCV), volume 00, pages 764–773, Oct. 2017.
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. **Deep residual learning for image recognition**. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

# 1.3 Propagated instance segmentation

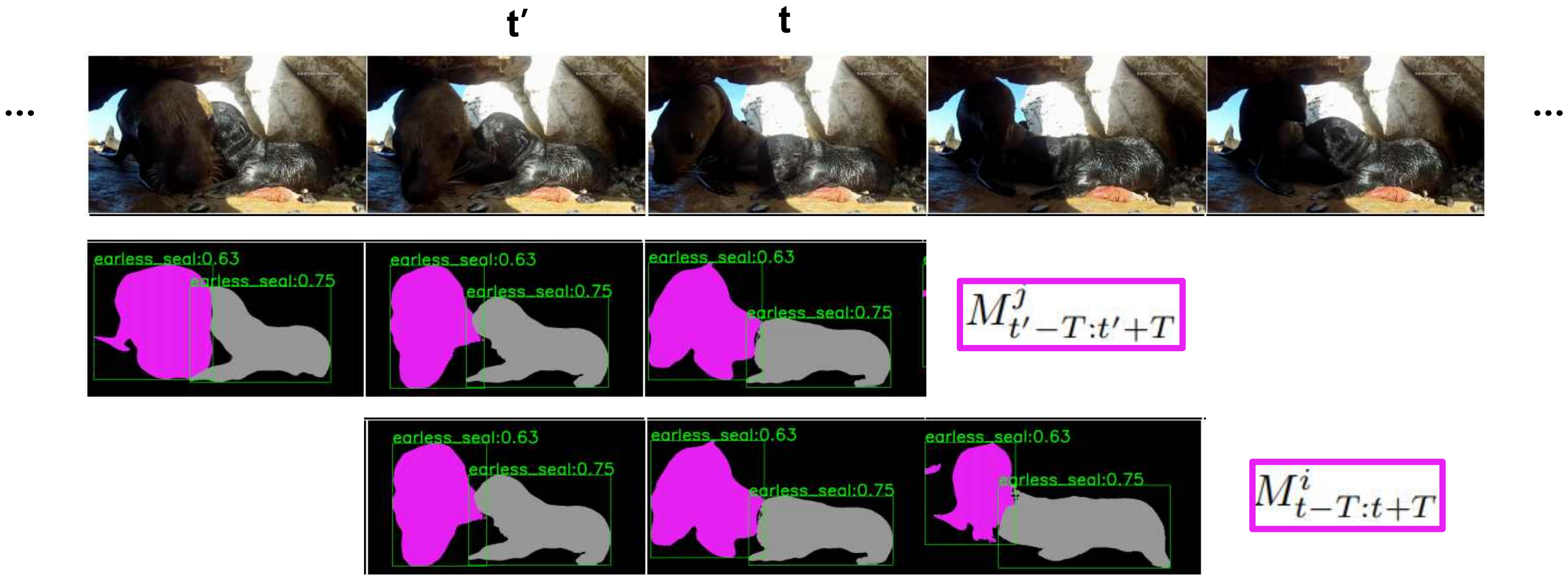
- Construct a new feature tensor $\phi^i_{t,t+\delta} = g^i_{t,t+\delta} + f_{t+\delta}$.
  - If the object instance <u>prediction is consistent</u> with the feature computation, the feature tensors will <u>reinforce each other</u> in the predicted region
- Compute an instance-agnostic attention map $A_{t+\delta}$
  - Zero-out pixels not belong to any object instance


1) Instance Feature Computation (time t)
Instance $i$
Instance $j$
3) Propagated Instance Segmentation (time t+$\delta$)
$M^i_{t+\delta}$
Instance $i$
Instance $j$
Attention
$A_{t+\delta}$
Softmax
2D Conv.
$\phi^i_{t,t+\delta}$
$g^i_{t,t+\delta}$
Tensor t
Offsets
Residual Block
Deform. Conv.
$f_{t+\delta}$
Tensor t+$\delta$
2) Instance Feature Propagation (from time t to time t+$\delta$)
feature tensors for each frame in a video clip
time t-10
time t-5
time t
time t+5
time t+10
instance feature computation at time t
feature tensor
time t
frame-level instance mask
time t
temporally propagated instance features
time t-10
time t-5
time t
time t+5
time t+10
propagated instance segmentation
clip-level instance track
time t-10
time t-5
time t
time t+5
time t+10

Attention : 3x3 conv
2D Conv. : 1x1 conv

# 2. Video-level instance segmentation

- Produce video-level segmentation instances $M^i \in \mathcal{R}^{L \times H \times W}$
  - Matching Clip-level instance tracks
  - Video-level instance ID assignment

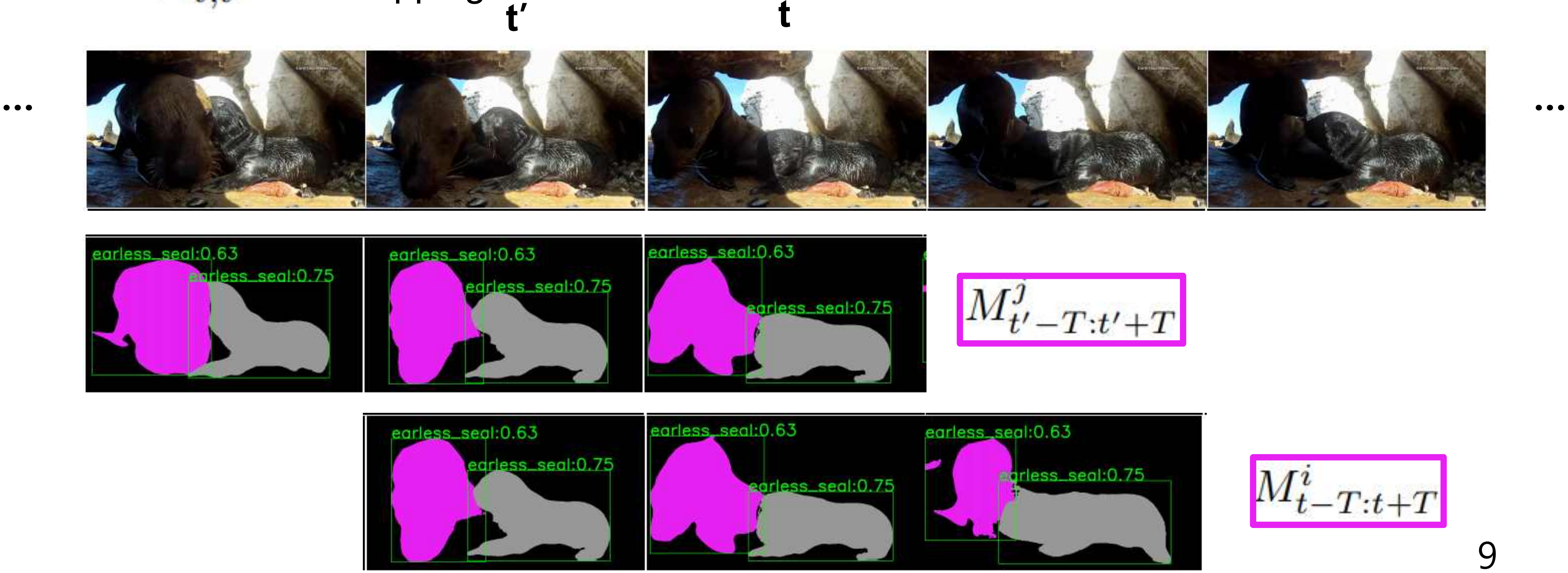


# Matching Clip-level instance tracks

- If |t − t′| < 2T + 1, then $M^i_{t-T:t+T}$ and $M^j_{t'-T:t'+T}$ overlap in time
  - Compute a matching score between two clip-level instance tracks in the overlapping frames

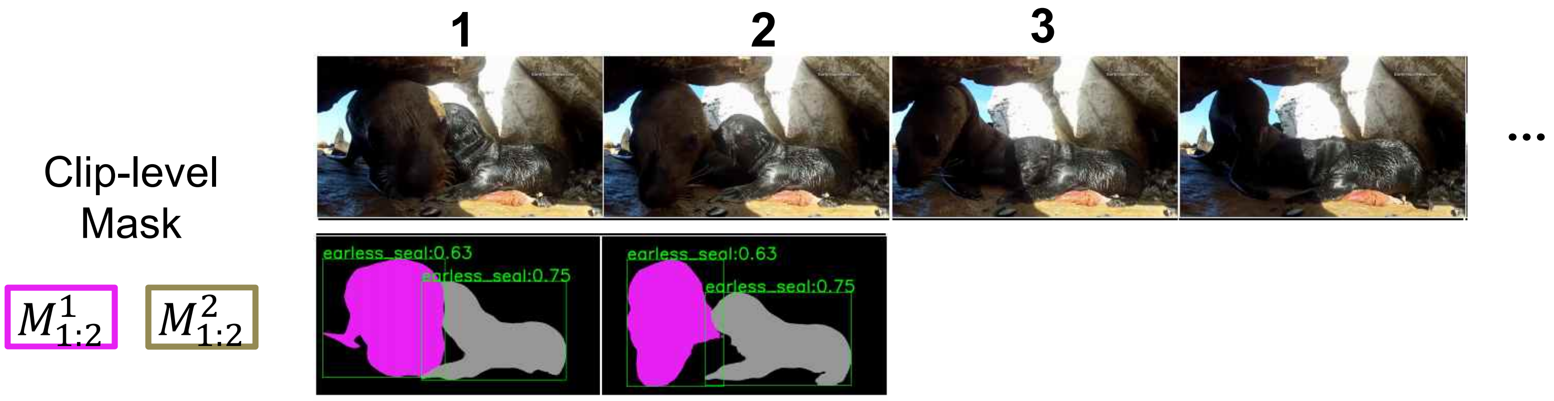
$$m^{i,j}_{t,t'} = \frac{1}{|\cap_{t,t'}|} \sum_{\tilde{t} \in \cap_{t,t'}} sIoU(M^i_{t-T:t+T}(\tilde{t}), M^j_{t'-T:t'+T}(\tilde{t}))$$

$\cap_{t,t'}$ : Overlapping time interval



# Video-level instance ID assignment(1)

- Initialize the set of video-level instance IDs $\mathcal{Y}$



# Video-level instance ID assignment(2)

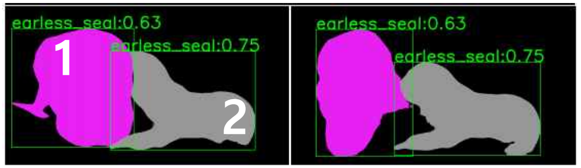
- Initialize the set of video-level instance IDs $\mathcal{Y}$

Assigned video-level instance ID

Clip-level Mask

$y_1 = \{1, 2\}$

$M_{1:2}^{1}$ $M_{1:2}^{2}$



Video-level Instance ID $\mathcal{Y} = \{1, 2\}$

# Video-level instance ID assignment(3)

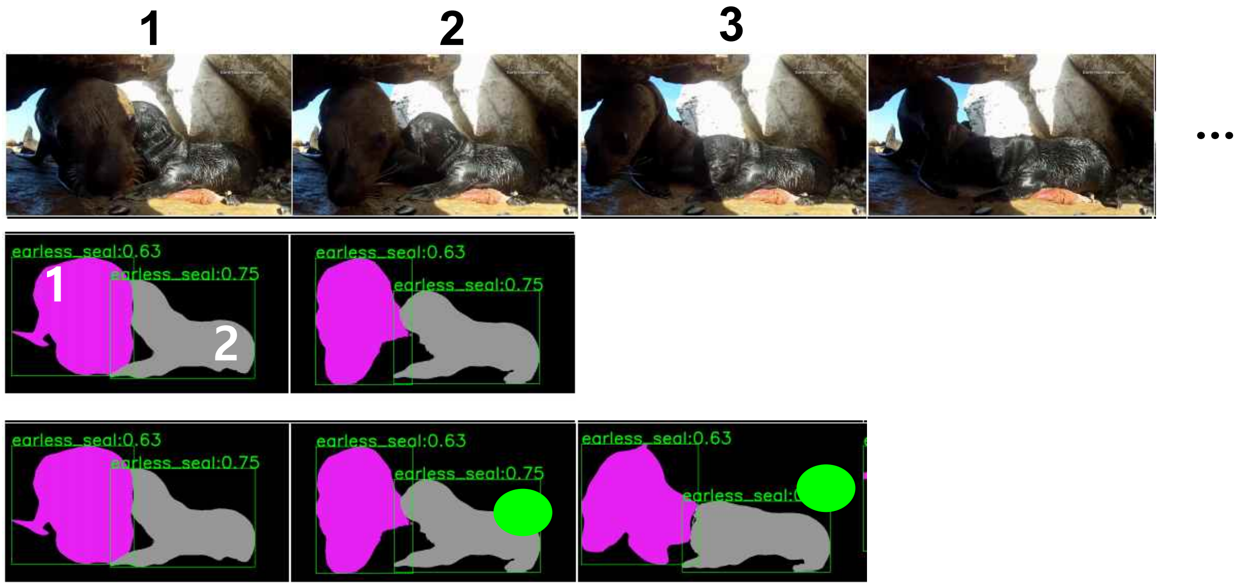
Assigned video-level instance ID

Clip-level Mask

$y_1 = \{1,2\}$

$M^1_{1:2}$ $M^2_{1:2}$

$M^1_{1:3}$ $M^2_{1:3}$ $M^3_{1:3}$



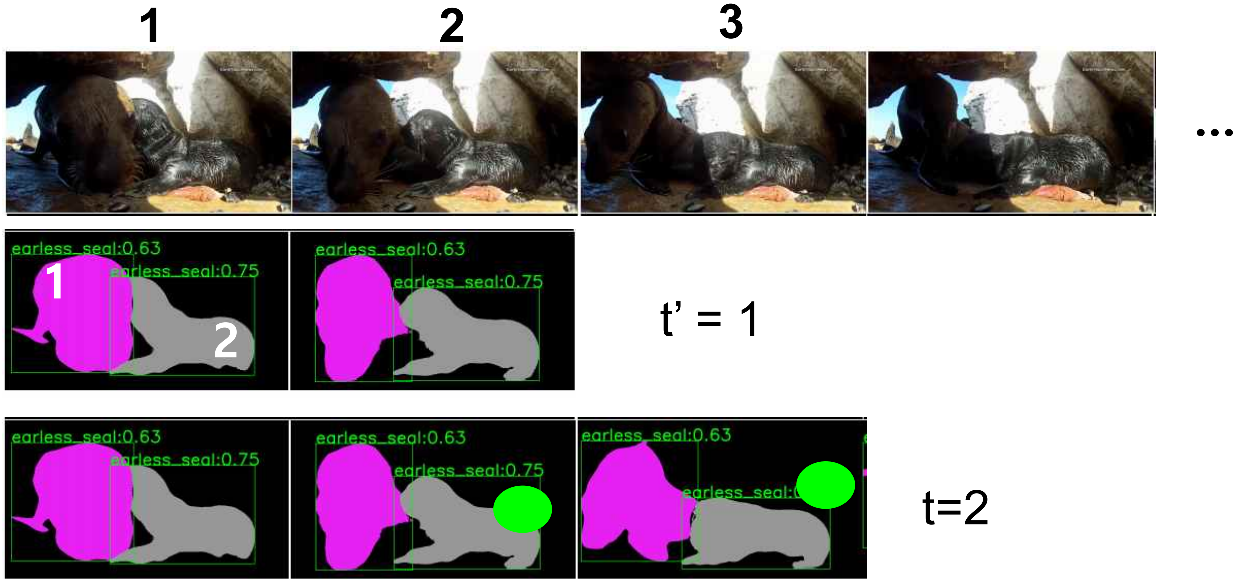
Video-level Instance ID $y = \{1, 2\}$

?

# Video-level instance ID assignment(4)

Assigned video-level instance ID

Clip-level Mask



$y_1 = \{1,2\}$ $M^1_{1:2}$ $M^2_{1:2}$ t' = 1

$M^1_{1:3}$ $M^2_{1:3}$ $M^3_{1:3}$ t=2

| | $M^1_{1:3}$ | $M^2_{1:3}$ | $M^3_{1:3}$ |
|---|---|---|---|
| y=1 | $q_2^1(1) = \frac{m_{2,1}^{1,1}}{2}$ | $q_2^2(1) = 0$ | $q_2^3(1) = 0$ |
| y=2 | $q_2^1(2) = 0$ | $q_2^2(2) = \frac{m_{2,1}^{2,2}}{2}$ | $q_2^3(2) = \frac{m_{2,1}^{3,2}}{2}$ |

$$q_t^i(y) = \frac{\sum_{t' \text{s.t. } \cap_{t,t'} \neq \emptyset} \sum_{j=1}^{N_{t'}} 1\{y_{t'}^j = y\} \cdot m_{t,t'}^{i,j}}{\sum_{t' \text{s.t. } \cap_{t,t'} \neq \emptyset} \sum_{j=1}^{N_{t'}} 1\{y_{t'}^j = y\}}$$

→ Video-level Instance ID $y = \{1,2\}$ → ?

Method

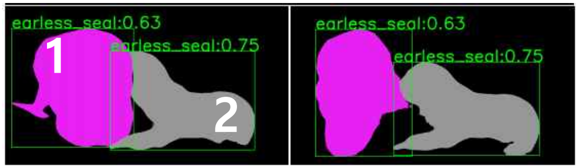
# Video-level instance ID assignment(5)

Assigned video-level instance ID

Clip-level Mask

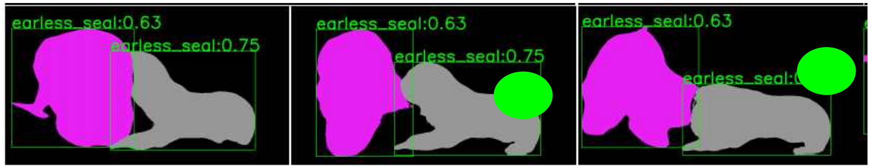
...

$y_1 = \{1,2\}$ $\boxed{M^1_{1:2}}$ $\boxed{M^2_{1:2}}$



t' = 1

$\boxed{M^1_{1:3}}$ $\boxed{M^2_{1:3}}$ $\boxed{M^3_{1:3}}$



t=2

| | $\boxed{M^1_{1:3}}$ | $\boxed{M^2_{1:3}}$ | $\boxed{M^3_{1:3}}$ |
|---|---|---|---|
| y=1 | $q_2^1(1) = \frac{m_{2,1}^{1,1}}{2}$ | $q_2^2(1) = 0$ | $q_2^3(1) = 0$ |
| y=2 | $q_2^1(2) = 0$ | $q_2^2(2) = \frac{m_{2,1}^{2,2}}{2}$ | $q_2^3(2) = \frac{m_{2,1}^{3,2}}{2}$ |

$$q_t^i(y) = \frac{\sum_{t' \text{s.t. } \cap_{t,t'} \neq \emptyset} \sum_{j=1}^{N_{t'}} 1\{y_{t'}^j = y\} \cdot m_{t,t'}^{i,j}}{\sum_{t' \text{s.t. } \cap_{t,t'} \neq \emptyset} \sum_{j=1}^{N_{t'}} 1\{y_{t'}^j = y\}}$$

$\max(q_t^i)$ $q_2^1(1)$ $q_2^2(2)$ $q_2^3(2)$

Assigned Video-level instance ID $y^* = 1$ $y^* = 2$ X

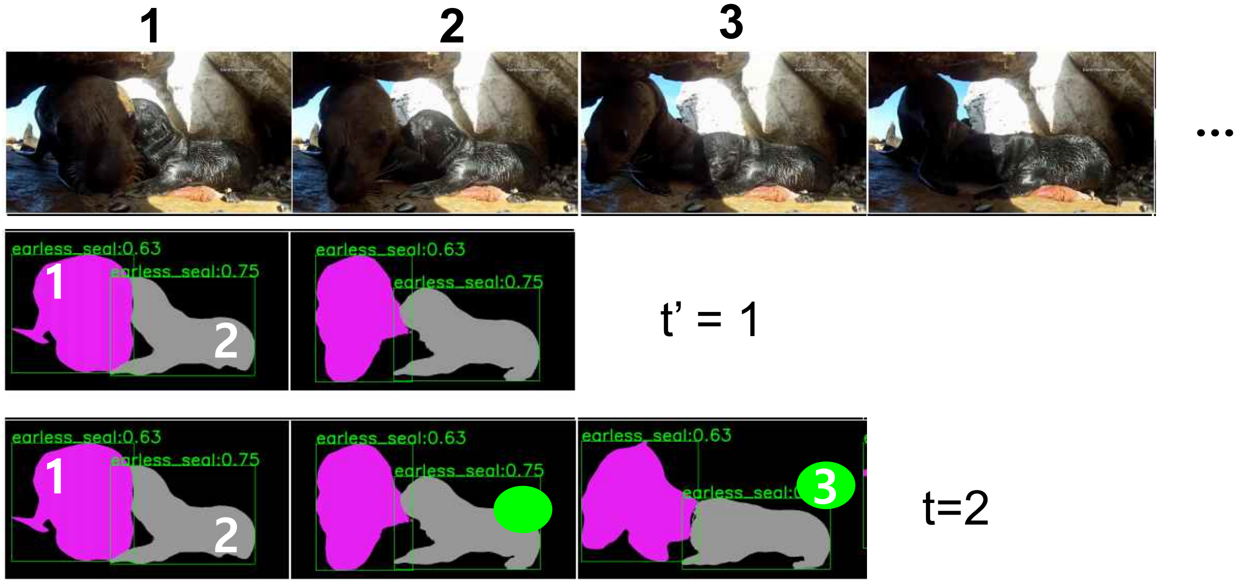
>TH

Video-level Instance ID $y = \{1, 2\}$ ?

# Video-level instance ID assignment(6)

Assigned video-level instance ID

Clip-level Mask

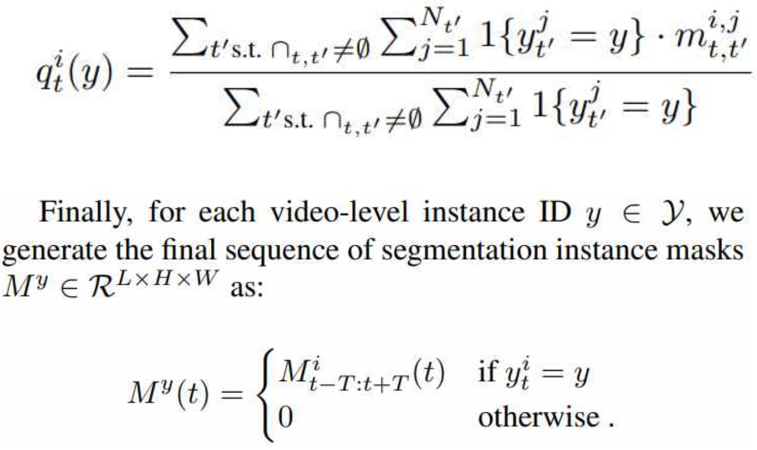


$y_1 = \{1,2\}$ $M^1_{1:2}$ $M^2_{1:2}$ t’ = 1

$y_2 = \{1,2,3\}$ $M^1_{1:3}$ $M^2_{1:3}$ $M^3_{1:3}$ t=2

| | $M^1_{1:3}$ | $M^2_{1:3}$ | $M^3_{1:3}$ |
|---|---|---|---|
| y=1 | $q_2^1(1) = \frac{m^{1,1}_{2,1}}{2}$ | $q_2^2(1) = 0$ | $q_2^3(1) = 0$ |
| y=2 | $q_2^1(2) = 0$ | $q_2^2(2) = \frac{m^{2,2}_{2,1}}{2}$ | $q_2^3(2) = \frac{m^{3,2}_{2,1}}{2}$ |
| $\max(q_t^i)$ | $q_2^1(1)$ | $q_2^2(2)$ | $q_2^3(2)$ |
| Assigned Video-level instance ID | $y^* = 1$ | $y^* = 2$ | 3 |

>TH



$$q_t^i(y) = \frac{\sum_{t' \text{s.t.} \cap_{t,t'} \neq \emptyset} \sum_{j=1}^{N_{t'}} 1\{y_{t'}^j = y\} \cdot m_{t,t'}^{i,j}}{\sum_{t' \text{s.t.} \cap_{t,t'} \neq \emptyset} \sum_{j=1}^{N_{t'}} 1\{y_{t'}^j = y\}}$$

Finally, for each video-level instance ID $y \in \mathcal{Y}$, we generate the final sequence of segmentation instance masks $M^y \in \mathcal{R}^{L \times H \times W}$ as:

$$M^y(t) = \begin{cases} M^i_{t-T:t+T}(t) & \text{if } y_t^i = y \\ 0 & \text{otherwise .} \end{cases}$$

Video-level Instance ID $y = \{1,2\}$ $\boxed{y = \{1,2,3\}}$

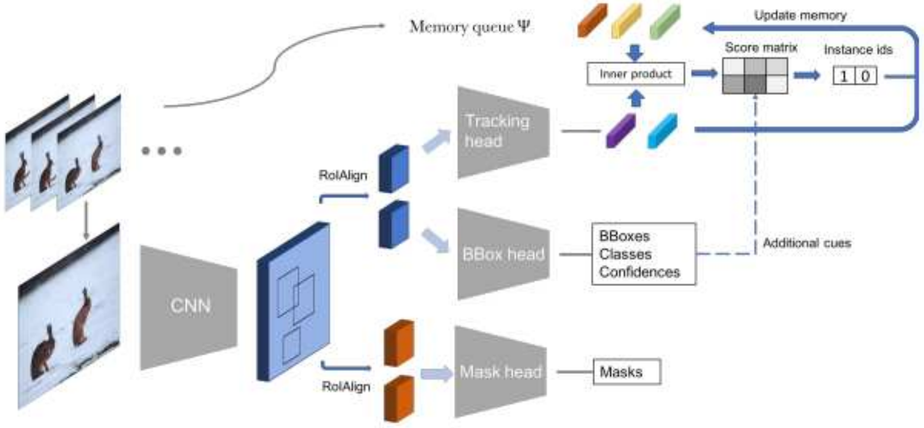
# Experimental condition(1)

- YouTube-VIS
  - 2, 238 training, 302 validation, and 343 test videos
  - Annotation : per-pixel segmentation, category, and instance labels
  - Contain 40 object categories
  - Since the evaluation on the test set is currently closed, we perform our evaluations on the validation set
- Assess each method according to
  - Mean video average precision (mAP),
  - Video average precision at IoU threshold of 75%
  - Average recall given 1 and 10 highest scored instances per video

# Experimental condition(2)

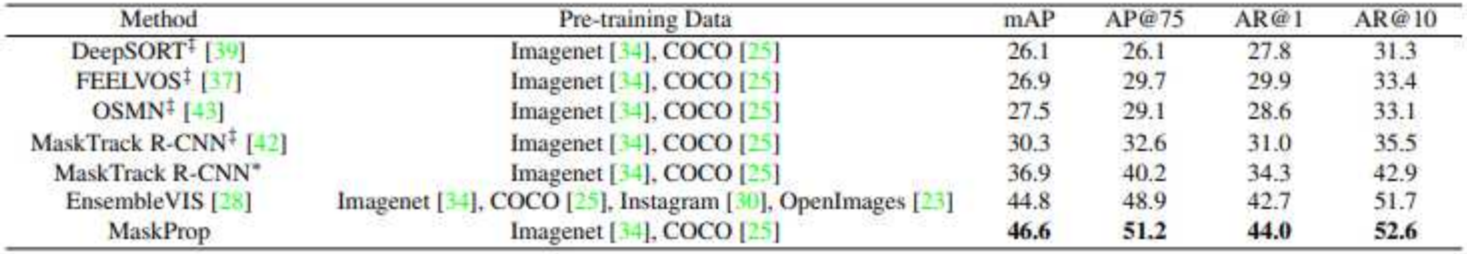
- EnsembleVIS
  - ICCV 2019 video instance segmentation challenge winner
  - Tackle video instance segmentation by dividing it into four individual problems: 1) detection, 2) classification, 3) segmentation, and 4) tracking

- MaskTrack R-CNN
  - Augments the original Mask R-CNN with a tracking branch that establishes associations among object instances segmented in separate frames.



▲ An overview of **MastTrack R-CNN**. A tracking head is embedded in the MaskRCNN framework to facilitate identity tracking of object instances through interaction with a memory queue. The memory queue is used to maintain all the existing object instances in the video.

# Experimental results(1)



| Method | Pre-training Data | mAP | AP@75 | AR@1 | AR@10 |
|---|---|---|---|---|---|
| DeepSORT‡ [39] | Imagenet [34], COCO [25] | 26.1 | 26.1 | 27.8 | 31.3 |
| FEELVOS‡ [37] | Imagenet [34], COCO [25] | 26.9 | 29.7 | 29.9 | 33.4 |
| OSMN‡ [43] | Imagenet [34], COCO [25] | 27.5 | 29.1 | 28.6 | 33.1 |
| MaskTrack R-CNN‡ [42] | Imagenet [34], COCO [25] | 30.3 | 32.6 | 31.0 | 35.5 |
| MaskTrack R-CNN* | Imagenet [34], COCO [25] | 36.9 | 40.2 | 34.3 | 42.9 |
| EnsembleVIS [28] | Imagenet [34], COCO [25], Instagram [30], OpenImages [23] | 44.8 | 48.9 | 42.7 | 51.7 |
| MaskProp | Imagenet [34], COCO [25] | **46.6** | **51.2** | **44.0** | **52.6** |

Table 2: The results of video instance segmentation on the YouTube-VIS [42] validation dataset. We evaluate the performance of each method according to mean average precision (mAP), average precision at 75% IoU threshold (AP@75), and average recall given top 1 (AR@1) and top 10 (AR@10) detections. The baselines denoted with ‡ were implemented by the authors in [42], whereas the methods marked with * were implemented by us, and use the same backbone and detection networks as our approach. Despite its simplicity, our MaskProp outperforms all prior video instance segmentation methods. Furthermore, we note that compared to EnsembleVIS [28], our approach uses orders of magnitude less labeled data for pre-training.

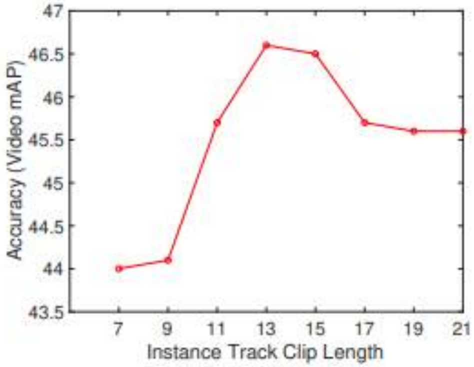
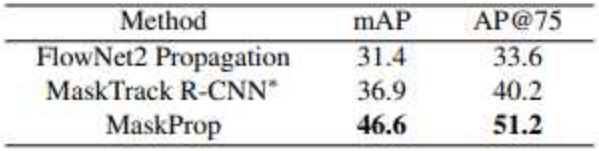


Figure 6: We plot video mAP as a function of an instance track clip length (denoted as $2T+1$ in the paper). Based on these results, we observe that optimal video instance segmentation performance is achieved when we propagate instance masks to $T = 6$ previous and subsequent frames.



| Method | mAP | AP@75 |
|---|---|---|
| FlowNet2 Propagation | 31.4 | 33.6 |
| MaskTrack R-CNN* | 36.9 | 40.2 |
| MaskProp | **46.6** | **51.2** |

Table 3: Here, we study the effectiveness of our mask propagation branch. If we replace it with the FlowNet2 propagation scheme, where masks are propagated using the optical flow predicted by a FlowNet2 network [20], the accuracy drops from 46.6 mAP to 31.4 mAP. Similarly, if we replace our mask propagation branch with the tracking branch from MaskTrack R-CNN, the accuracy drops to 36.9 mAP. Note that all of these baselines are implemented using the same backbone and detection networks.

# Experimental results(2)

- The performance of MaskProp is especially noticeable when a video contains large object motion, occlusions, or overlapping objects.

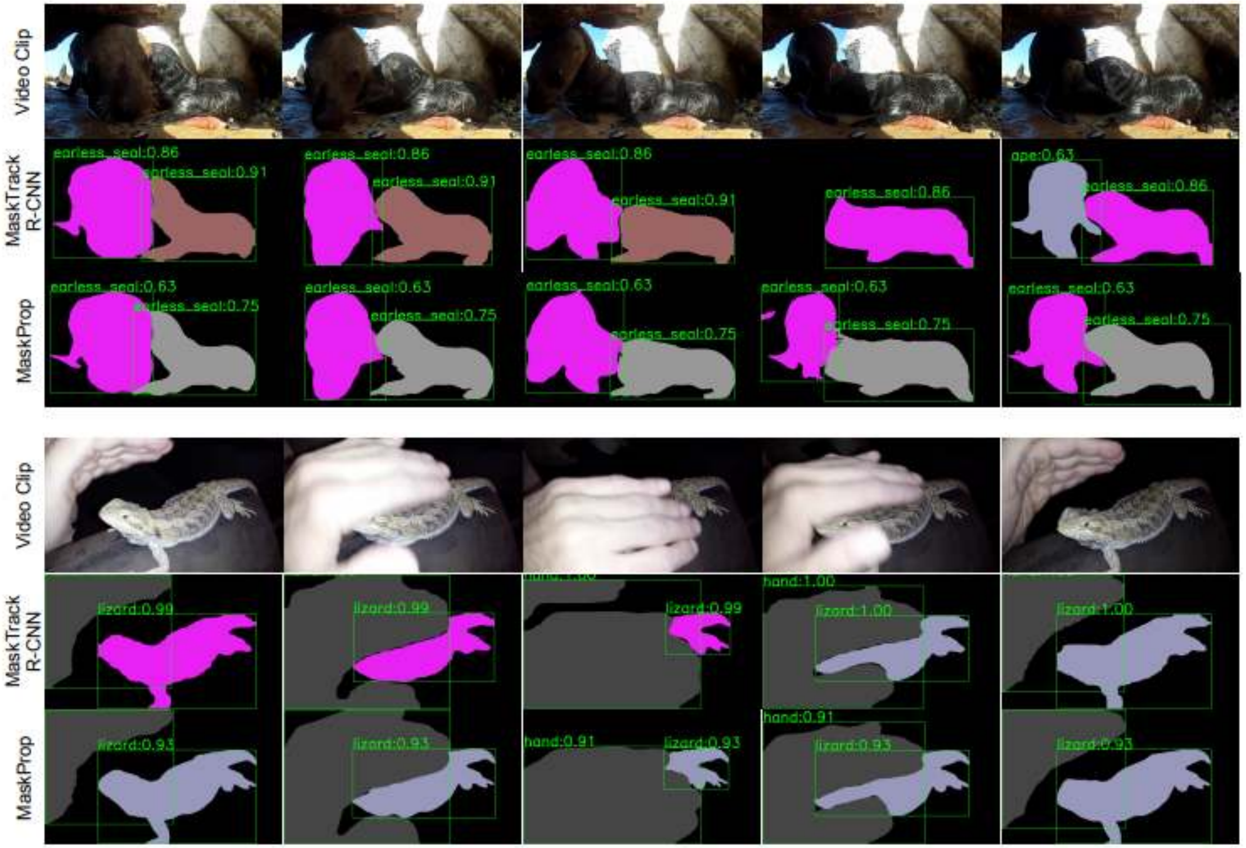


Figure 7: We compare our video instance segmentation results with MaskTrack R-CNN [42] predictions. Different object instances are encoded with different colors. The first row for each video shows the original frames. The second row illustrates the mask predictions of MaskTrack R-CNN and the third row those obtained with our MaskProp. Compared to MaskTrack R-CNN, our MaskProp tracks object instances more robustly even when they are occluded or overlap with each other. Additional video instance segmentations produced by our method are included in our supplementary video[1].

# Experimental results(3)

- MaskProp reliably propagates features that are specific to each instance despite motion blur, object deformations and large variations in object appearance

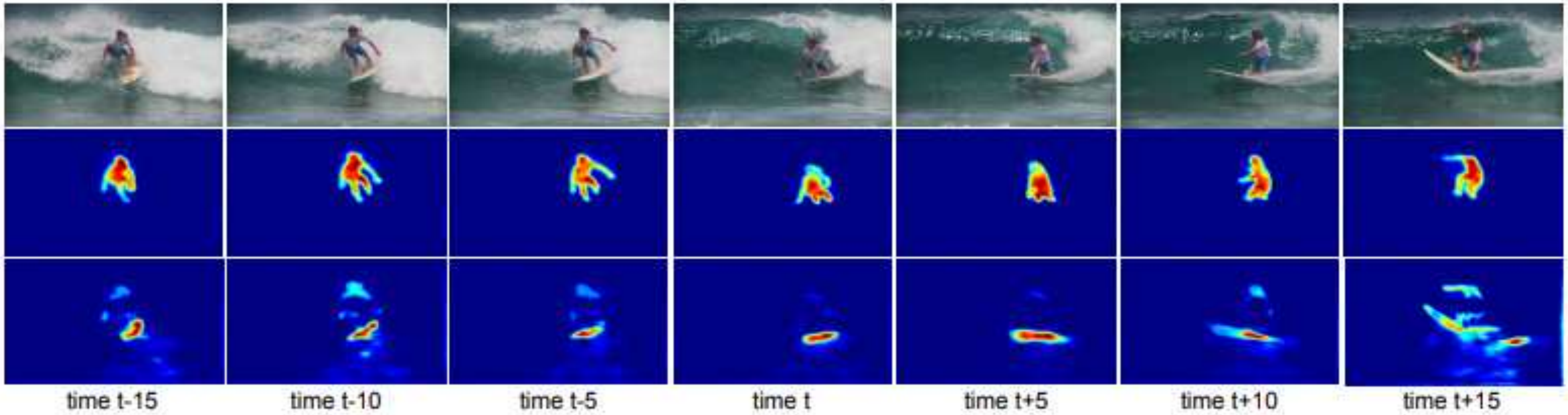


Figure 5: An illustration of instance-specific features propagated from frame $t$ to other frames in the given video clip. Here, we visualize propagated activations from one randomly selected feature channel. The activations in the two rows correspond to two different object instances detected at time $t$. Our visualizations suggest that MaskProp reliably propagates features that are specific to each instance even when instances appear next to each other, and despite the changes in shape, pose and the nuisances effects of deformation and occlusion.

The end

DAVIAN