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

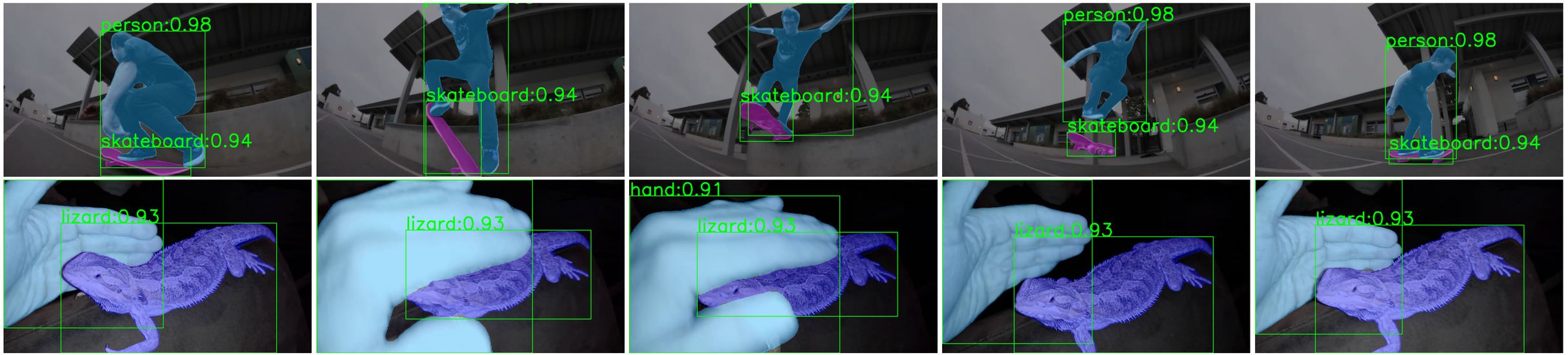
FACEBOOK AI

Gedas Bertasius

Lorenzo Torresani

Problem Overview

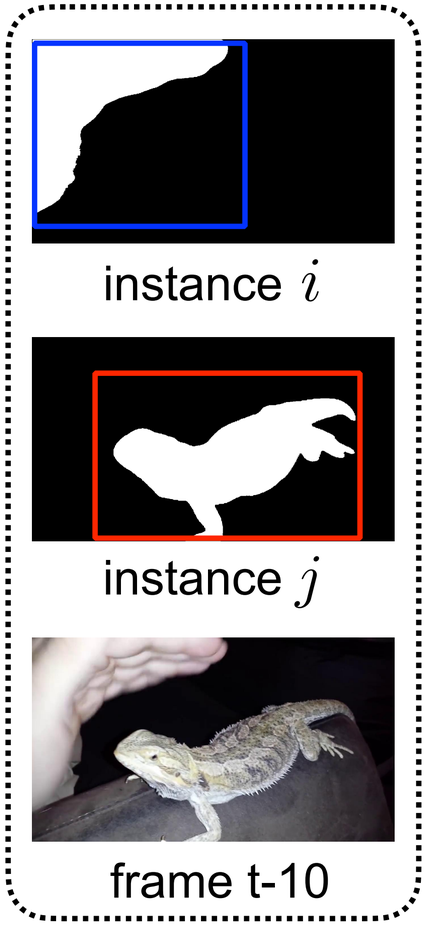
- Video instance segmentation task requires classifying, segmenting, and tracking object instances in video.

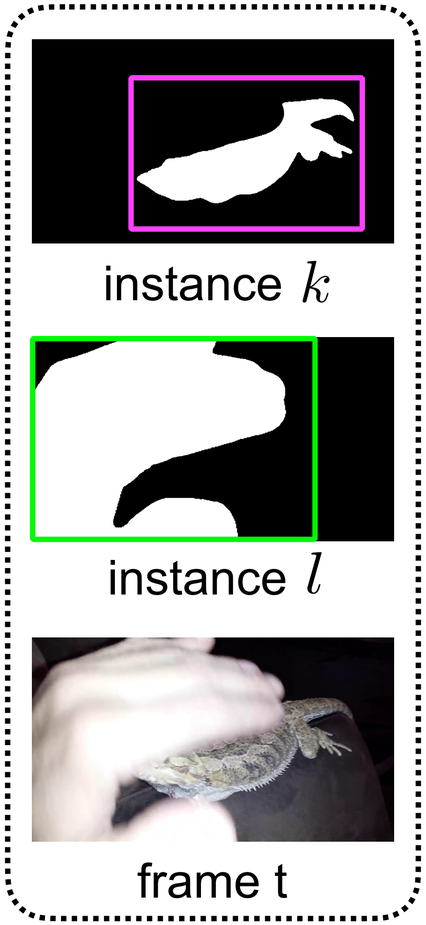


Frame t-20 Frame t-10 Frame t Frame t+10 Frame t+20

Challenges

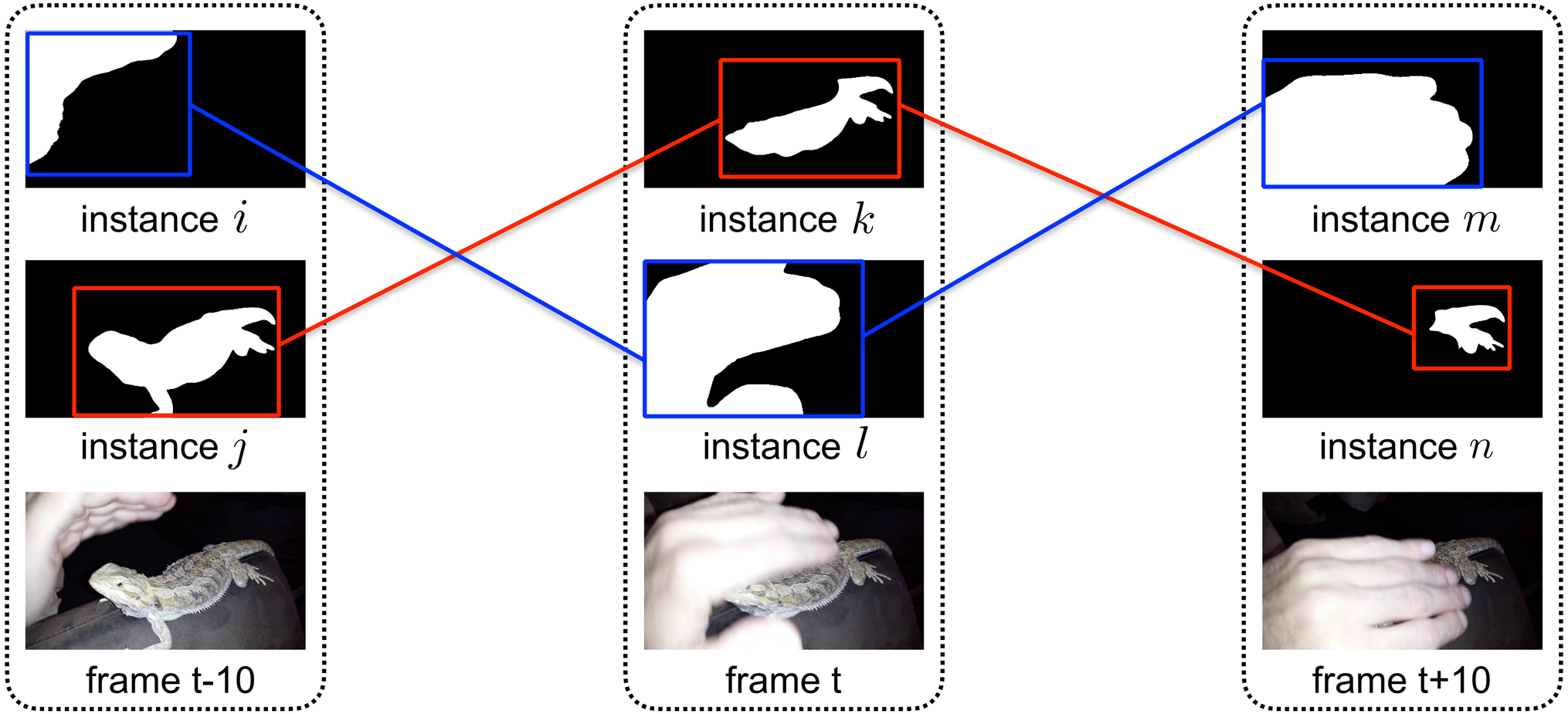
- How can we link instances across frames in an end-to-end learnable fashion?





Challenges

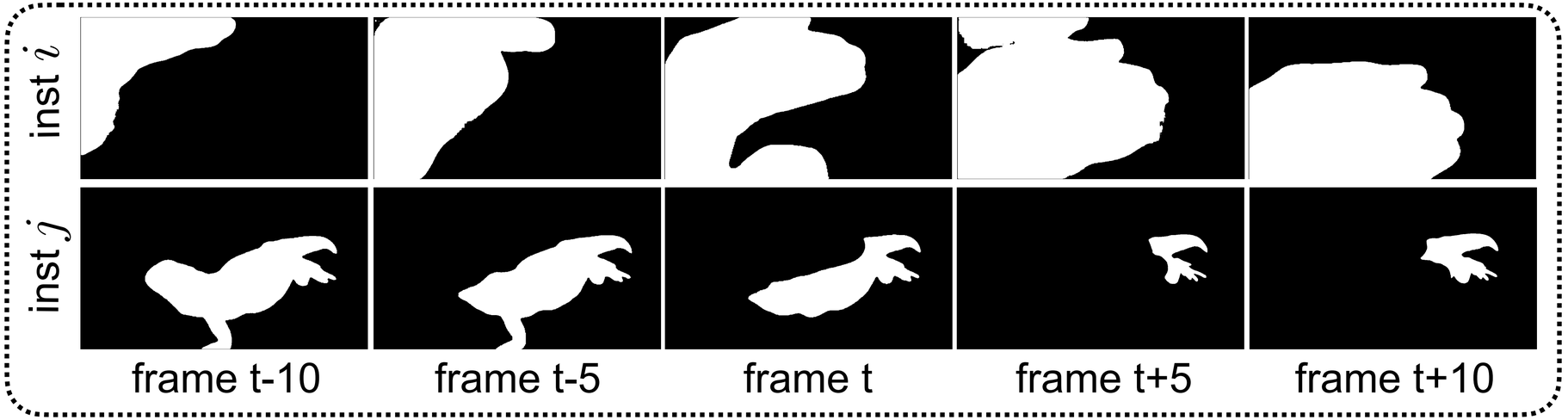
- How can we link instances across frames in an end-to-end learnable fashion?



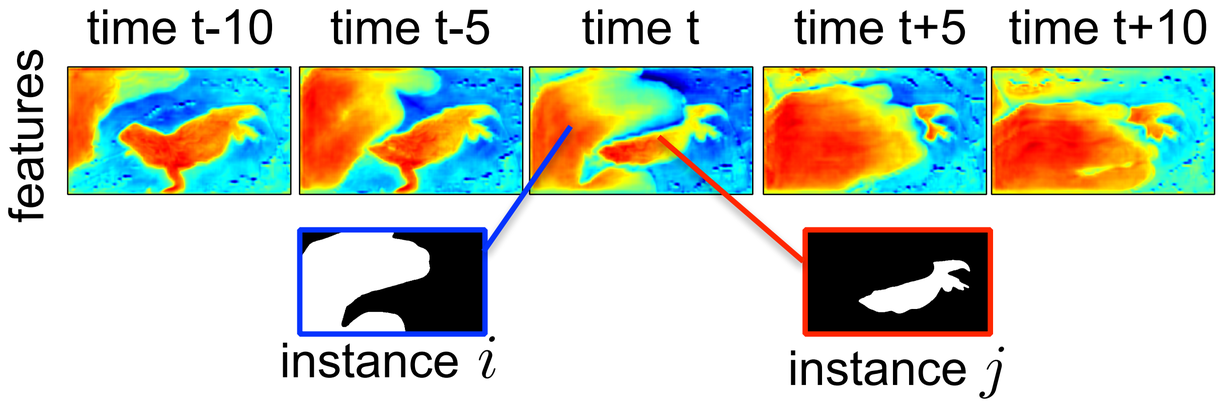
Mask Propagation

Input: a video clip of ~15 frames (~3 seconds).

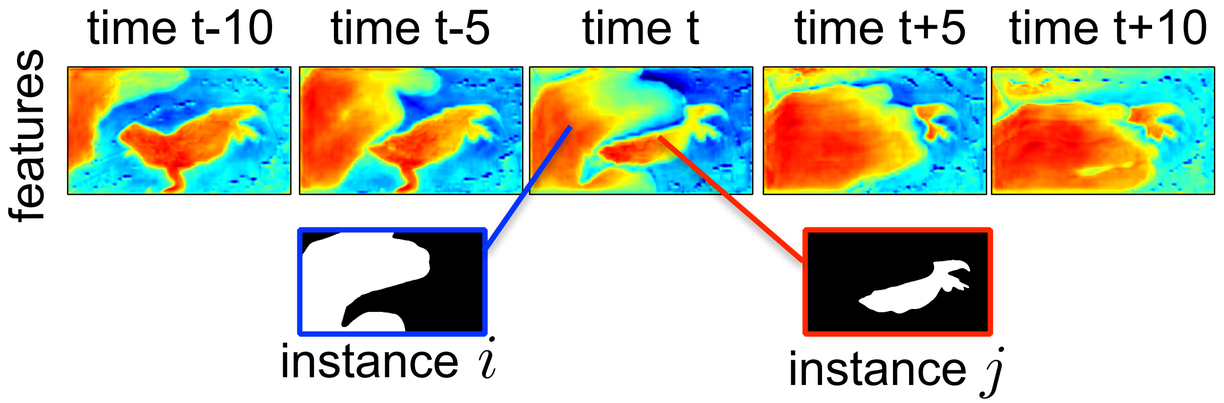
Output: clip-level instance tracks.



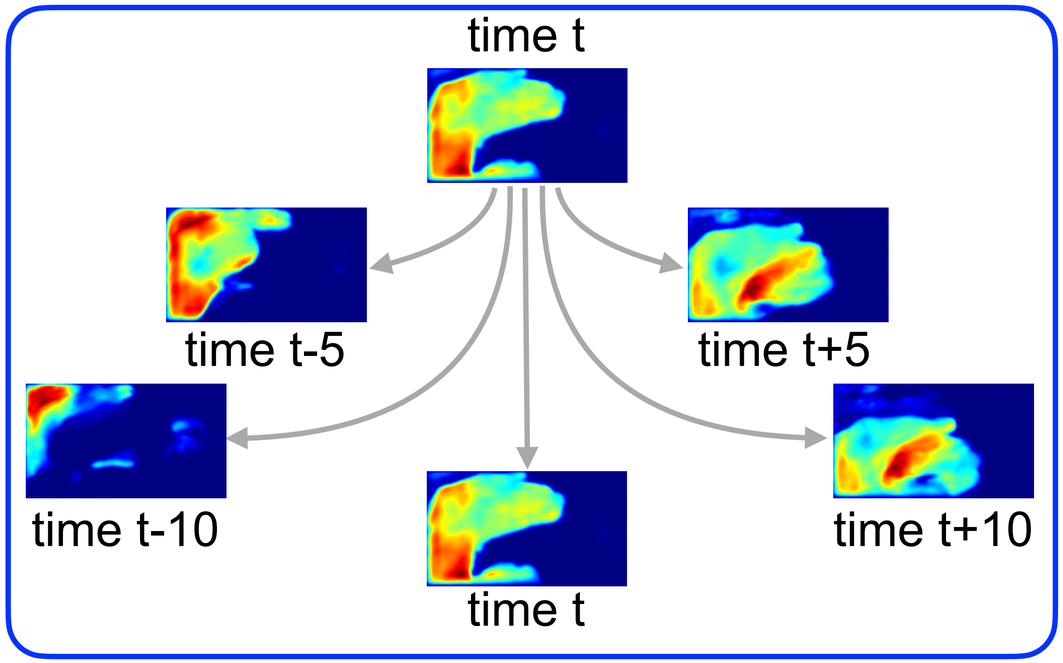
Mask Propagation



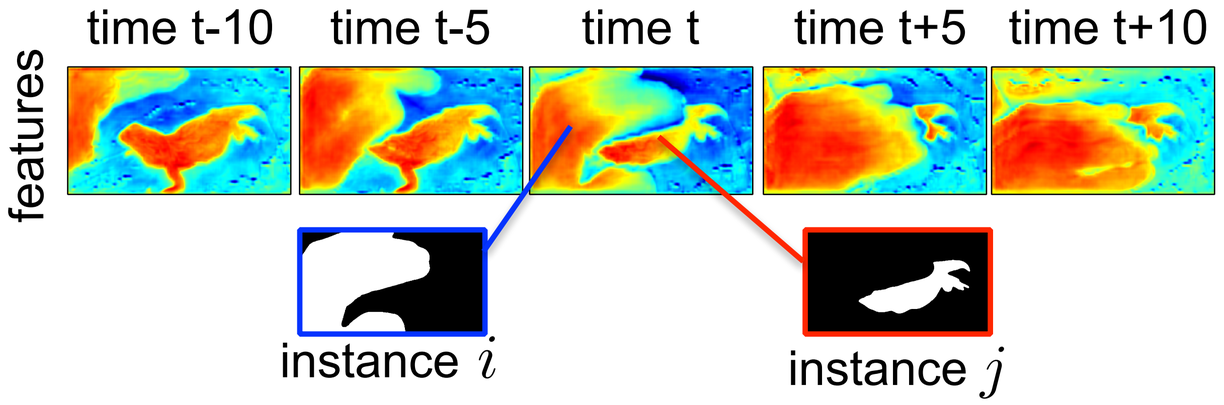
Mask Propagation



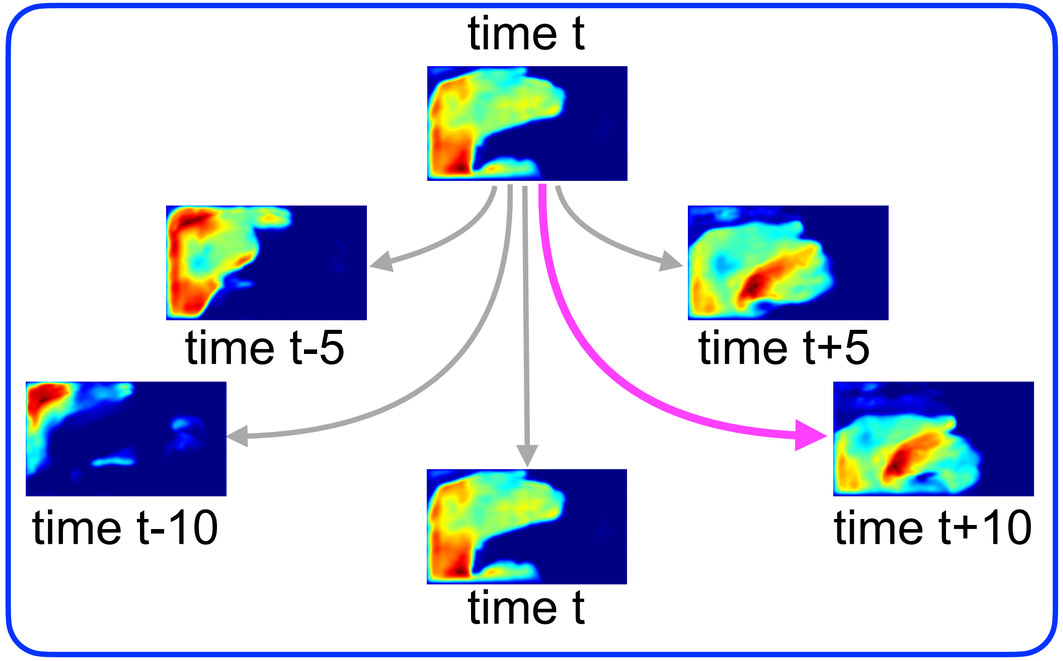
temporal feature propagation for instance i



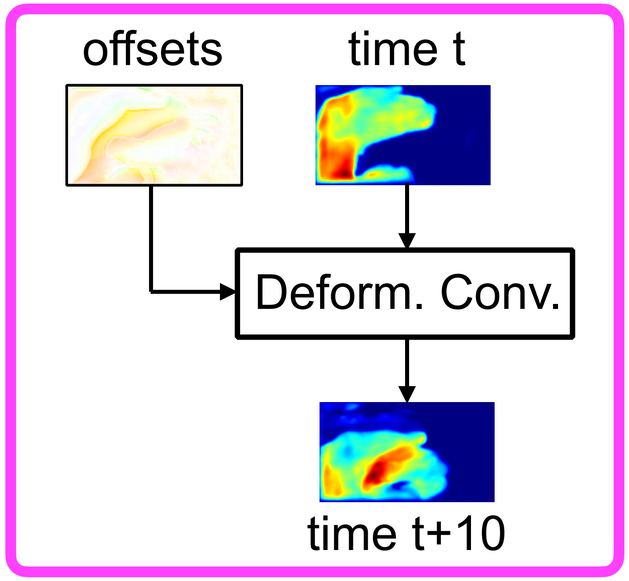
Mask Propagation



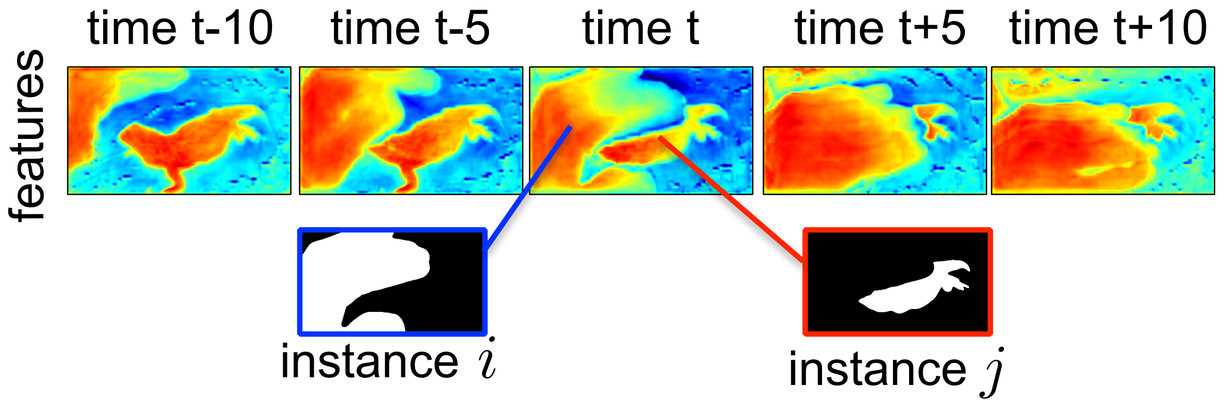
temporal feature propagation for instance i



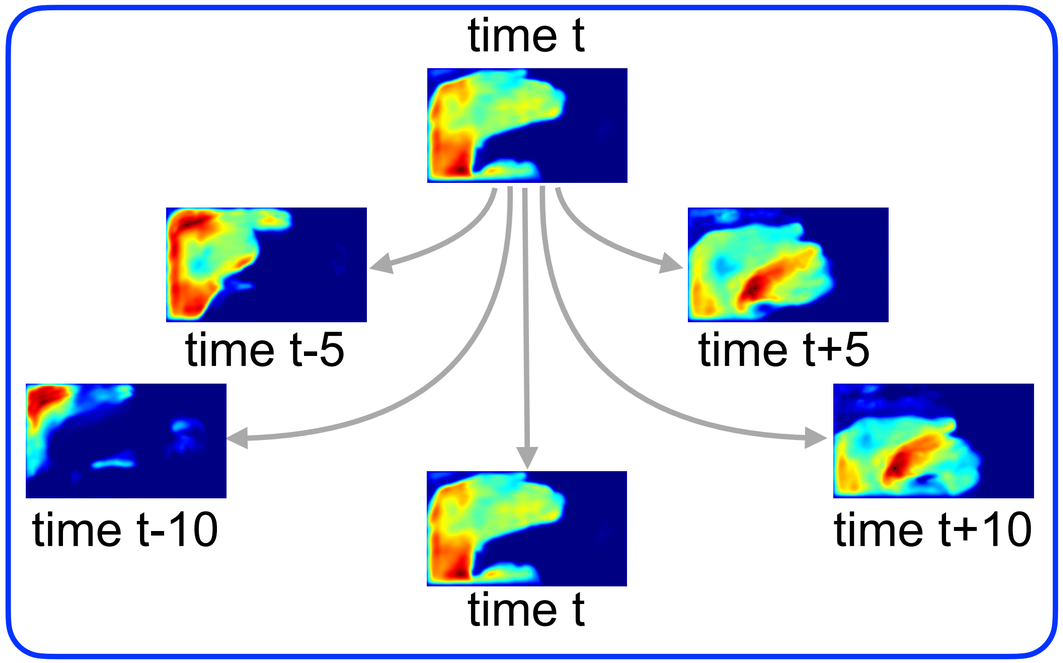
feature warping using deformable convolution



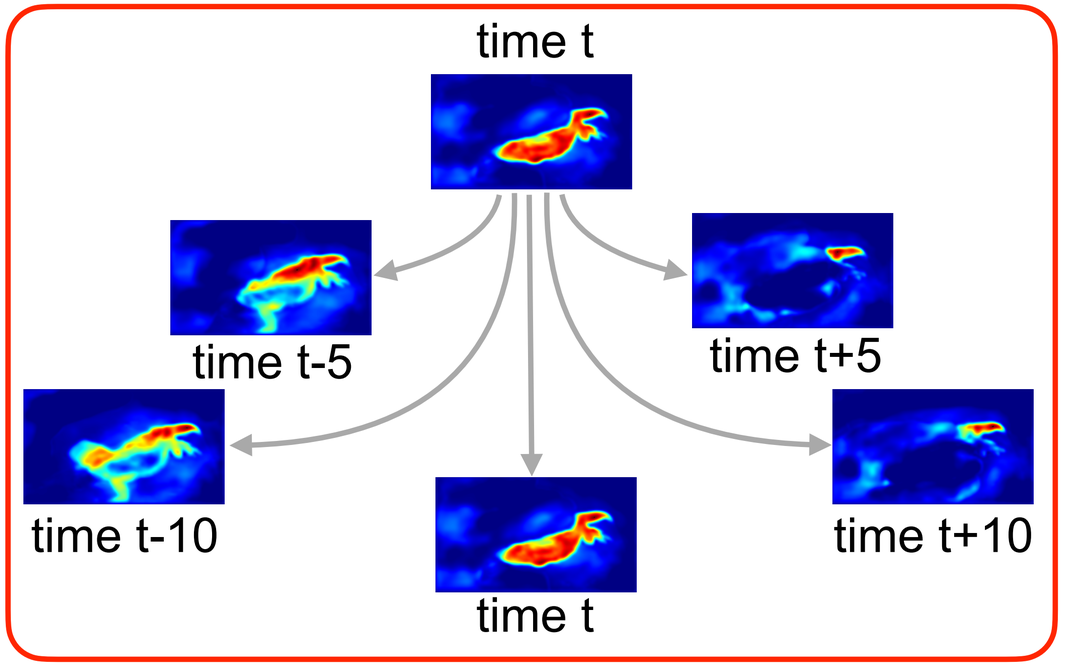
Mask Propagation



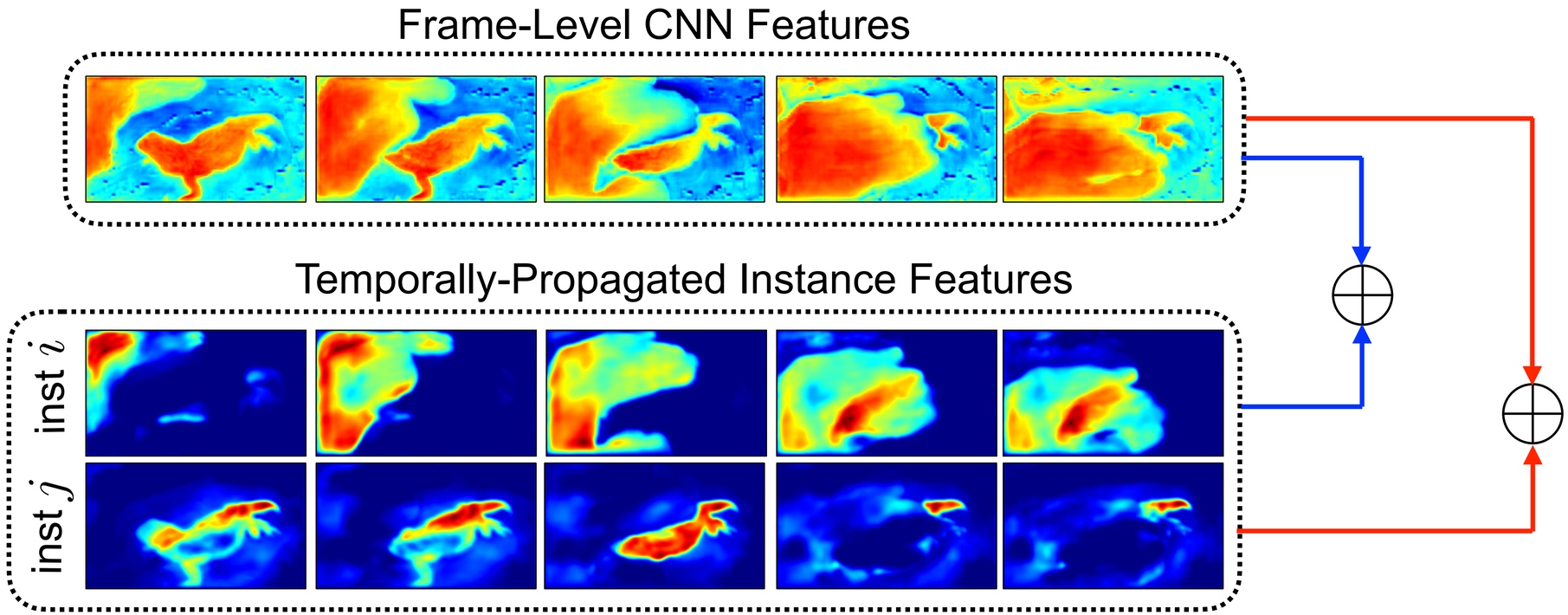
temporal feature propagation for instance i



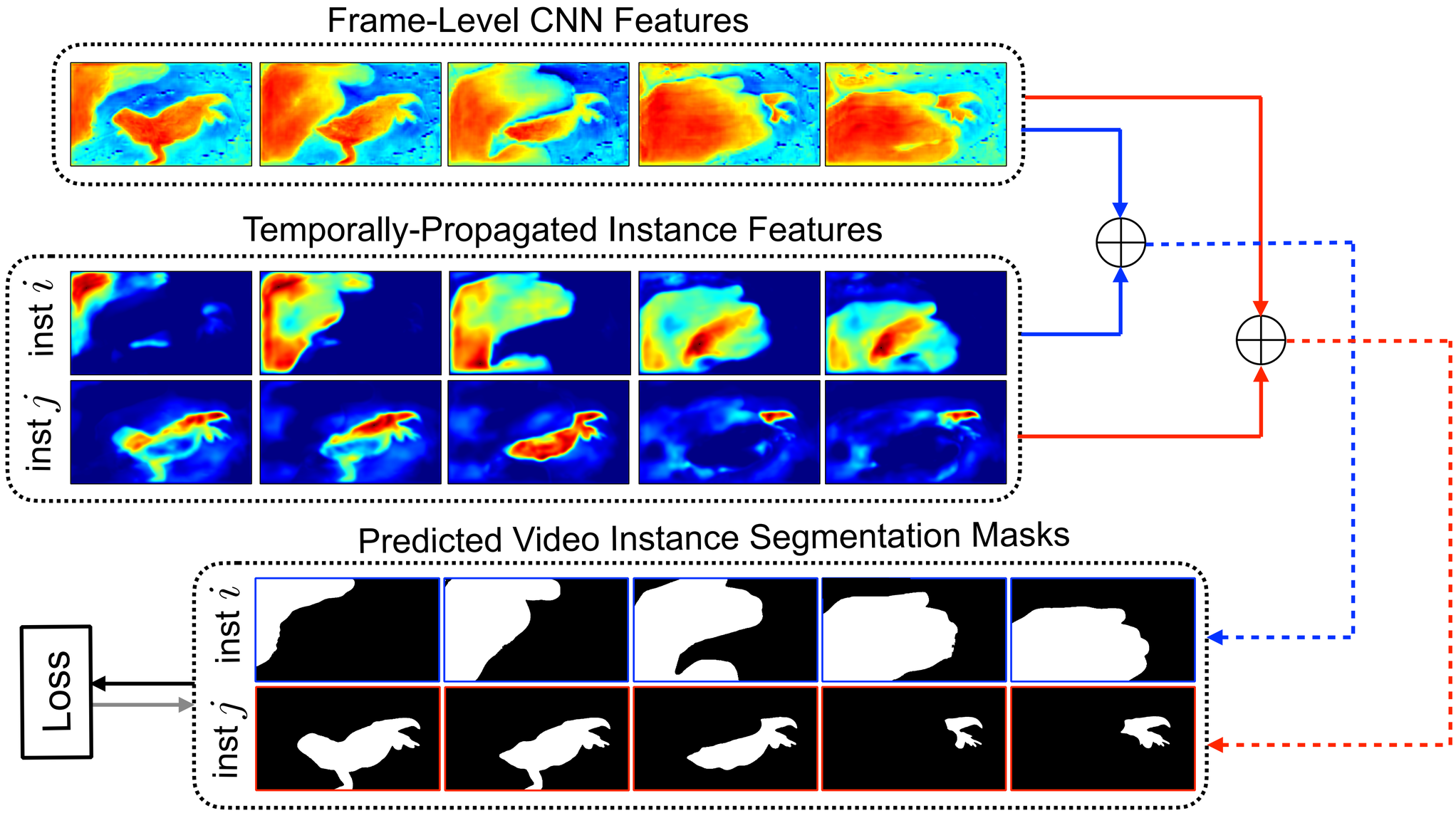
temporal feature propagation for instance j



Mask Propagation



Mask Propagation

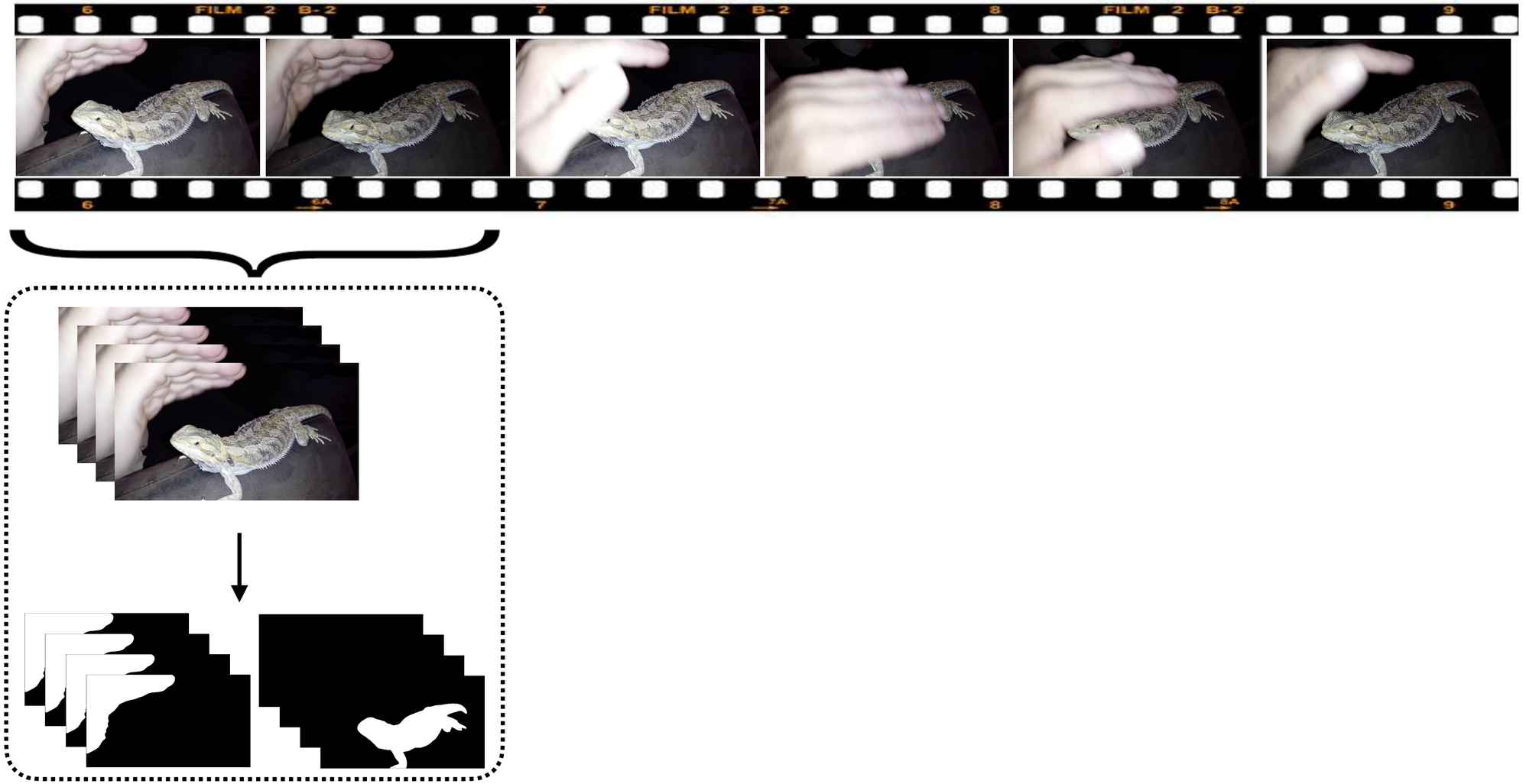


Linking Clip-Level Tracks

A video of arbitrary length

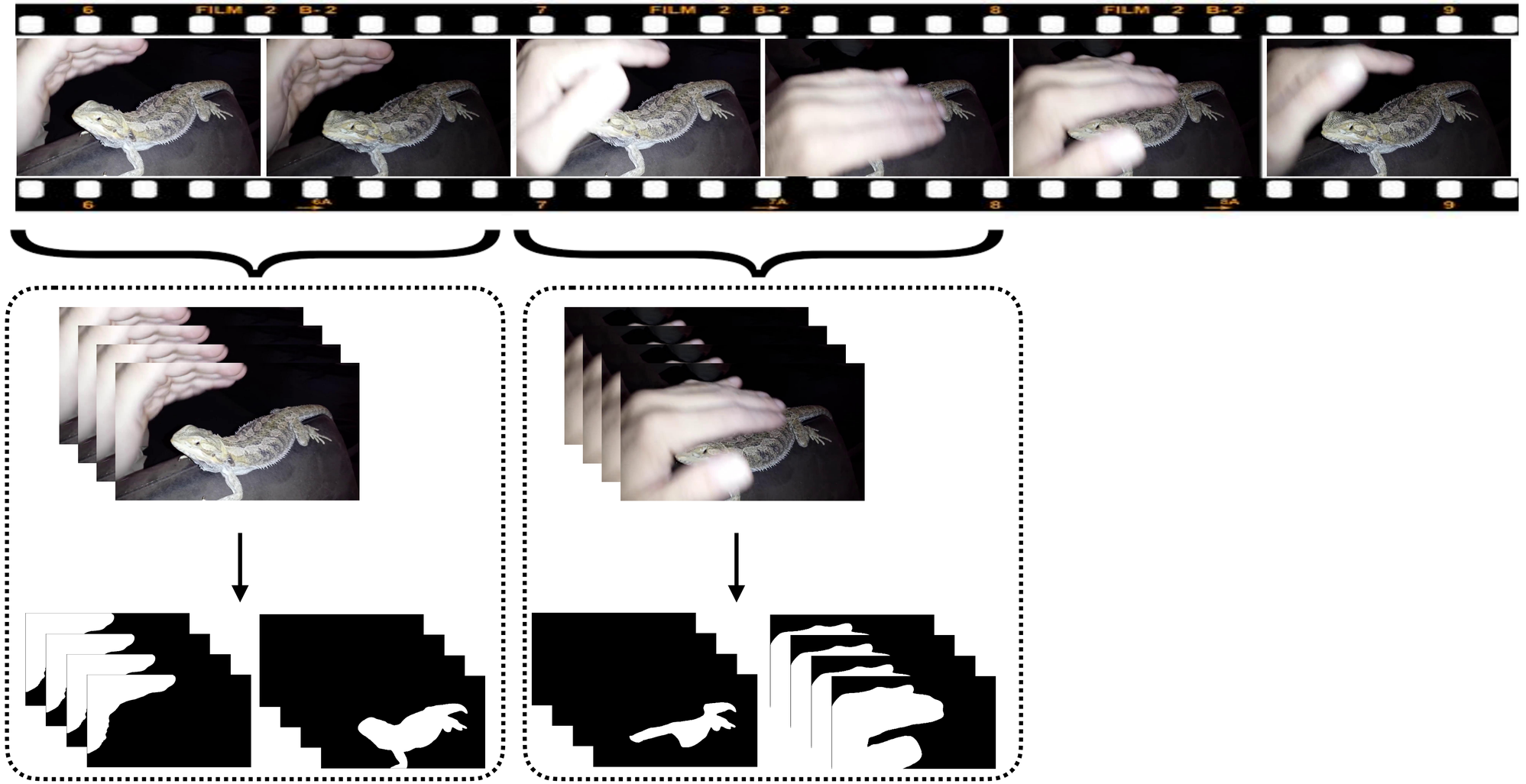
Linking Clip-Level Tracks

A video of arbitrary length



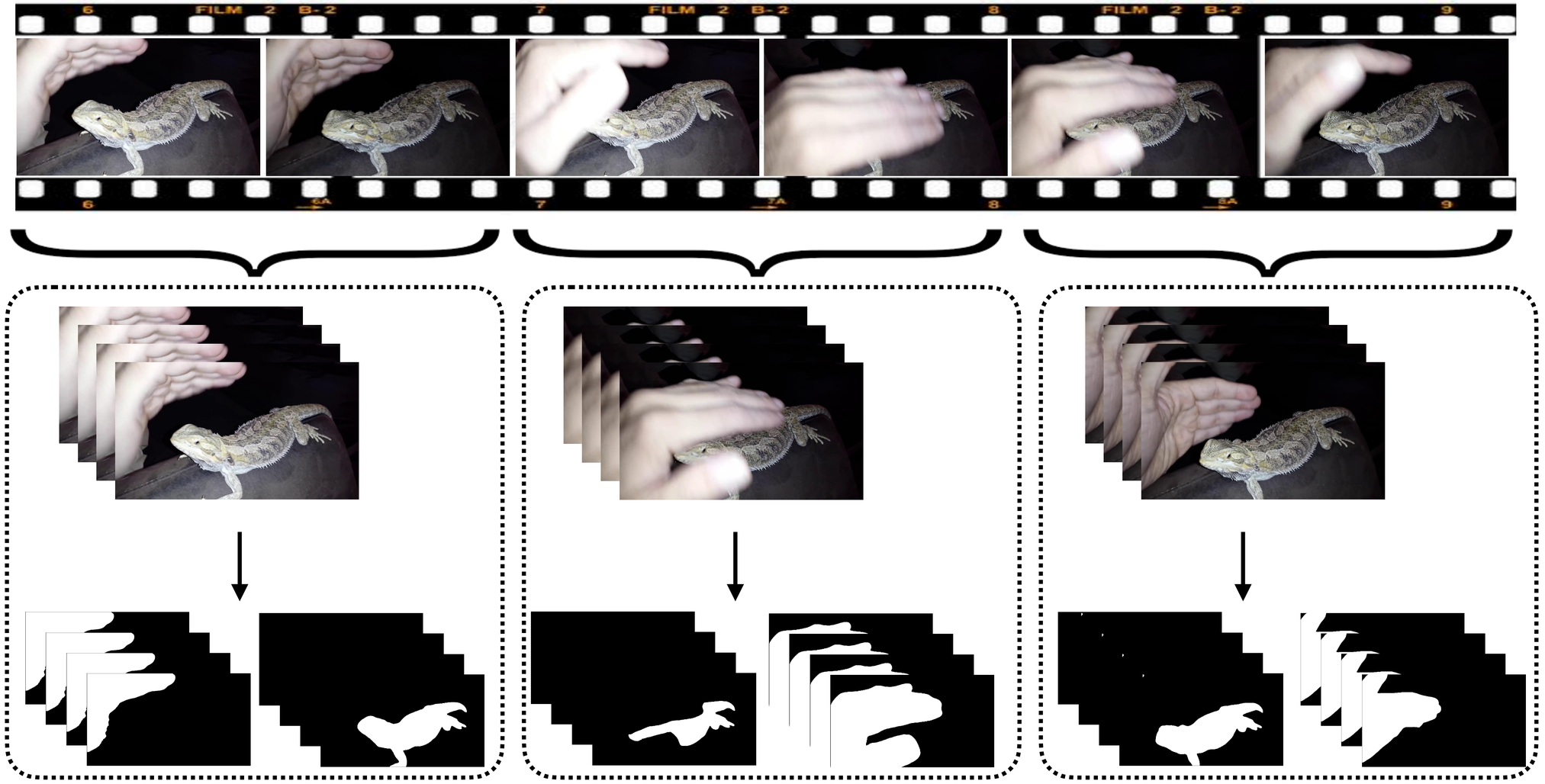
Linking Clip-Level Tracks

A video of arbitrary length

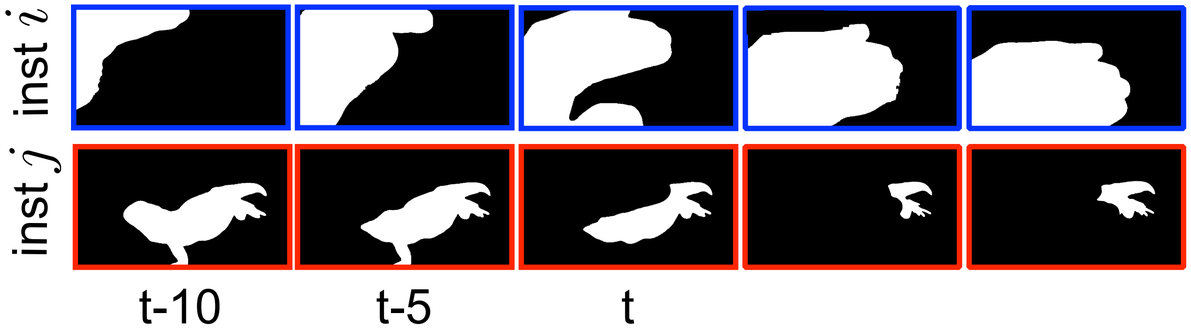


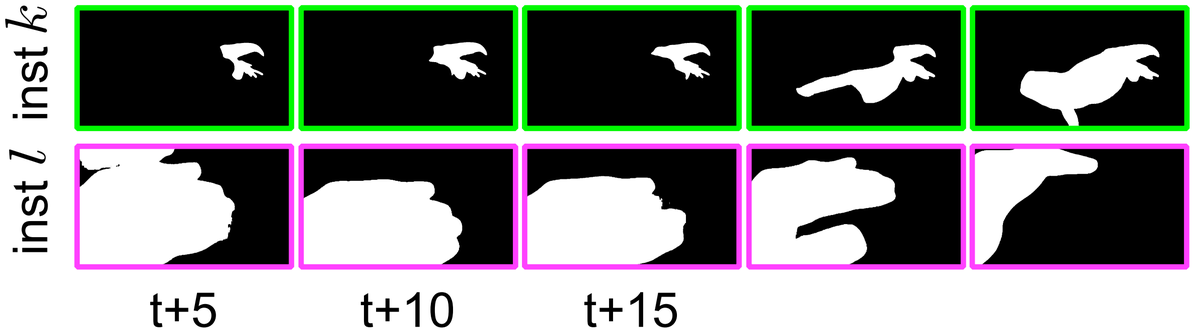
Linking Clip-Level Tracks

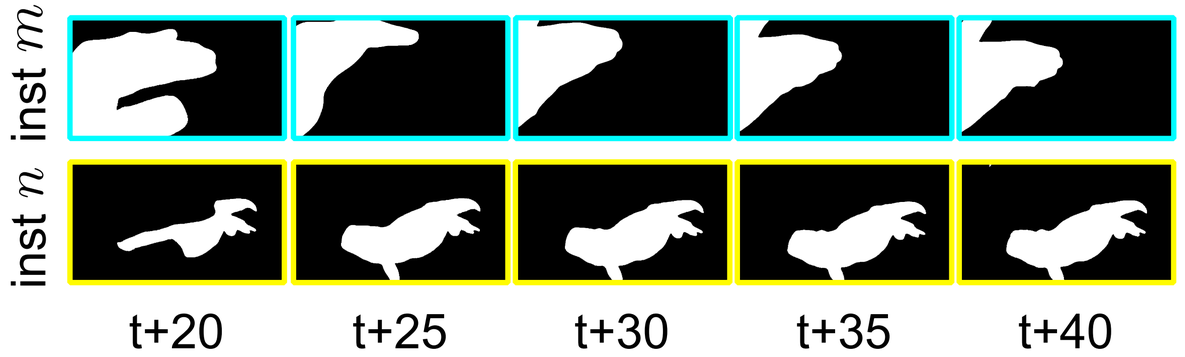
A video of arbitrary length



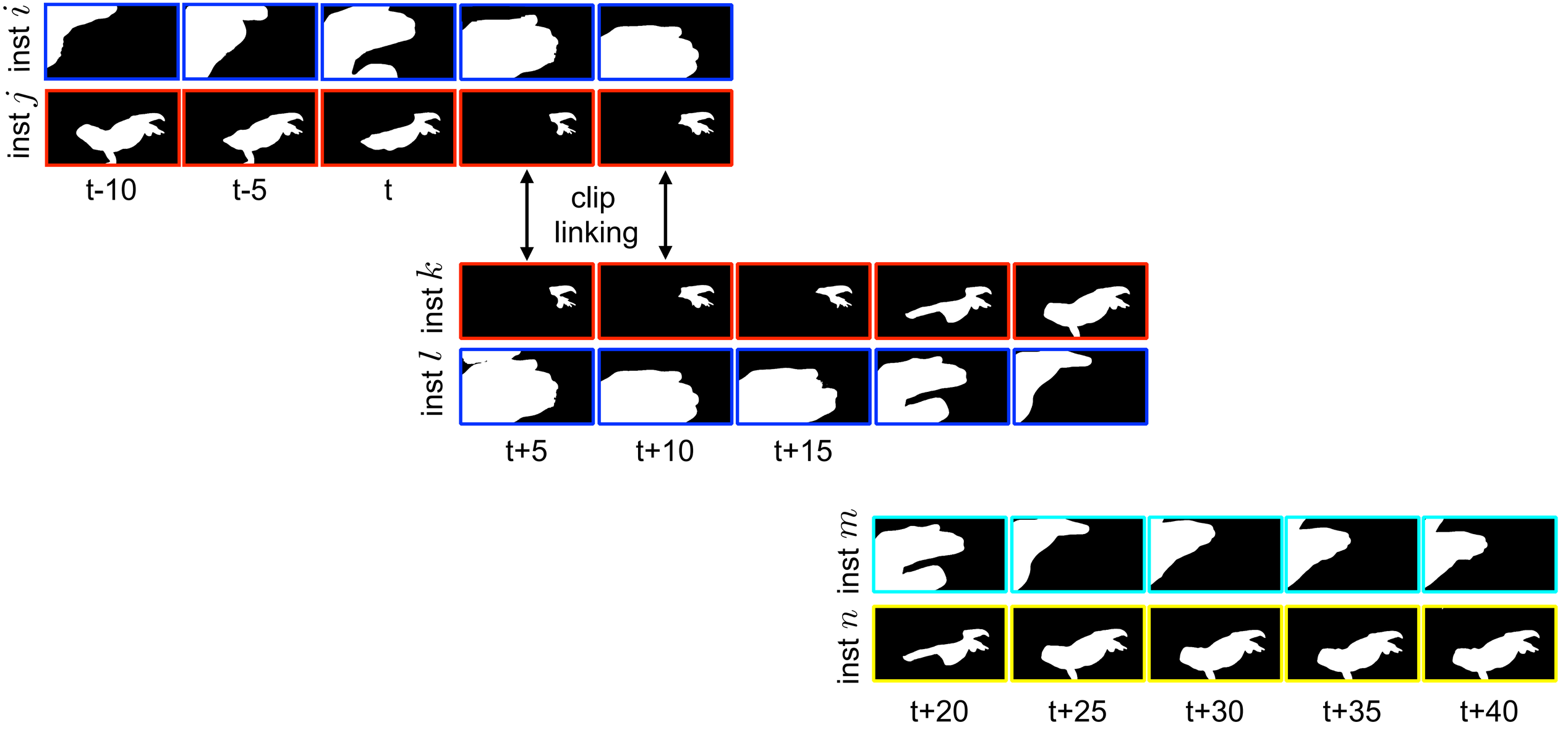
Linking Clip-Level Tracks



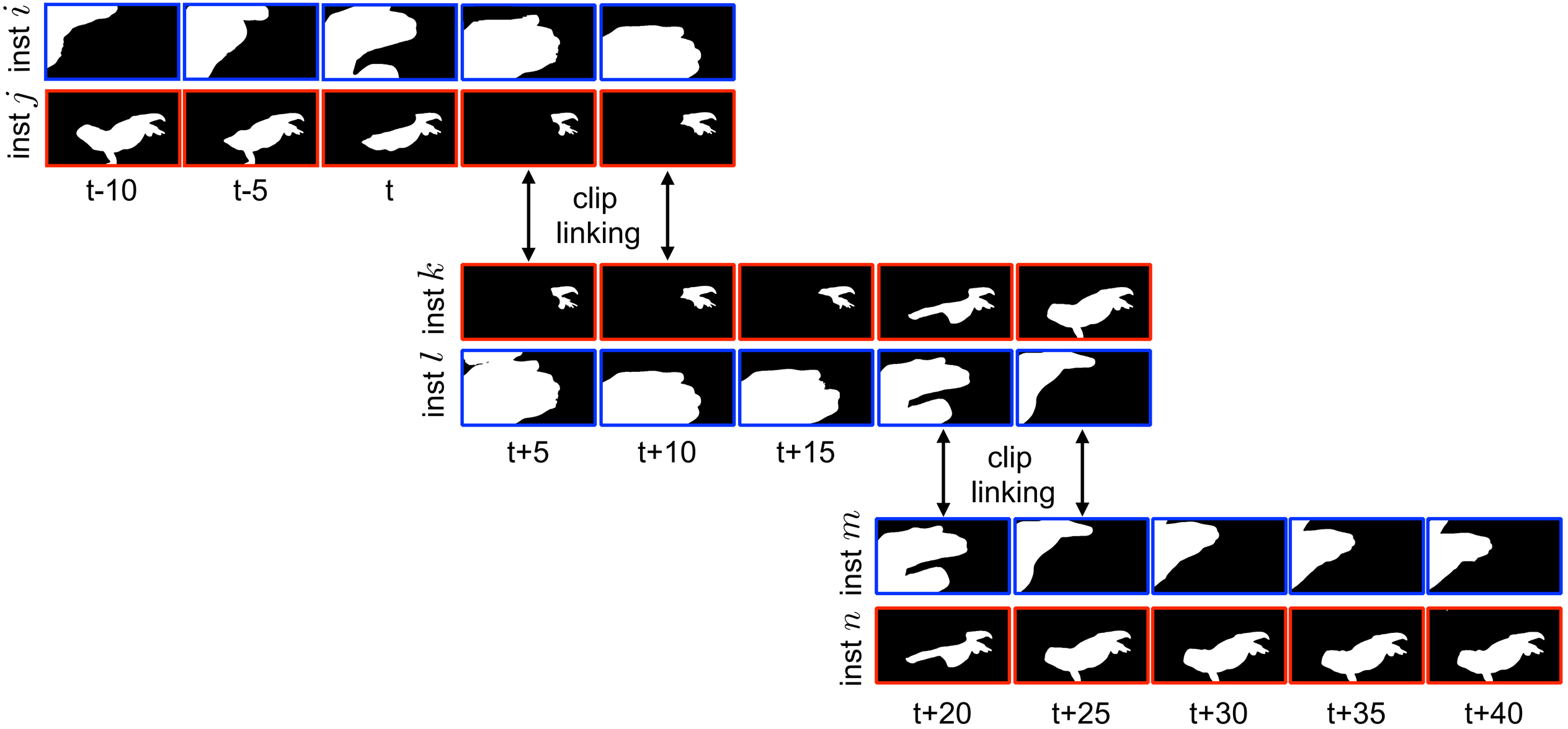




Linking Clip-Level Tracks



Linking Clip-Level Tracks

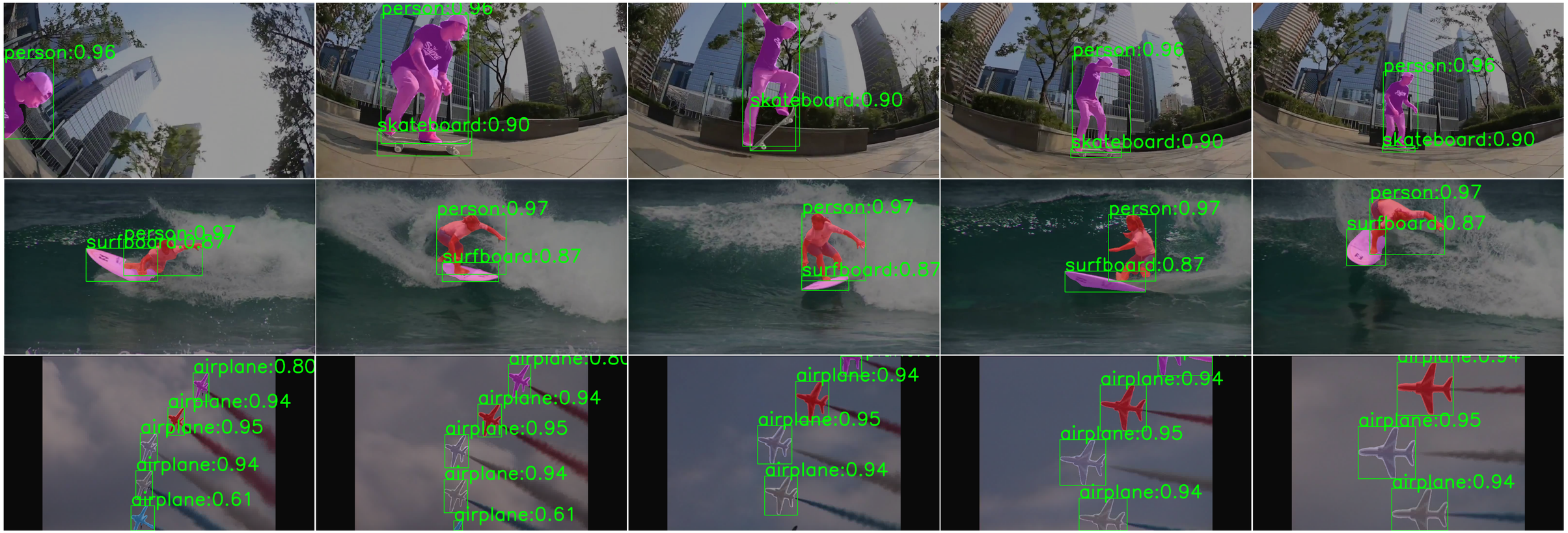


Results on YouTube-VIS

		MaskTrack R-CNN [42]	ICCV19 Challenge Winner [28]	MaskProp
Model	Classification	cls head	Mask R-CNN [16], ResNeXt-101 32x48d [30]	cls head
	Localization	bbox head	Mask R-CNN [16]	bbox head
	Segmentation	mask head	DeepLabv3 [9], Box2Seg [29]	mask head
	Tracking	tracking head	UnOVOST [47], ReID Net [18, 31]	mask propagation head
	Optical Flow	-	PWC-Net [35]	-
Pre-training Datasets	ImageNet [34] (1.3M images)	✓	✓	✓
	COCO [25] (860K bboxes)	✓	✓	✓
	Instagram [30] (1B images)	-	✓	-
	OpenImages [23] (14M bboxes)	-	✓	-
Performance	video mAP	30.3	44.8	**46.6**
	video AP@75	32.6	48.9	**51.2**

Results on YouTube-VIS

- MaskProp tracks objects robustly even when they are occluded or overlap with each other.



Project Page

https://gberta.github.io/maskprop/