SENSETIME
商 汤 科 技

Video Object Segmentation with Re-identification

Xiaoxiao Li, Yuankai Qi, Zhe Wang, Kai Chen, Ziwei Liu, Jianping Shi
Ping Luo, Chen Change Loy, Xiaoou Tang

The Chinese University of Hong Kong, SenseTime Group Limited

Semi-supervised Segmentation

- Input :Video sequence, ground-truth label of the first frame

- Output : Masks of all instances

Challenge

- Instance Segmentation
  - Small objects and fine structures
  - Scale & pose-variations
- Tracking
  - Frequent occlusions

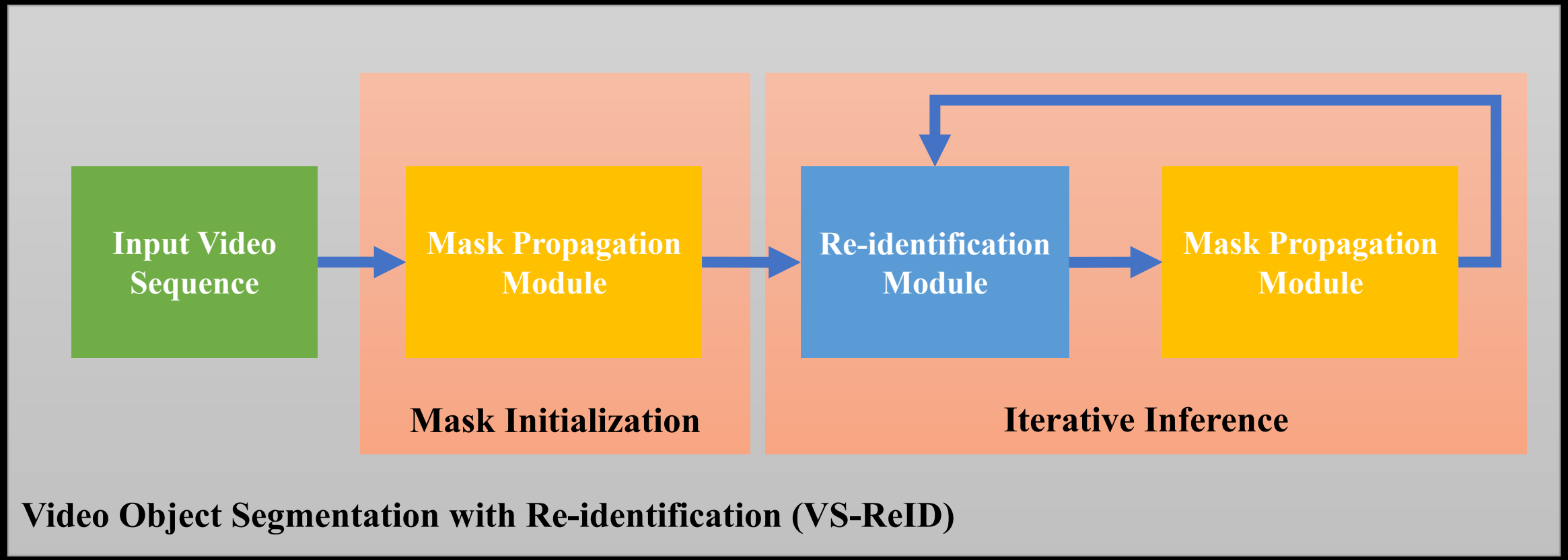
Challenge

- Instance Segmentation
  - Small objects and fine structures
  - Scale & pose-variations
- Tracking
  - Frequent occlusions

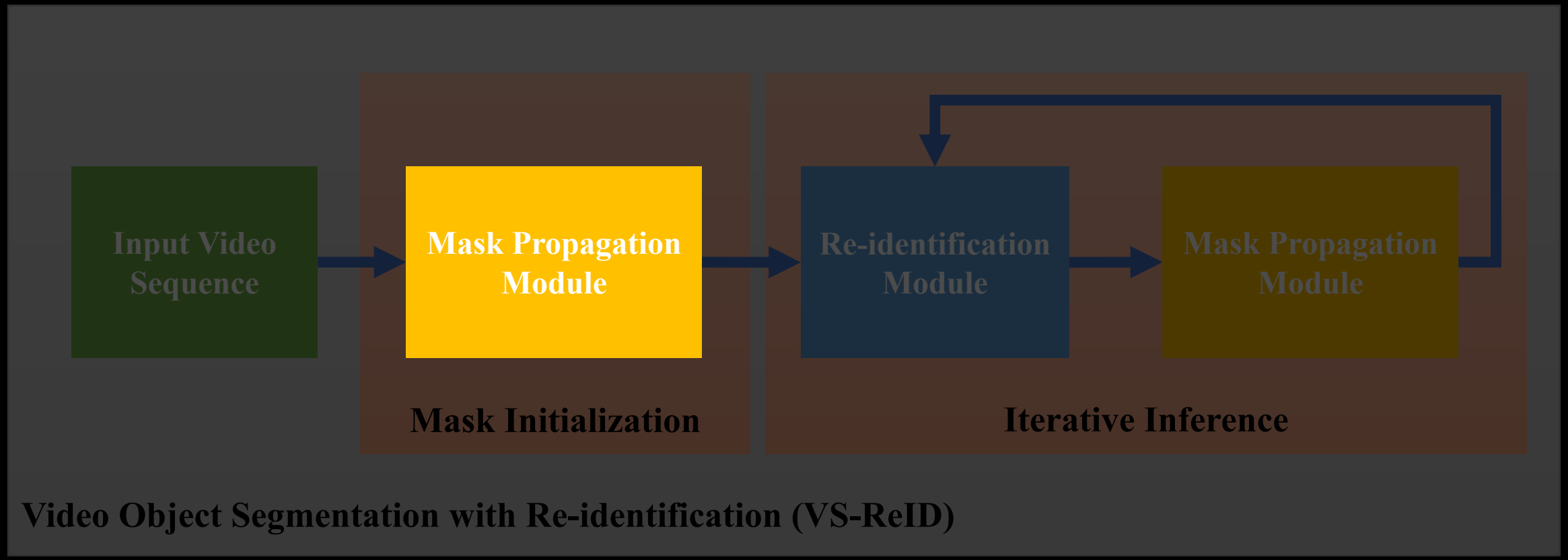
Mask Propagation Module — Short Term

Re-identification Module — Long Term

Proposed Framework



Mask Propagation Module

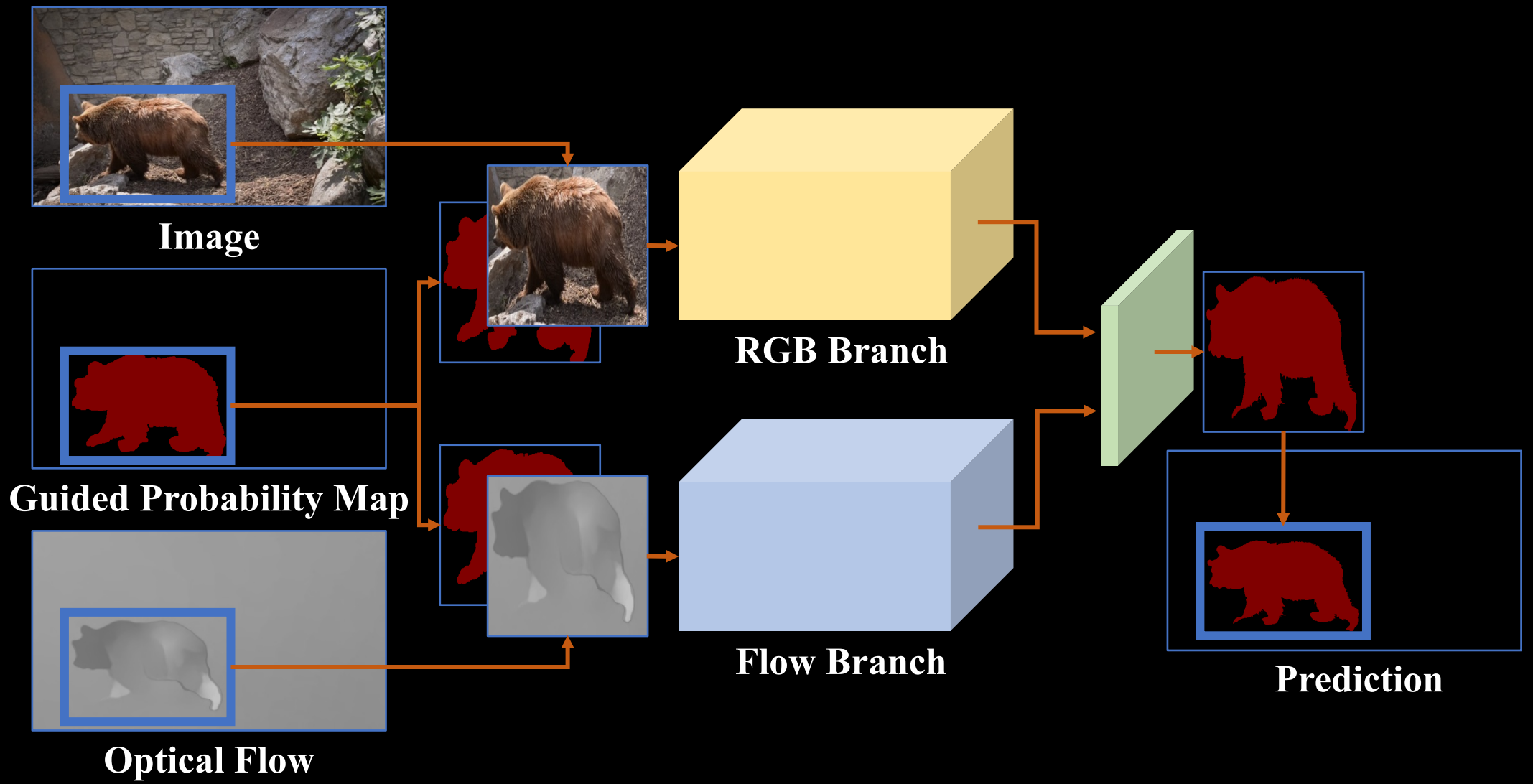


Mask Propagation Module

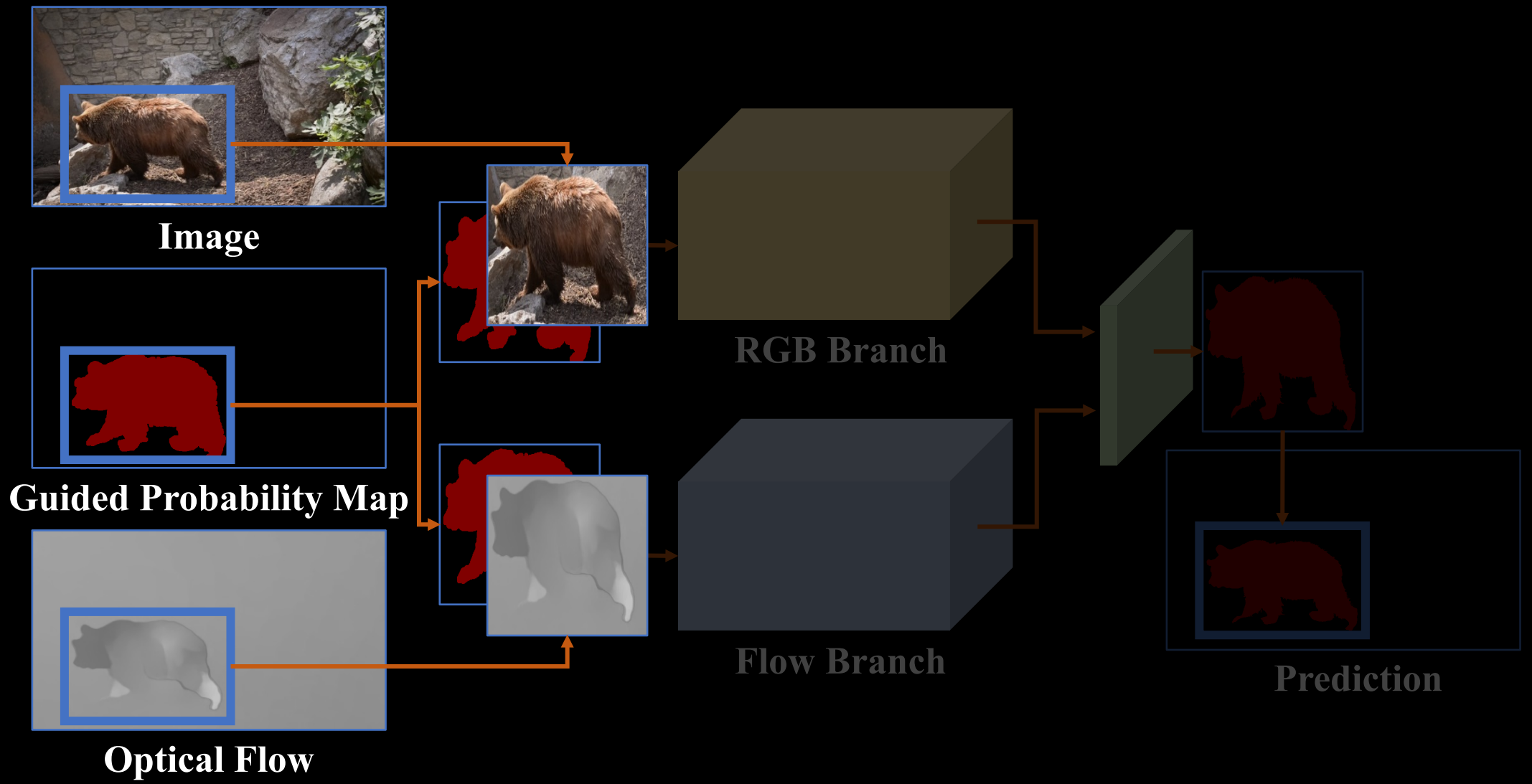
- Inspired by MSK[1] and LucidTracker[2]
- Use the temporal continuity property of the video sequence
- Propagate the mask from the previous frame to the current frame

[1] Perazzi F, Khoreva A, Benenson R, et al. Learning video object segmentation from static images[C]. CVPR, 2017.
[2] Khoreva A, Benenson R, Ilg E, et al. Lucid Data Dreaming for Object Tracking[J]. arXiv preprint arXiv:1703.09554, 2017.

Mask Propagation Module

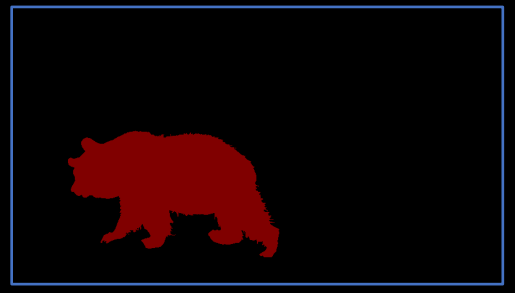


Mask Propagation Module



Mask Propagation Module

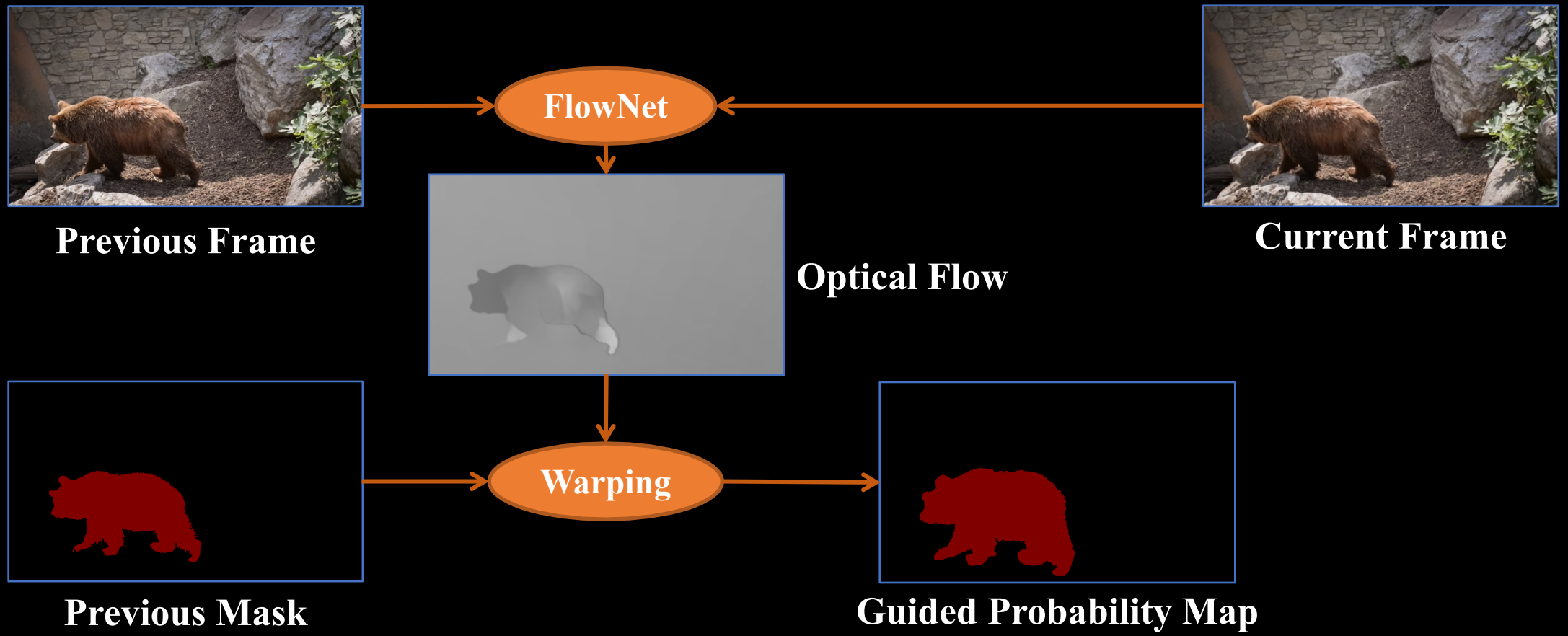
Previous Frame



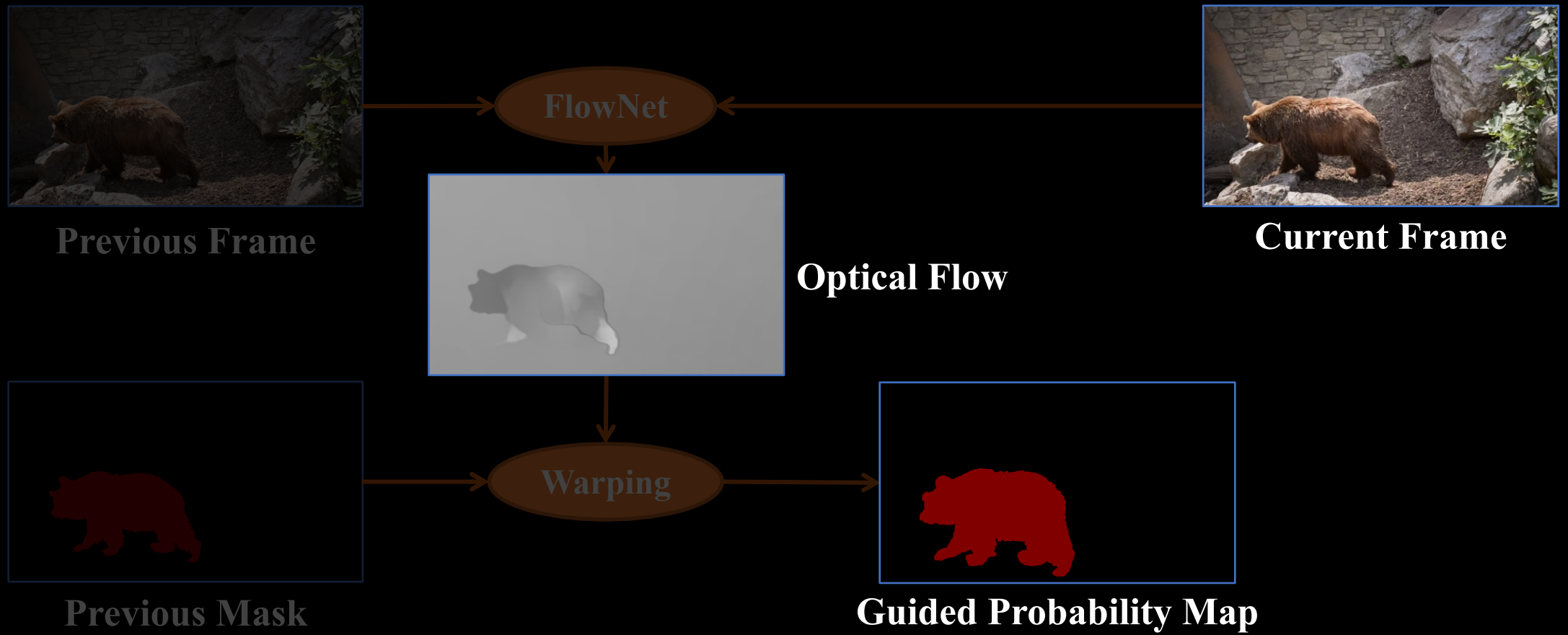
Previous Mask

Current Frame

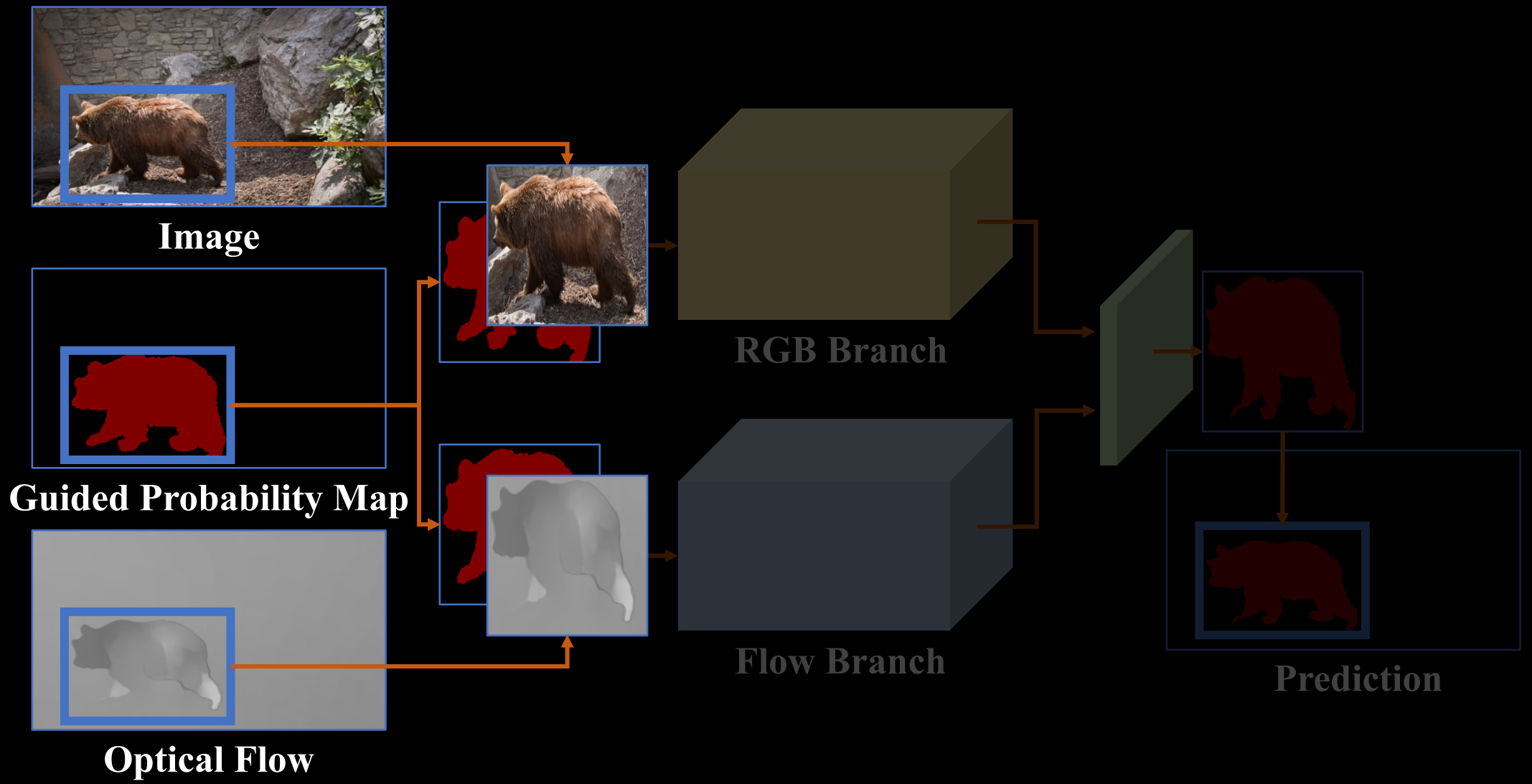
Mask Propagation Module



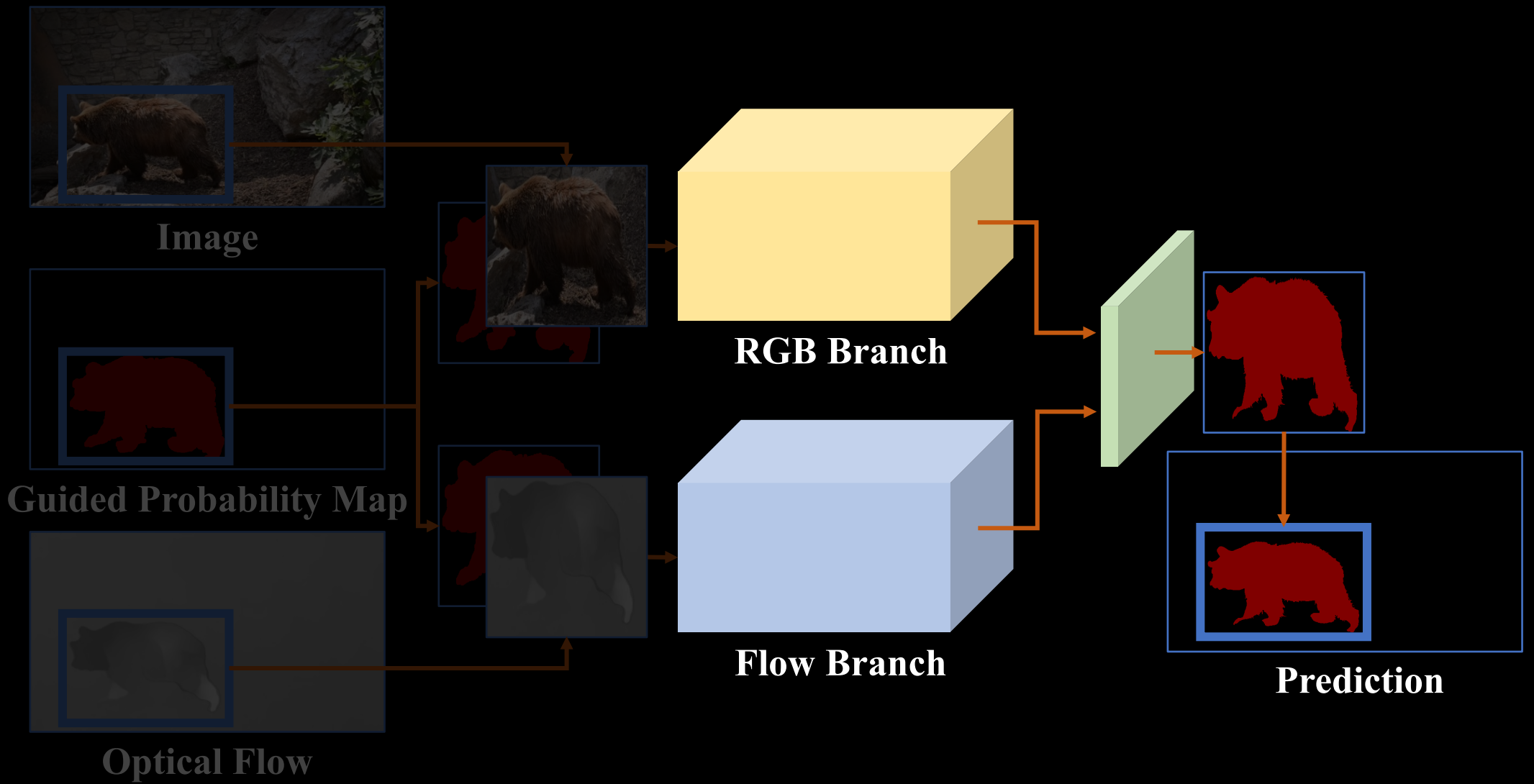
Mask Propagation Module



Mask Propagation Module



Mask Propagation Module



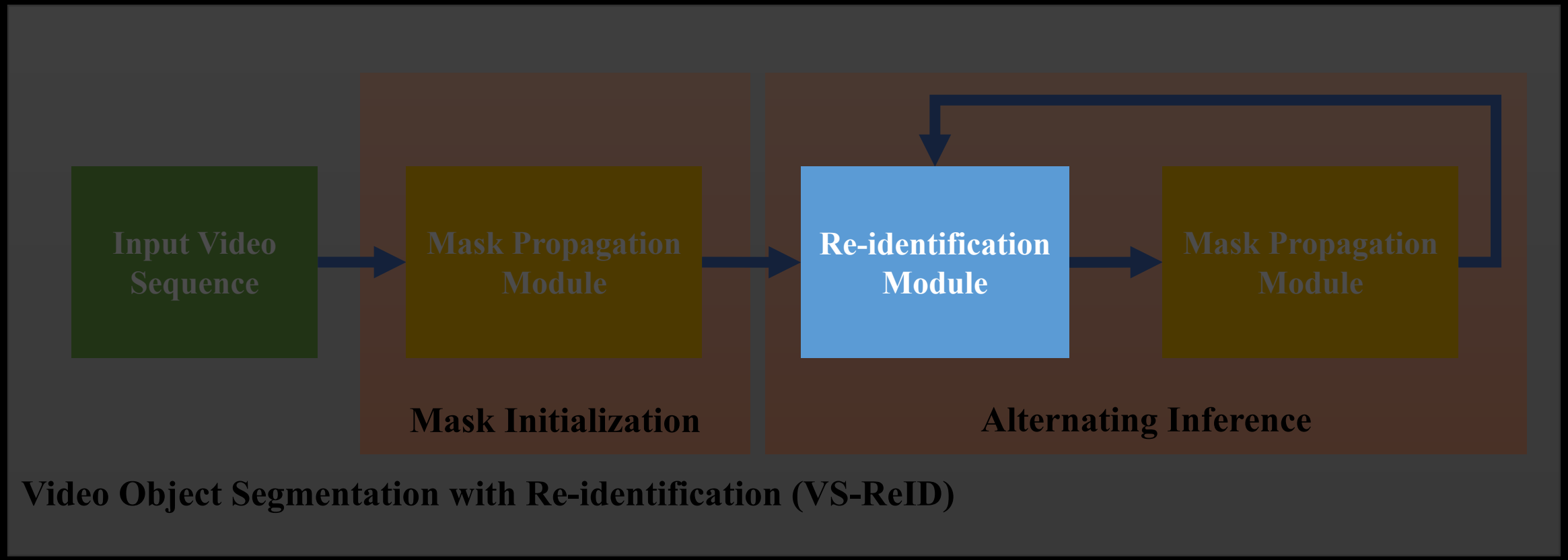
Video Frame
Guided Probability Map
Prediction
Warping
Mask Propagation Module

Mask Propagation Module

- Deeper Backbone Network
  - ResNet101
- RGB-branch
  - Pre-trained on the MS-COCO and PASCAL VOC dataset
    - Augmented ground-truth label as the guided probability map
  - Fine-tuned on the DAVIS dataset
- Flow-branch
  - Initialized with RGB-Branch's weights
  - Trained on the DAVIS dataset
- Multi-instance
  - Inference on each instance individually

Mask Propagation Module

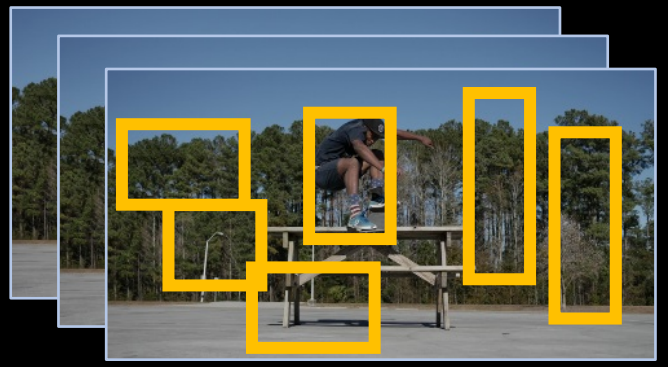
Proposed Framework



Re-identification Module

- Detection and re-identification

First Frame



Rest Frames

Template

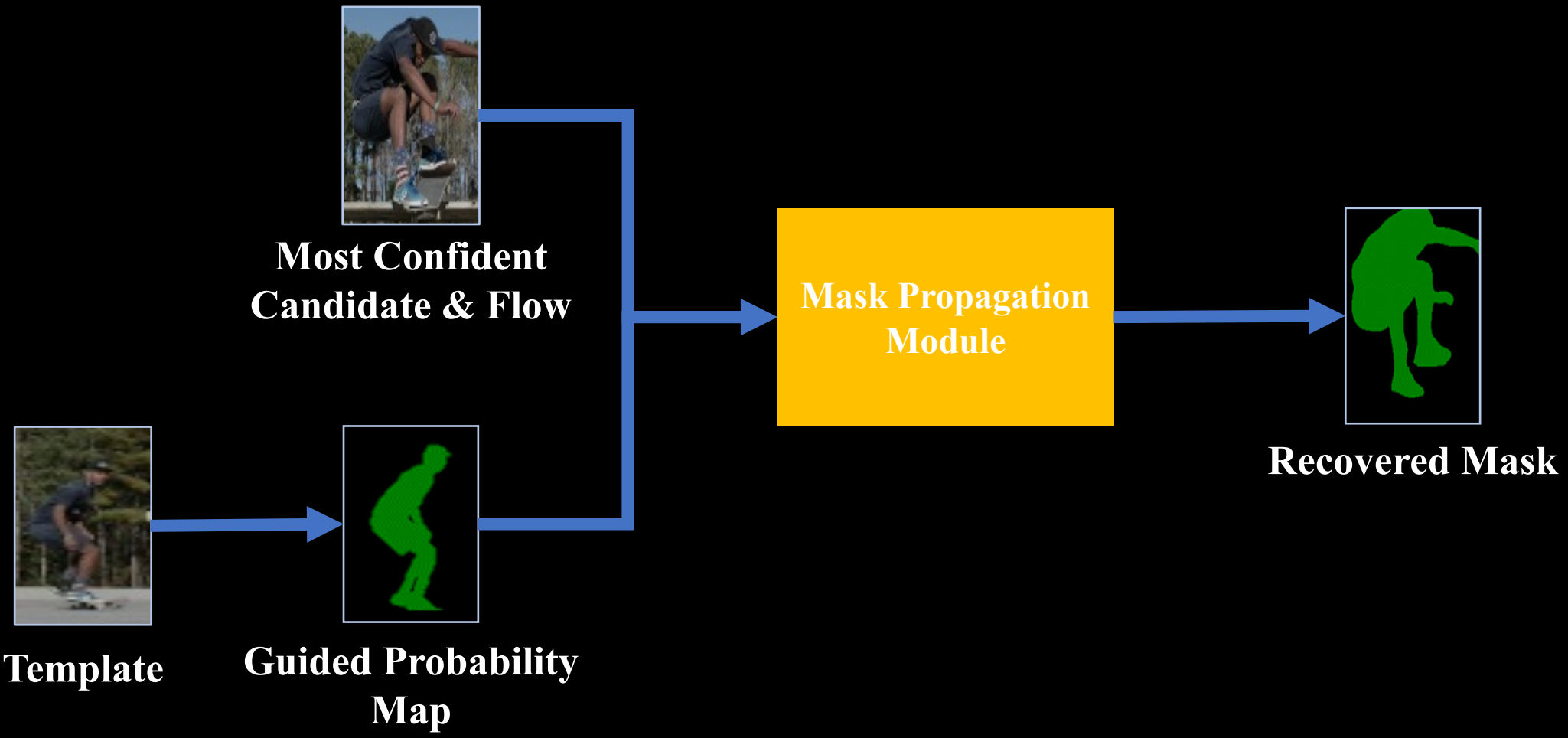
Re-identification

Candidate Bounding Boxes

Most Confident Candidate

Re-identification Module

- Recover the mask from a bounding box



Re-identification Module

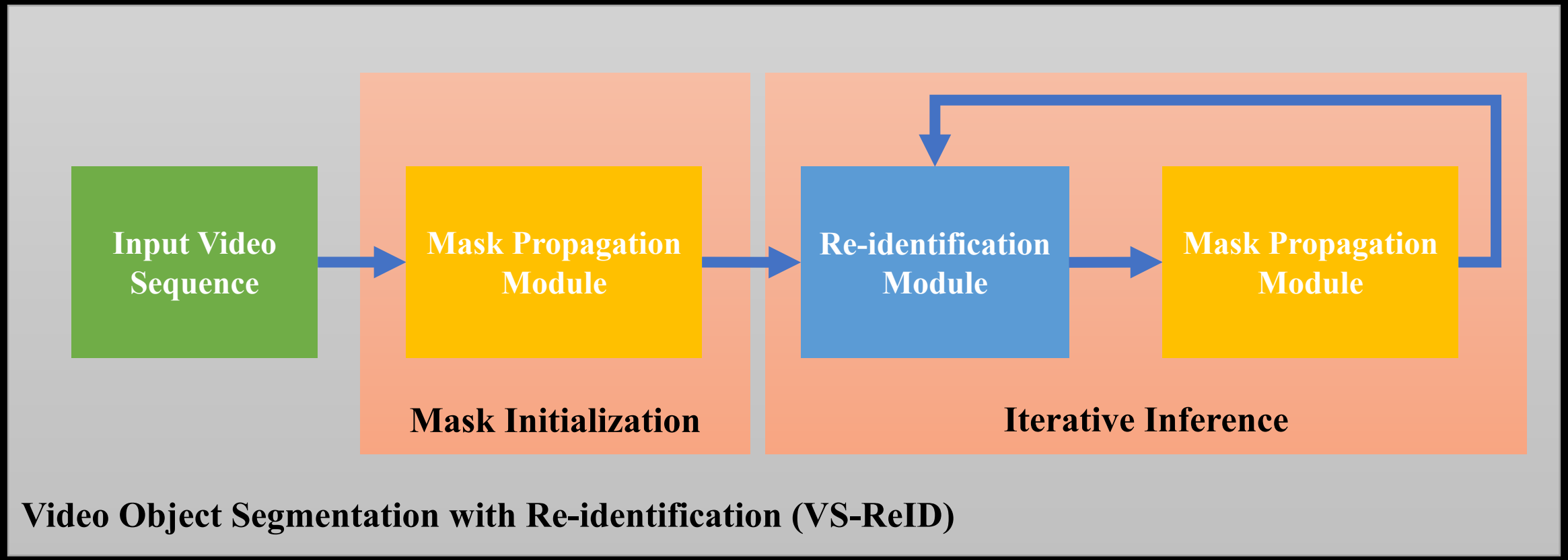
- Detection Model
  - Faster RCNN
  - Trained on the ImageNet

- Re-identification Model
  - 'Identification Net' in Person Search[1]
  - For the person category, we directly use the 'Identification Net' in Person Search[1]
  - Trained on the ImageNet VID

- Retrieve an instance in a single frame each time

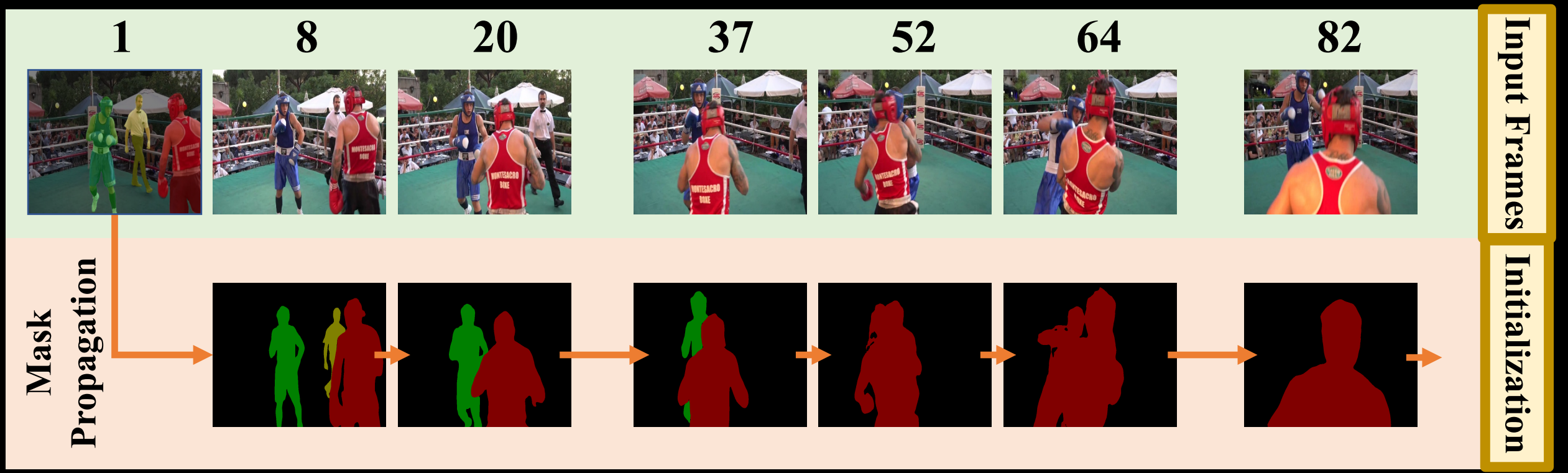
[1] Xiao T, Li S, Wang B, et al. Joint detection and identification feature learning for person search[C] CVPR. 2017.

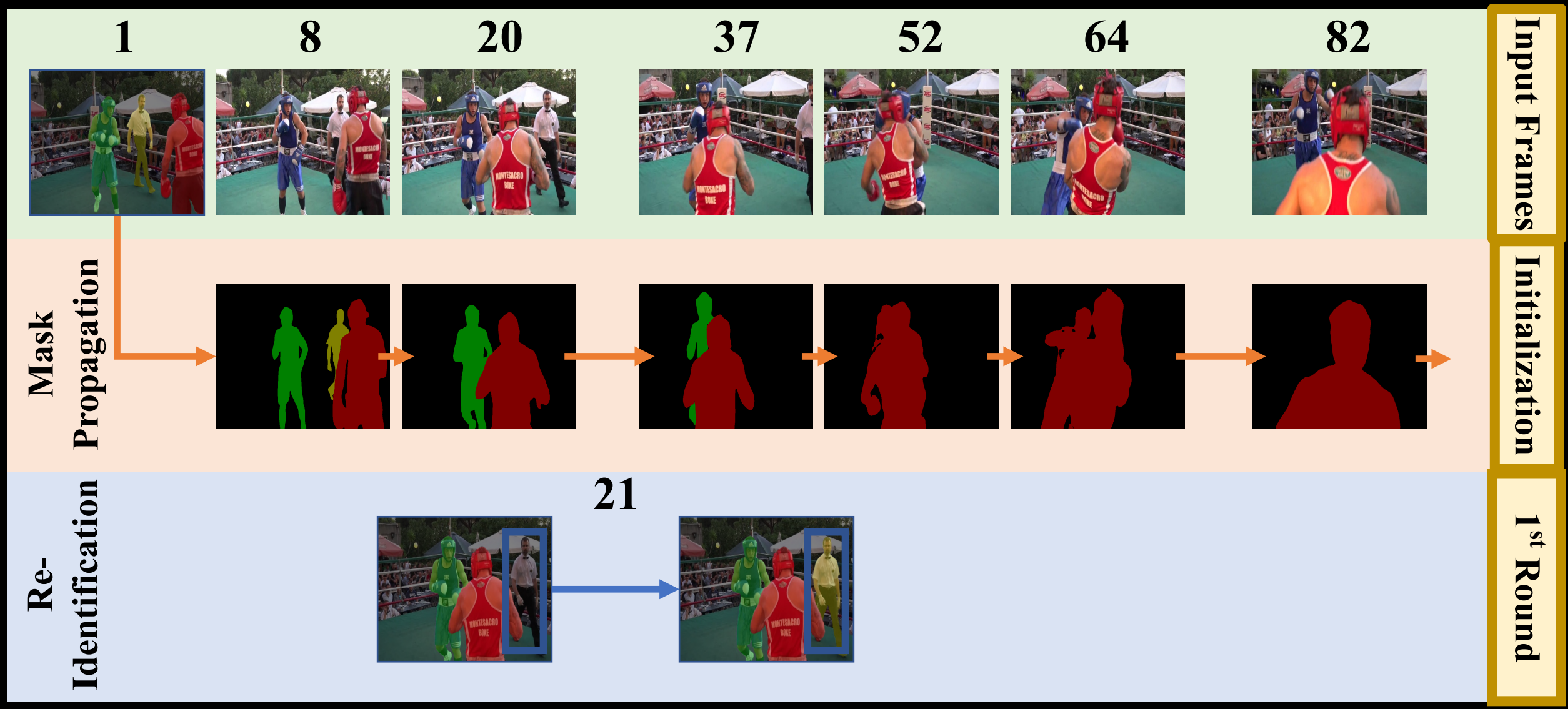
Mask Propagation Module



VS-ReID

- **Mask Initialization**




1
8
20
37
52
64
82
Input Frames
Mask Propagation
Initialization
21
Re-Identification
1st Round

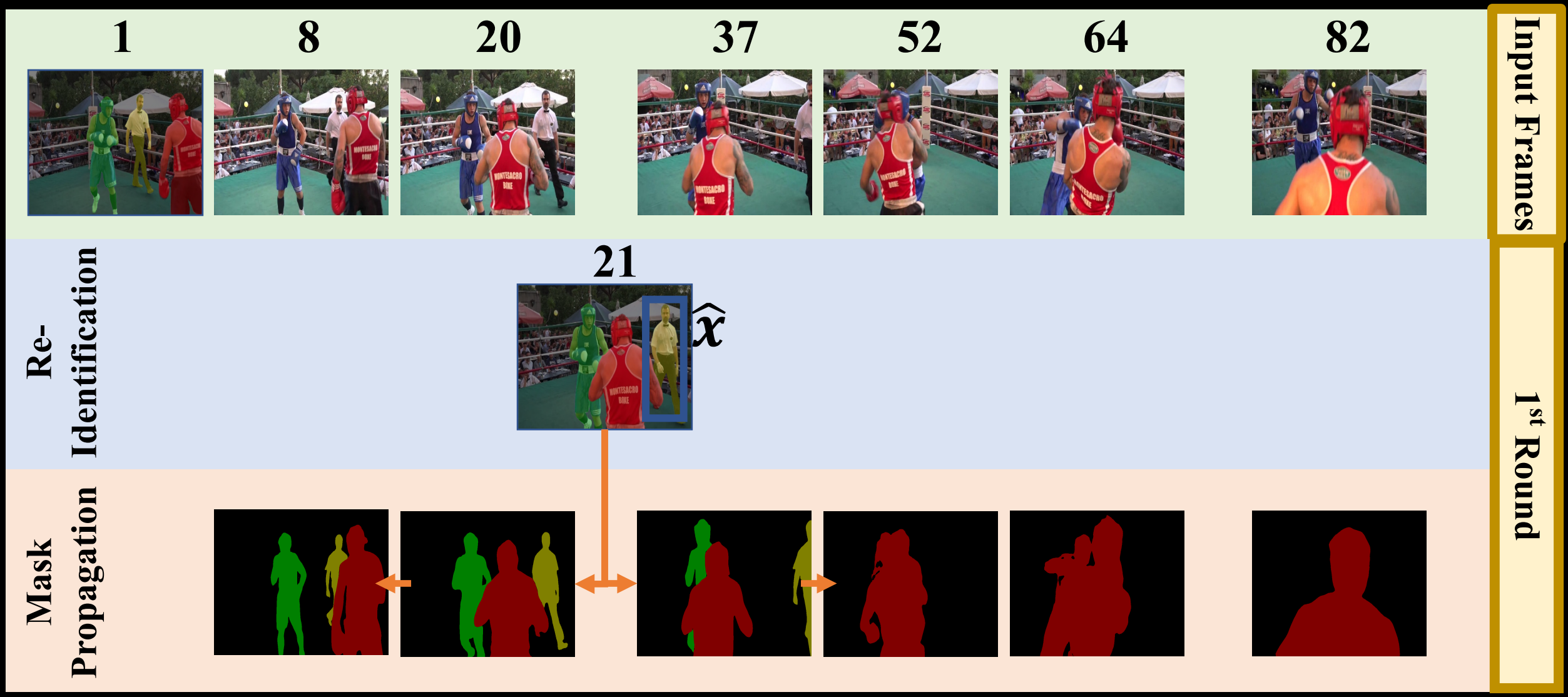
Input Frames

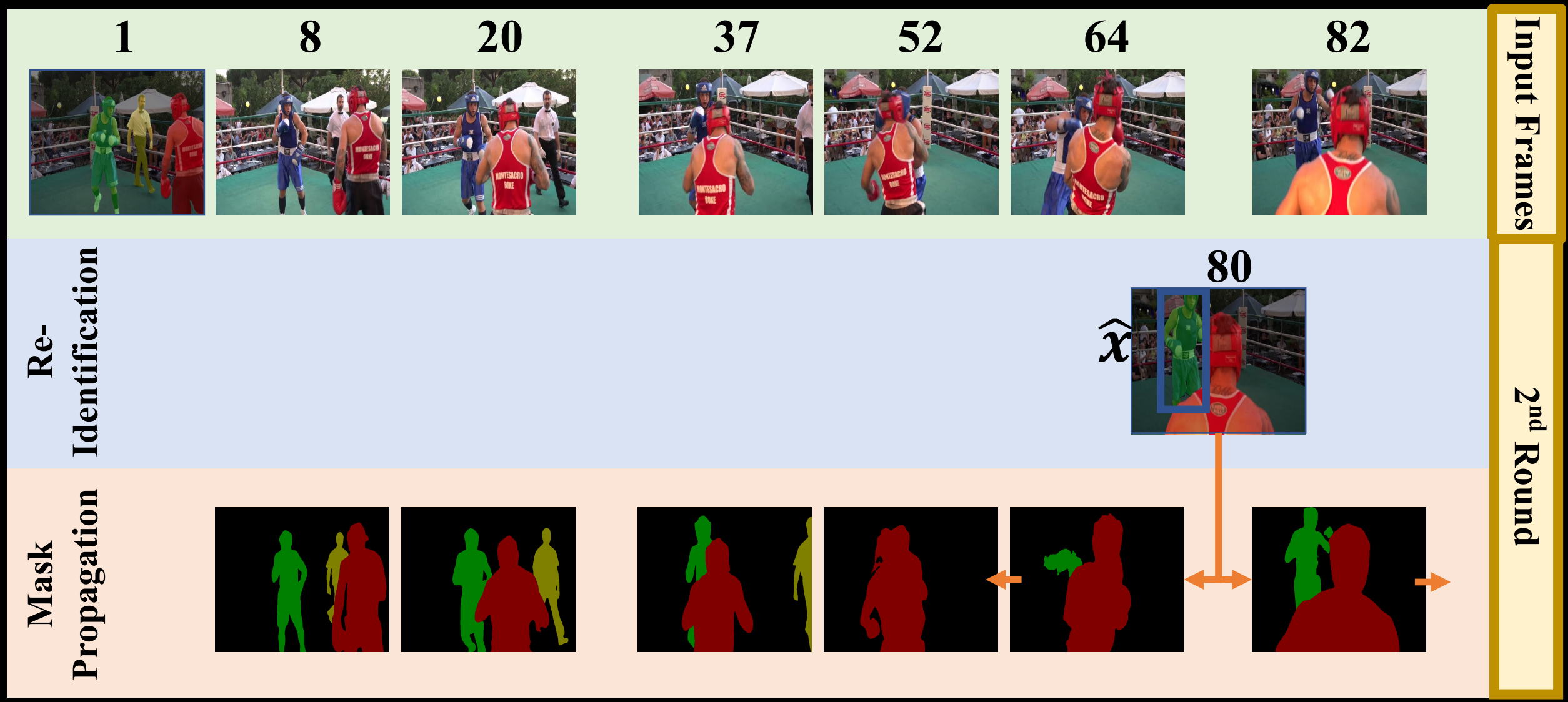
1 8 20 37 52 64 82

1st Round

Re-Identification

21


1
8
20
37
52
64
82
21
$\hat{x}$
Input Frames
Re-Identification
Mask Propagation
1st Round


1
8
20
37
52
64
82
80
$\hat{x}$
Input Frames
2nd Round
Re-Identification
Mask Propagation

Performance

	J Mean	F Mean	Global Mean
Voigt	54.8	60.5	57.7
Haamo	59.8	63.2	61.5
Vanta	61.5	66.2	63.8
Apata	65.1	70.6	67.8
Ours	**67.9**	**71.9**	**69.9**

(DAVIS 2017 Challenge test-challenge set)

Visualization

Thanks!