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   Harbin Institute of Technology

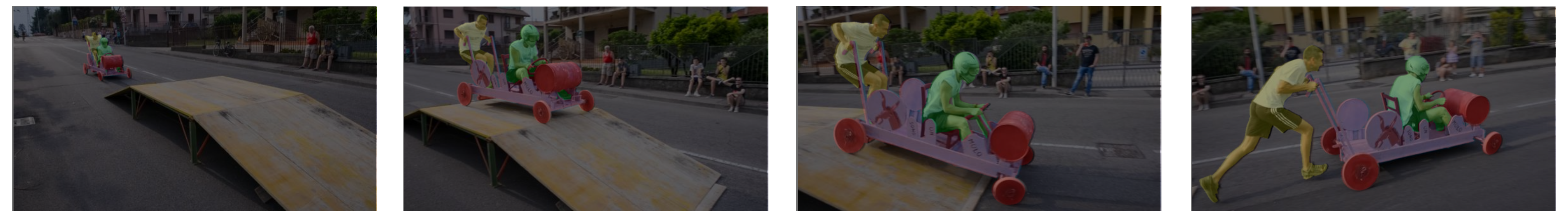
SenseTime Group Limited

IEEE 2017 Conference on Computer Vision and Pattern Recognition

1. Introduction

- **Problem**
  - Semi-supervised video object segmentation
  - Input: video sequence and masks in the first frame

  - Output: masks in the rest of the frame



- **Challenges**
  - Small objects and fine structures
  - Scale & pose-variations
  - Frequent occlusions and fast motion

- **Our Idea**
  - Two modules:
    - Mask propagation module   Short term
    - Re-identification module   Long term
  - Alternating updating algorithm

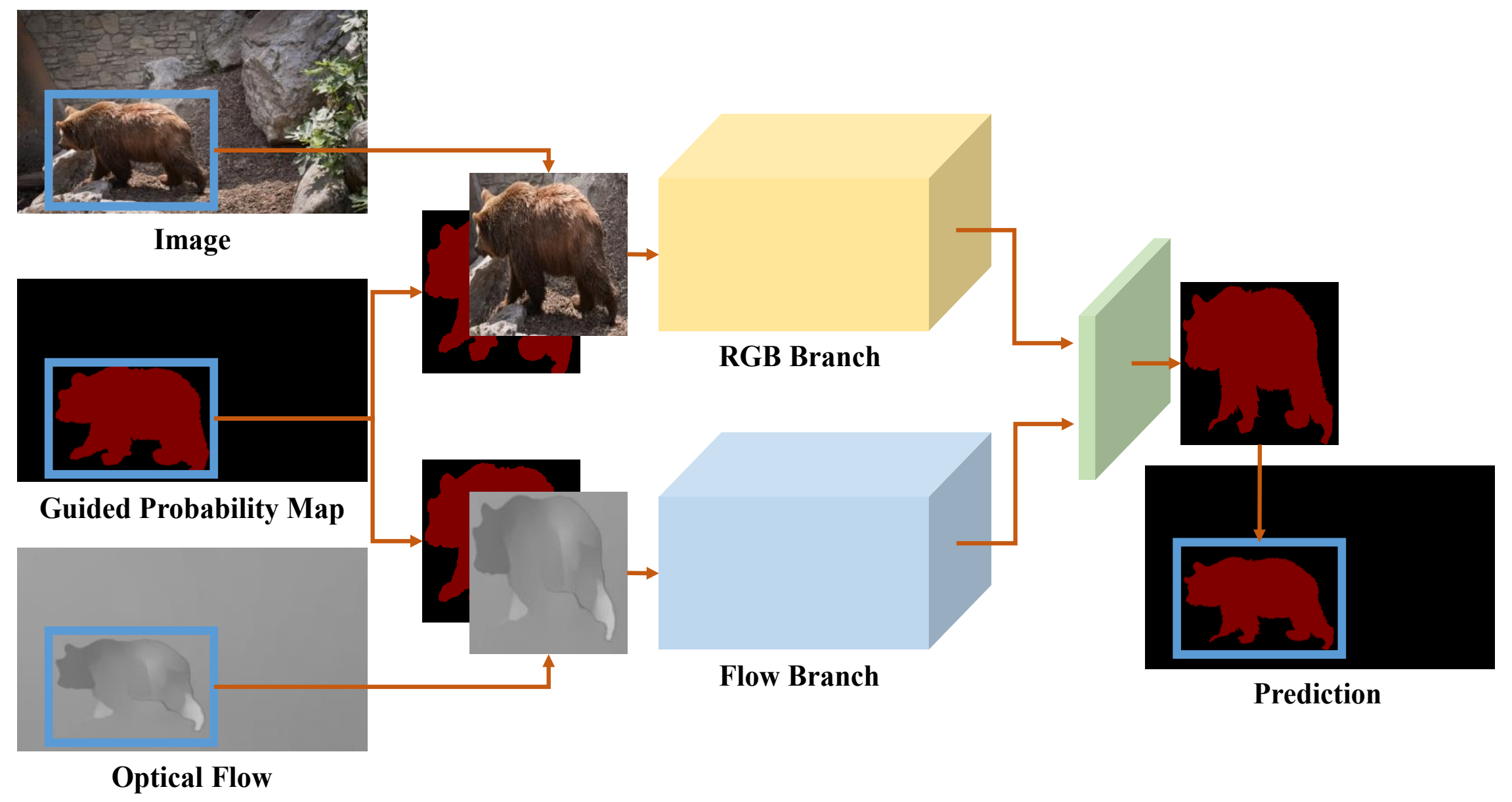
5. Overall Performance

Results on 2017 DAVIS Challenge test-challenge set.

Measure	Ours	Apata	Vanta	Haamo	Voigt	Lalal	Cjc	YXLKJ	Wasid	Froma	Zwrq0	Drbea	Anews	Ilanv	Koh	Make	Kozab	Xn881	Zpd	Griff	Nitin	Team5
Ranking	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Global Mean ↑	69.9	67.8	63.8	61.5	57.7	56.9	56.9	55.8	54.8	53.9	53.6	51.9	50.9	49.7	49.1	48.0	47.8	47.6	47.1	42.0	25.6	11.2
J Mean ↑	67.9	65.1	61.5	59.8	54.8	54.8	53.6	53.8	51.6	50.7	50.5	50.5	49.0	46.0	45.9	46.3	43.9	47.8	44.9	40.6	24.9	11.8
J Recall ↑	74.6	72.5	68.6	71.0	60.8	60.7	59.5	60.1	56.3	54.9	54.9	56.4	55.1	49.3	50.2	50.0	45.8	56.3	48.0	42.1	12.3	7.3
J Decay ↓	22.5	27.7	17.1	21.9	31.0	34.4	25.3	37.7	26.8	32.5	28.0	34.1	21.3	33.1	36.1	40.2	33.0	16.7	31.8	37.4	13.1	14.0
F Mean ↑	71.9	70.6	66.2	63.2	60.5	59.1	60.2	57.8	57.9	57.1	56.7	53.3	52.8	53.3	52.3	49.7	51.6	47.3	49.3	43.3	26.3	10.6
F Recall ↑	79.1	79.8	79.0	74.6	67.2	66.7	67.9	62.1	64.8	63.2	63.5	57.9	58.3	58.4	57.1	52.8	56.0	53.0	54.4	43.2	9.1	3.0
F Decay ↓	24.1	30.2	17.6	23.7	34.7	36.1	27.6	42.9	28.8	33.7	30.4	39.5	23.7	36.4	39.2	44.8	36.3	21.6	36.2	40.2	13.0	12.6

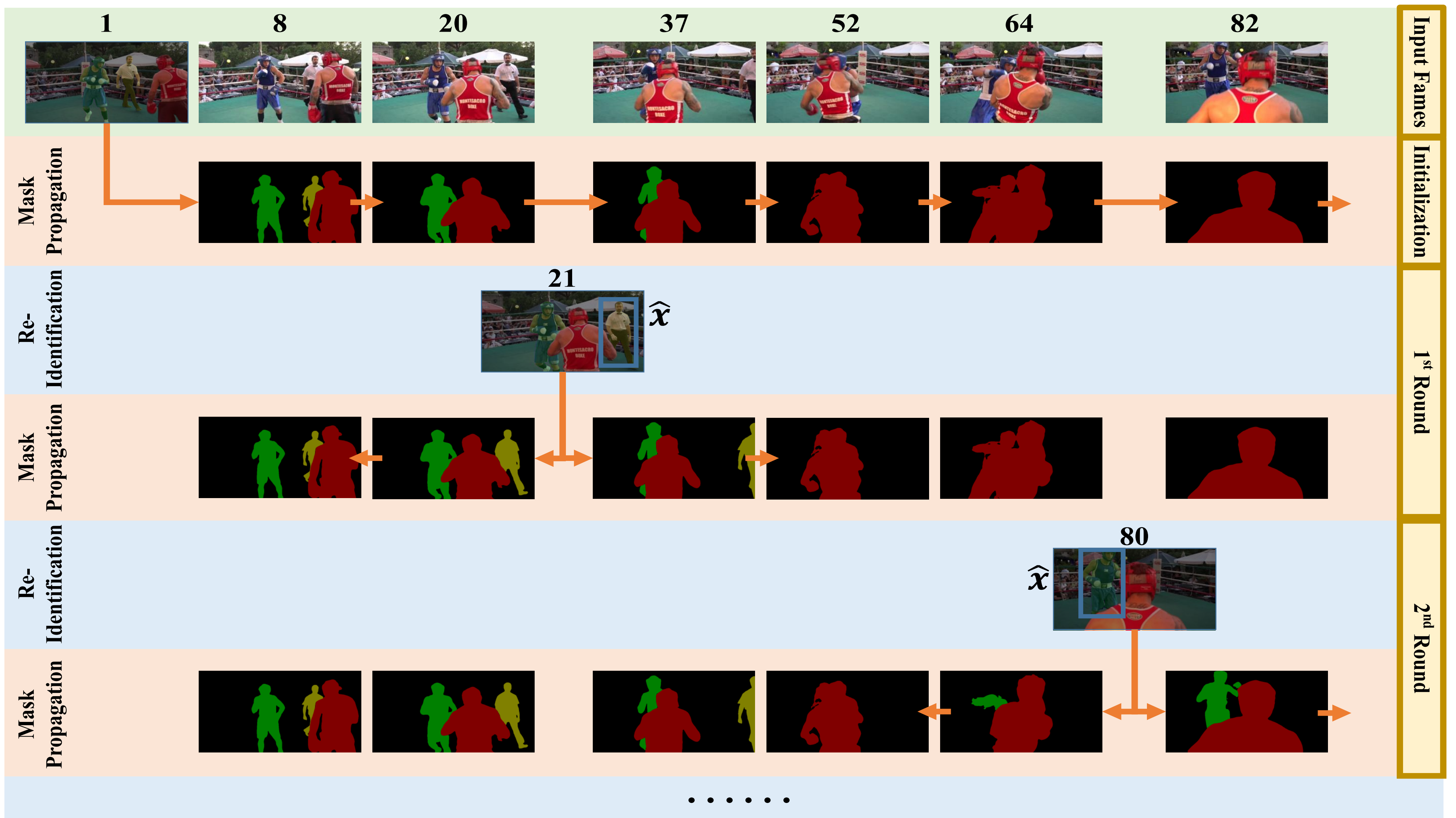
2. Approach

- **Mask Propagation Module**
  - Inspired by LucidTracker
  - Several important modifications
    - Deeper network (ResNet101)
    - Bounding box input (Handle scale-variations)
    - Two streams are jointly fine-tuned



- **Re-identification Module**
  - Detection and re-identification network (similar with person search)
  - Choose the instances in the first frame as the templates
  - Scan the whole video sequence and recover the most confident instance

- **Video Object Segmentation with Re-identification (VS-ReID)**

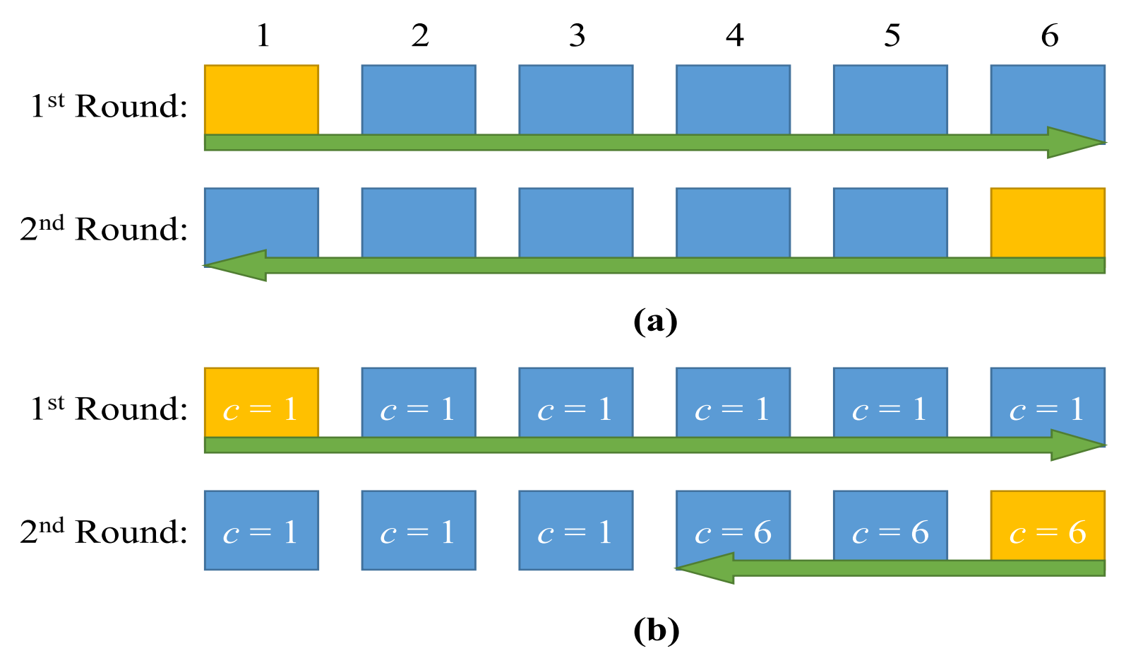


3. Implementation Details

- **Training Set**
  - Mask propagation module: COCO, PASCAL VOC, DAVIS
  - Re-identification Module: ImageNet, ImageNet-VID

- **Recover Mask from a Bounding Box**
  - Employ the mask in the first frame as the guided probability map
  - Execute the mask propagation module

- **Checkpoint Mechanism**

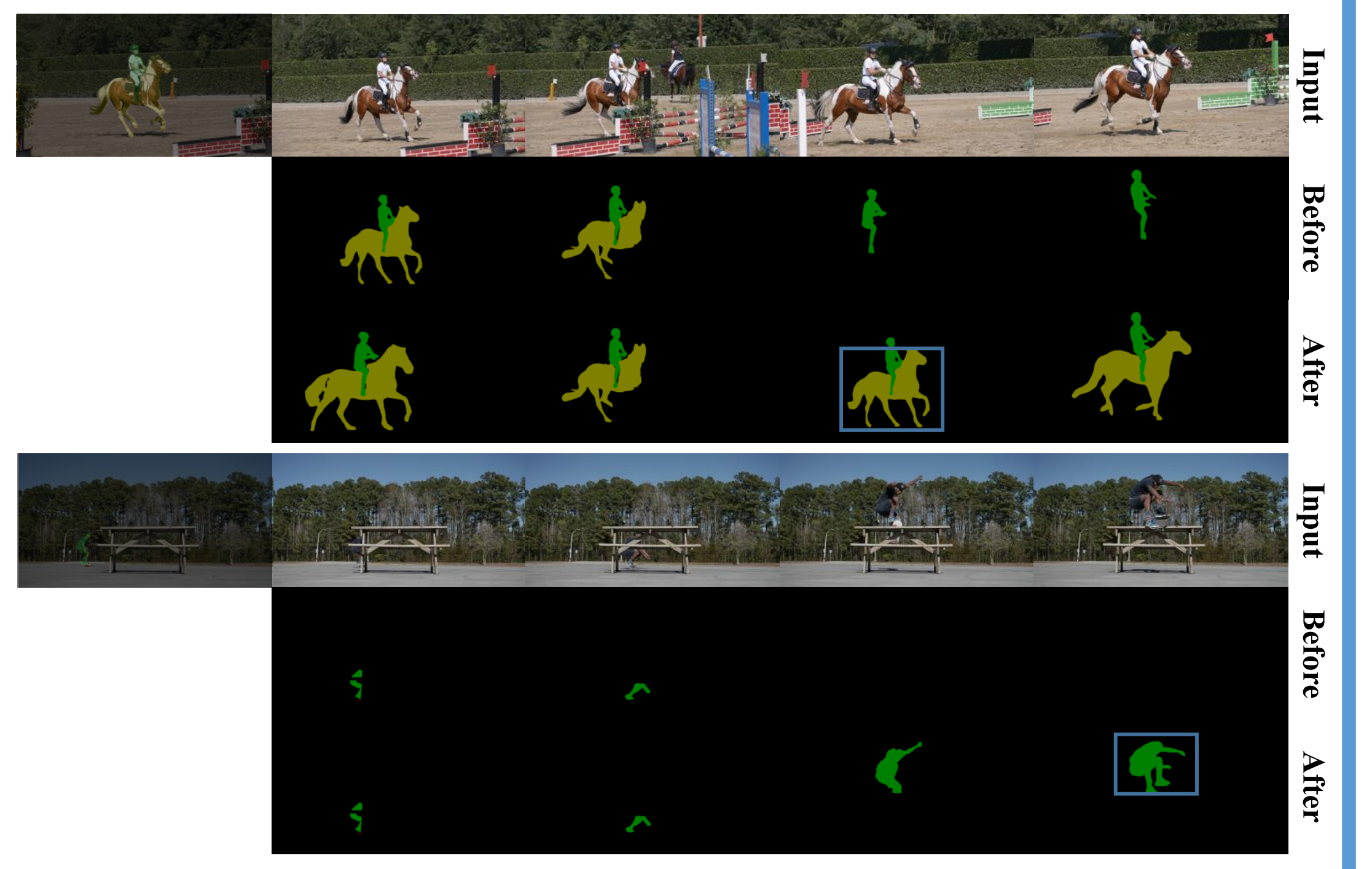


4. Experiments

- **Ablation Study of Each Module**

	J -mean	F -mean	global-mean	boost
baseline (MSK)	0.509	0.526	0.517	-
+ full-image to bbox	0.532	0.577	0.555	+ 0.038
+ f ow-stream	0.568	0.600	0.584	+ 0.007
+ re-id module	0.633	0.670	0.652	+ 0.068
+ multi-scale testing	0.644	0.678	0.661	+ 0.009

- **Missing Instances Are Retrieved**



- **Results**