facebook JOHNS HOPKINS UNIVERSITY

CVPR VIRTUAL JUNE 19-25

# Weakly Supervised Instance Segmentation for Videos with Temporal Mask Consistency

Qing Liu[1], Vignesh Ramanathan[2], Dhruv Mahajan[2],
Alan Yuille[1], Zhenheng Yang[2]
[1]Johns Hopkins University [2]Facebook

# Goal and Motivation

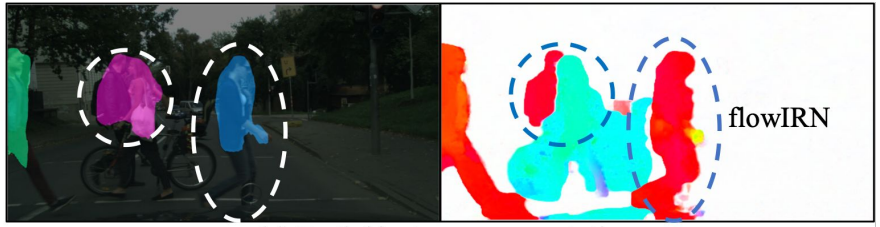
**Our goal:**
Weakly supervised instance segmentation for videos

- Supervision: frame level class labels
- Evaluation: FIS & VIS

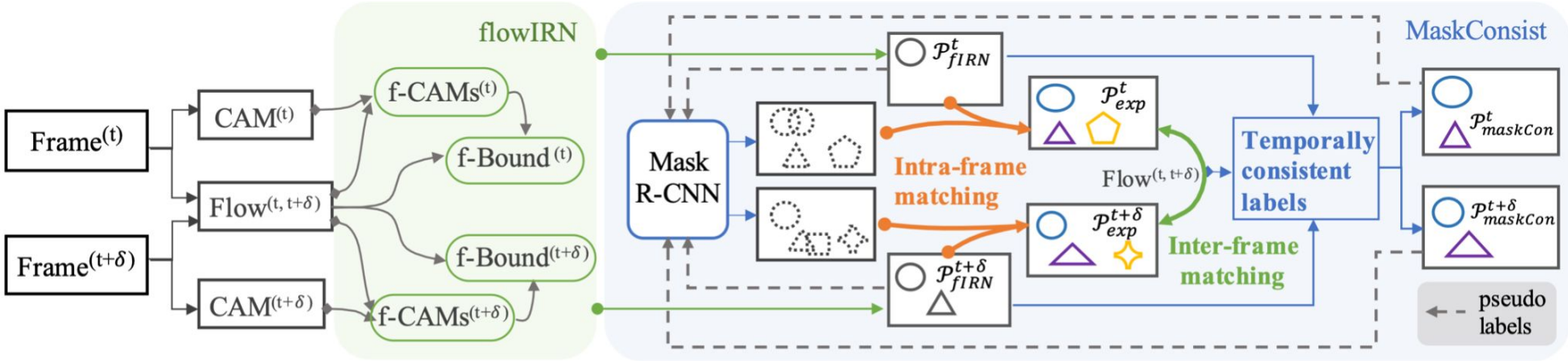
**Motivation:**

- Existing methods suffer from two problems:
  - Partial segmentation
  - Missing object

- Video data can help



# Overall Framework




1. **FlowIRN**: Introduce motion into weakly supervised instance segmentation training
2. **Mask-Consist**: Add cross-frame temporal consistency to Mask-RCNN training

# flowIRN

- f-CAM: Use flow to amplify CAMs

- Objects of interest tend to be close to the camera and have large motion

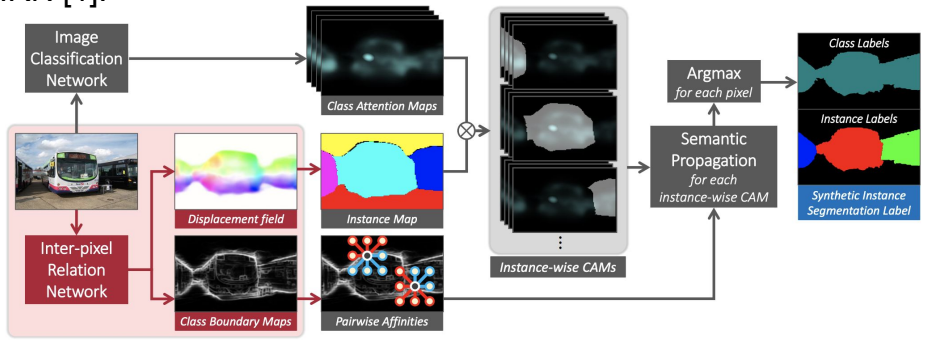
$$\text{f-CAM}_c(\boldsymbol{x}) = \text{CAM}_c(\boldsymbol{x}) \times A^{\mathbb{I}(||\mathcal{F}(\boldsymbol{x})||_2 > T)}$$

- f-boundary: Use Flow to guide the learning of instance boundary

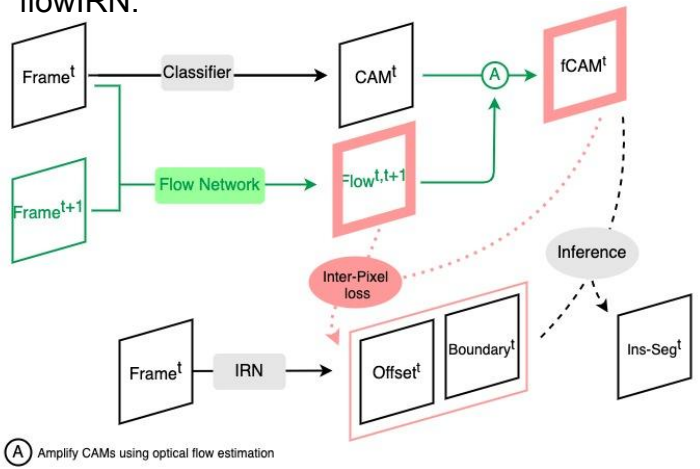
- Pixels of the same instance tend to move together

$$\mathcal{L}_{\mathcal{F}}^{\mathcal{B}} = \sum_{\boldsymbol{x}_j \in \mathcal{N}_i} ||\mathcal{F}'(\boldsymbol{x}_i) - \mathcal{F}'(\boldsymbol{x}_j)||\alpha_{i,j} + \lambda|1 - \alpha_{i,j}|$$

IRN [1]:



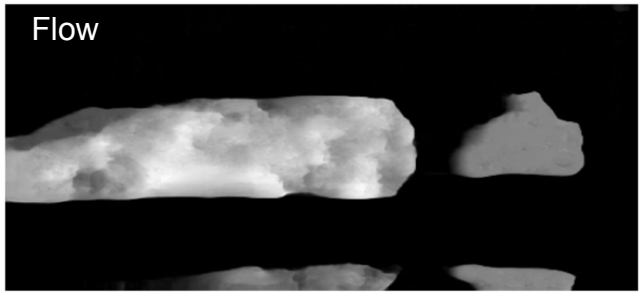

flowIRN:

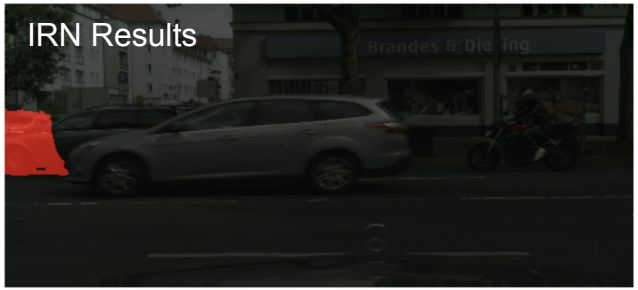



[1] Ahn, Jiwoon, Sunghyun Cho, and Suha Kwak. "Weakly supervised learning of instance segmentation with inter-pixel relations." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

# flowIRN Results: Fix Missing Object

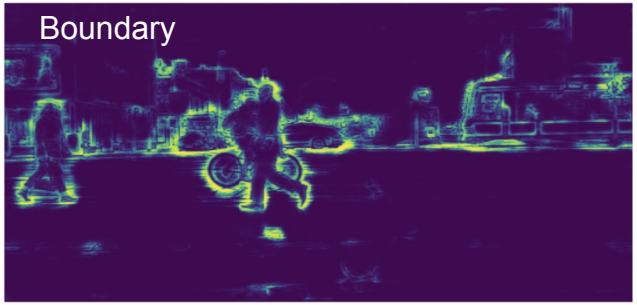
Amplify CAM using flow


# flowIRN Results: Fix Incorrect Boundary

Add flow into boundary learning



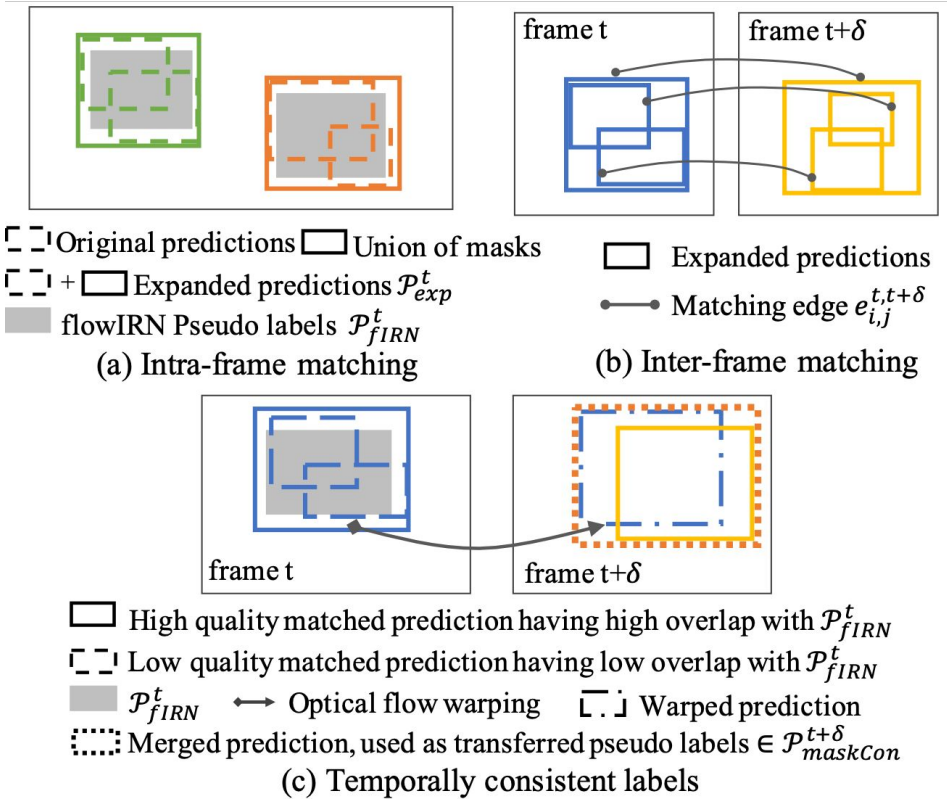



# MaskConsist

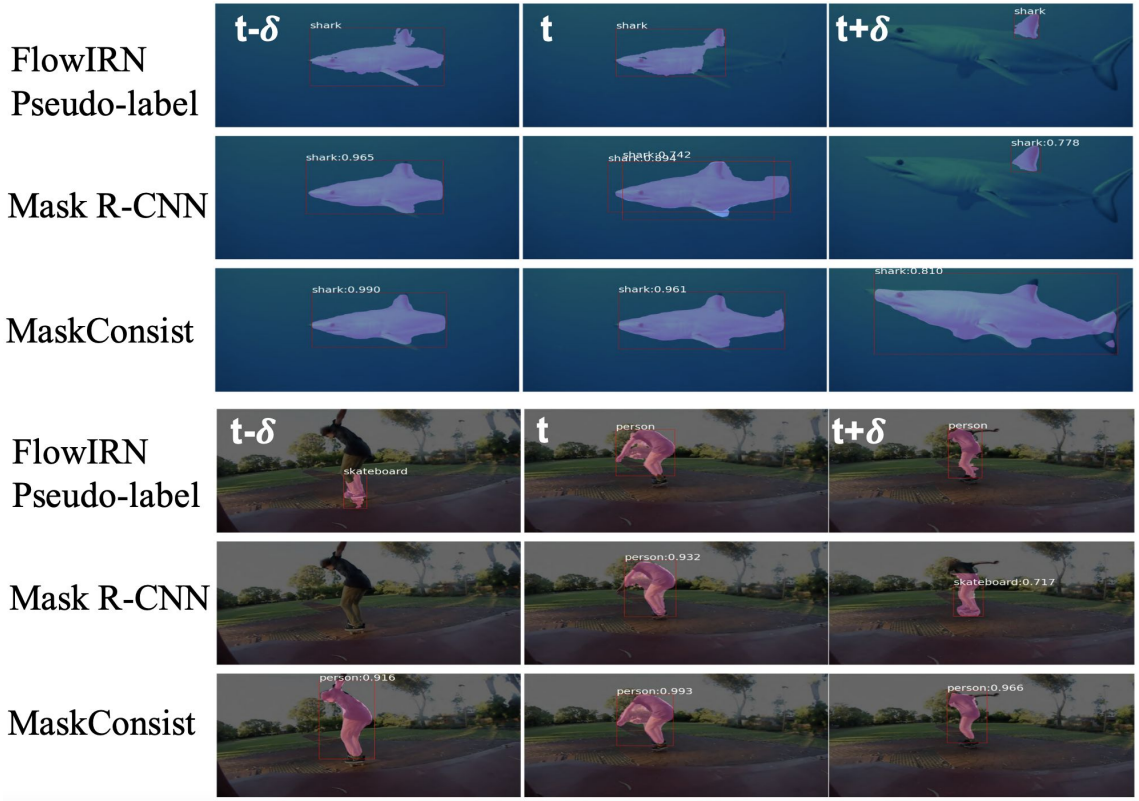
Goal: making Mask R-CNN training more robust to noisy pseudo-labels

Solution:
- find “high-quality” mask predictions
- transfer them to neighboring frames as new pseudo-labels

“High-quality” prediction:
- overlapped with flowIRN pseudo-labels
- temporally stable



# MaskConsist Results



# Frame Instance Segmentation Results

| Methods | Video Info | Supervision | $AP_{50}$ |
|---|---|---|---|
| Mask R-CNN [17] | ✗ | Mask | 78.24 |
| WSIS-BBTP [20] | ✗ | Bbox | 46.80 |
| WISE [27] | ✗ | Class | 24.54 |
| F2F [29]+MCG [41] | ✓ | Class | 26.31 |
| IRN [6] | ✗ | Class | 29.64 |
| IRN [6]+F2F[29] | ✓ | Class | 30.27 |
| Ours | ✓ | Class | 34.66 |
| Ours (self-training) | ✓ | Class | **36.00** |

Table 1. Frame-level instance segmentation performance ($AP_{50}$) on YTVIS train_val split.

| Methods | Supervision | Instance seg | Semantic seg |
|---|---|---|---|
| Mask R-CNN [17] | Mask | 38.73 | 79.23 |
| WISE [27] | Class | 10.51 | 35.82 |
| F2F [29]+MCG [41] | Class | 10.73 | 33.26 |
| IRN [6] | Class | 12.33 | 33.48 |
| IRN [6]+F2F[29] | Class | 12.53 | 34.17 |
| Ours | Class | 16.05 | 39.88 |
| Ours (self-training) | Class | **16.82** | **41.31** |

Table 2. Frame-level instance segmentation ($AP_{50}$) and semantic segmentation ($IoU$) on Cityscapes validation split.

# Video Instance Segmentation Results

| Methods | | Train_Val Split | | | | | Validation Split | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $mAP$ | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ | $mAP$ | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ |
| Fully supervised learning methods | IoUTracker+ [58] | - | - | - | - | - | 23.6 | 39.2 | 25.5 | 26.2 | 30.9 |
| | DeepSORT [57] | - | - | - | - | - | 26.1 | 42.9 | 26.1 | 27.8 | 31.3 |
| | MaskTrack [58] | - | - | - | - | - | 30.3 | 51.1 | 32.6 | 31.0 | 35.5 |
| Weakly supervised learning methods | WISE [27] | 8.7 | 22.1 | 5.5 | 9.8 | 10.7 | 6.3 | 17.5 | 3.5 | 7.1 | 7.8 |
| | IRN [6] | 10.8 | 26.4 | 7.7 | 12.6 | 14.4 | 7.3 | 18.0 | 3.0 | 9.0 | 10.7 |
| | Ours | 14.1 | 34.4 | 9.4 | 16.0 | 17.9 | 10.5 | 27.2 | 6.2 | 12.3 | 13.6 |

Table 3. Video instance segmentation results on Youtube-VIS dataset.

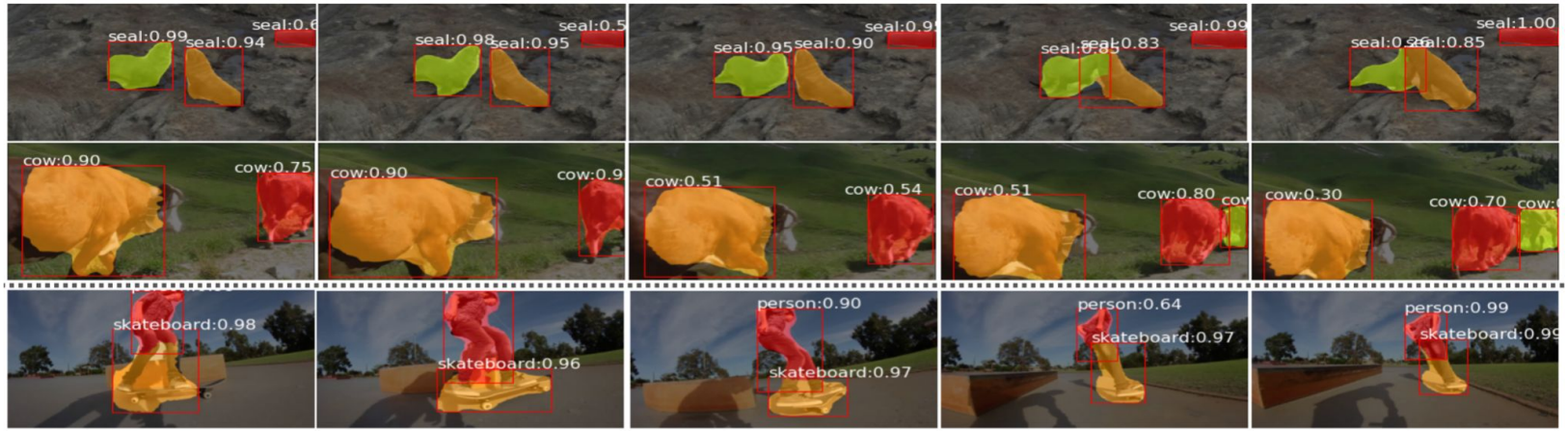



Figure 4. Example Video instance segmentation results from our method on Youtube-VIS dataset.

The end.
Thank you for your attention.