Beijing JiaoTong University

Weakly Supervised Instance Segmentation for Videos with Temporal Mask Consistency

Qing Liu*
Johns Hopkins University
qingliu@jhu.edu

Vignesh Ramanathan
Facebook
vigneshr@fb.com

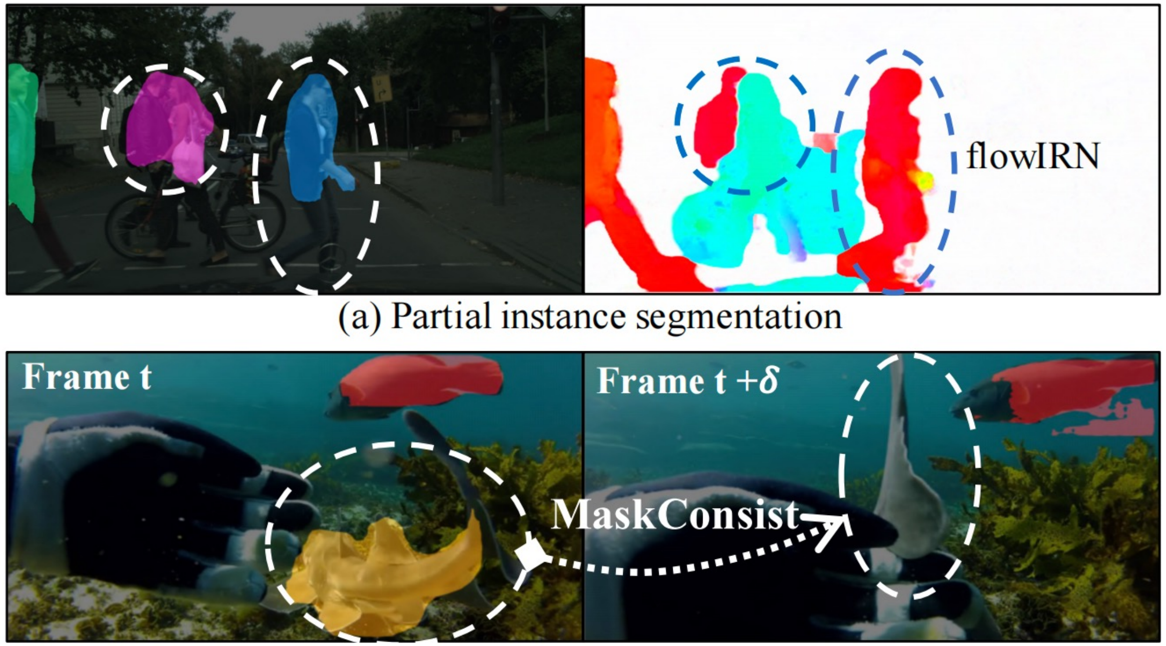
Dhruv Mahajan
Facebook
dhruvm@fb.com

Alan Yuille
Johns Hopkins University
alan.l.yuille@gmail.com

Zhenheng Yang
Facebook
zhenheny@gmail.com

1 Background Knowledge

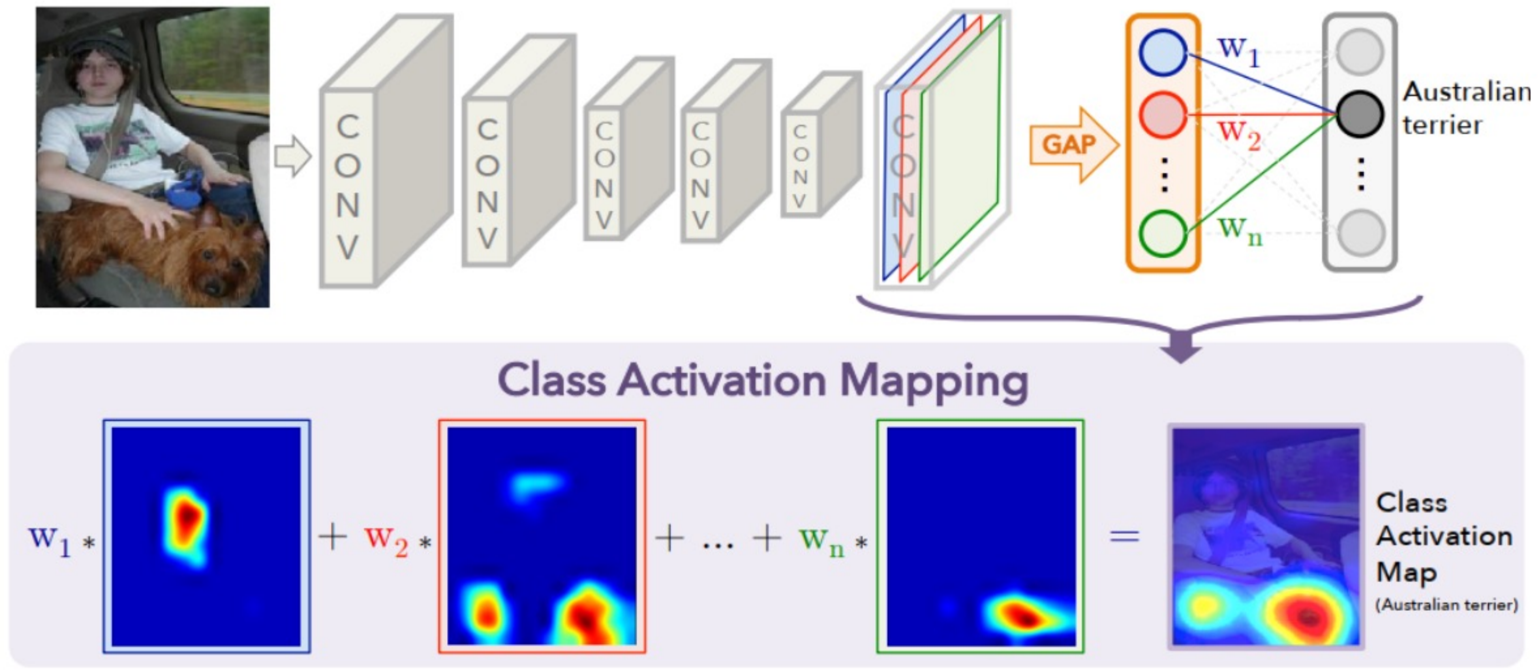
Semantic segmentation:
rely on CAMs, leverages motion and temporal consistency in videos
Instance segmentation:
mining other information to distinguish instance
Typical pipeline: a) generating pseudo label. b) trainning a supervised model.



(a) Partial instance segmentation

(b) Missing object instance

1 Class attention maps(CAMs)



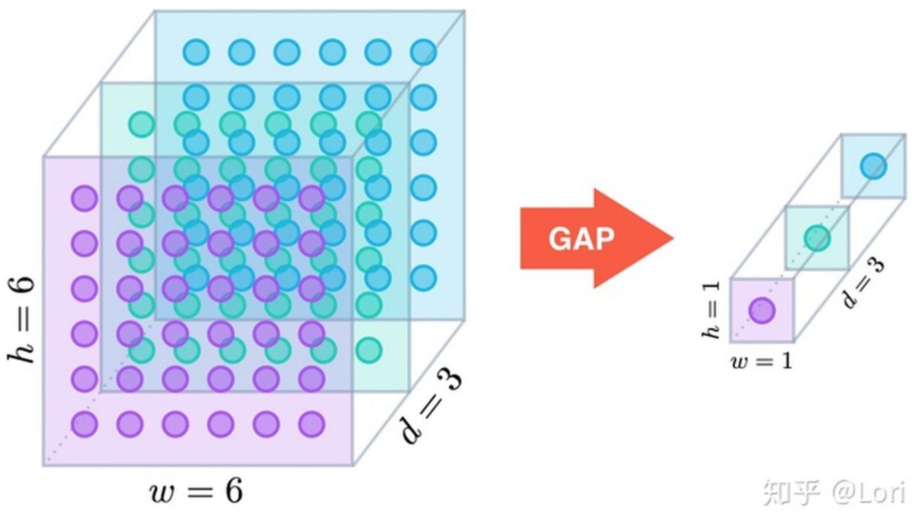


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

1 Inter-pixel relation network(IRN)

Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations

Jiwoon Ahn
DGIST, Kakao Corp.
jyun@dgist.ac.kr

Sunghyun Cho*
DGIST
scho@dgist.ac.kr

Suha Kwak*
POSTECH
suha.kwak@postech.ac.kr

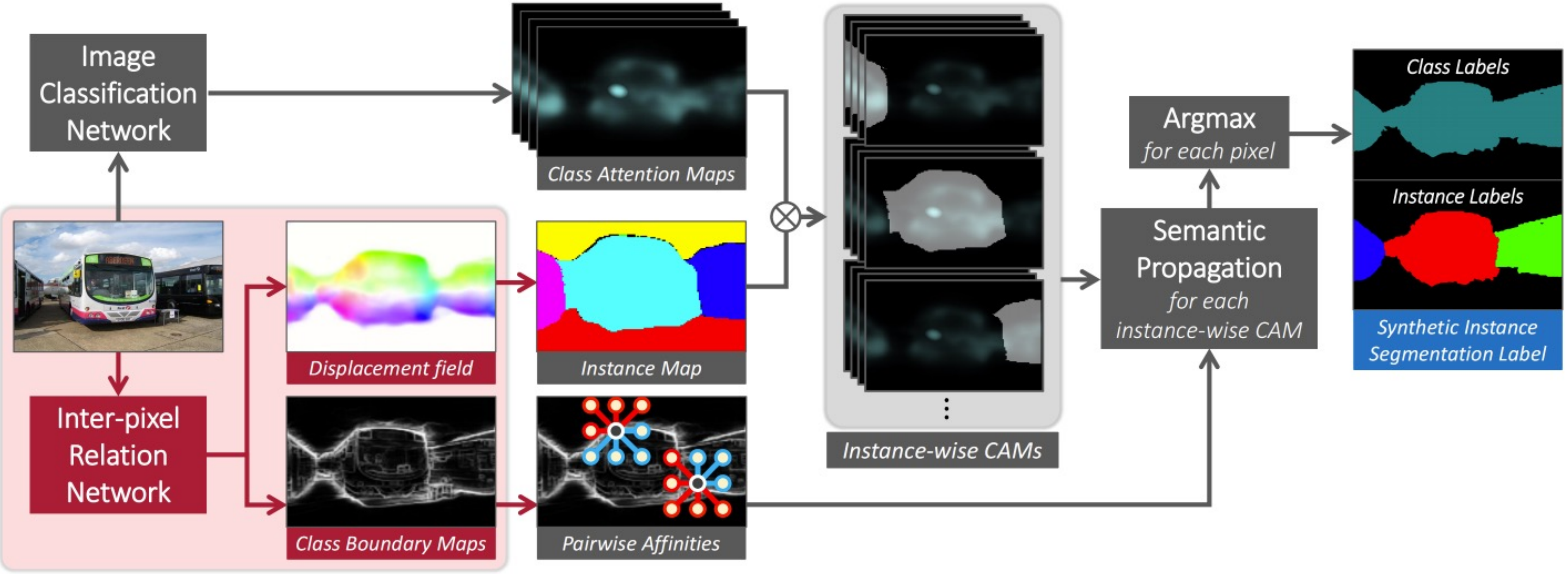


Figure 1. Overview of our framework for generating pseudo instance segmentation labels.

1 IRN -- architecture

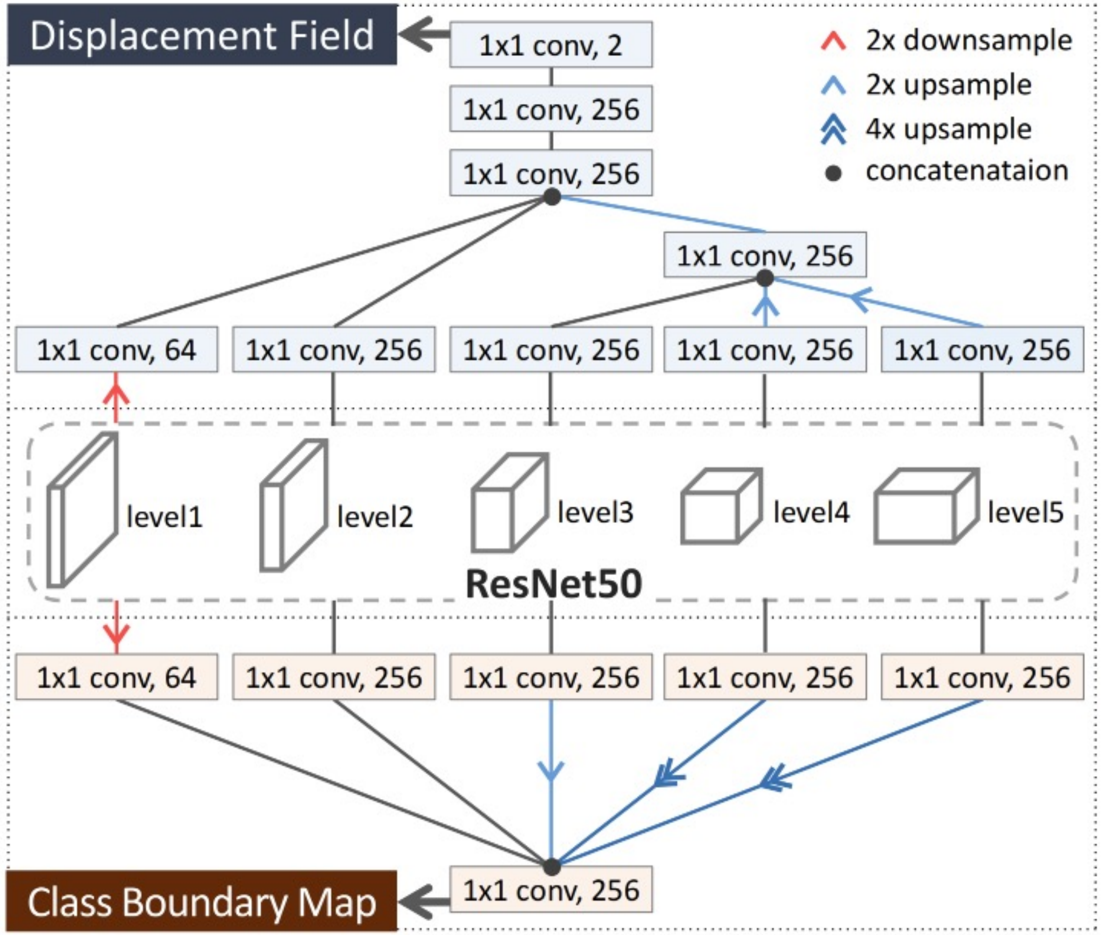


Figure 2. Overall architecture of IRNet.

1 IRN -- inter-pixel relations

Two kinds of inter-pixel relations:

displacement between a pair of pixels

class equivalence

$$\mathcal{P} = \{(i,j) \mid \|\mathbf{x}_i - \mathbf{x}_j\|_2 < \gamma, \forall i \neq j\},$$

$$\mathcal{P}^+ = \{(i,j) \mid \hat{M}(\mathbf{x}_i) = \hat{M}(\mathbf{x}_j), (i,j) \in \mathcal{P}\}$$

$$\mathcal{P}^- = \{(i,j) \mid \hat{M}(\mathbf{x}_i) \neq \hat{M}(\mathbf{x}_j), (i,j) \in \mathcal{P}\}$$

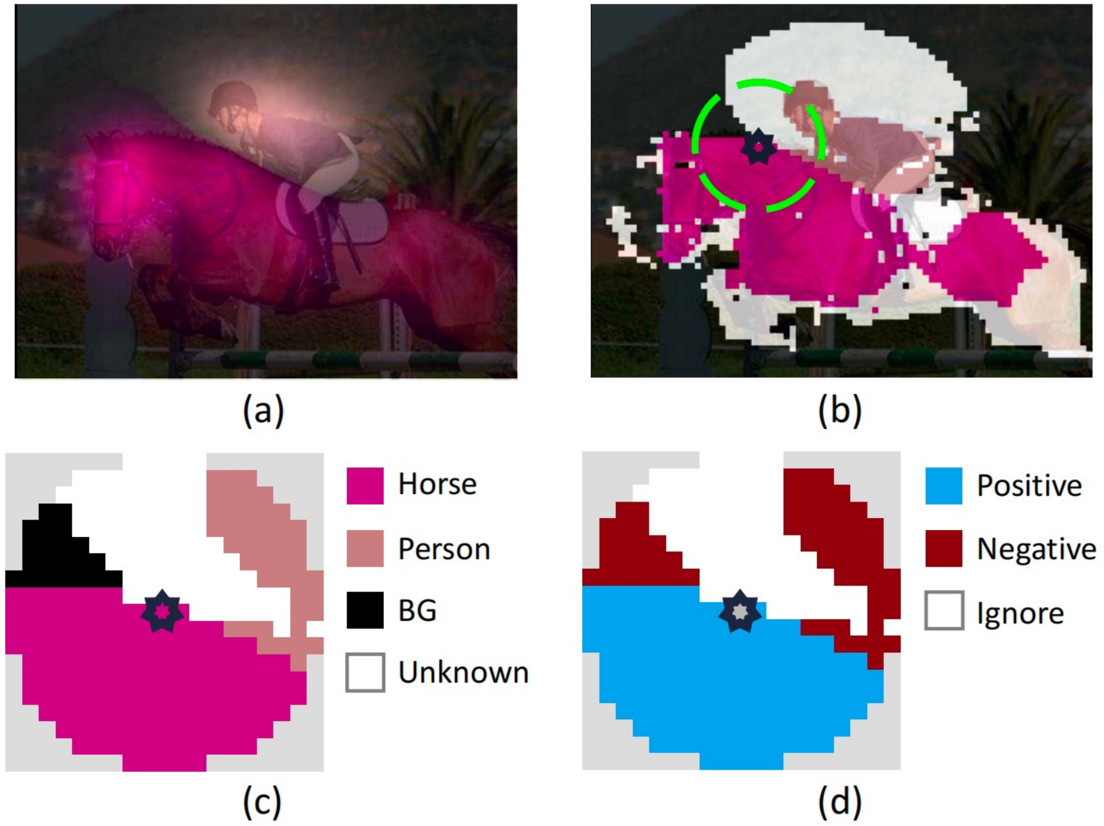


Figure 3. Visualization of our inter-pixel relation mining process. (a) CAMs. (b) Confident areas of object classes. (c) Pseudo class label map within a local neighborhood. (d) Class equivalence relations between the center and the others.

1 IRN -- Displacement Field

$$D \in \mathbb{R}^{w \times h \times 2}$$

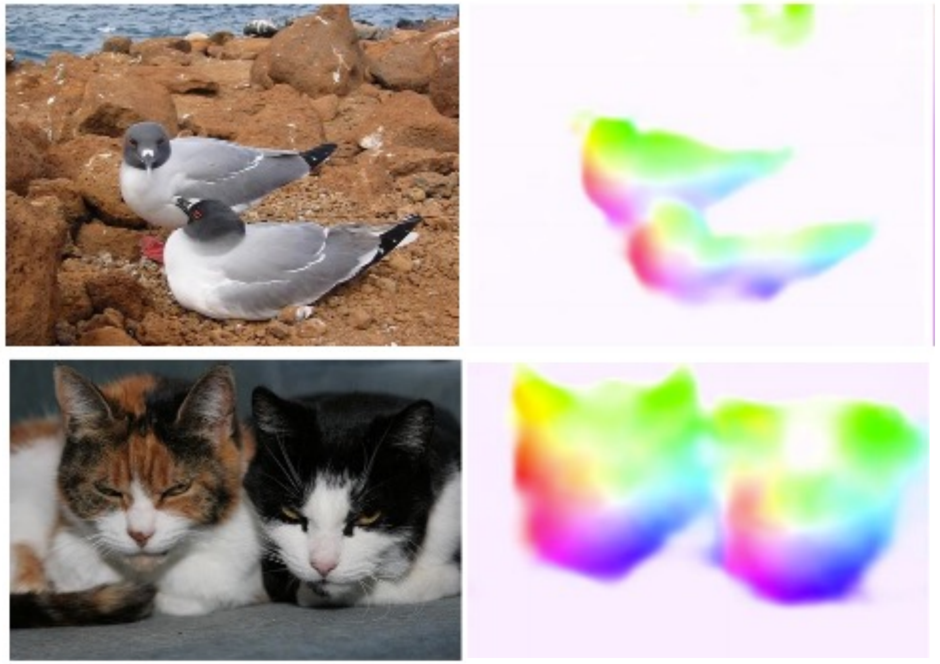
For pixels belonging to the same instance:

$$X_i + D(X_i) = X_j + D(X_j)$$

$$\sum_{\forall X} D(X) = 0$$

$$\mathcal{L}_{\mathrm{fg}}^{\mathcal{D}} = \frac{1}{|\mathcal{P}_{\mathrm{fg}}^{+}|} \sum_{(i,j) \in \mathcal{P}_{\mathrm{fg}}^{+}} \left| \delta(i,j) - \hat{\delta}(i,j) \right|.$$

$$\mathcal{L}_{\mathrm{bg}}^{\mathcal{D}} = \frac{1}{|\mathcal{P}_{\mathrm{bg}}^{+}|} \sum_{(i,j) \in \mathcal{P}_{\mathrm{bg}}^{+}} |\delta(i,j)|.$$



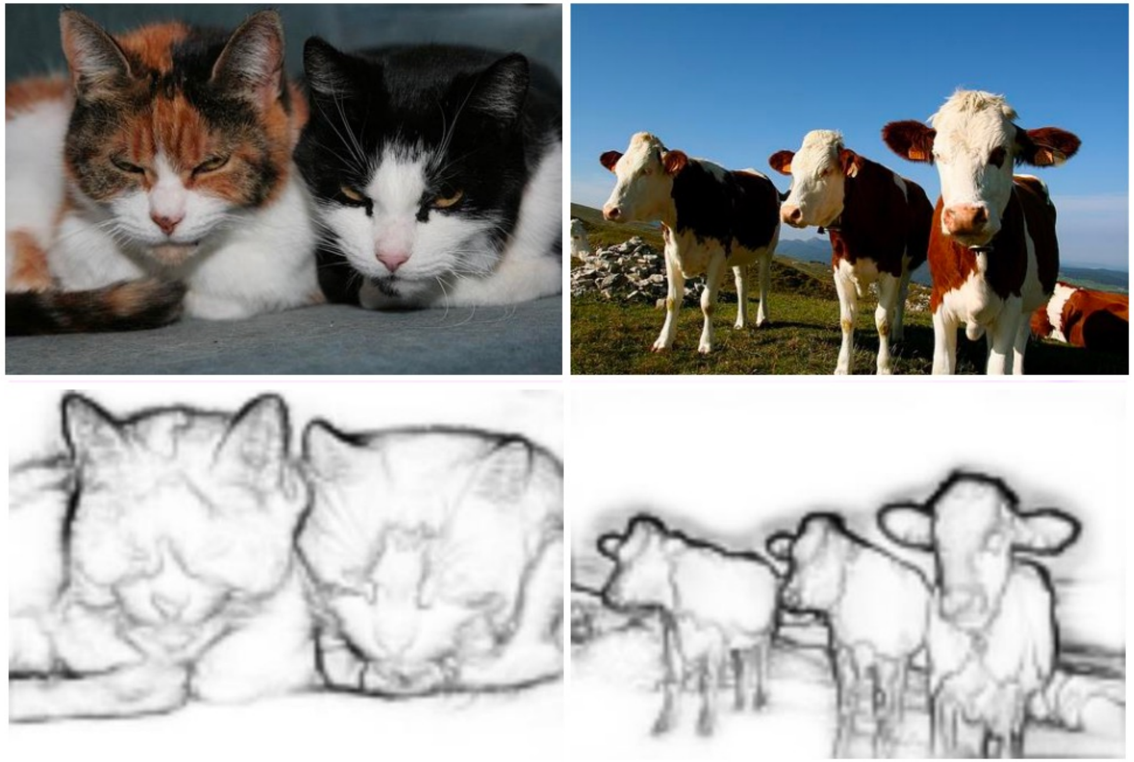
1 IRN -- Class Boundary Detection

$$B \in [0,1]^{w \times h}$$

$$a_{ij} = 1 - \max_{k \in \Pi_{ij}} \mathcal{B}(\mathbf{x}_k)$$

$$\mathcal{L}^{\mathcal{B}} = -\sum_{(i,j)\in\mathcal{P}^+_{\mathrm{fg}}} \frac{\log a_{ij}}{2|\mathcal{P}^+_{\mathrm{fg}}|} - \sum_{(i,j)\in\mathcal{P}^+_{\mathrm{bg}}} \frac{\log a_{ij}}{2|\mathcal{P}^+_{\mathrm{bg}}|} - \sum_{(i,j)\in\mathcal{P}^-} \frac{\log(1-a_{ij})}{|\mathcal{P}^-|}$$

$$\mathcal{L} = \mathcal{L}^{\mathcal{D}}_{\mathrm{fg}} + \mathcal{L}^{\mathcal{D}}_{\mathrm{bg}} + \mathcal{L}^{\mathcal{B}}.$$



1 IRN -- Generating pseudo labels

Stage 1: Generating Class-agnostic Instance Map:

$$\mathcal{D}_{u+1}(\mathbf{x}) = \mathcal{D}_u(\mathbf{x}) + \mathcal{D}\left(\mathbf{x} + \mathcal{D}_u(\mathbf{x})\right) \quad \forall \mathbf{x},$$

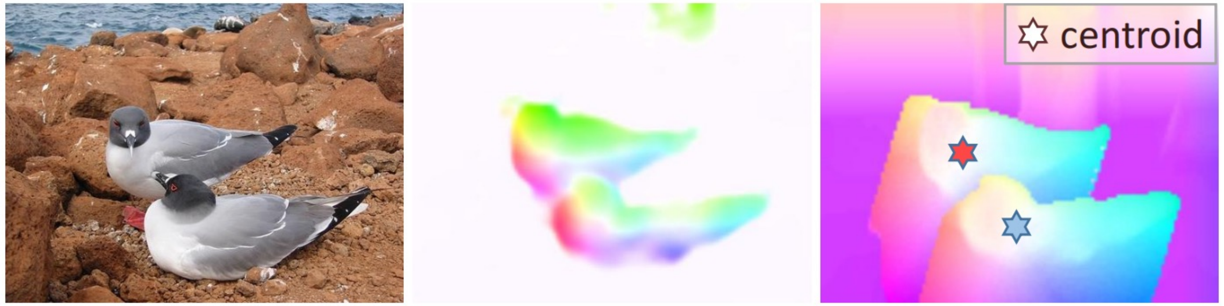


Figure 5. Detecting instance centroids. (left) Input image. (center) An initial displacement field. (right) A refined displacement field and detected centroids.

$$class - agnostic\ instance\ map:\ I \in [1, k]^{w \times h}$$

1 IRN -- Generating pseudo labels

BEIJING JIAOTONG UNIVERSITY

Stage 2: Synthesizing Instance Segmentation Labels

$$\bar{M}_{ck}(\mathbf{x}) = \begin{cases} M_c(\mathbf{x}) & \text{if } I(\mathbf{x}) = k, \\ 0 & \text{otherwise,} \end{cases}$$

$$A = [a_{ij}] \in \mathbb{R}^{wh \times wh}$$

$$T = S^{-1} A^{\circ\beta}, \quad \text{where} \quad S_{ii} = \sum_j a_{ij}^{\beta}$$

$$\text{vec}(\bar{M}^*_{ck}) = T^t \cdot \text{vec}(\bar{M}_{ck} \odot (1 - \mathcal{B})),$$

Method	mIoU
CAM	8.6
CAM + Class Boundary	34.1
CAM + Displacement Field + Class Boundary (Ours)	37.7

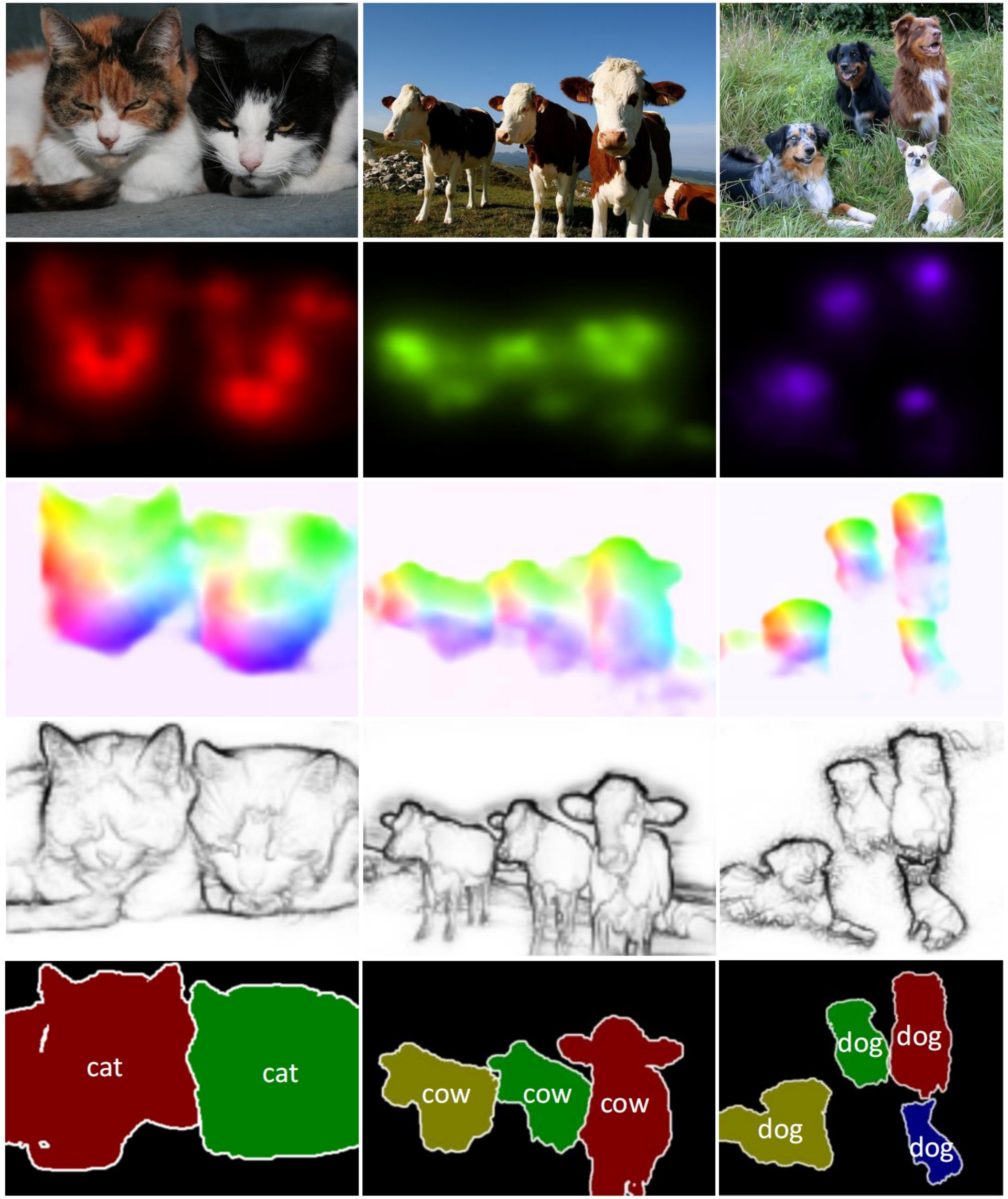
Table 1. Quality of our pseudo instance segmentation labels in AP^r_{50}, evaluated on the PASCAL VOC 2012 *train* set.

CAM	Prop. w/ AffinityNet [1]	Prop. w/ IRNet (Ours)
48.3	59.3	66.5

Table 2. Quality of pseudo semantic segmentation labels in mIoU, evaluated on the PASCAL VOC 2012 *train* set. "Prop" means the semantic propagation using predicted affinities.

Method	Sup.	Extra data / Information	AP^r_{50}	AP^r_{70}
PRM [50]	$\mathcal{I}$	MCG [2]	26.8	-
SDI [22]	$\mathcal{B}$	BSDS [33]	44.8	-
SDS [16]	$\mathcal{F}$	MCG [2]	43.8	21.3
MRCNN [17]	$\mathcal{F}$	MS-COCO [29]	69.0	-
Ours-ResNet50	$\mathcal{I}$	-	46.7	23.5

Table 3. Instance segmentation performance on the PASCAL VOC 2012 *val* set. The supervision types (Sup.) indicate: $\mathcal{I}$–image-level label, $\mathcal{B}$–bounding box, and $\mathcal{F}$–segmentation label.



2 Overall architecture

Beijing JiaoTong University

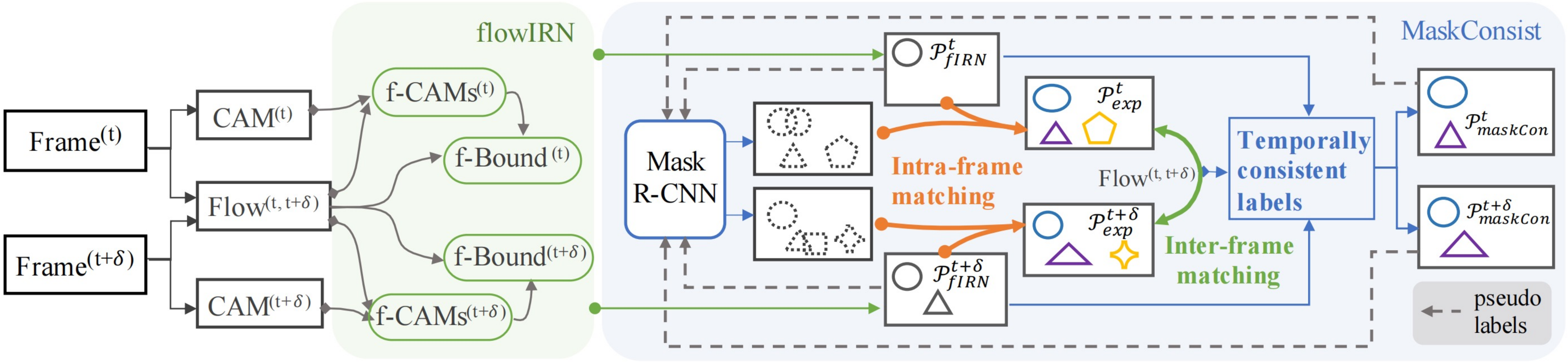


Figure 2. Our pipeline mainly consists of two modules: flowIRN and MaskConsist. FlowIRN adapts IRN [6] by incorporating optical flow to modify CAMs (f-CAMs), as well as introducing a new loss function: flow-boundary loss (f-Bound loss). MaskConsist matches the predictions from two successive frames and transfers high-quality predictions from one frame as pseudo-labels to another. It has three components: intra-frame matching, inter-frame matching and temporally consistent labels, shown in orange, green and blue, respectively. First, flowIRN is trained with frame-level class labels. Next, MaskConsist is trained with the pseudo-labels generated by flowIRN.

2 FlowIRN

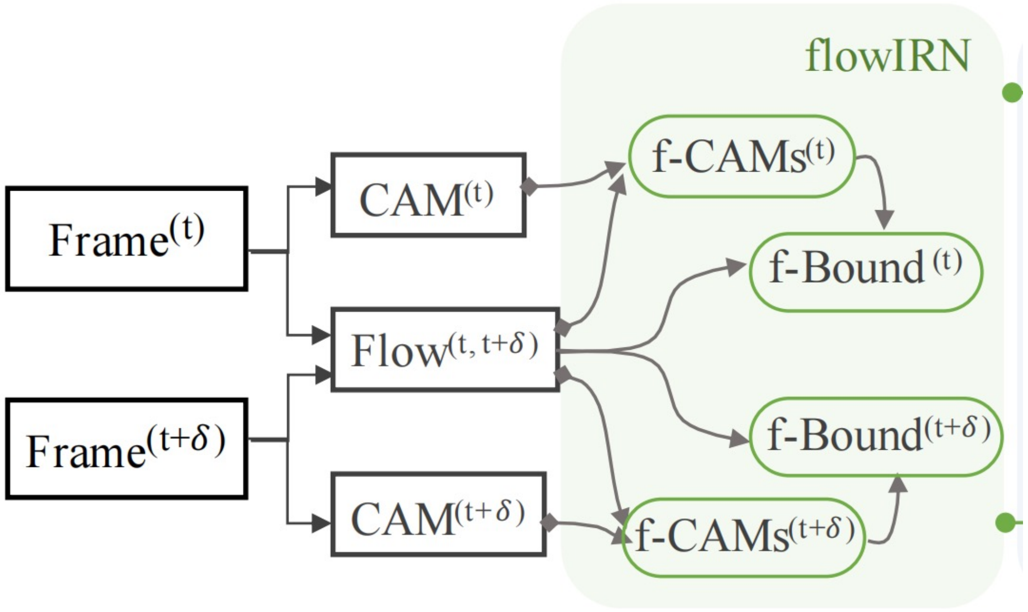
Flow-Amplified CAMs:

$Optical\ Flow\ \mathcal{F} \in \mathbb{R}^{H\times W\times 2}$

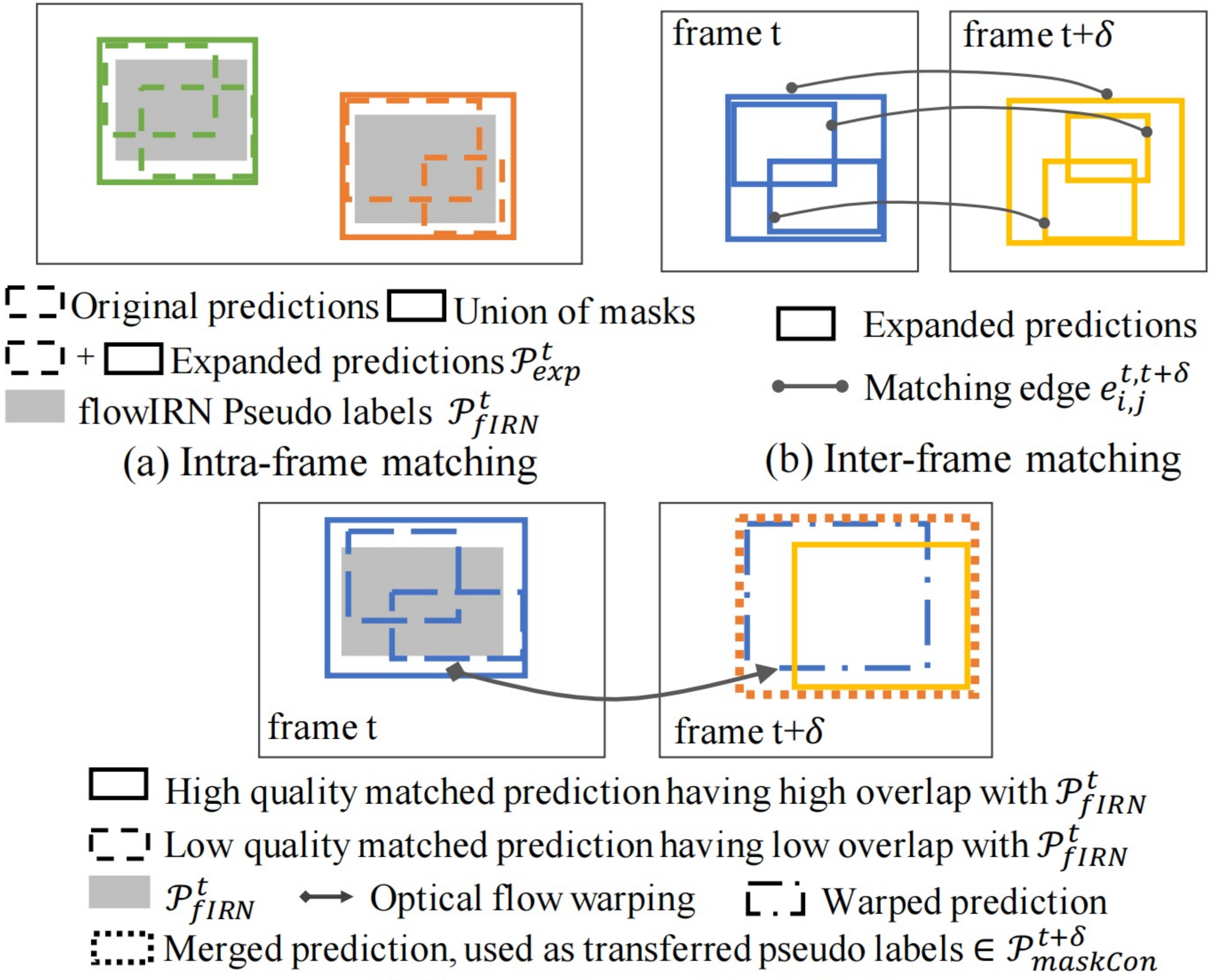
$$\text{f-CAM}_c(x) = \text{CAM}_c(x) \times A^{\mathbb{I}(||\mathcal{F}(x)||_2 > T)}$$

Flow-boundary loss:

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



2 MaskConsist



(a) Intra-frame matching

(b) Inter-frame matching

(c) Temporally consistent labels

3 Experiments

Optical flow network: self-supervised DDFlow trained on "Flying Chairs" dataset and then fine-tuned on YTVIS in an unsupervised way training 120 hours on 4 P100

FlowIRN: first trained on PASCAL VOC 2012 before training with MaskConsist

MaskConsist: 90K iter for YTVIS, 75K for Cityscapes

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.

4 Ablation study

	YTVIS	Cityscapes
IRN [6]	25.42	8.46
IRN+f-Bound	26.60	9.51
IRN+f-CAMs	27.47	10.55
flowIRN	**28.45**	**10.75**

Table 4. Ablation study of flowIRN components. Results are reported on training data to evaluate pseudo-label quality. No second-step Mask R-CNN or MaskConsist training is applied here.

MaskConsist Components			AP_{50}	
Intra-F	Inter-F	IoM-NMS	YTVIS	Cityscapes
✗	✗	✗	31.43	14.66
✗	✓	✓	33.75	14.92
✓	✗	✓	31.08	14.43
✓	✓	✗	33.65	15.27
✓	✓	✓	**34.66**	**16.05**

Table 5. Ablation study of MaskConsist components. The numbers in this table are generated by models with two-step training.

4 Ablation study

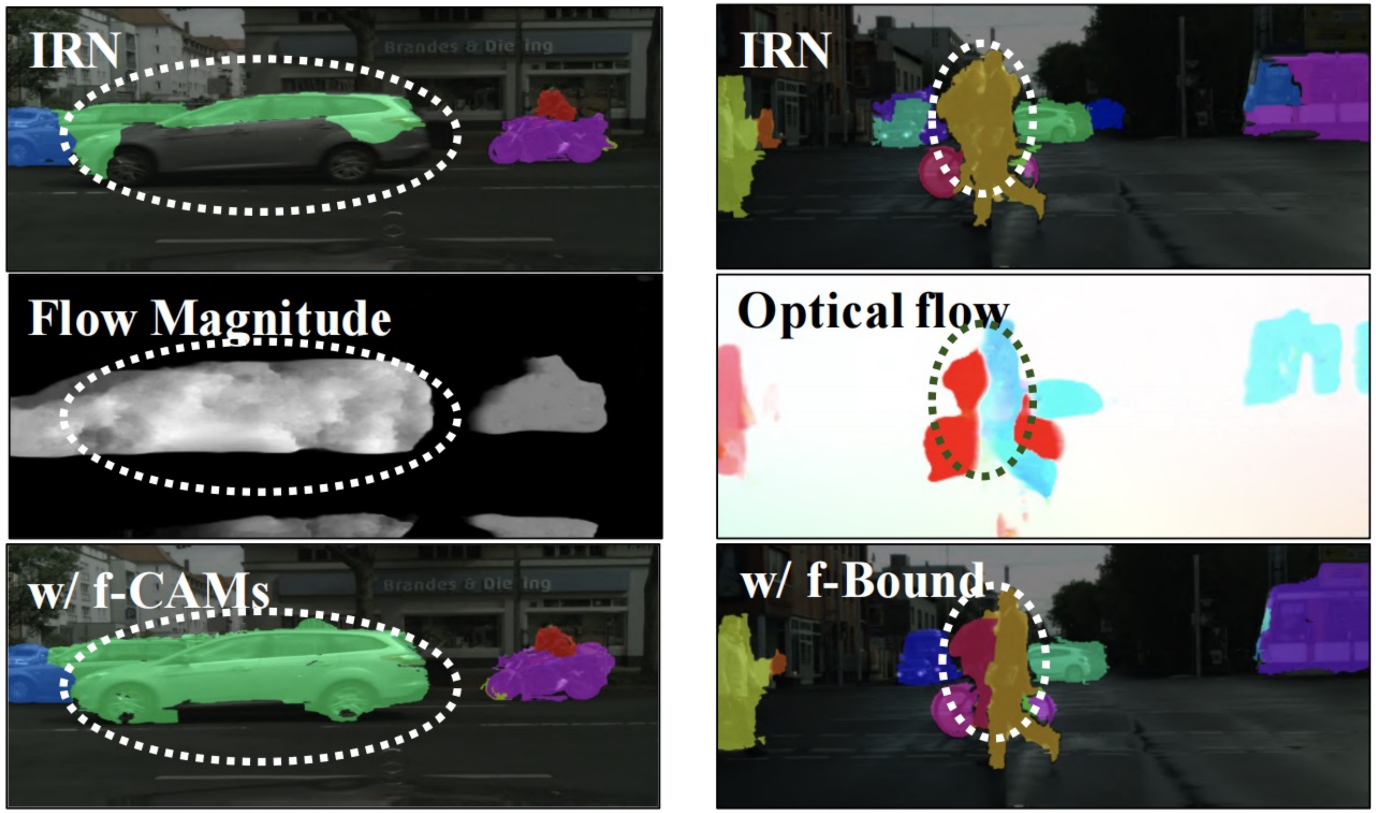


Figure 5. Improvement introduced by f-CAMs and f-Bound. Top: output of IRN. Middle: optical flow extracted for the input frame. Bottom: output after incorporating f-CAMs or f-Bound.

BEIJING JIAOTONG UNIVERSITY

4 Ablation study

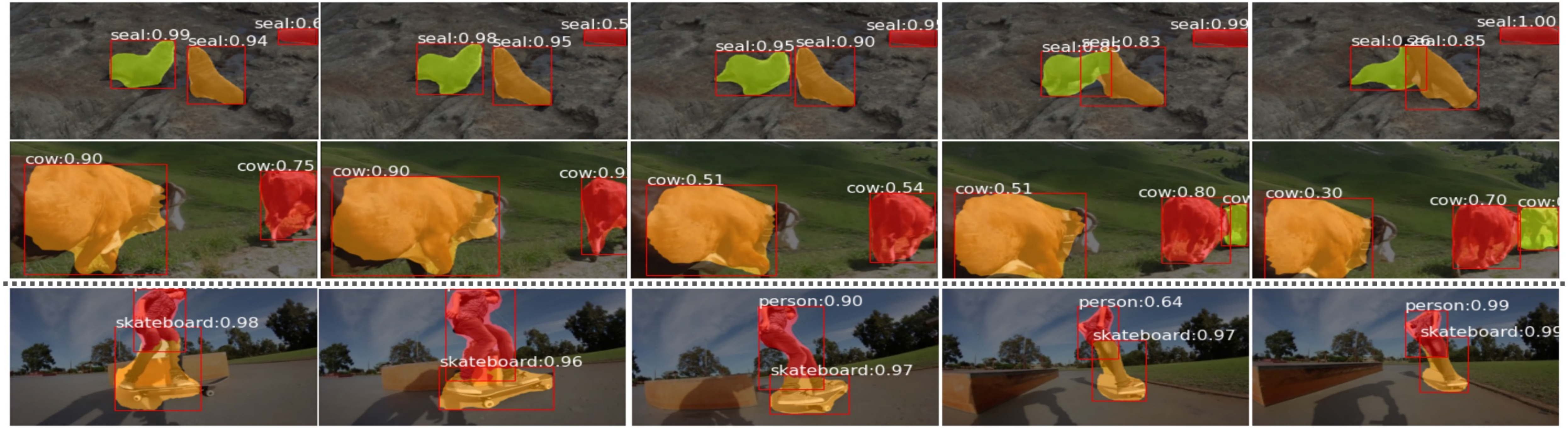


Figure 4. Example Video instance segmentation results from our method on Youtube-VIS dataset.