Holistic, Instance-level Human Parsing

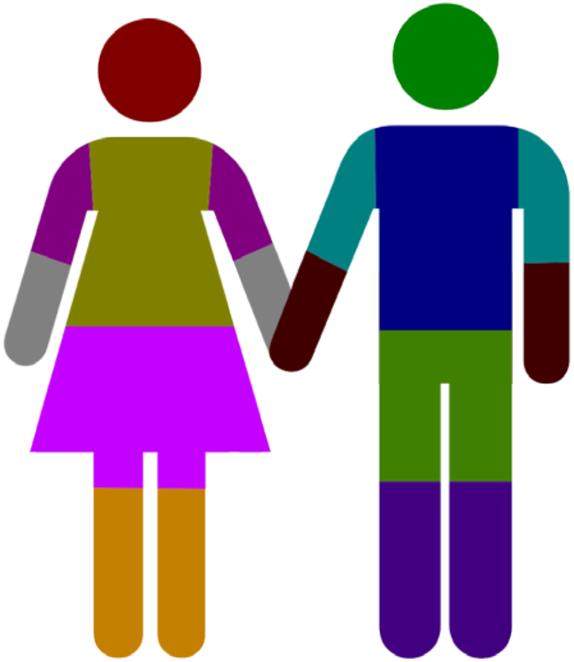
Qizhu Li*, Anurag Arnab*, Philip Torr

* Indicates equal contribution by the authors

05 September 2017

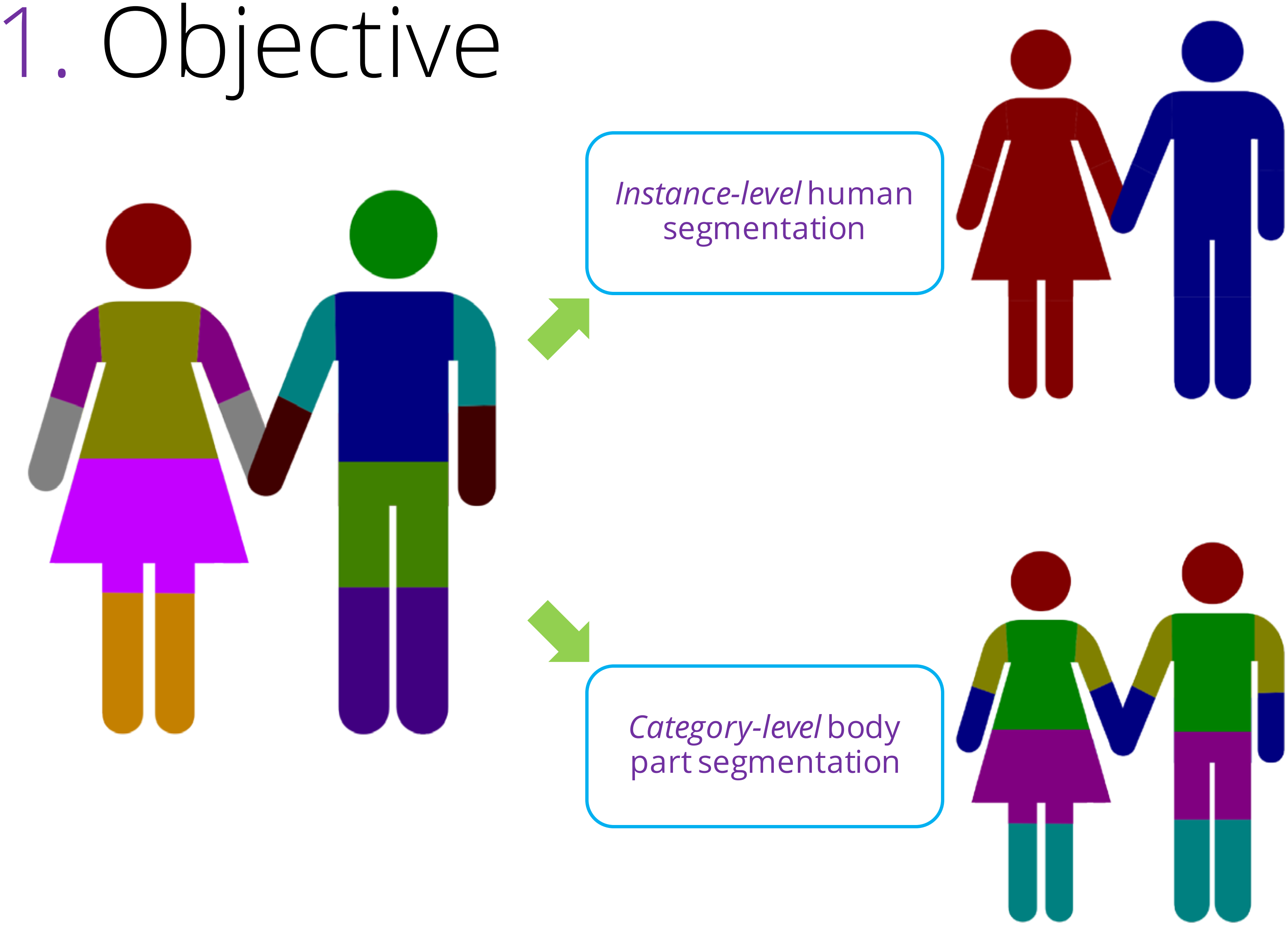
1. Objective

1. Objective



Instance-aware body part segmentation of humans

1. Objective



Instance-level human segmentation

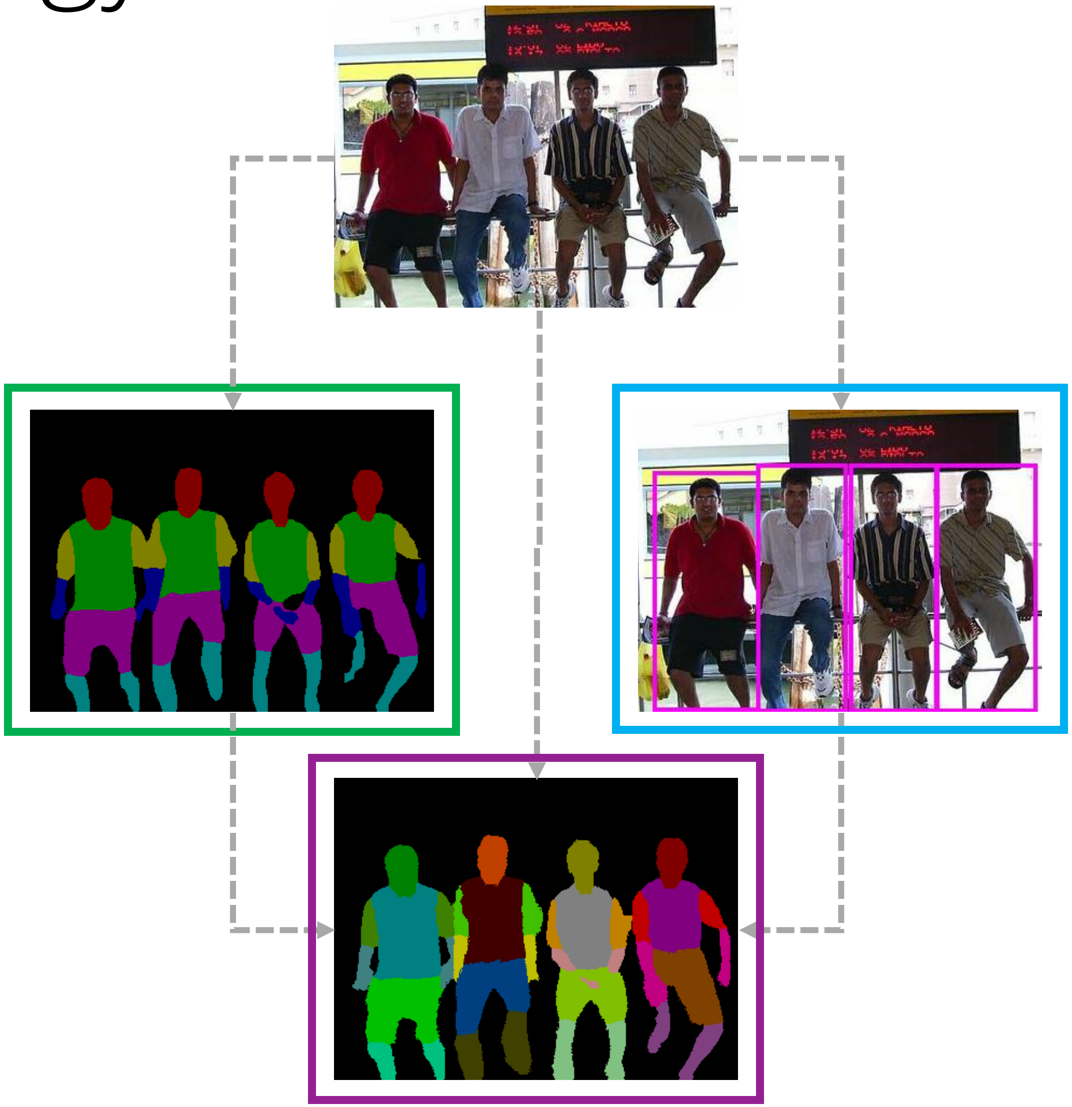
Category-level body part segmentation

2. Methodology

1. Do category-level body part segmentation

2. Detect humans

3. Use the instance-level segmentation module to assign instance labels.



2.1 Methodology: Instance-level segmentation module

$$E(\boldsymbol{V}=\boldsymbol{v}) = -\overbrace{\sum_{i}^{N} \ln(\overbrace{w_1\psi_{Box}(v_i)}^{\text{Box term}} + \overbrace{w_2\psi_{Global}(v_i)}^{\text{Global term}} + \varepsilon) + \sum_{i<j}^{N} \psi_{Pairwise}(v_i, v_j)}^{\text{Instance CRF}}$$

Where $V = \{V_1, V_2, \dots V_N\}$ is a multinormial variable at all $\boldsymbol{N}$ pixels,
$V_i \in \{1, 2, \dots D\} \times \{1, 2, \dots P\} \cup \{0, 0\}$
$E(\boldsymbol{V}=\boldsymbol{v})$ is the energy of $\boldsymbol{V}$ taking a particular value $\boldsymbol{v}$

2.1.1 Methodology: Box term

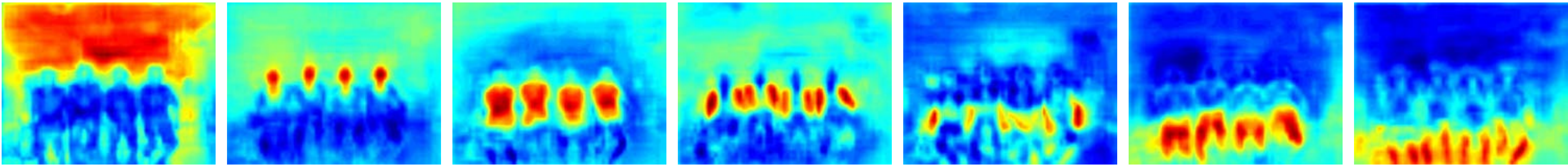
- Input 1: human detections:
  - scores s_i and bounding boxes B_i for $1 \le i \le D$
- Input 2: semantic segmentation network output:
  - Feature map Q with $P + 1$ channels ($P = 6$ here)
- Output:

$$\psi_{Box}\left(V_k = (i, j)\right) = \begin{cases} s_i Q_k(j), & k \in B_i \\ 0, & k \notin B_i \end{cases}$$

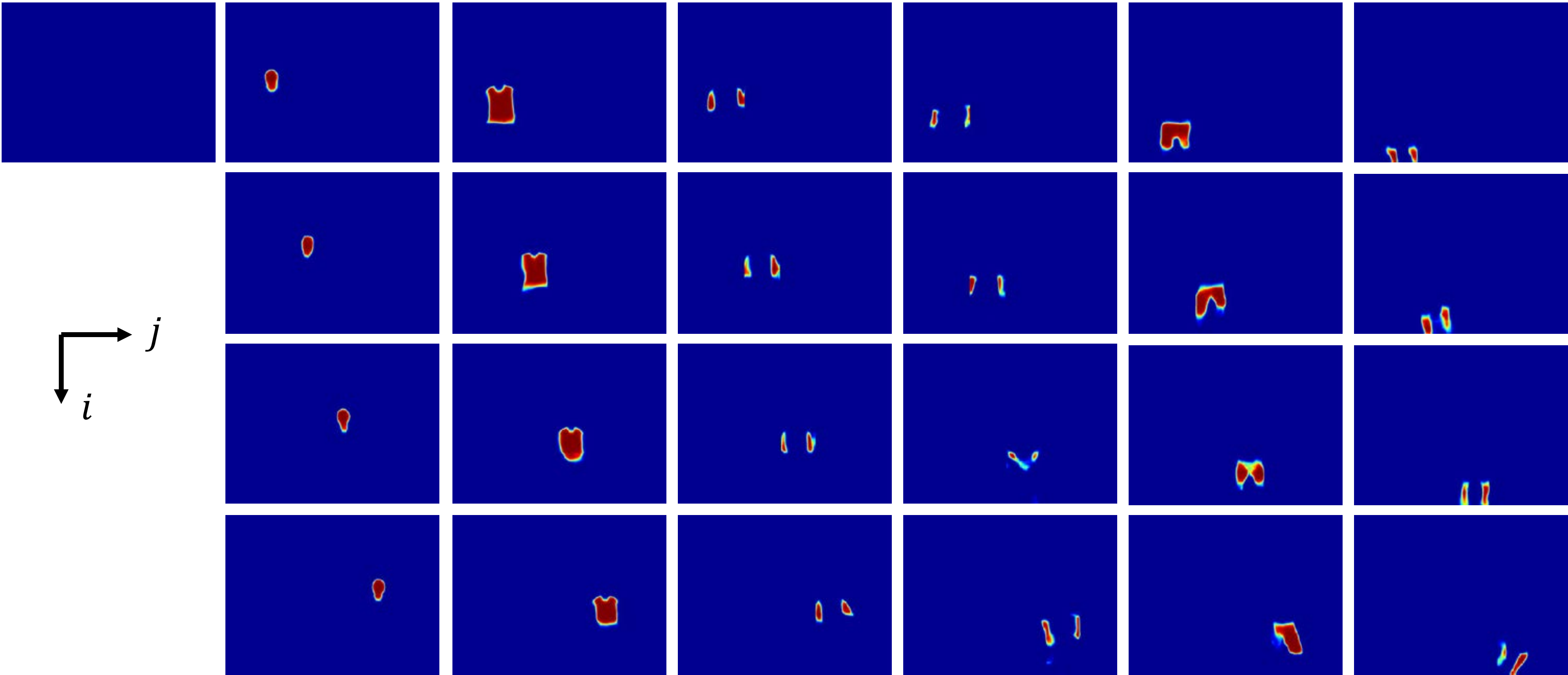
for $(i, j) \in \{1, 2, \ldots D\} \times \{1, 2, \ldots P\}$

Input 1: Human detections

Input 2: Semantic body part segmentation probability maps



Output: Box terms



2.1.2 Methodology: Global term

- Input: semantic segmentation network output:
  - Feature map Q with $P+1$ channels ($P=6$ here)
- Output:

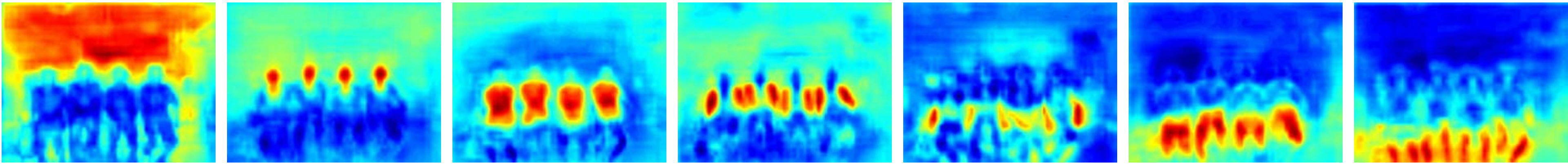
$$\psi_{Global}\big(V_k=(i,j)\big)=Q_k(j)$$

for $(i,j)\in\{1,2,\ldots D\}\times\{1,2,\ldots P\}\cup\{0,0\}$

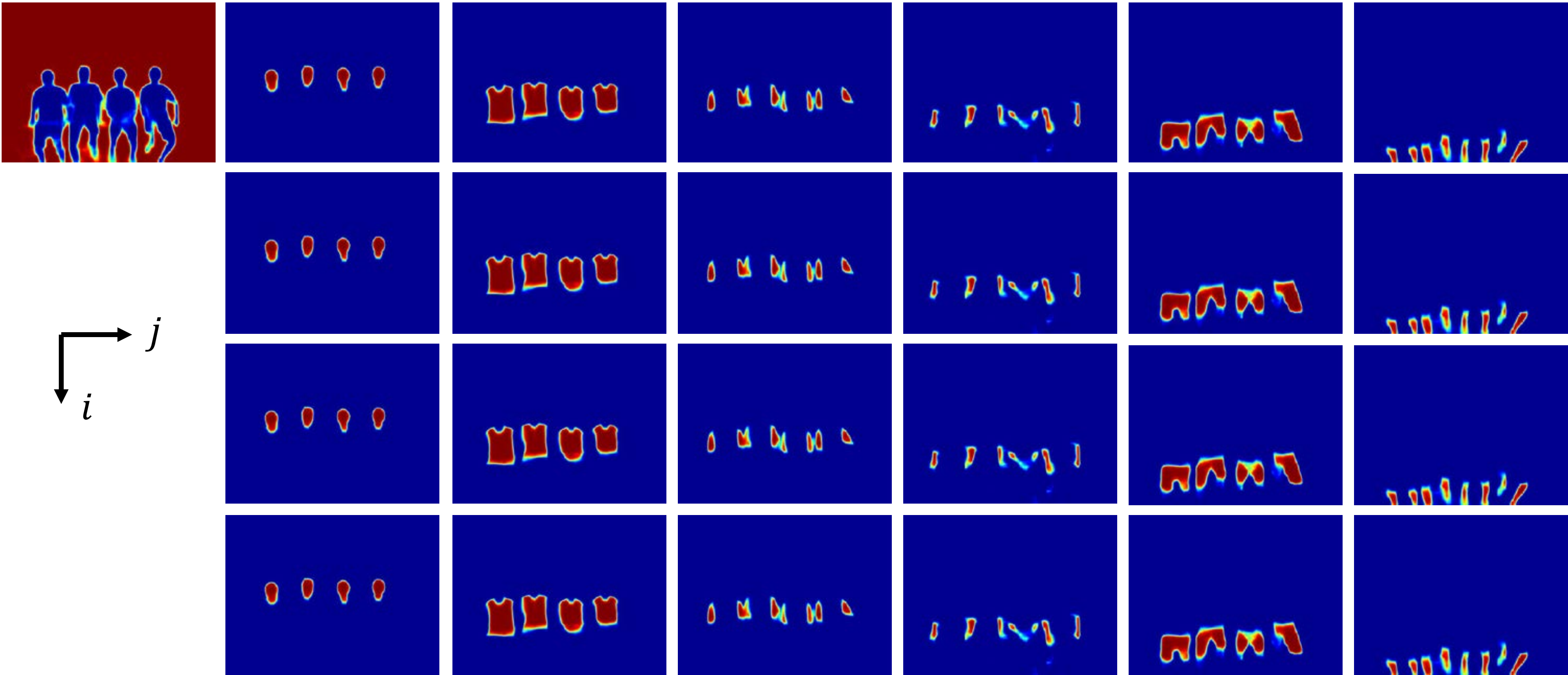
Input 1: **Number** of human detections

4

Input 2: Semantic body part segmentation probability maps



Output: Global terms



2.2 Methodology: Loss function

- Observation: permuting the label IDs in an instance segmentation ground truth produces an equally valid ground truth.

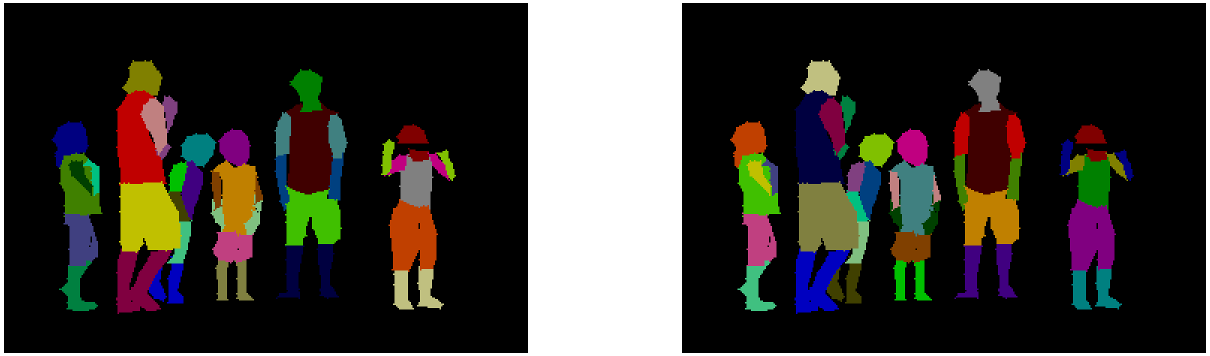


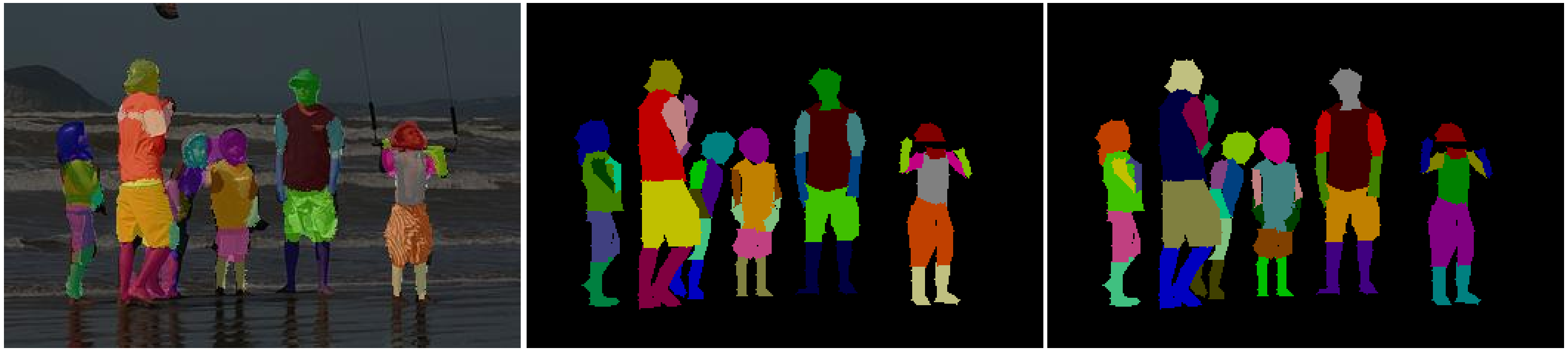
Figure 1. Permutations of ground truth labels are equally valid

2.2 Methodology: Loss function

- We match ground truth $\mathcal{Y}$ to prediction $\mathcal{P}$ before we carry out loss calculation.
- Matched ground truth is given by:

$$\mathcal{Y}^* = \underset{\mathcal{Z} \in \pi(\mathcal{Y})}{\mathbf{argmax}}\, \mathbf{IoU}(\mathcal{Z}, \mathcal{P})$$

- Then cross-entropy loss is calculated as per normal

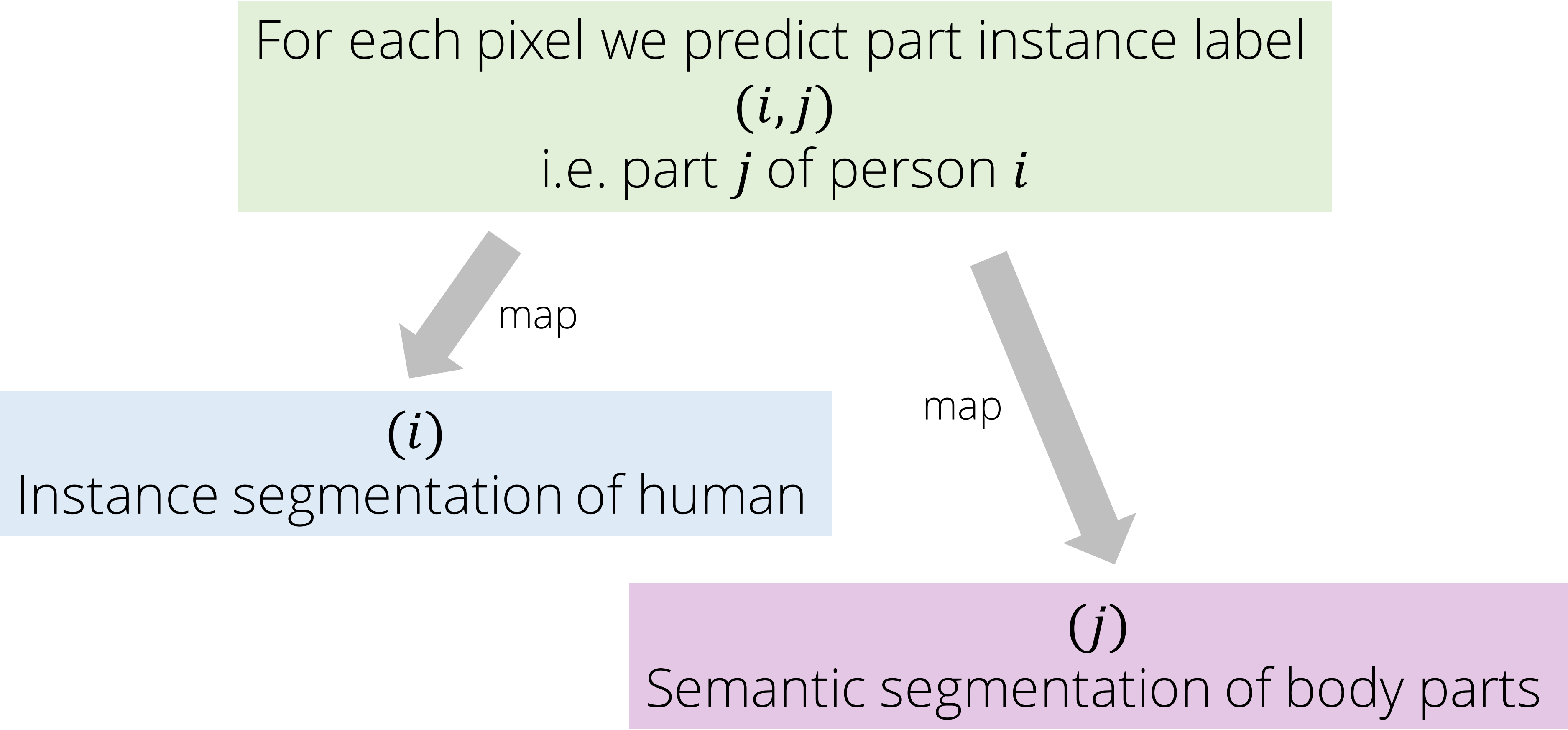


(a) Prediction (a) Matched GT (a) Ground Truth

Figure 2. Ground truth is matched to our prediction before calculating loss

2.3 Methodology:

Obtaining segmentation at other granularities



For each pixel we predict part instance label

(i, j)

i.e. part j of person i

map

(i)

Instance segmentation of human

map

(j)

Semantic segmentation of body parts

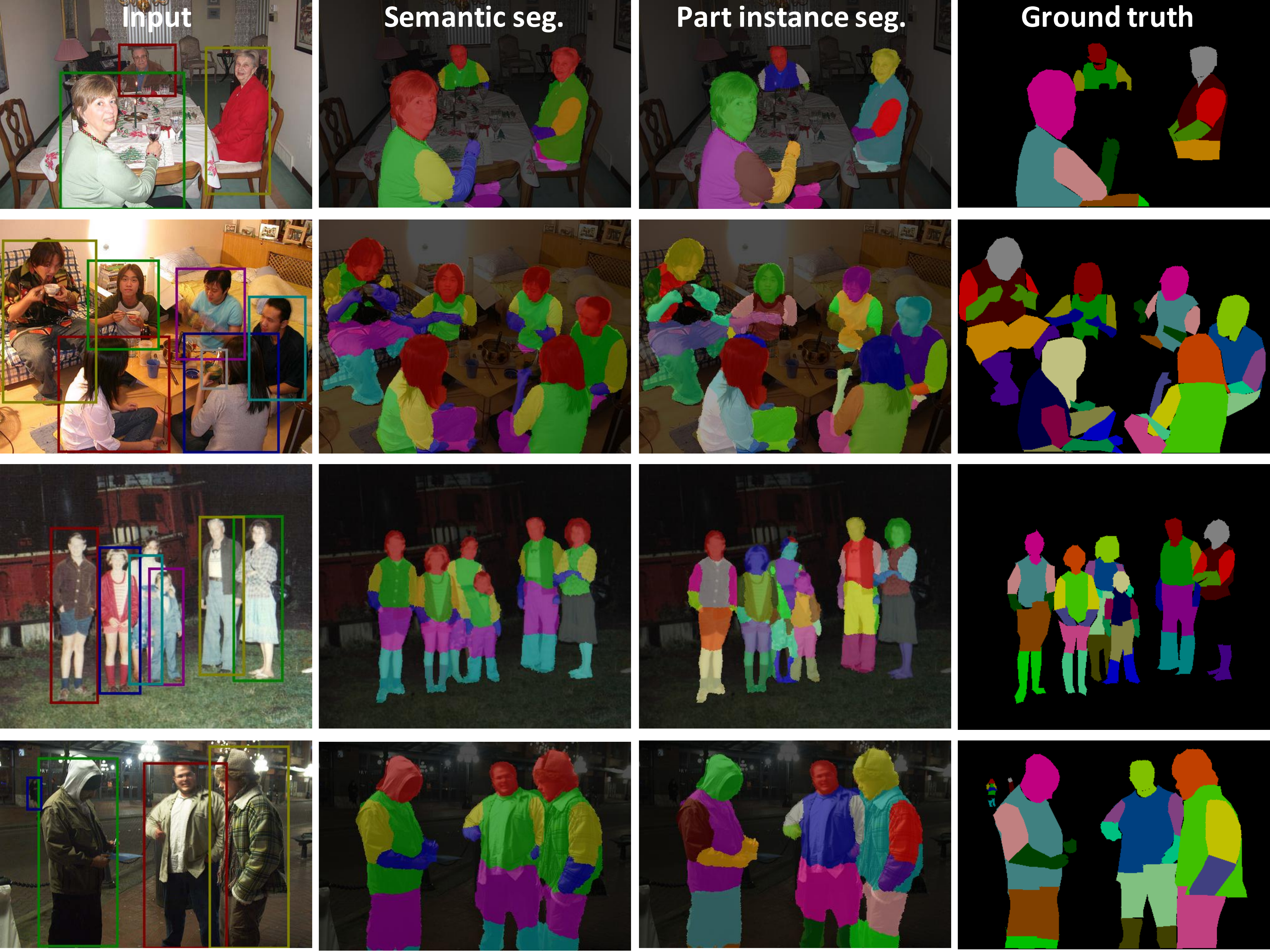
3.1 Results: Part instance segmentation

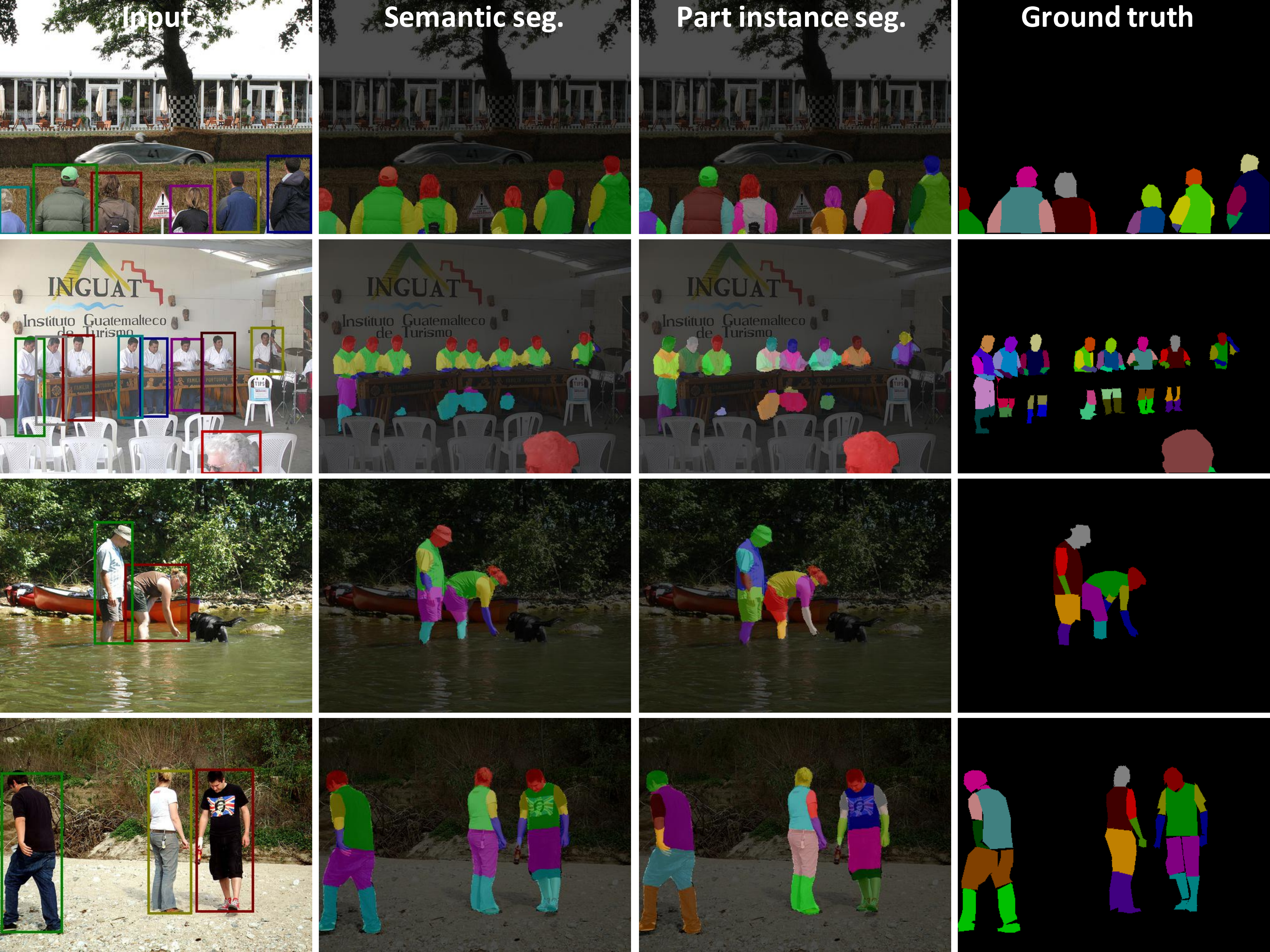
Method	IoU Threshold			AP^r_{vol}
	0.5	0.6	0.7	
MNC	38.8	28.1	**19.3**	36.7
Ours, piecewise, box term only	38.0	27.4	16.7	36.6
Ours, piecewise	38.8	28.5	17.6	37.3
Ours, end-to-end	39.0	28.6	17.4	37.7
Ours, piecewise, box term only, OHEM	38.7	28.9	17.5	36.7
Ours, piecewise, OHEM	39.7	29.7	18.7	37.4
Ours, end-to-end, OHEM	**40.6**	**30.4**	19.1	**38.4**

Table 2. Ablation study and comparison of AP^r at various thresholds to MNC on Pascal Person-Parts test set. $AP^r_{vol} = \frac{1}{9}\sum_{t=1}^{9} AP^r_{t/10}$.

3.1 Results:
Part instance segmentation

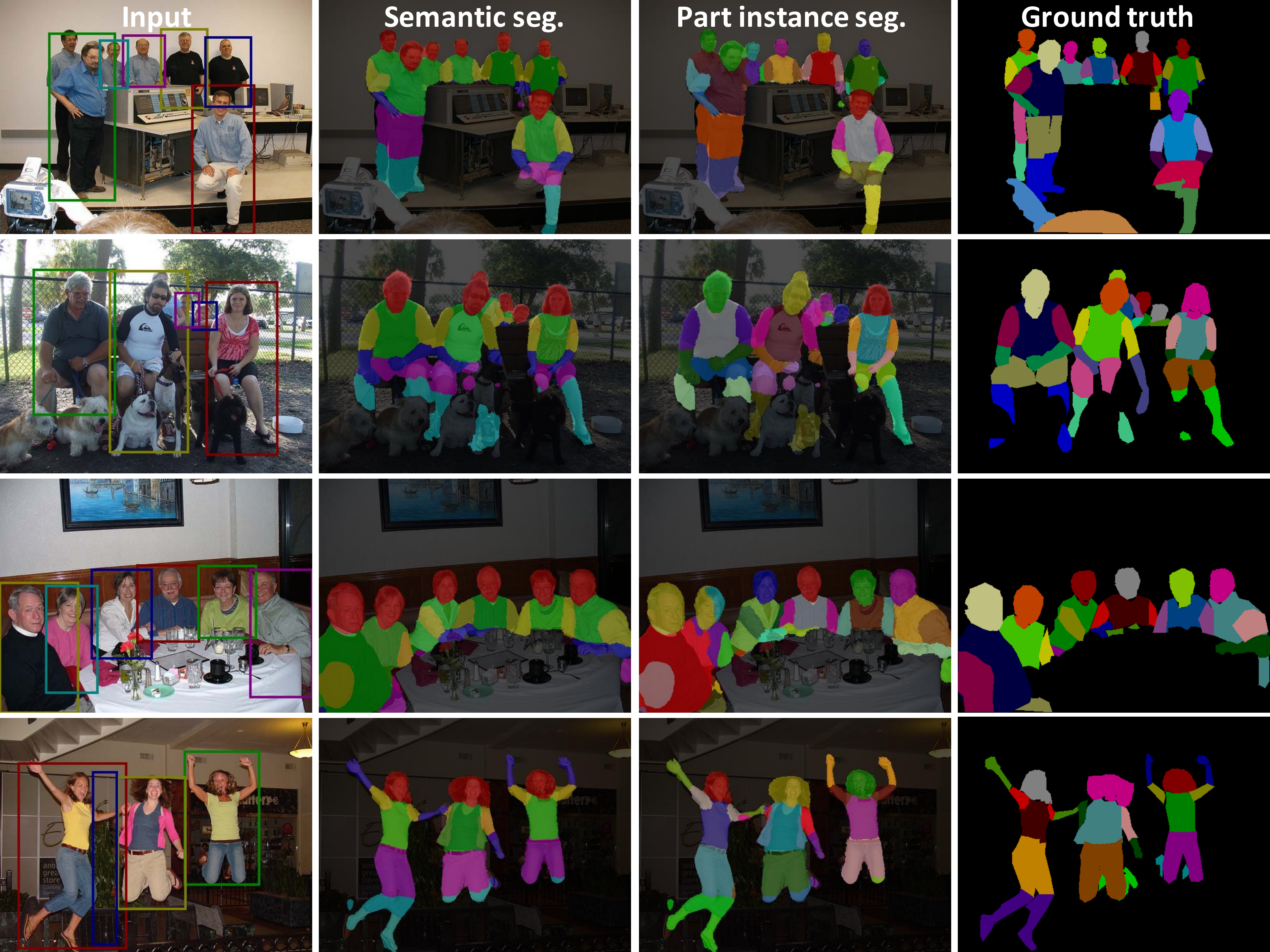
Success cases…


Input
Semantic seg.
Part instance seg.
Ground truth


Input
Semantic seg.
Part instance seg.
Ground truth
INGUAT
Instituto Guatemalteco de Turismo

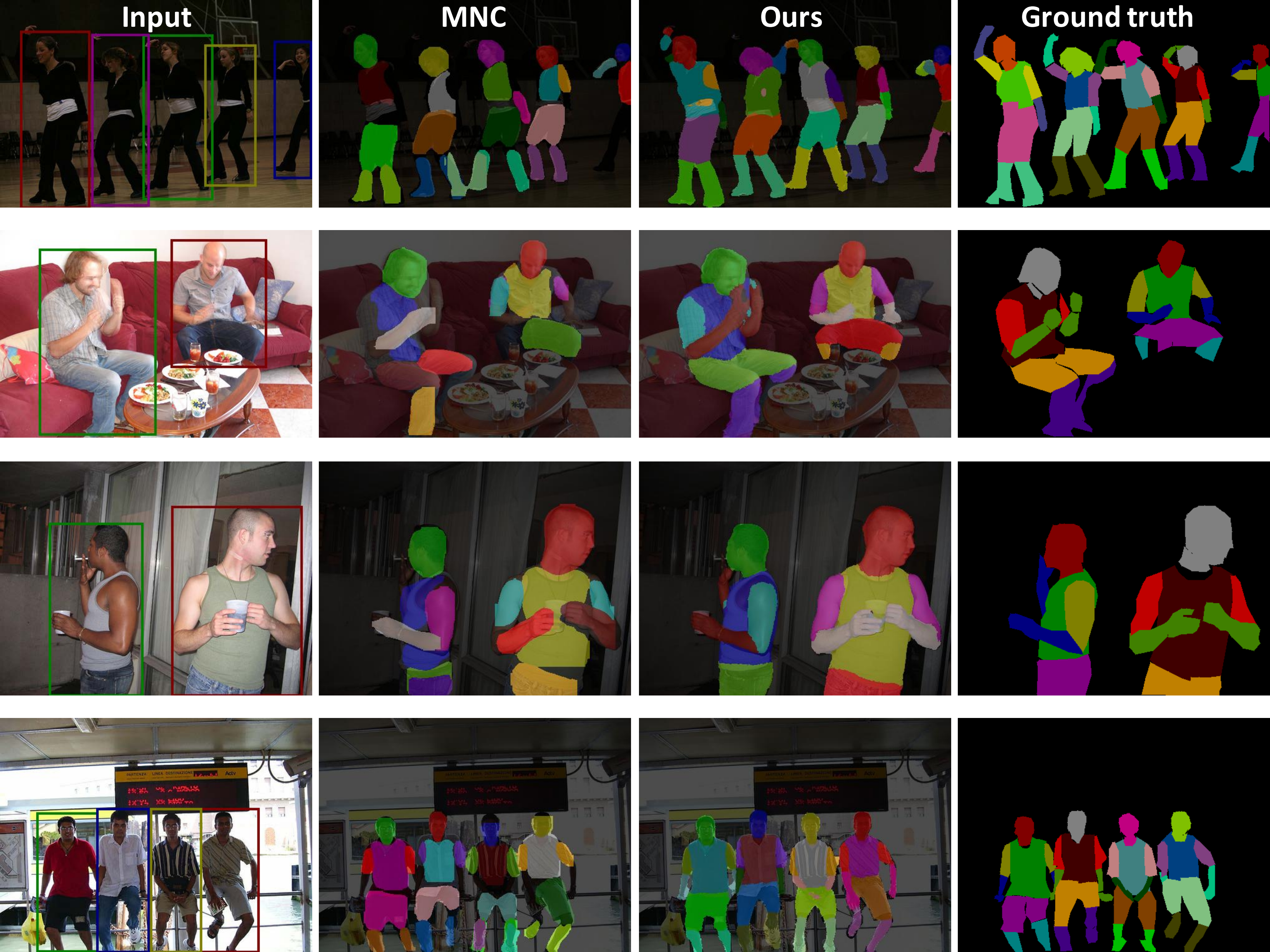
3.1 Results:
Part instance segmentation

Failure cases…


Input
Semantic seg.
Part instance seg.
Ground truth

3.1 Results:
Part instance segmentation

Comparison to MNC…


Input
MNC
Ours
Ground truth

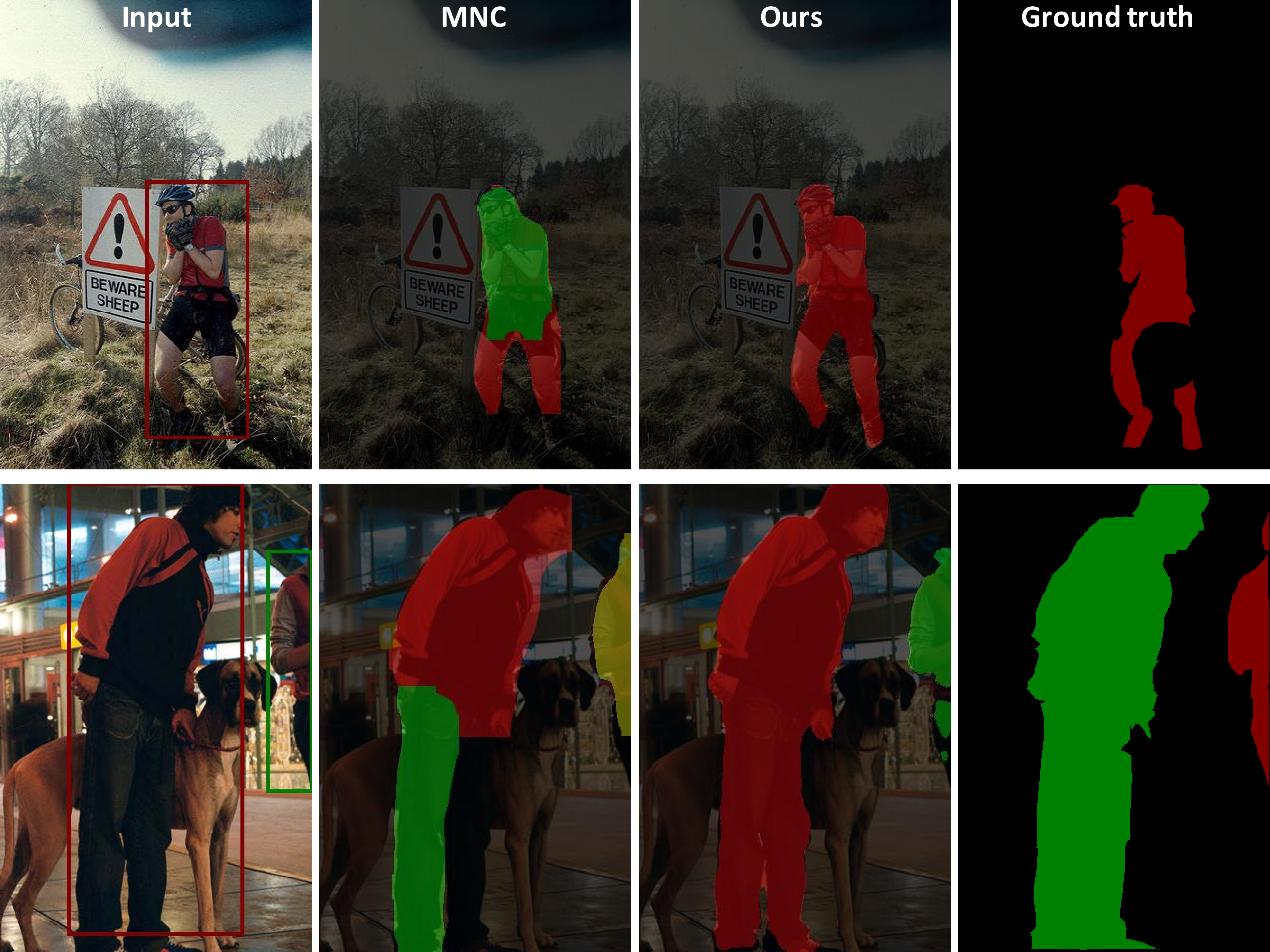
3.2 Results: Human instance segmentation

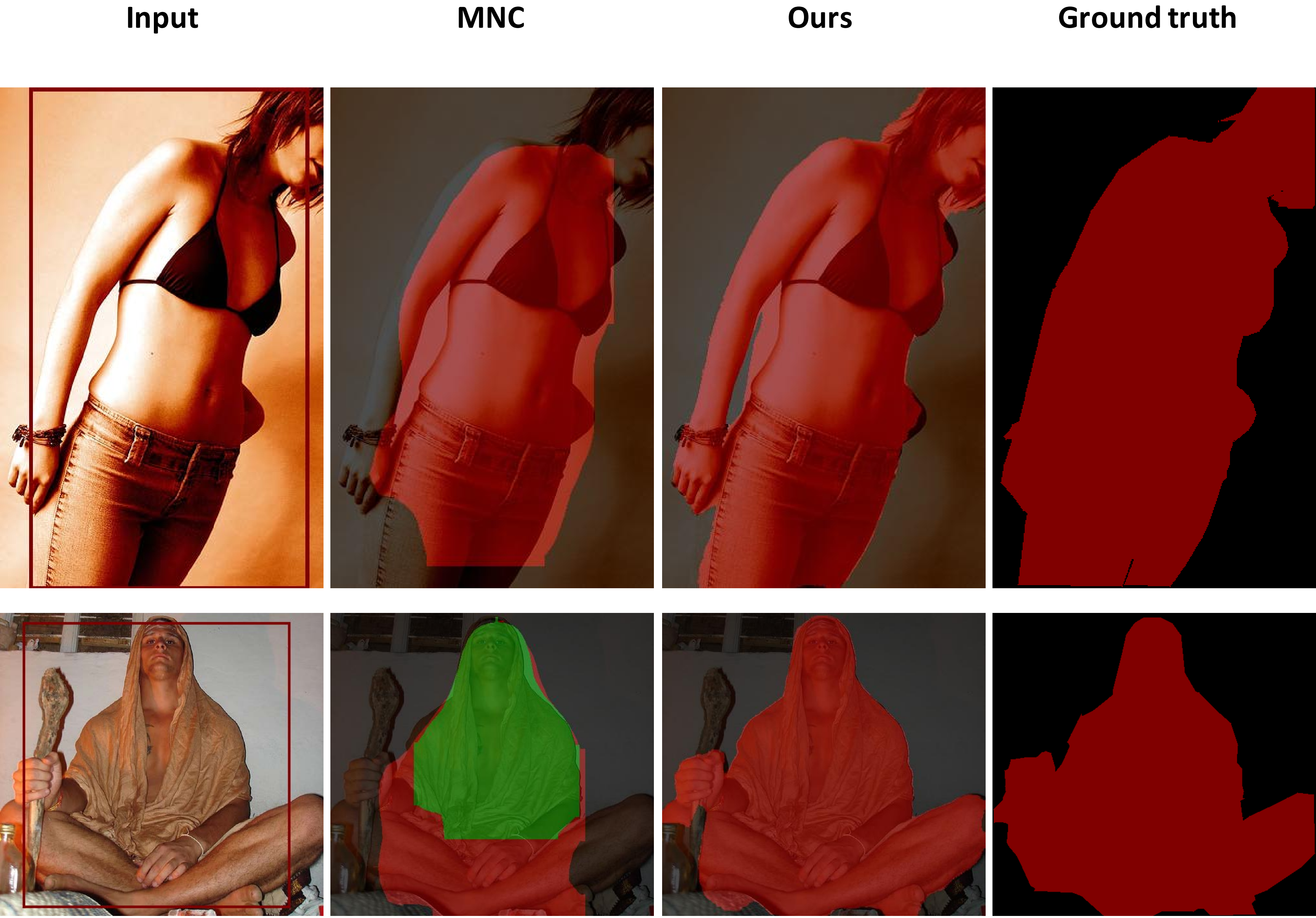
Methods	IoU Thresholds					AP^r_{vol}
	0.5	0.6	0.7	0.8	0.9	
SDS	47.9	31.8	15.7	3.3	0.1	-
Chen et al.	48.3	35.6	22.6	6.5	0.6	-
PFN	48.4	38.0	26.5	16.5	5.9	41.3
Arnab et al.	58.6	52.6	41.1	30.4	10.7	51.8
R2-IOS	60.4	51.2	33.2	-	-	-
Arnab et al.	65.6	58.0	46.7	33.0	14.6	57.4
Ours, piecewise	64.0	59.8	51.0	38.3	**20.1**	57.2
Ours, end-to-end	**70.2**	**63.1**	**54.1**	**41.0**	19.6	**61.0**

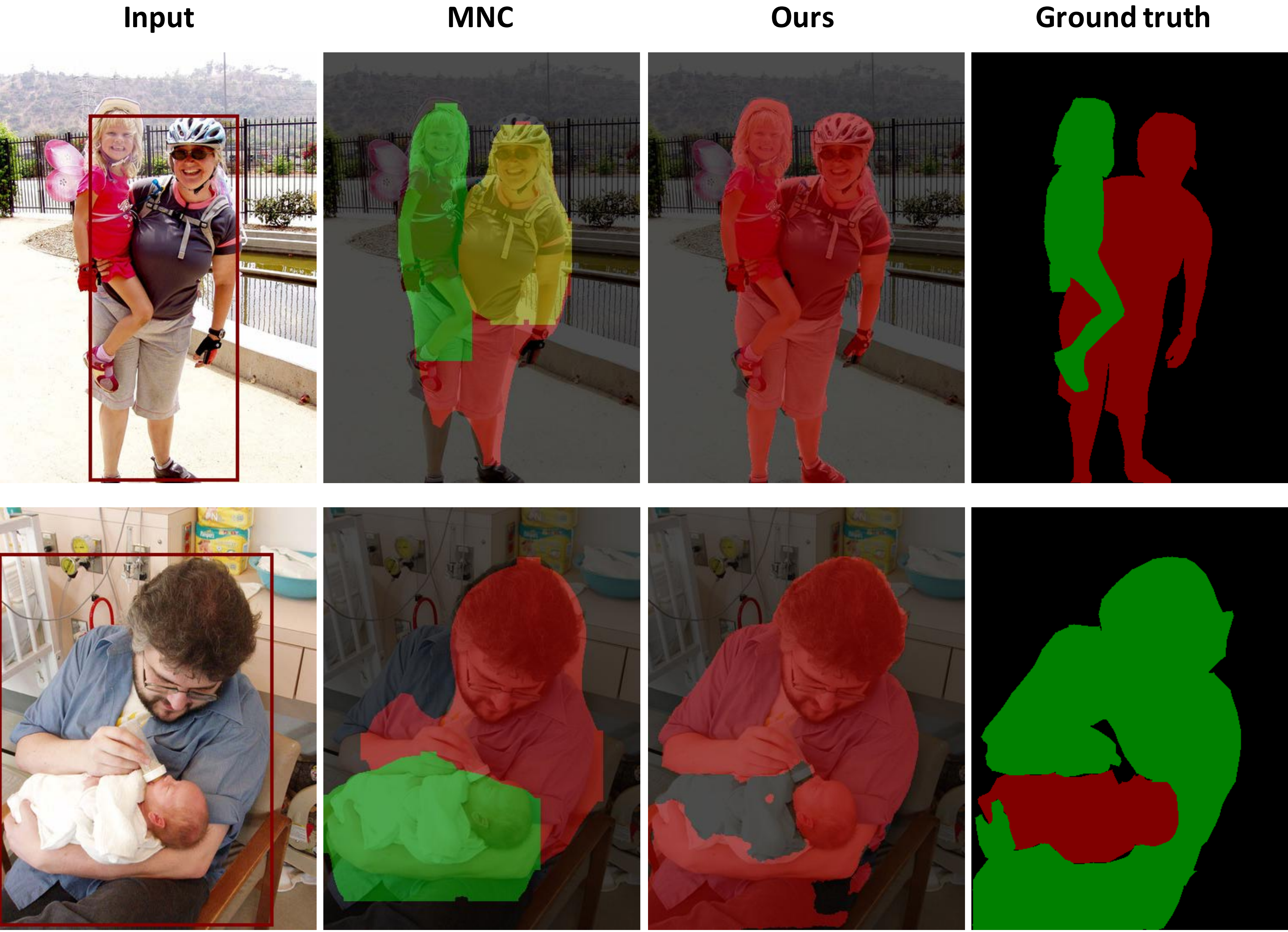
Table 3. Comparison of AP^r at various thresholds for instance-level human segmentation on the VOC 2012 validation set. $AP^r_{vol} = \frac{1}{9}\sum_{t=1}^{9} AP^r_{t/10}$.

3.2 Results: Human instance segmentation

Comparison to MNC…
(We run the public MNC model on
Pascal Person-Parts test set)


Input
MNC
Ours
Ground truth
BEWARE SHEEP


Input
MNC
Ours
Ground truth


Input
MNC
Ours
Ground truth

3.3 Results:

Semantic segmentation of body parts

Method	IoU [%]
Deeplab	53.0
Attention	56.4
HAZN	57.5
LG-LSTM	58.0
Graph LSTM	60.2
Deeplab-v2	64.9
RefineNet	68.6
Ours, pre-trained	65.9
Ours, final network	66.3

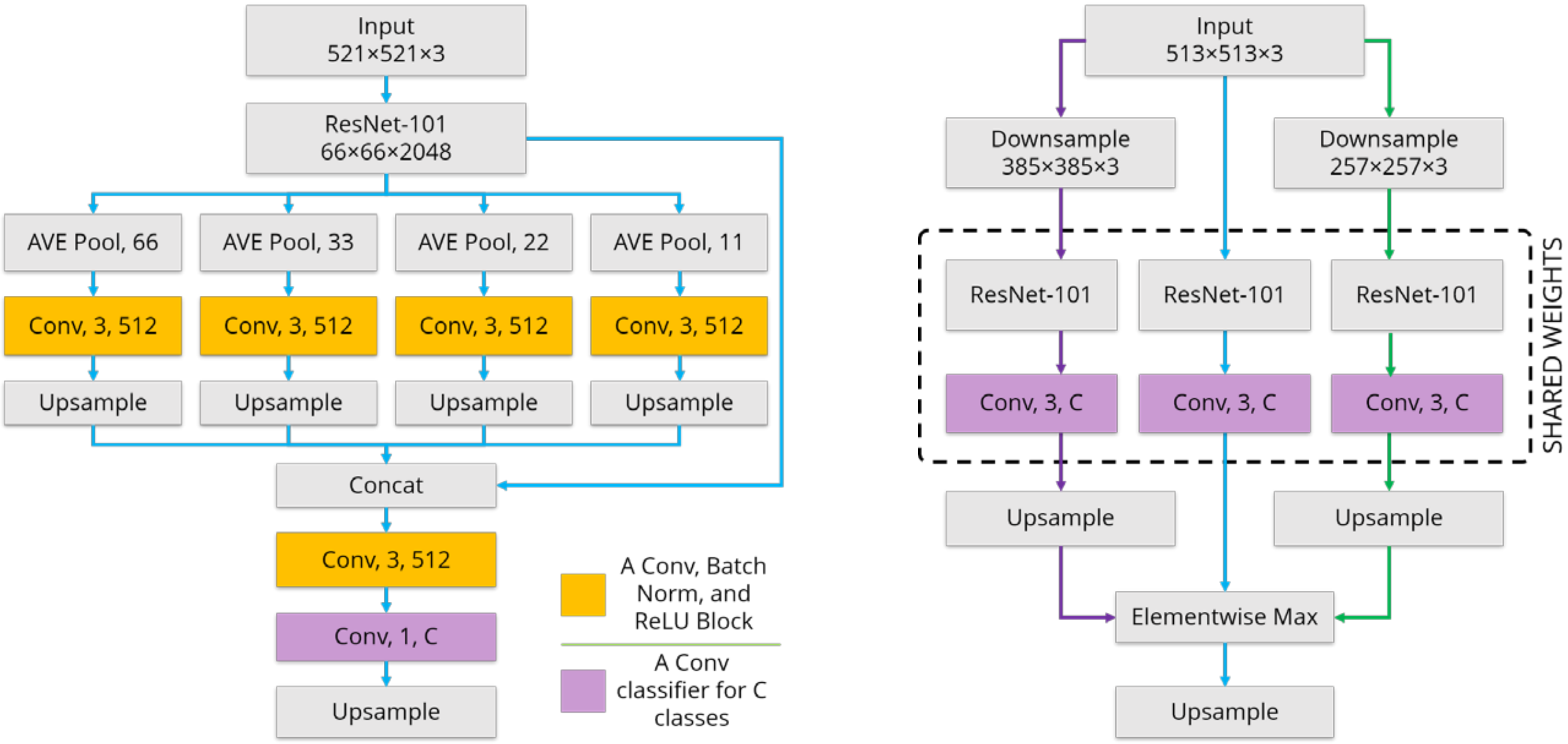
Table 4. Comparison of semantic part segmentation results on the Pascal Person-Parts test set.

The End

Thank you!

Appendix

Methodology:
Category-level segmentation module



(a) Ours

(b) Deeplab-v2

Figure 2. Comparison of our category-level segmentation module to Deeplab-v2

Methodology:
Category-level segmentation module

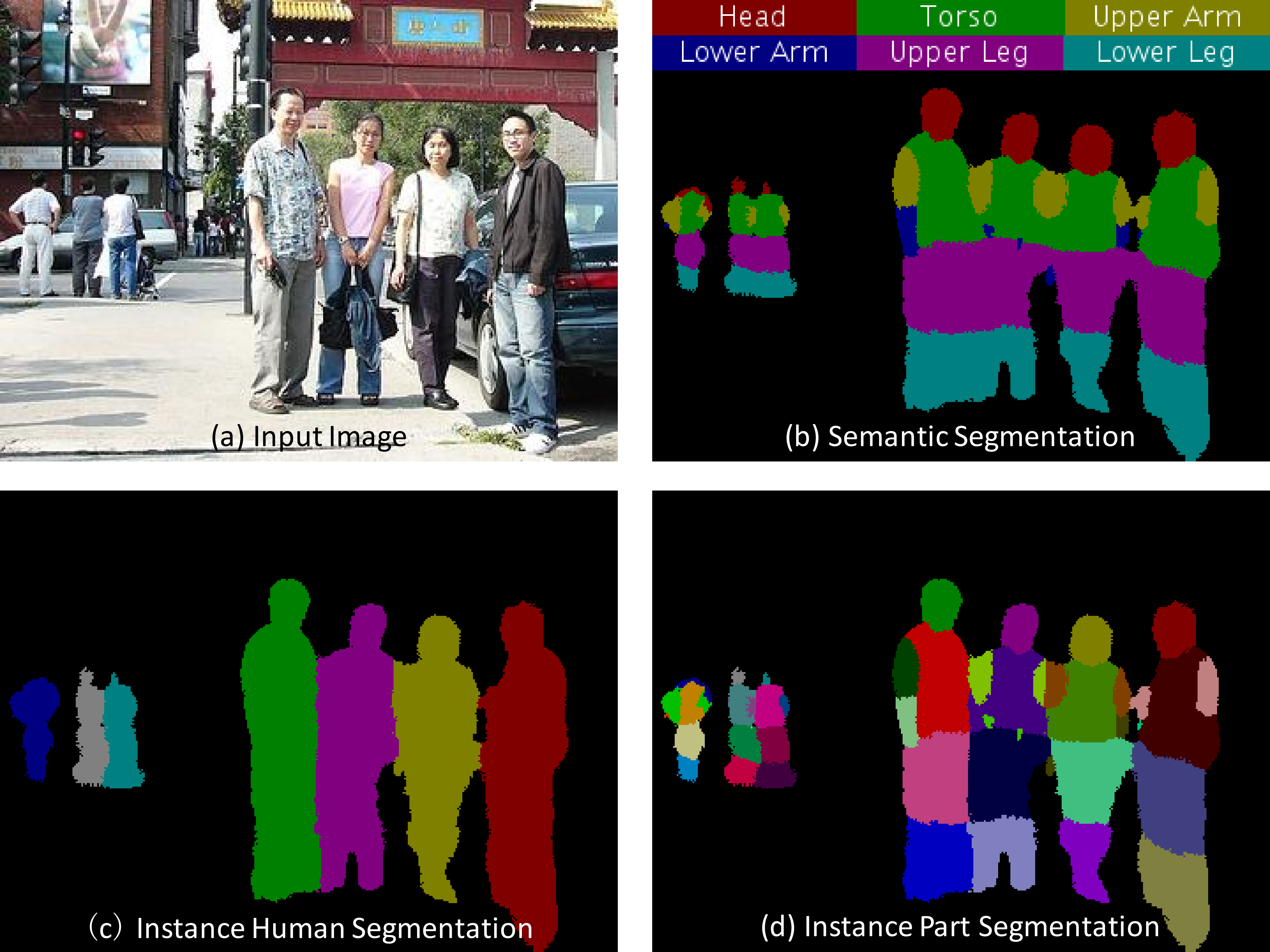
	Test IoU [%]	Memory [GB]	Time [s] (fps [s^{-1}])
Deeplab-v2	64.4	9.5	0.396 (2.5)
Deeplab-v2+CRF	64.9	11.2	0.960 (1.0)
Ours	65.9	4.3	0.255 (3.9)

Table 1. Comparison of our category-level segmentation module to Deeplab-v2. Tests done on the Pascal Person-Parts dataset. Memory and time requirements are for a single forward pass of the network.

Experiments

Training steps:

1. Pretrain the **semantic segmentation network** on all VOC 2012 train and SBD images minus VOC 2012 val and Pascal Person-Parts test to learn the VOC 21 classes.
2. Finetune the model on Pascal Person-Parts training set to predict the 7 body part classes (including the background class).
3. Train a **human detector** on VOC 07+12 trainval images minus VOC 2012 val and Pascal Person-Parts test images. We use the publicly available R-FCN framework.
4. Finetune the **full instance model** end-to-end.



(a) Input Image

(b) Semantic Segmentation

(c) Instance Human Segmentation

(d) Instance Part Segmentation