Supplementary material: A Sentence is Worth a Thousand Pixels

Sanja Fidler
TTI Chicago
fidler@ttic.edu

Abhishek Sharma
University of Maryland
bhokaal@cs.umd.edu

Raquel Urtasun
TTI Chicago
rurtasun@ttic.edu

In the supplementary material for paper [3] we provide more quantitative and qualitative results. Table 1 shows the results obtained with our approach when using GT cardinality potentials instead of text cardinality potentials. Note that GT cardinality only affects the potential on z and potential on the number of detection boxes. The remaining potentials (along with the score of the detector) are kept the same. "GT noneg" denotes the experiment, where we encourage at least as many boxes as dictated by the cardinality to be on in the image. With "GT - neg" we denote the experiment where we also **suppress** the boxes for classes with cardinality 0. This means that for images where GT cardinality for a class is 0, we simply do not input any boxes of that class into the model.

Interestingly, by using GT instead of extracted cardinalities from text, we do not observe a significant boost in performance. This means that our holistic model is able to (close to) fully exploit the available information about the scene.

Note, that conditioned on whether a noun is mentioned in text for a particular image, our approach uses additional boxes which originally did not pass the detector's threshold. In Table 2, we compare our approach with that of Yao et al. [5], where we use as many boxes as used in our approach. To do this, we reduce the thresholds of the detector, so that for each class the number of boxes exactly matches the number of boxes that our approach uses for the class. Notice that the original approach improved its performance by only 0.4%, while our performance is much higher (12.5% improvement). This means that the success of our approach is not due to an increased number of used boxes, but in how they are used in the model by exploiting additional text information.

In the future we plan to incorporate more powerful segmentation features [1] and detector [2].

	back.	aerop.	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	monitor	**averg.**
Oracle Z - noneg	76.8	36.7	28.3	34.8	21.9	30.9	56.1	47.6	36.8	10.0	58.2	28.8	33.4	54.8	42.6	41.8	15.1	28.1	16.3	35.7	48.7	37.3
Oracle Z - neg	76.8	31.2	28.2	34.7	21.7	31.1	56.0	50.3	36.7	10.5	57.4	29.5	34.4	55.1	42.5	41.9	16.9	32.2	18.8	35.7	49.2	37.7
ours	76.9	31.3	29.7	37.3	27.7	29.5	52.1	40.0	38.0	6.6	55.9	25.2	33.2	38.2	44.3	42.5	15.2	32.0	20.2	40.7	48.4	36.4

Table 1. Comparison to oracle Z (see text for details).

	back.	aerop.	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	monitor	**averg.**
Textonboost (unary) [4]	77.8	14.1	3.4	0.7	11.3	3.3	25.5	30.9	10.3	0.7	13.2	10.8	5.2	15.1	31.8	41.0	0.0	3.7	2.4	17.1	33.7	16.8
Holistic Scene Understanding [5]	77.3	25.6	12.9	14.2	19.2	31.0	34.6	38.6	16.1	7.4	11.9	9.0	13.9	25.4	31.7	38.1	11.2	18.8	6.2	23.6	34.4	23.9
[5] num boxes from text	77.8	26.7	14.3	11.5	18.6	30.8	34.4	37.9	17.2	5.7	19.0	7.3	12.4	27.3	36.5	37.1	11.6	9.4	6.2	25.7	43.8	24.3
ours	76.9	31.3	29.7	37.3	27.7	29.5	52.1	40.0	38.0	6.6	55.9	25.2	33.2	38.2	44.3	42.5	15.2	32.0	20.2	40.7	48.4	36.4

Table 2. Comparison to the state-of-the-art that utilizes only image information in the UIUC sentence dataset. By leveraging text information our approach improves 12.5% AP. Note that this dataset contains only 600 PASCAL VOC 2008 images for training, and thus is significantly a more difficult task than recent VOC challenges which have up to 10K training images.

References

[1] J. Carreira, R. Caseiroa, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 1

[2] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In *CVPR*, 2013. 1

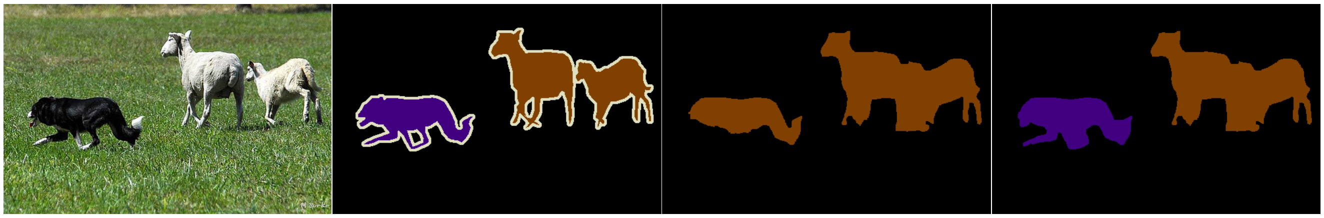
[3] S. Fidler, A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. In *CVPR*, 2013. 1

[4] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 1

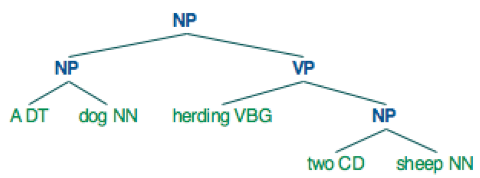
[5] Y. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

In Figure 1- 8 we show some examples of successful results, while Figure 9- 11 shows (partial) failures.

(a) or. image (b) GT (c) Yao et al. [5] **our approach**

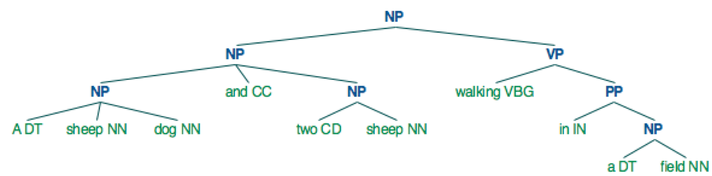


sent 1: A dog herding two sheep.

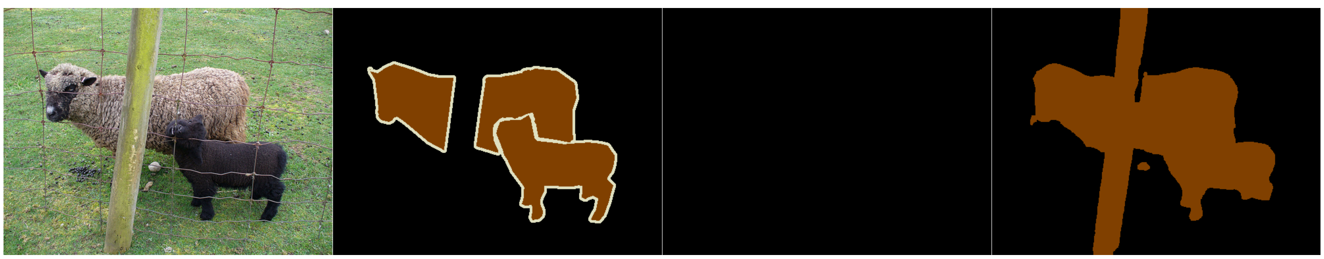


CARD	**dog**	1
CARD	**sheep**	2
PREP	**none**	

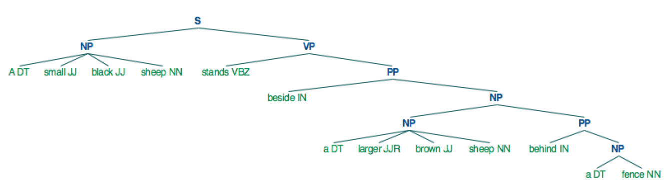
sent 2: A sheep dog and two sheep walking in a field.



CARD	**dog**	1
CARD	**sheep**	2
PREP	**none**	

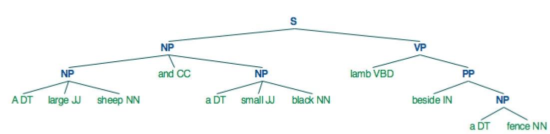


sent 1: A small black sheep stands beside a larger brown sheep behind a fence.



CARD	**sheep**	2
PREP	**beside**	(sheep, beside, sheep)

sent 2: A large sheep and a small black lamb beside a fence.



CARD	**sheep**	2
PREP	**none**	

Figure 1. Examples of results. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.

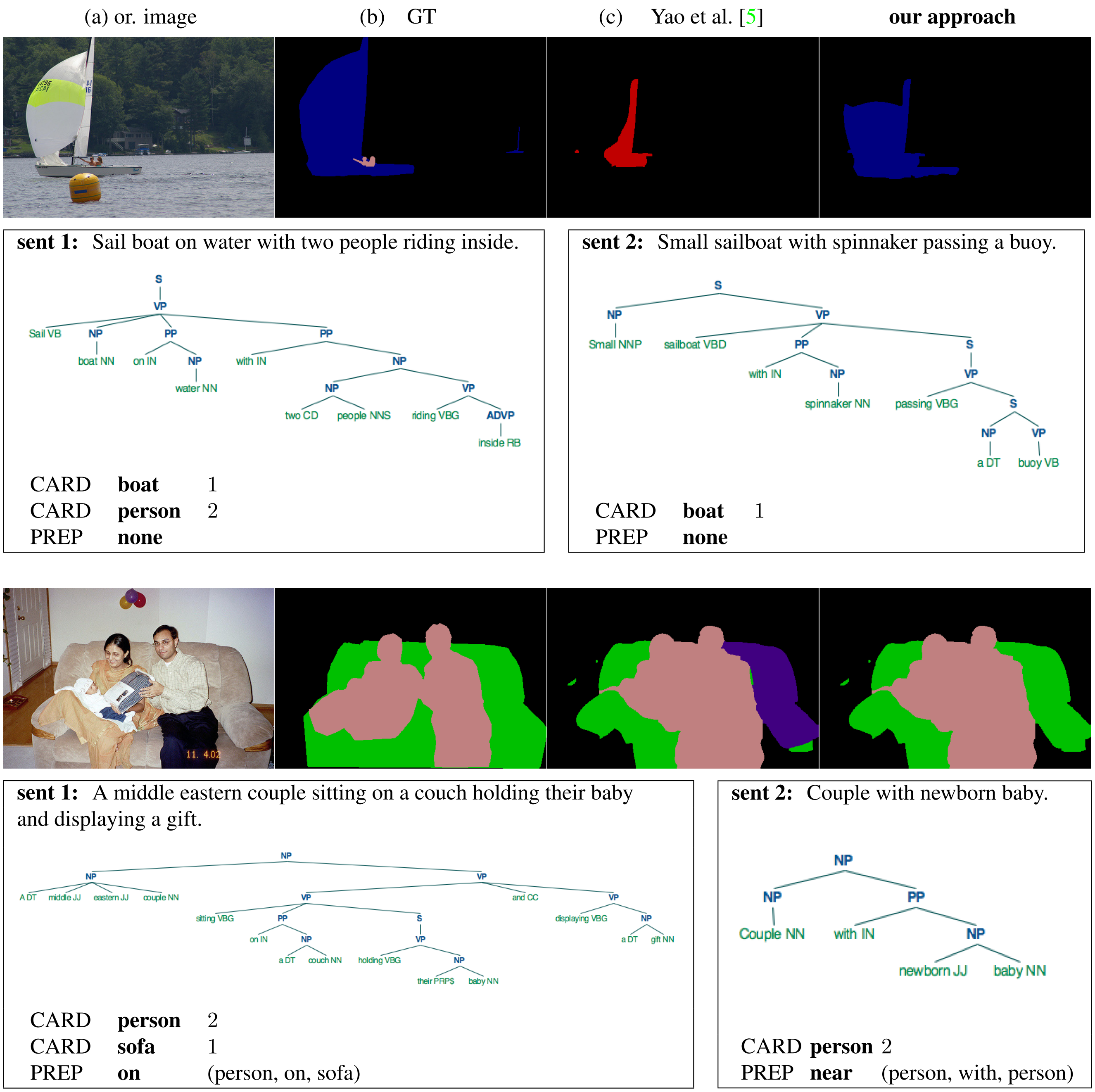


Figure 2. Examples of results. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.

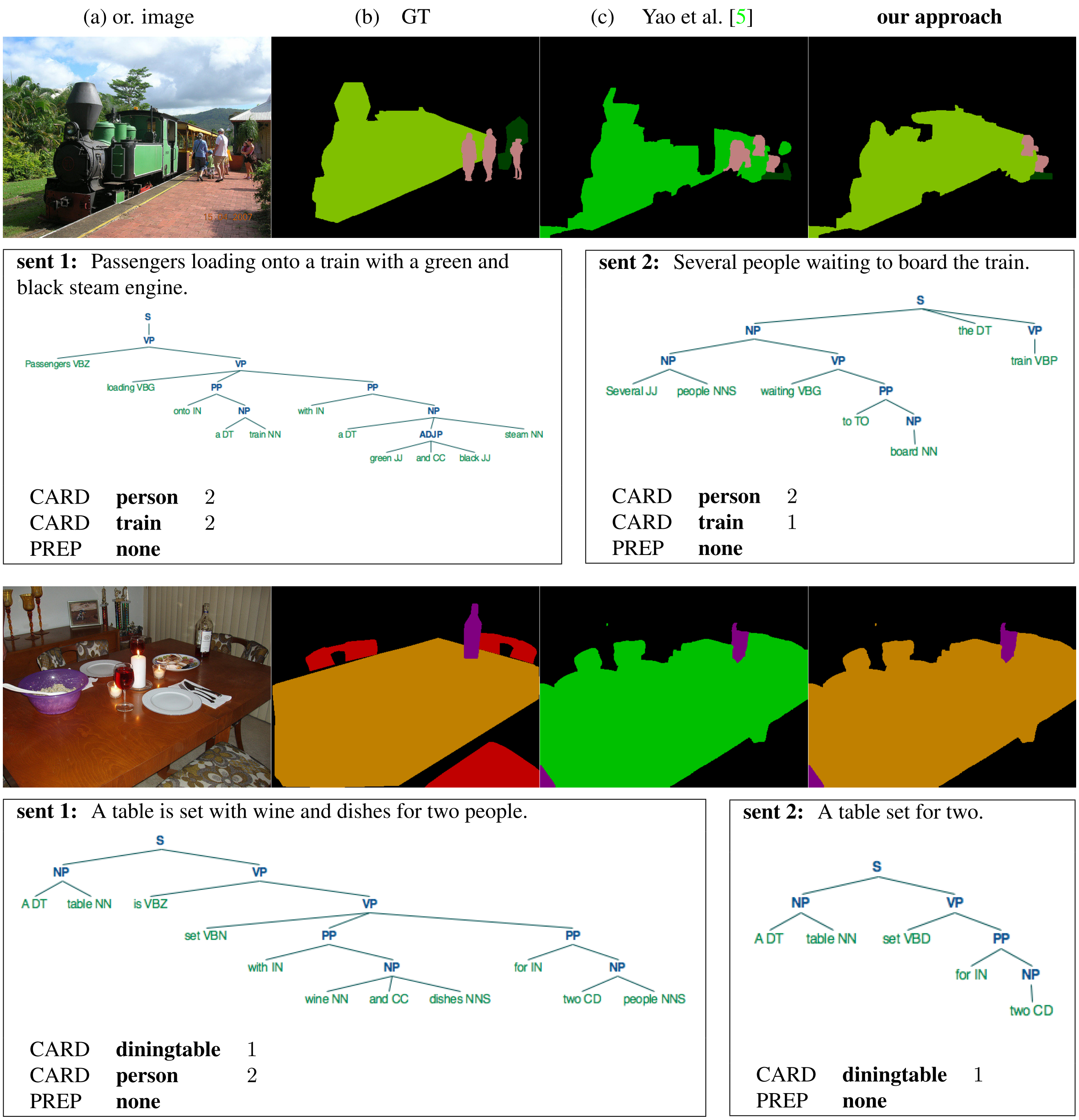


Figure 3. Examples of results. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.

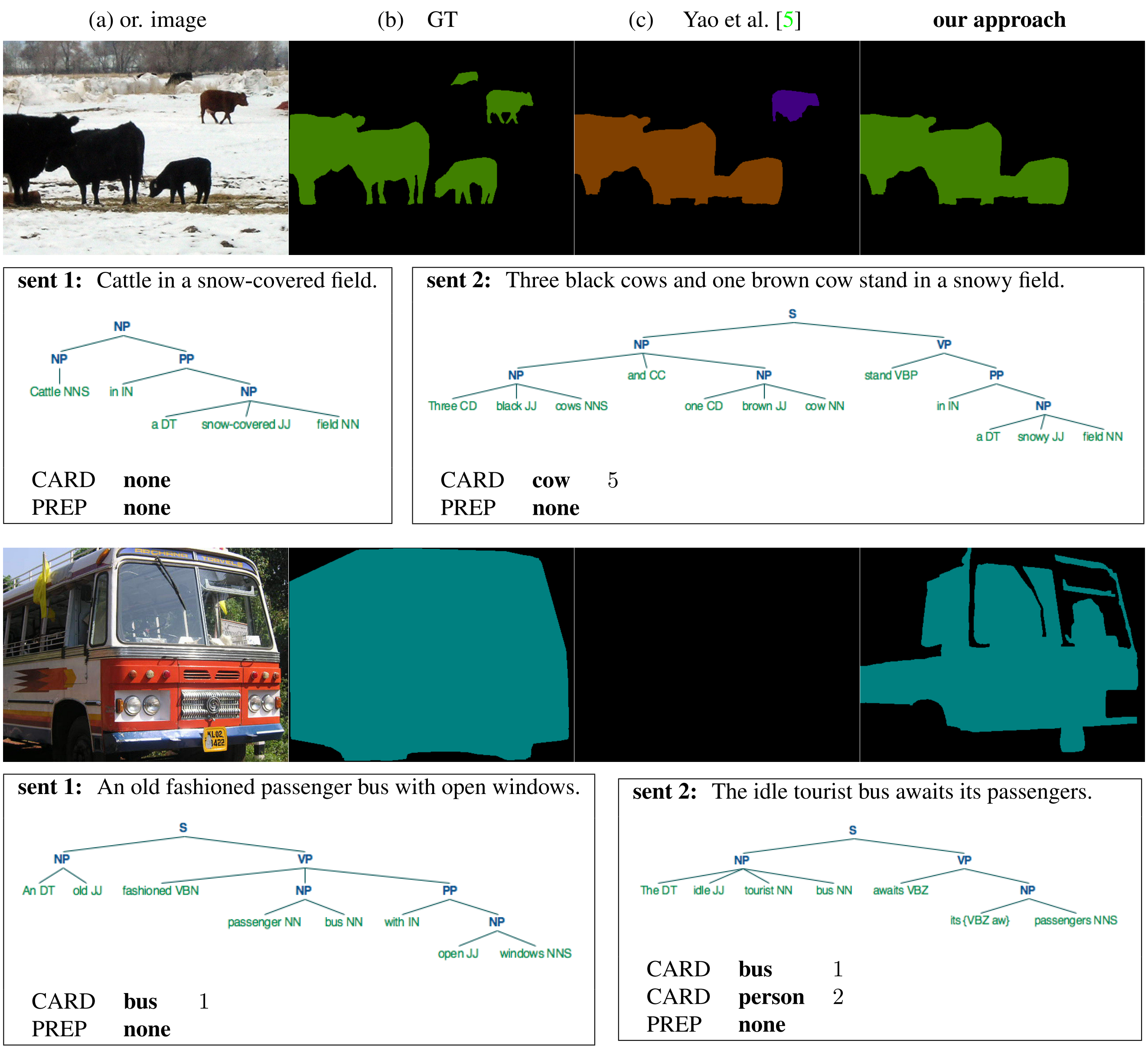


Figure 4. Examples of results. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.

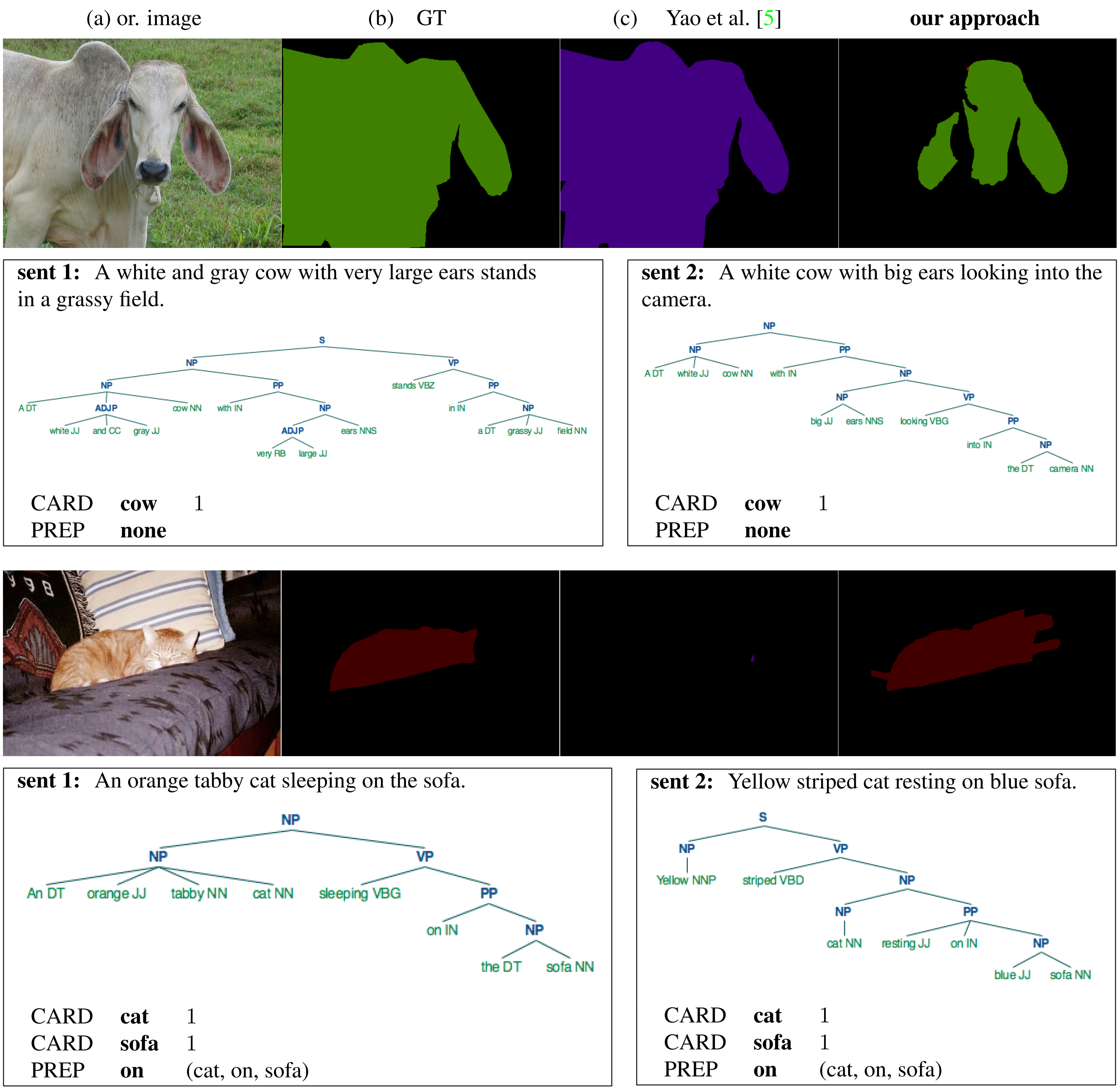


Figure 5. Examples of results. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.

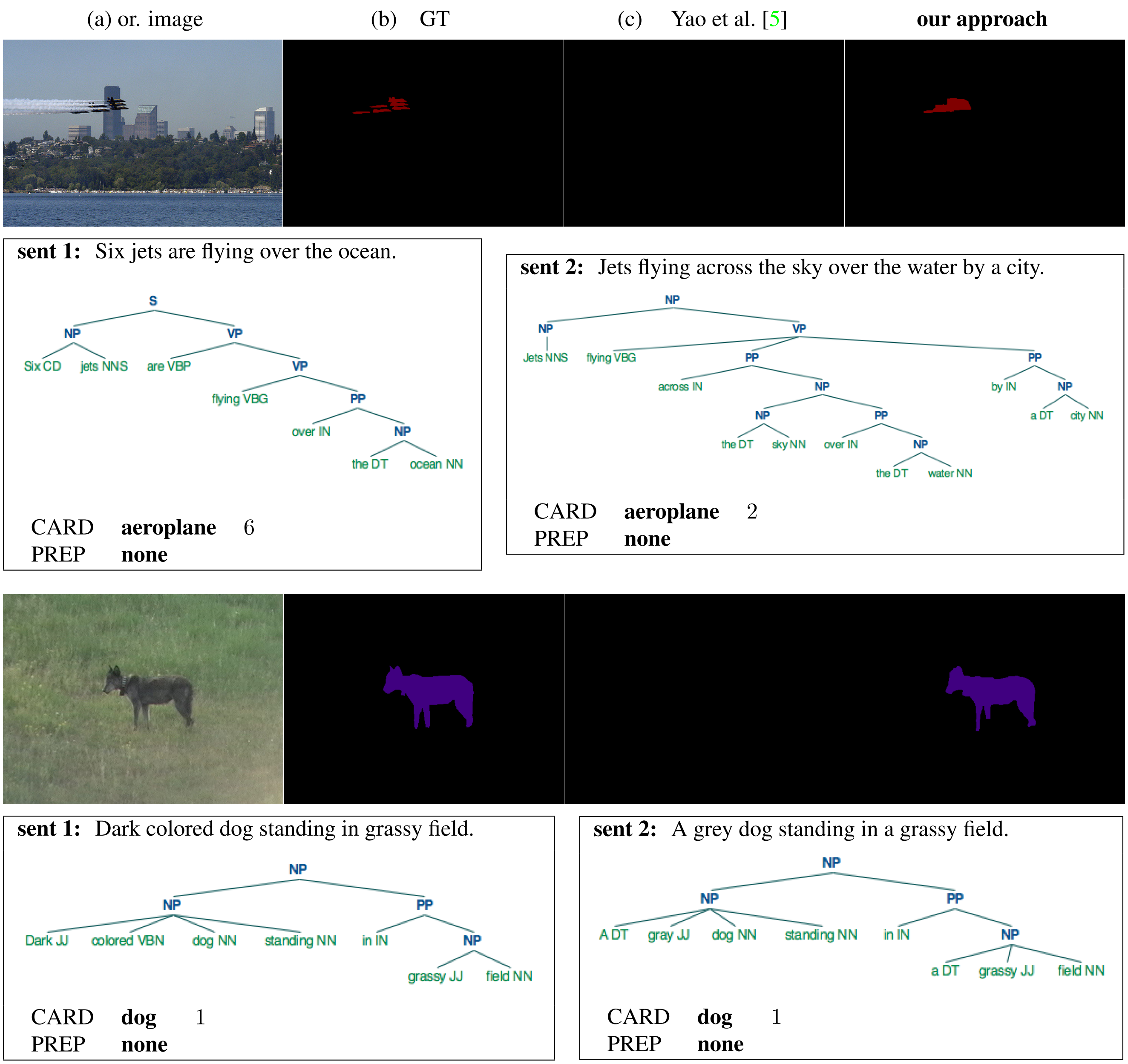
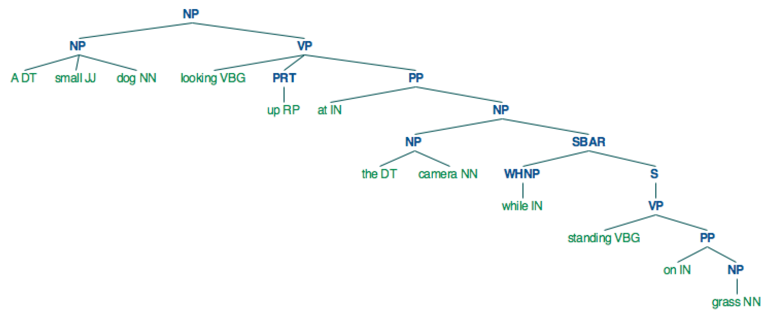


Figure 6. Examples of results. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.

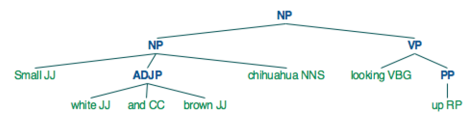
(a) or. image	(b) GT	(c) Yao et al. [5]	**our approach**

sent 1: A small dog looking up at the camera while standing on grass.

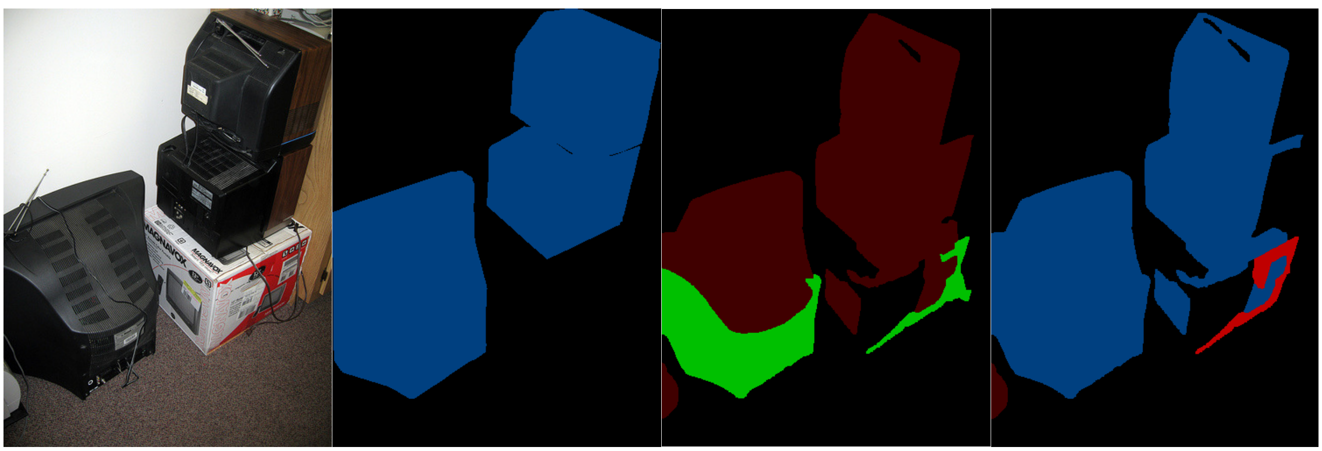


CARD **dog** 1
PREP **none**

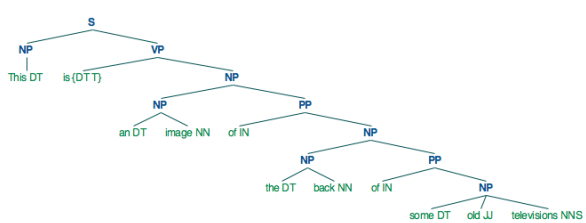
sent 2: Small white and brown chihuahua looking up.



CARD **dog** 1
PREP **none**

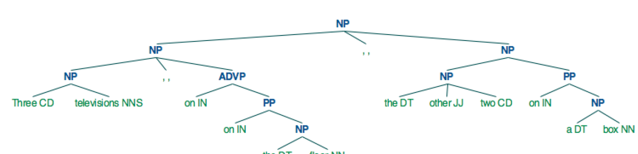


sent 1: This is an image of the back of some old televisions.



CARD **tvmonitor** 2
PREP **none**

sent 2: Three televisions, on on the floor, the other two on a box.



CARD **tvmonitor** 3
PREP **none**

Figure 7. Examples of results. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.

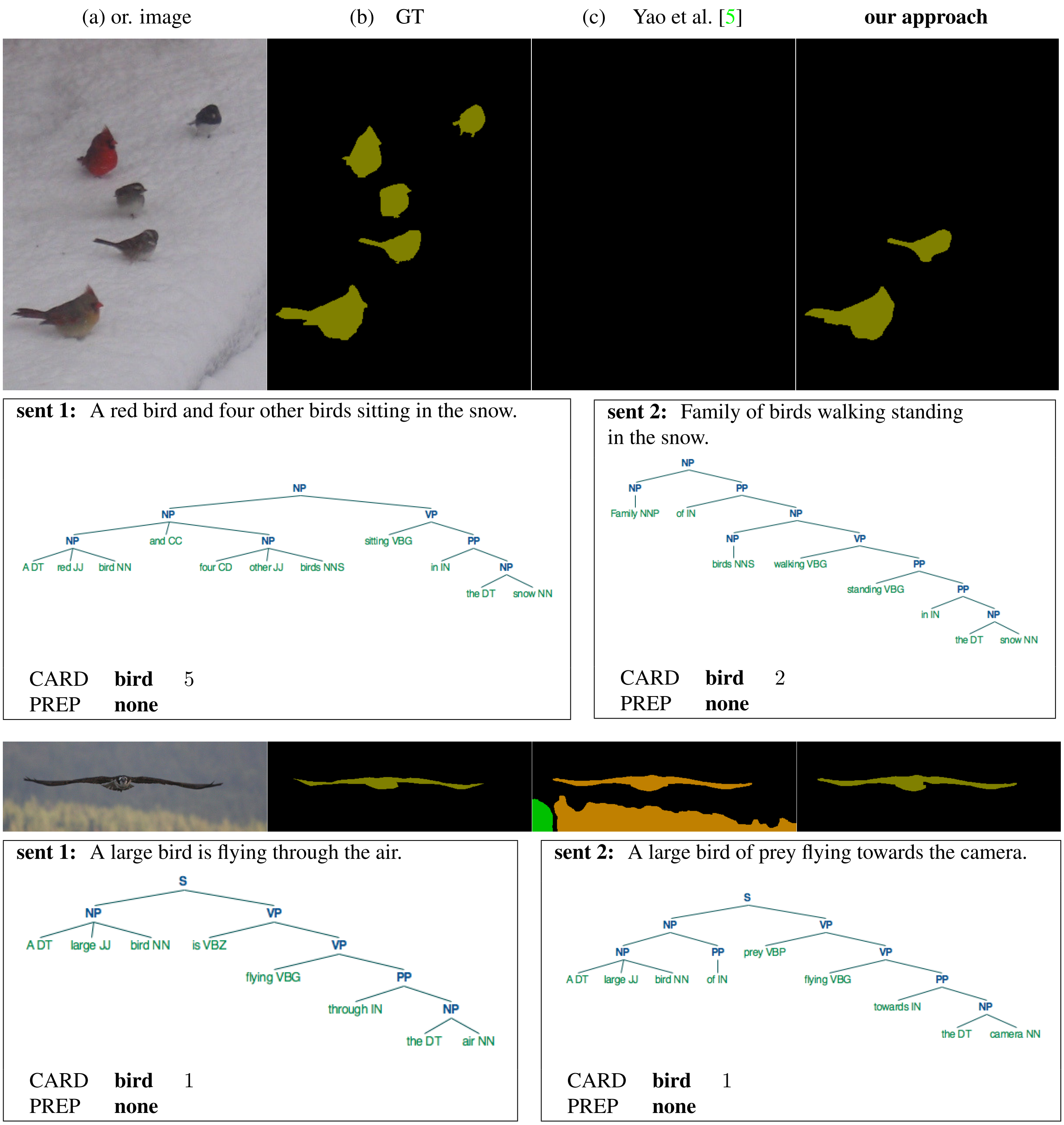


Figure 8. Examples of results. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.

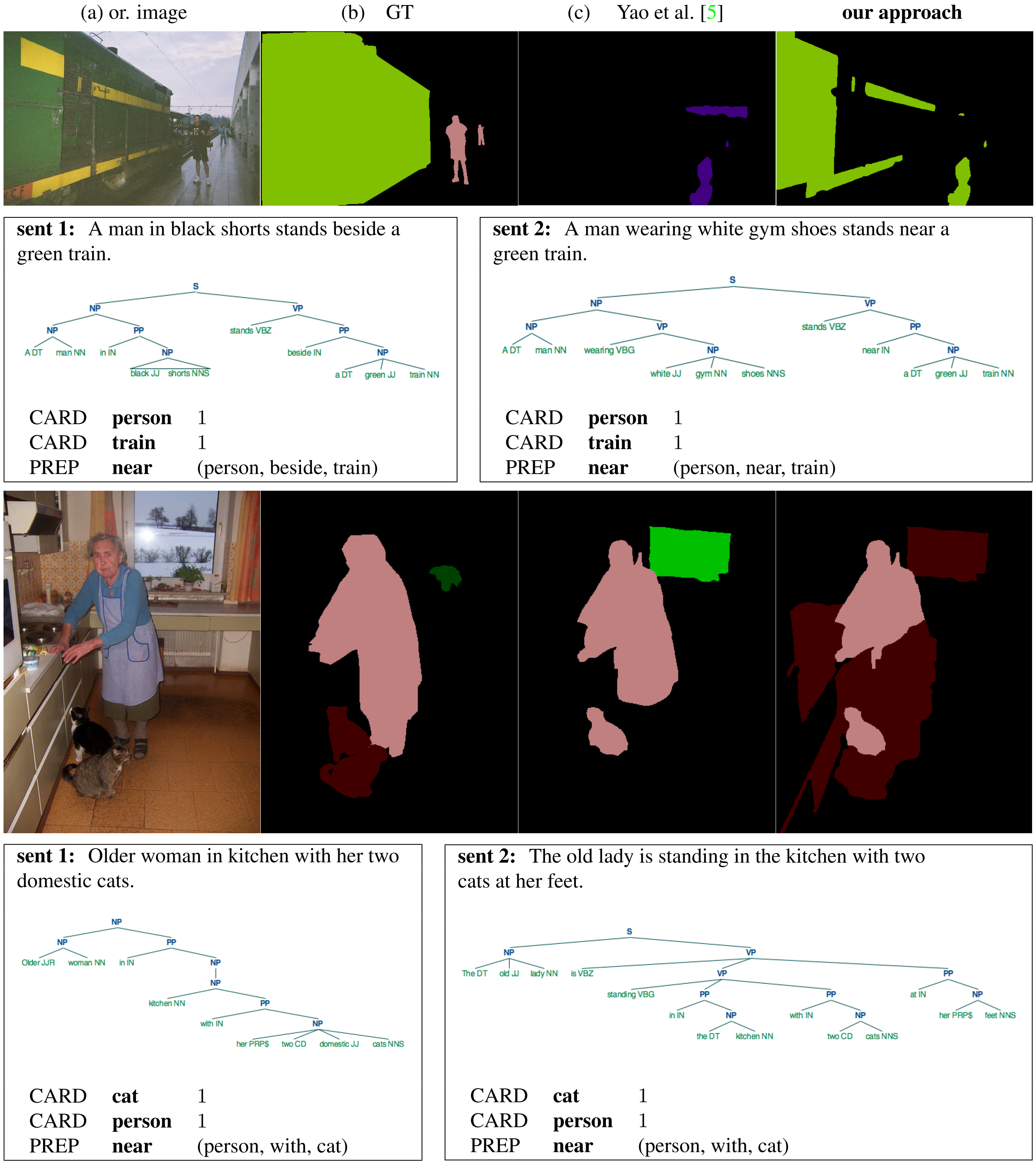
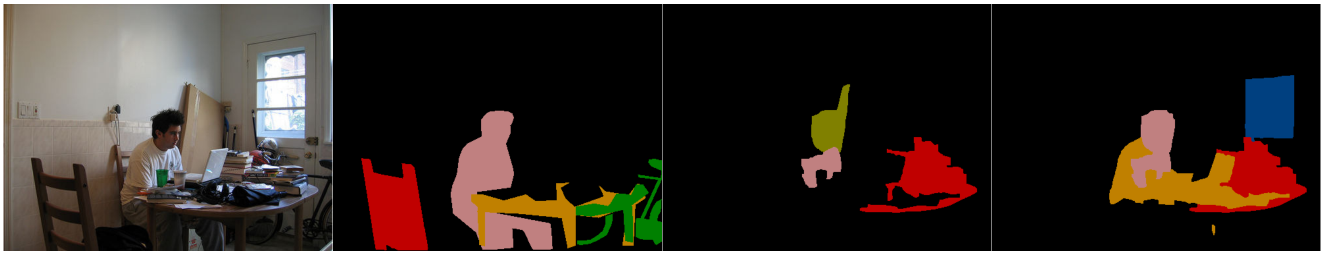
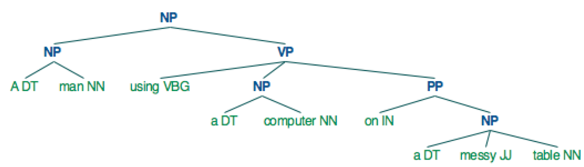


Figure 9. Examples of (partial) failures. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.

(a) or. image (b) GT (c) Yao et al. [5] **our approach**

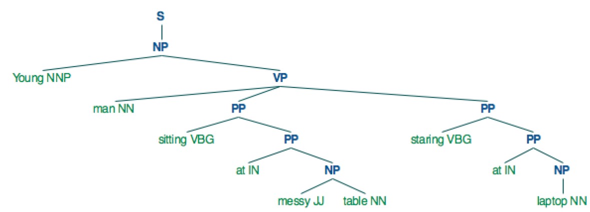


sent 1: A man using a computer on a messy table.

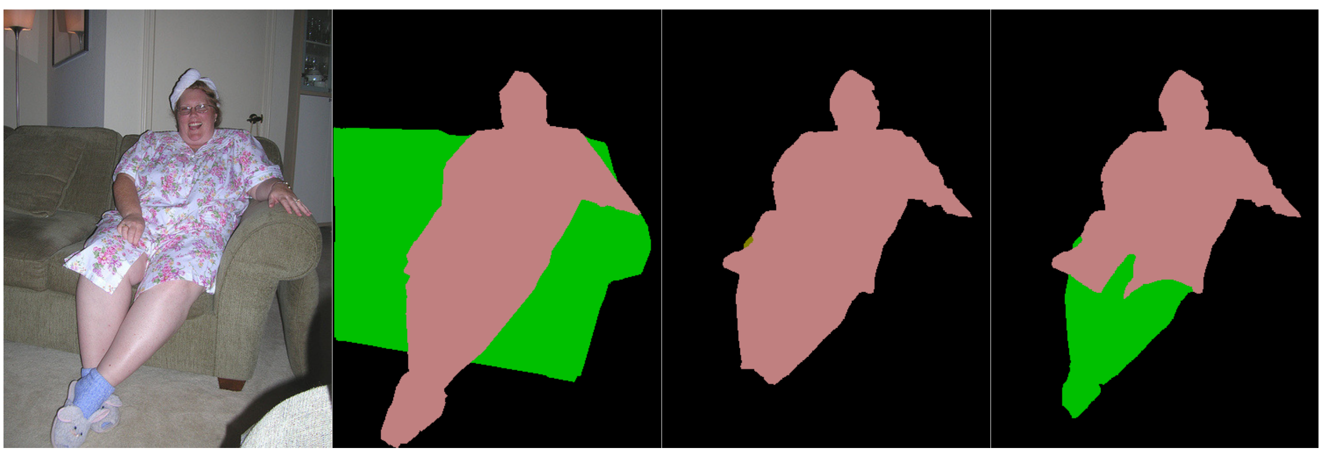


CARD	**diningtable**	1
CARD	**person**	1
CARD	**tvmonitor**	1
PREP	**on**	(person, on, diningtable)

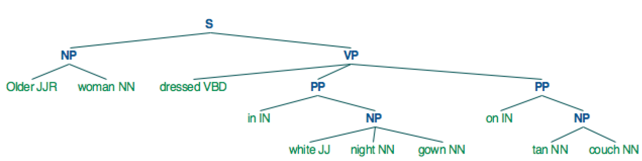
sent 2: Young man sitting at messy table staring at laptop.



CARD	**diningtable**	1
CARD	**person**	1
CARD	**tvmonitor**	1
PREP	**none**	

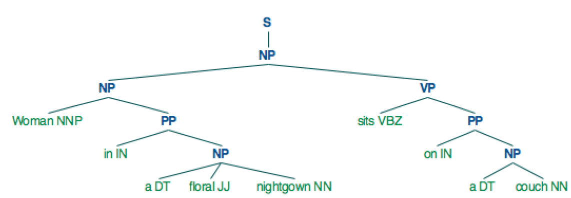


sent 1: Older woman dressed in white night gown on tan couch.



CARD	**person**	1
CARD	**sofa**	1
PREP	**on**	(person, on, sofa)

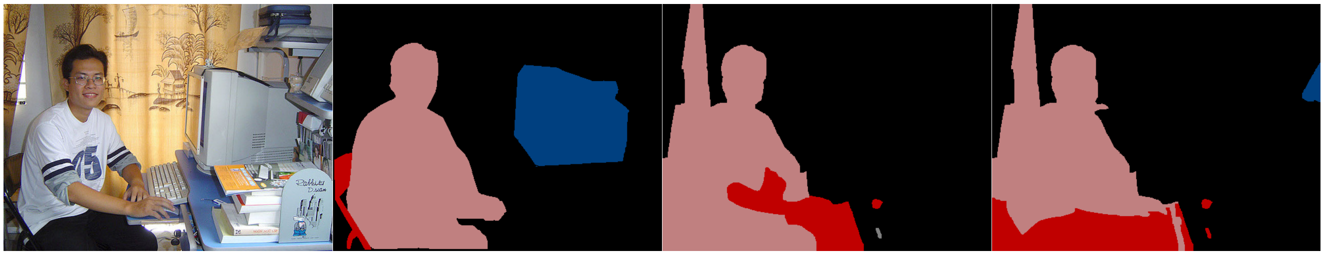
sent 2: Woman in a floral nightgown sits on a couch.



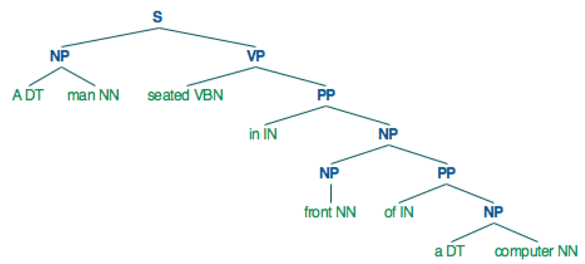
CARD	**person**	1
CARD	**sofa**	1
PREP	**on**	(person, on, sofa)

Figure 10. Examples of (partial) failures. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.

(a) or. image (b) GT (c) Yao et al. [5] **our approach**

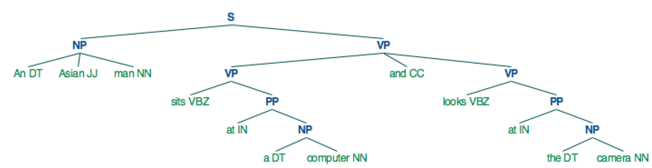


sent 1: A man seated in front of a computer.

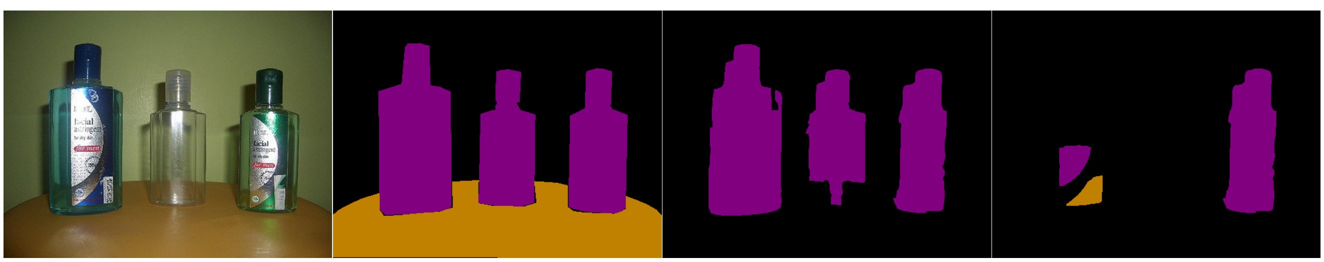


CARD	**person**	1
CARD	**tvmonitor**	1
PREP	**in-front-of**	(person, in-front-of, tvmonitor)

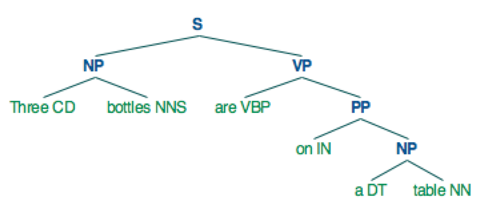
sent 2: A man wearing a sports jersey poses at a computer in front of curtains.



CARD	**person**	1
CARD	**tvmonitor**	1
PREP	**none**	

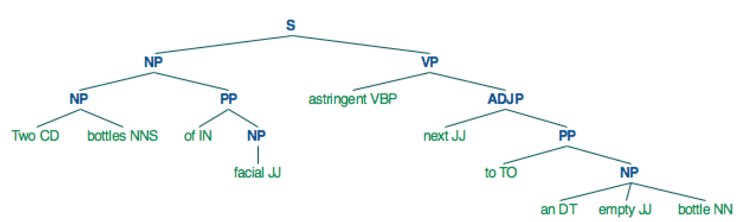


sent 1: Three bottles are on a table.



CARD	**bottle**	3
CARD	**diningtable**	1
PREP	**on**	(bottle, on, diningtable)

sent 2: Two bottles of facial astringent next to an empty bottle.



CARD	**bottle**	3
PREP	**near**	(bottle, next-to, bottle)

Figure 11. Examples of (partial) failures. We show the original image, GT segmentation, the output of [5], and our result. Below the images we show the exemplar sentences with parse tree, with extracted object classes, cardinalities and prepositions.