# Occlusion-Ordered Semantic Instance Segmentation

9/1/26

Soroosh Baselizadeh, Cheuk-To Yu, Olga Veksler, Yuri Boykov
David R. Cheriton School of Computer Science
University of Waterloo

# Outline

- **Introduction**
- Related Work
- Our Approach
- Experiments & Results
- Conclusion

# Introduction

- Single image scene understanding
  - Object classes? / Object locations? / Object spatial relations?

# Introduction

## Semantic Segmentation


- Assign class to each image pixel

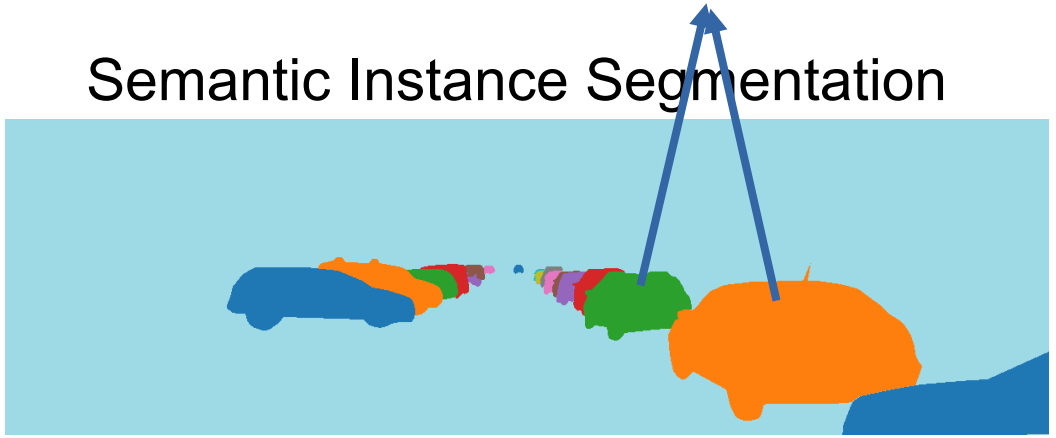
## Semantic ‘Instance’ Segmentation

- Mask + class for each object

# Introduction

Input Image

Semantic Instance Segmentation

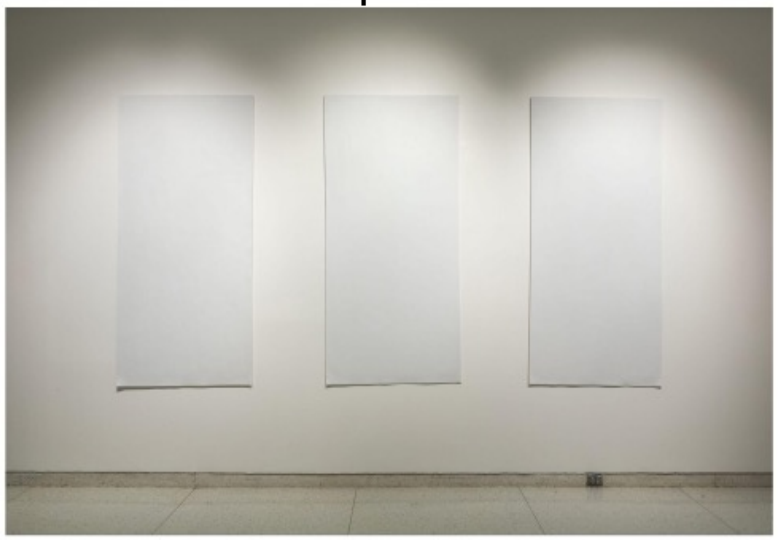



- Instance segmentation gives more scene details
- However, no depth information
- Depth info is important, e.g. in autonomous driving, scene editing

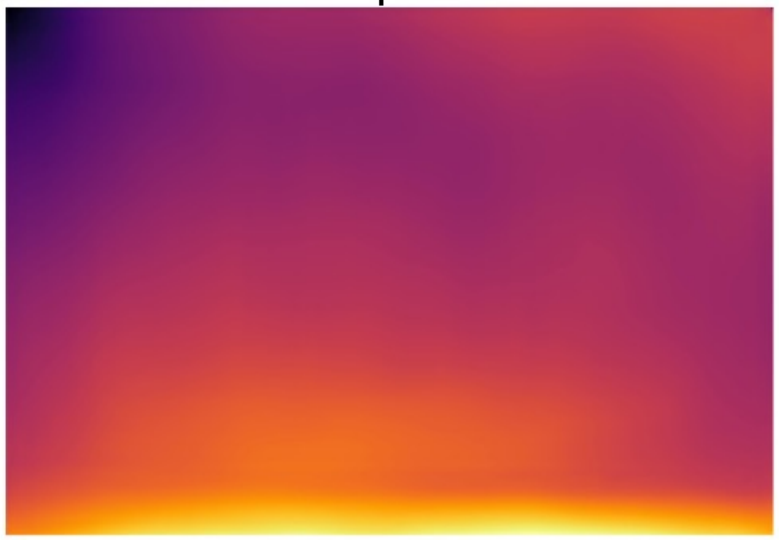
# Introduction

- Monocular Depth Estimation [15] predicts the absolute depth of pixels
  - Difficult problem, hard to estimate accurately
  - Lacks semantic information
  - Estimates may be too coarse to establish relative depth relations

Input



Monocular Depth Estimation

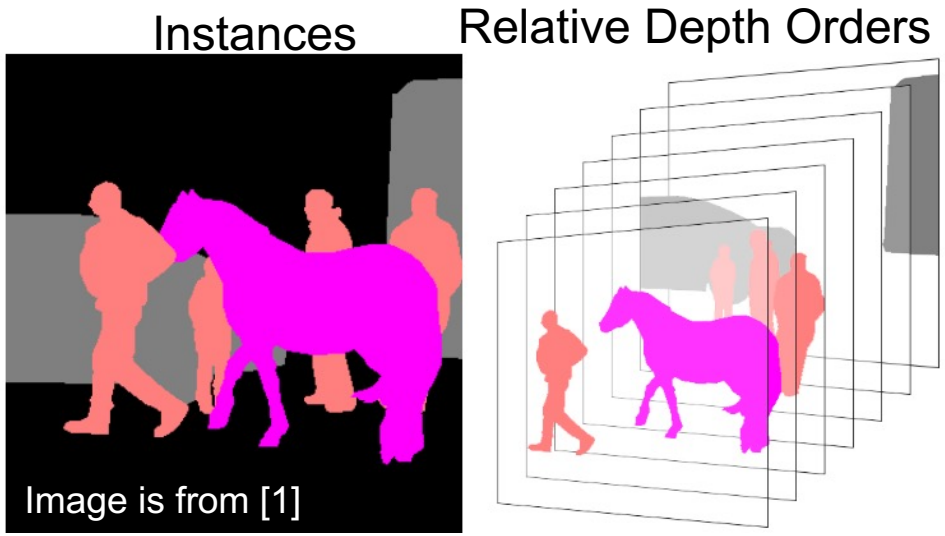


# Introduction

- Absolute depth is unnecessary for answering many questions


- *Relative* depth ordering can answer many questions

# Introduction




- Relative depth order → coarse 3D geometry useful in applications
  - Scene/ video editing [43]
  - Image retrieval [24]
  - Question answering [16]

# Introduction

- Unlike absolute depth, relative depth estimation is easy for humans [17]
- Humans rely on *occlusions* for relative depth [18, 19, 20]


5 occludes 4

...

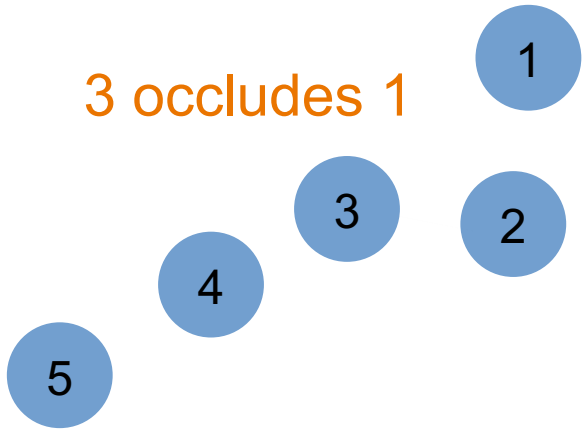
5 is closer than 4

...

# Introduction

Input Image


Occlusion Ordering Graph



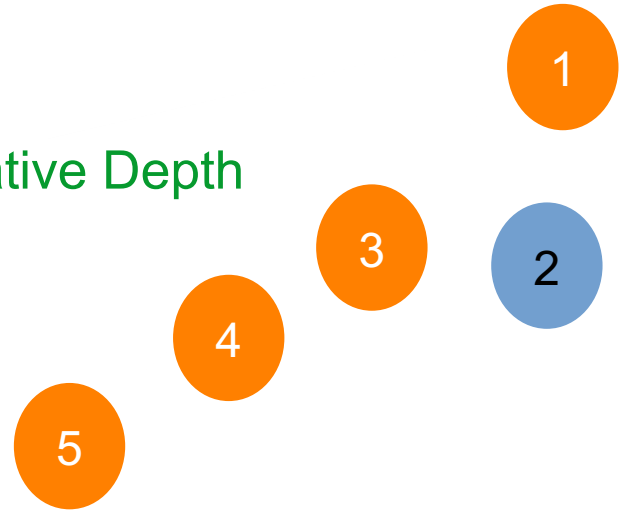

- Recovering relative depth from occlusion relations

# Introduction

- Occlusion ordering gives relative depth of objects in a ‘chain’
  - **Not** just for adjacent instances!

Input Image


Occlusion Ordering Graph



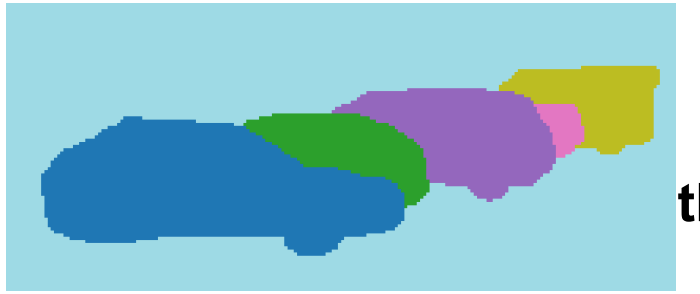
# Introduction

- We address

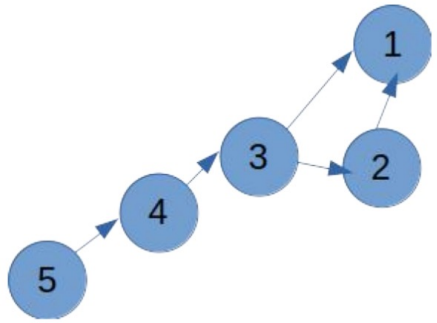
**Occlusion-Ordered Semantic Instance Segmentation (OOSIS)**

Input Image


Output



**Instance masks + their semantic classes**




**Occlusion odering of instances**

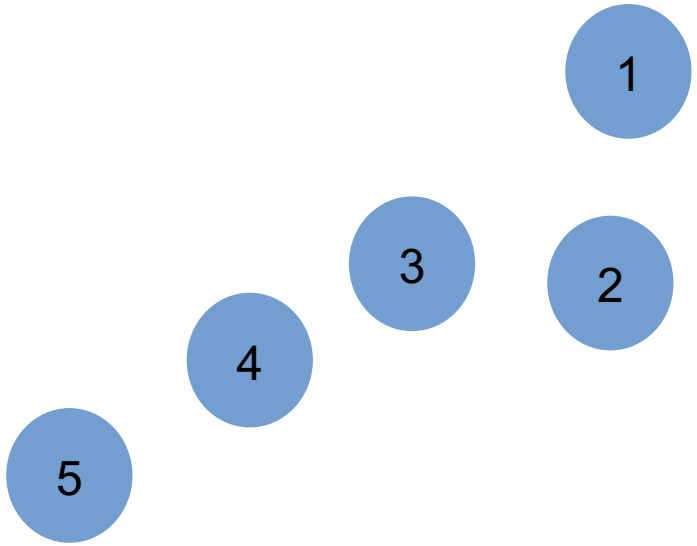
# Introduction

## Major Contributions:

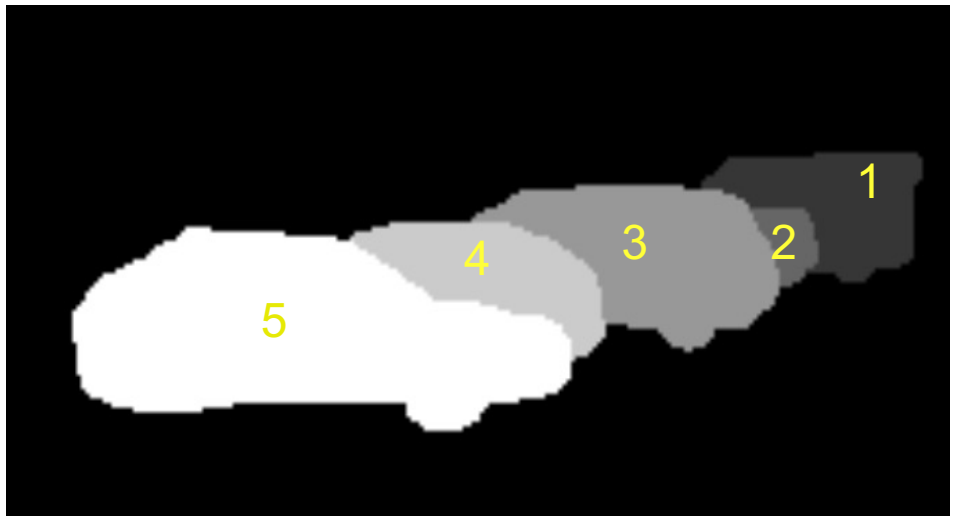
- First deep learning approach for OOSIS
- Novel state-of-the-art oriented occlusion boundary model
- Novel discrete optimization formulation for simultaneous instance segmentation and occlusion ordering
- New metric for OOSIS evaluation assessing instance segmentation and occlusion ordering simultaneously

# Visualization of Occlusion Ordering

Visualization by Graph




Visualization by Relative Depth Map




‘occluder’ is brighter
than its occludees

# Outline

- Introduction
- **Related Work**
- Our Approach
- Experiments & Results
- Conclusion

# Related Work

- Semantic Instance Segmentation
  - Lots of works
  - Two stages: *Detect Candidate Regions → Segment & Classify*
  - We use **mask-based**: *Mask-RCNN [13] and* **contour-based:** *E2EC [14]*

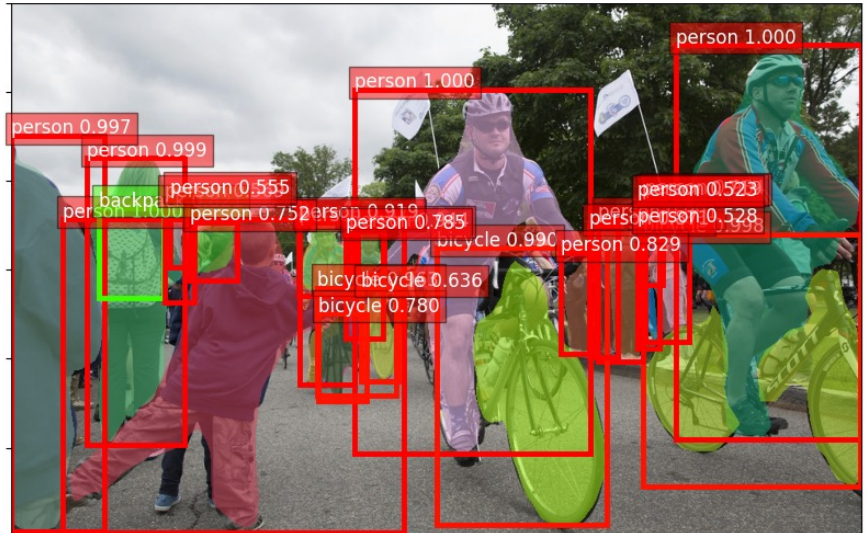
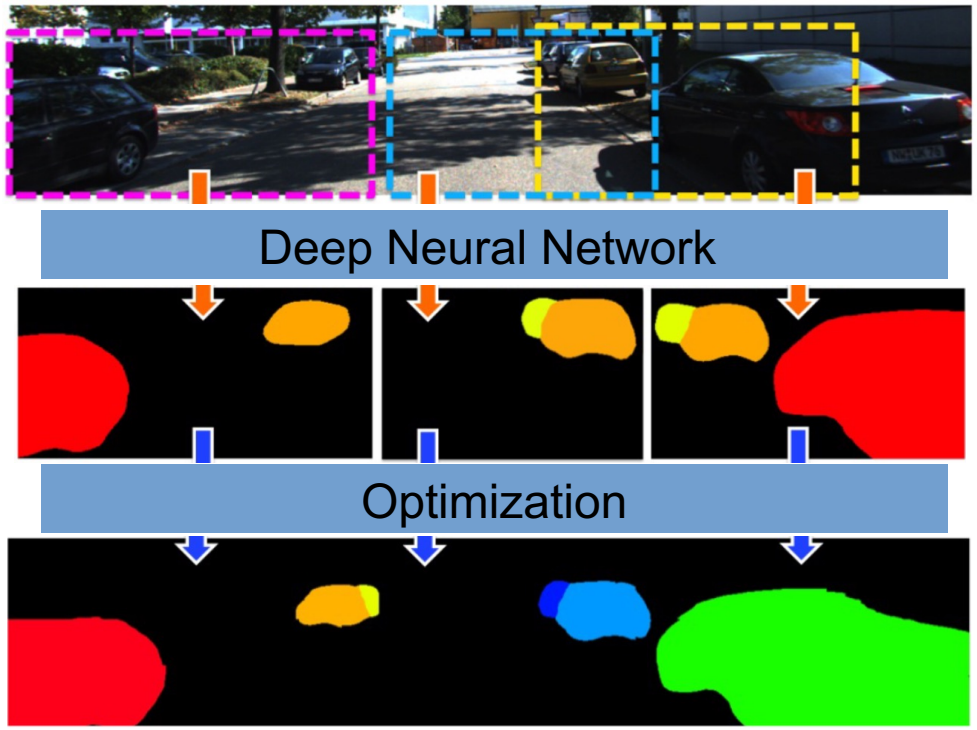



Image from: Gluon.ai

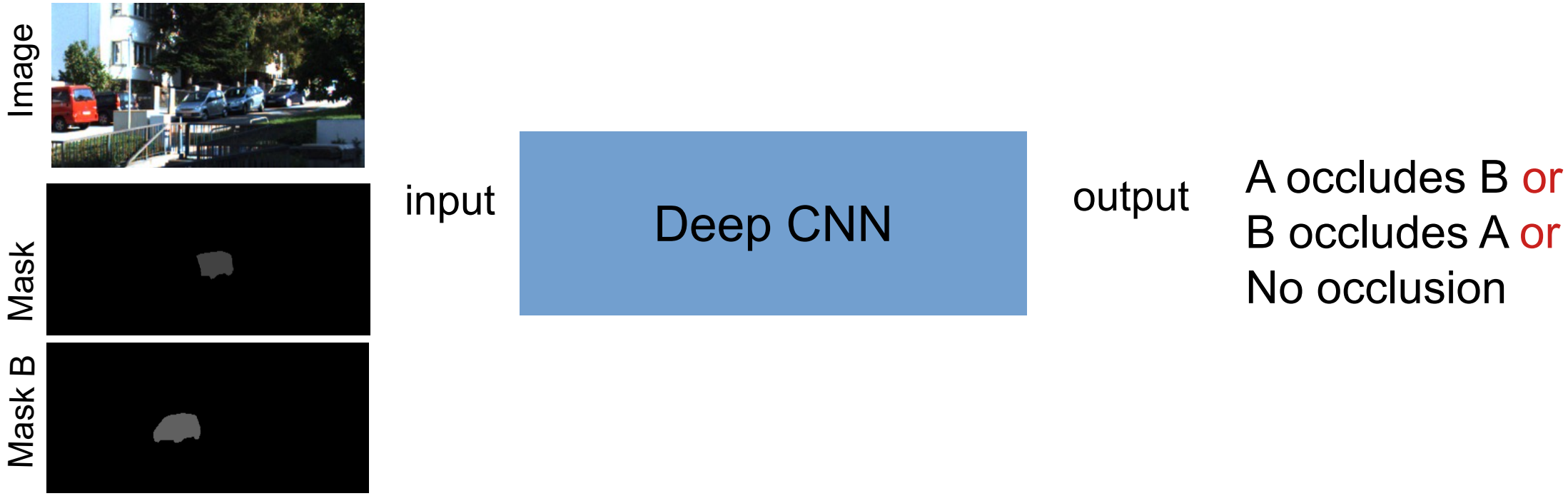
# Related Work

- 3D-Augmented Instance Segmentation
  - **Pre-deep learning**: [1, 2, 3]:
    - Simple heuristics,
      - e.g. *size* of objects
    - Weak performance
  - **Using deep learning**: only [4].
    - Depth ordering not occlusion
    - Only for ‘cars’
    - Simple cues



# Related Work

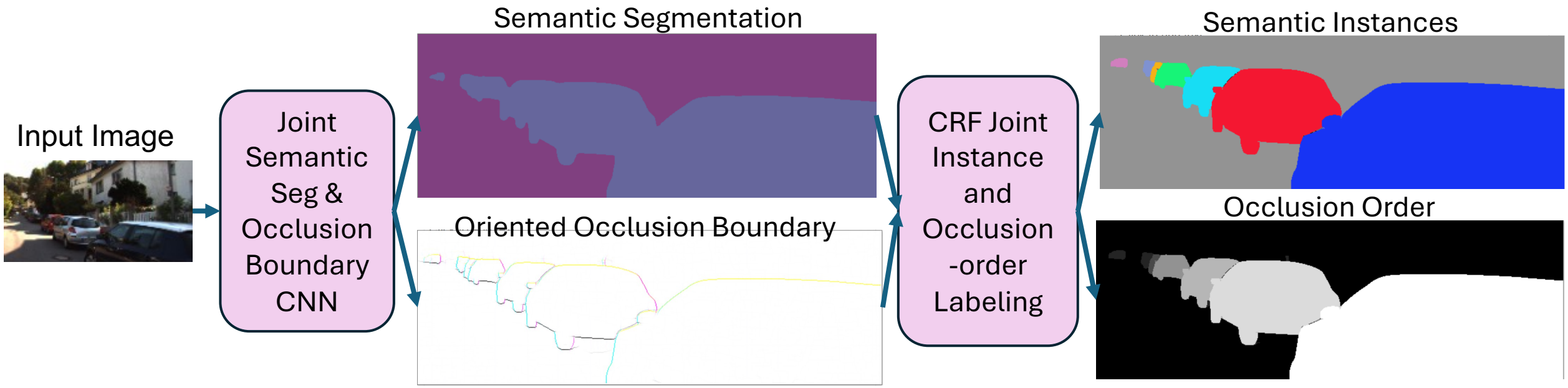
- ‘Pairwise’ Occlusion Ordering
  - *OrderNet* [5] and *InstaOrder* [6] are CNN classifiers
  - **No** instance mask generation



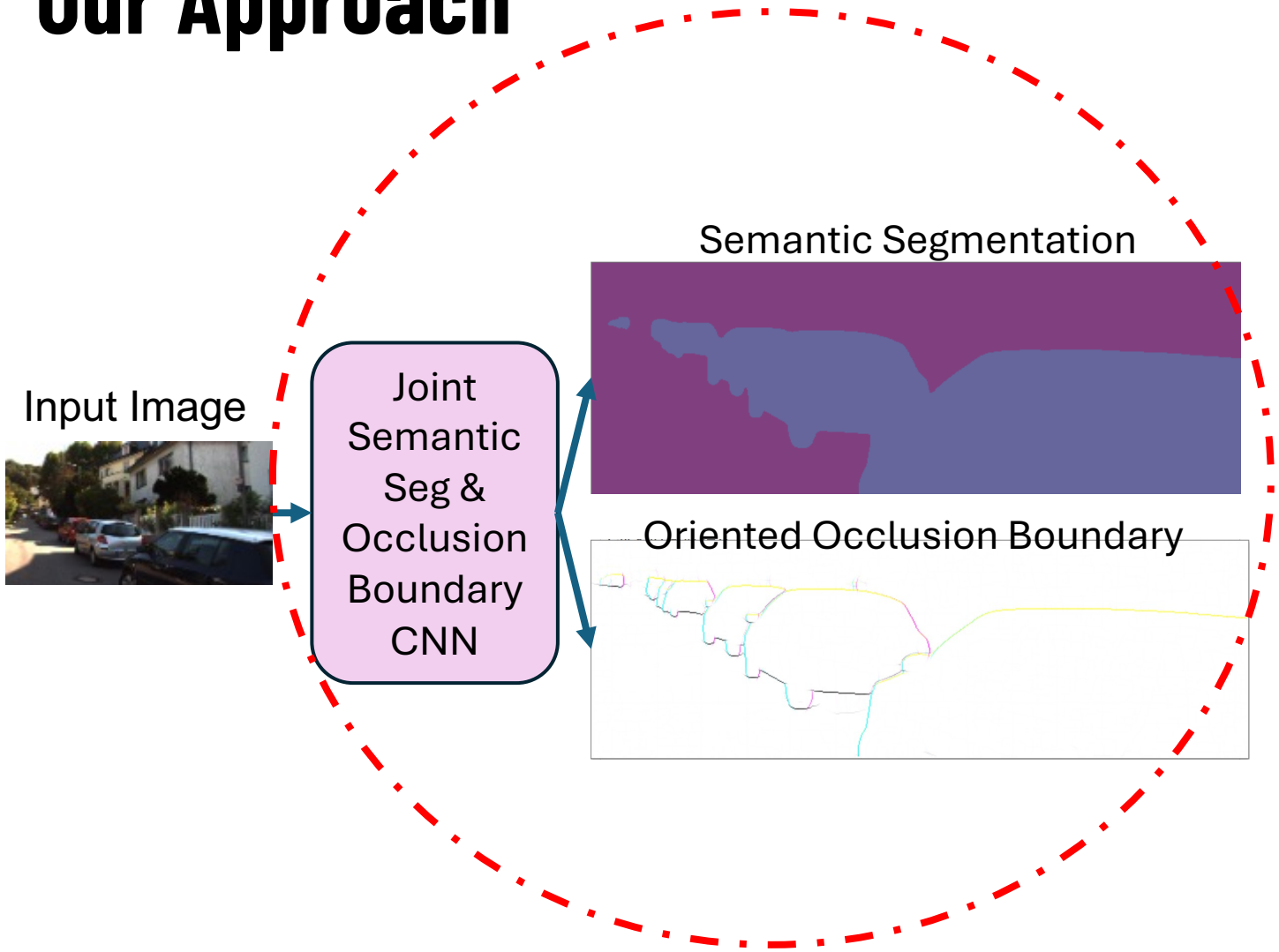
# Outline

- Introduction
- Related Work
- **Our Approach**
- Experiments & Results
- Conclusion

# Our Approach



# Our Approach



# Orientation at Occlusion Boundary

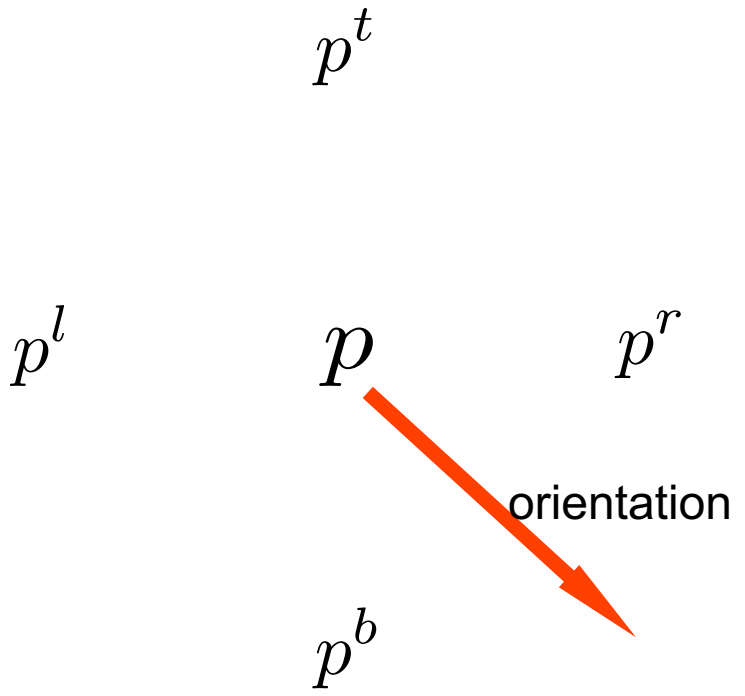
Normal vector's orientation at the boundary

- from 'occluder' to 'occludee'

# Orientation at Occlusion Boundary

- Representation at pixel level

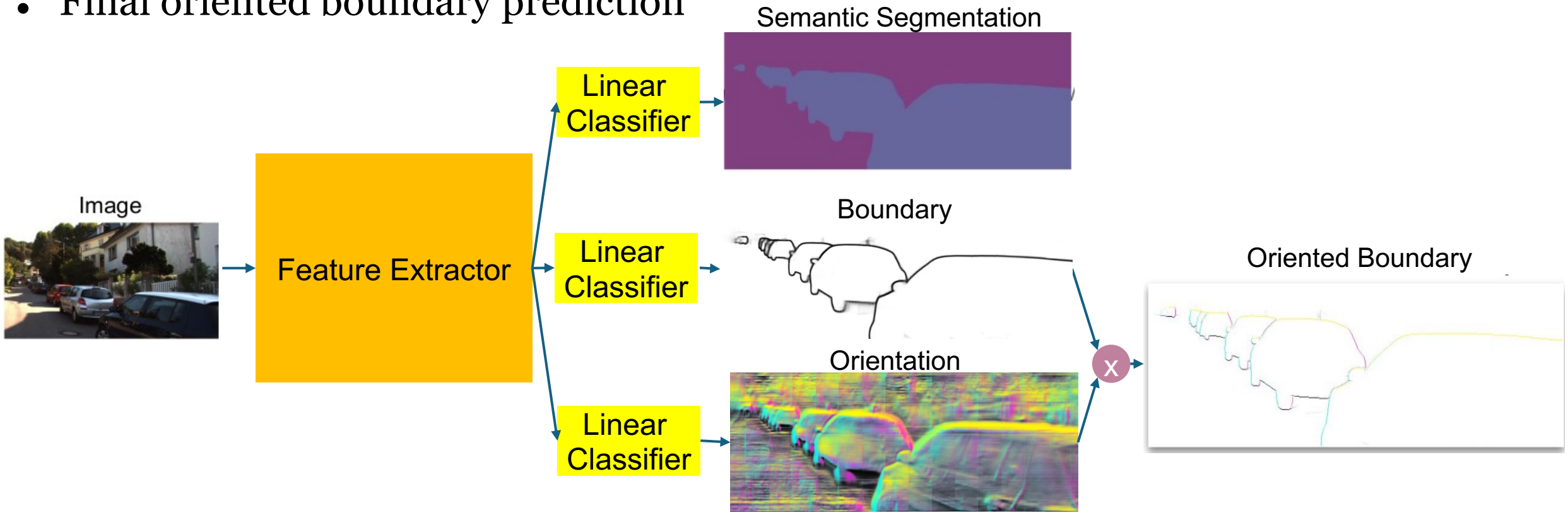
- Assume “p” is boundary
  - Orientation: [left, right, top, bottom]
    - “p”: [ 0, **1**, 0, **1** ]
  - L1 normalization
    - “p”: [ 0, 0.5, 0, 0.5 ]




Geometric Interpretation: Projecting the orientation vector on the axes (imprecise)

# Joint Semantic Seg & Occ Boundary CNN

- Final oriented boundary prediction



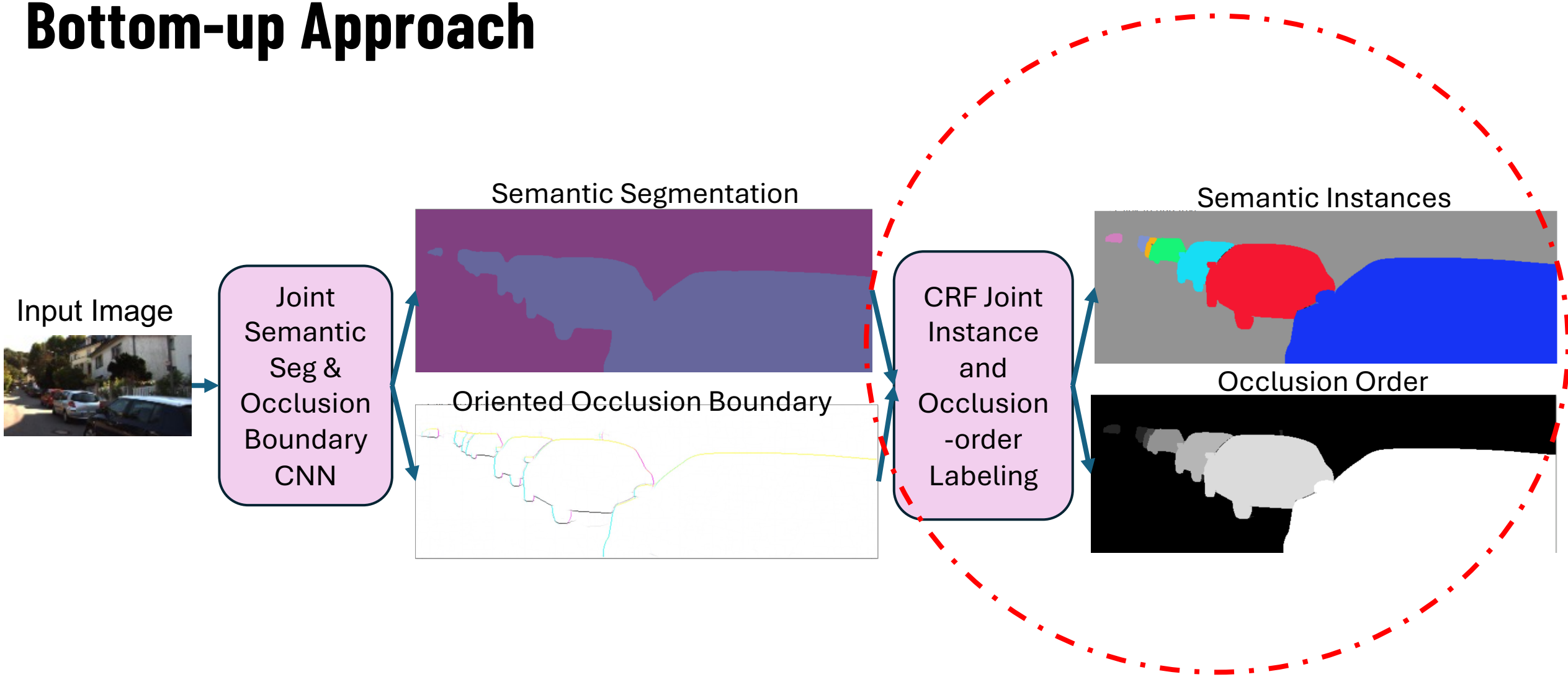
# Joint Semantic Seg & Occ Boundary CNN

- Final oriented boundary prediction

# Our Boundary Model vs. Prior Work


- Prior work:
  - Separates boundary and orientation, but uses angle regression
  - Or avoids angle regression, but does not separate the two tasks
- Our model:
  - Separates boundary prediction from orientation
  - No angle regression, so no special losses needed
  + Novel probabilistic interpretation
  + Novel upsampling architecture

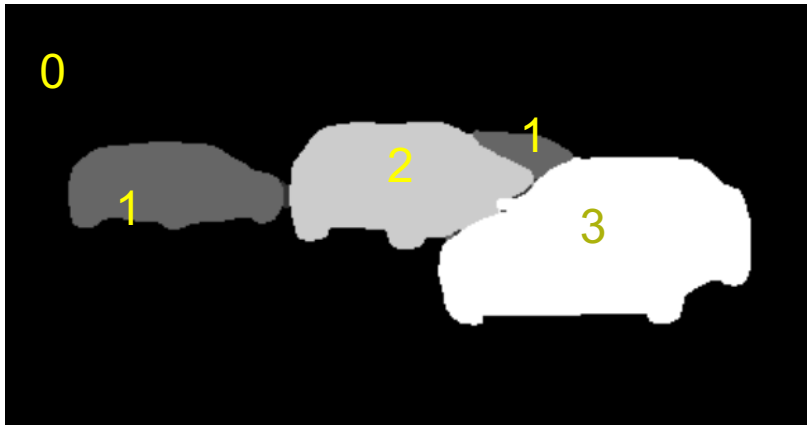
# Bottom-up Approach



# Optimization for Instance & Occlusion Order

- We solve a **labeling problem**
- Assign a label, $x_p$, to each pixel p

  $x_p \in \{0 = \text{background}, 1, 2, 3, \dots\}$
- Connected components represent instances
- $x_p$ represents the occlusion order



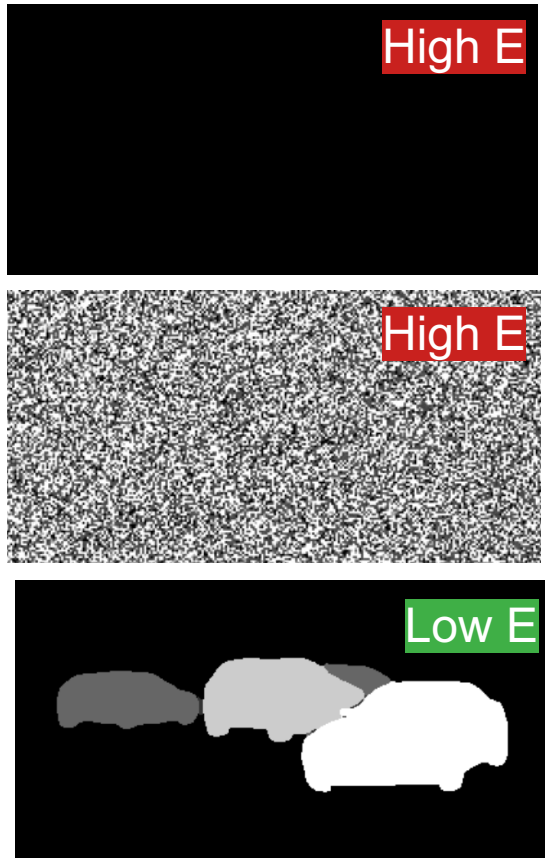

3 **>** 1 so 3 occludes 1

# Optimization for Instance & Occlusion Order

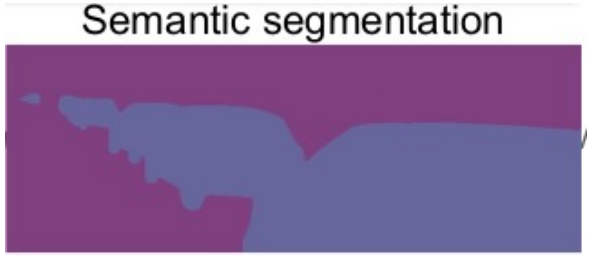
- Find the best such labeling!

- Define an **energy** function, ***E***
  - Takes a labeling → outputs its goodness
  - Lower $E$ = Better labeling
- **Minimize *E*** = find the best labeling

Possible Labelings



# Energy Function



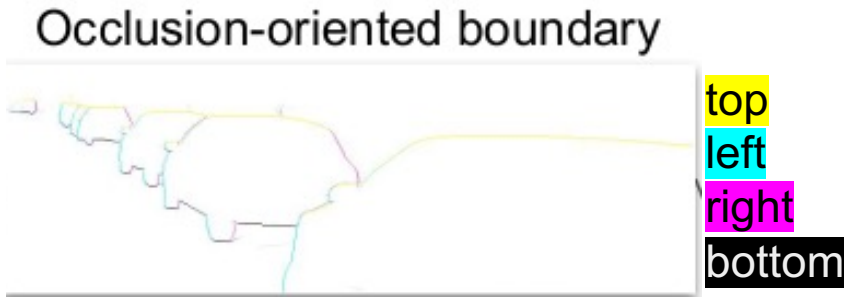

$$\sigma_p : \Pr(\mathrm{p} = \text{background})$$

1) Unary term for any pixel p

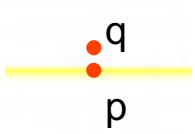
$$u_p(x_p) = \overbrace{(1-\sigma_p)\cdot[x_p = 0]}^{\text{Low } \sigma_p \quad \text{Penalize } x_p = 0} + \overbrace{\sigma_p \cdot [x_p \neq 0]}^{\text{High } \sigma_p \quad \text{Penalize } x_p \neq 0}$$

# Energy Function

2) Pairwise occlusion term for pixels p and q, where p occludes q




$$o(x_p, x_q) = \underbrace{c_\infty \cdot [x_p < x_q]}_{\text{Prohibit } x_p < x_q} \quad \underbrace{- [x_p > x_q]}_{\text{Encourage } x_p > x_q}$$



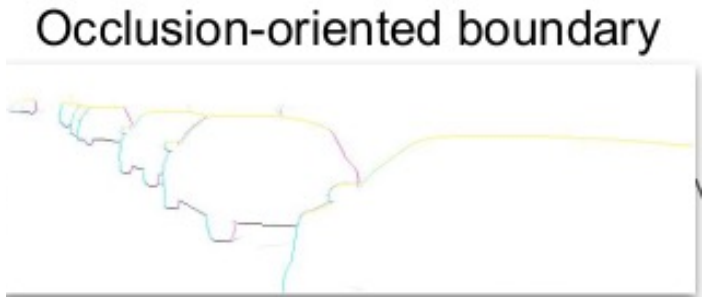
# Energy Function

3) Pairwise smoothing term: penalize if neighboring pixels, p and q, have different labels

$$v(x_p, x_q) = \underbrace{[x_p \neq x_q]}$$

Penalize if $x_p \neq x_q$

p q



# Energy Function

- The best labeling for all pixels, $\boldsymbol{x}$ , is found by minimizing the energy:

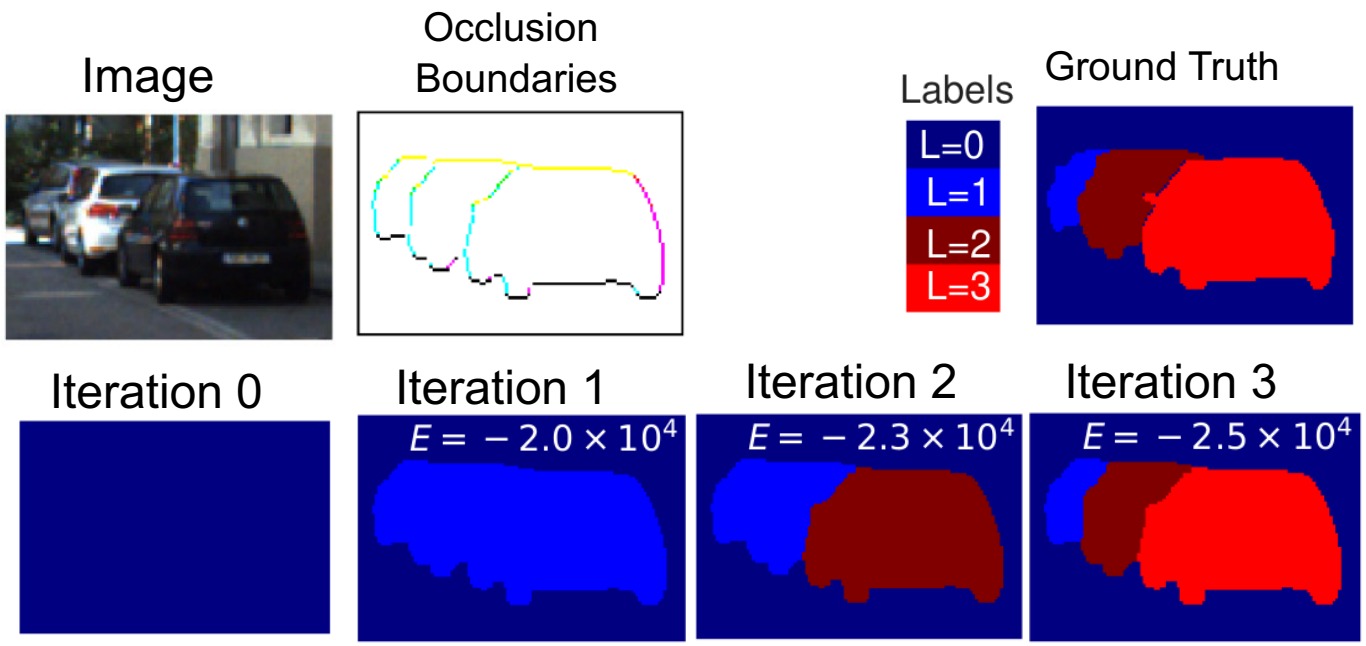
$$E(\boldsymbol{x}) = \sum_{p \in \mathcal{P}} u_p(x_p) + \lambda_v \sum_{(p,q) \in \mathcal{N}} v(x_p, x_q) + \lambda_o \sum_{(p,q) \in \mathcal{O}} o(x_p, x_q)$$

- NP-hard to minimize

# Minimizing Energy

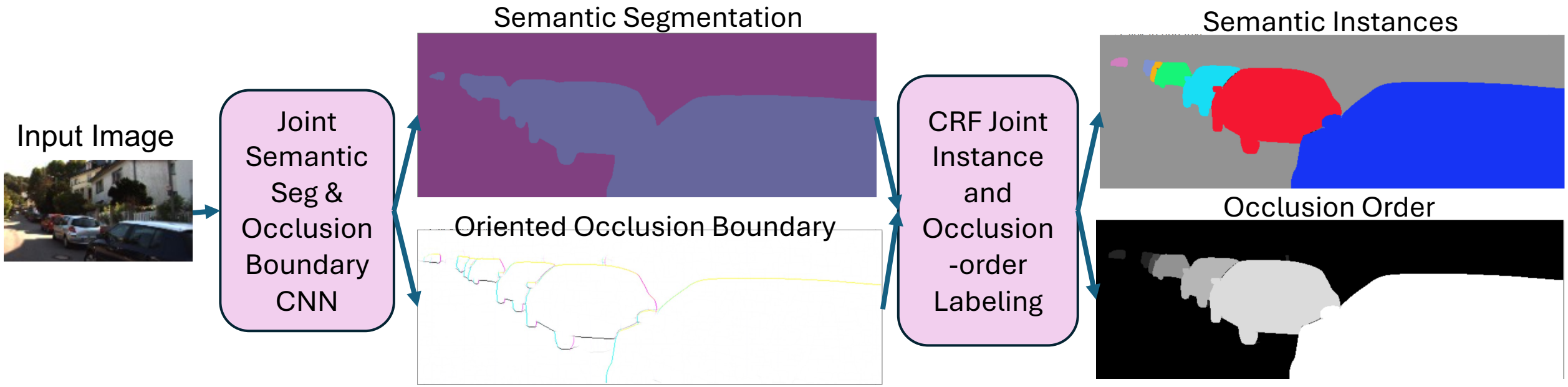
- We approximately minimize the energy using Jump moves [9]
  - Non-trivial quality guarantee

1) Start by all labels as 0
2) Make the best jump move
   - found by graph-cuts [8]
   - Jump move: any label can increase by 1 at most
3) Continue until no more moves can decrease $E$



# Bottom-up Approach

- Connected components → instances
- Labels → occlusion order
- Instance class : majority voting

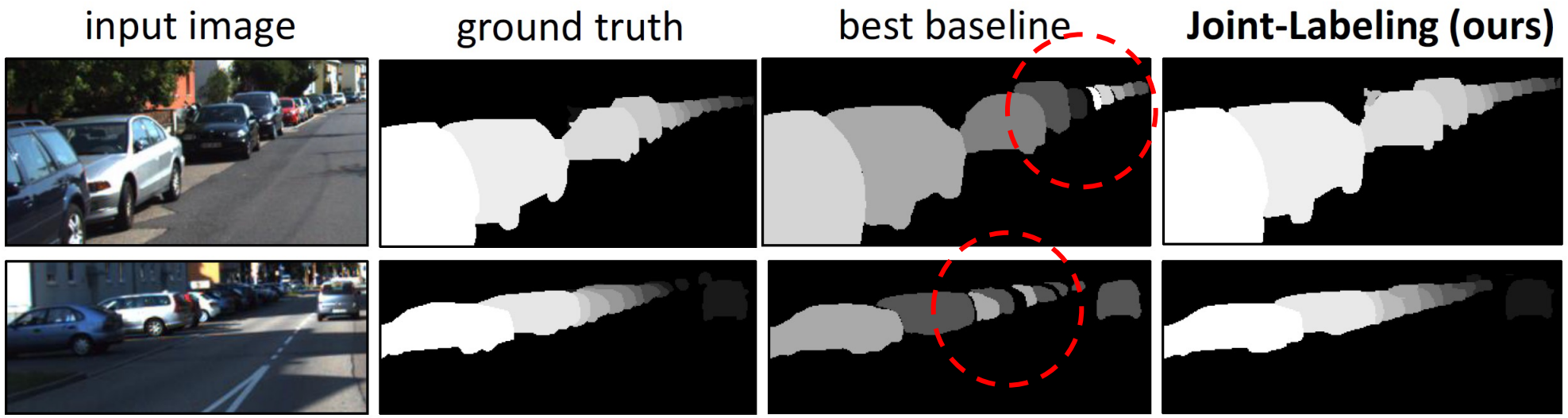


# Outline

- Introduction
- Related Work
- Our Approach
- **Experiments & Results**
- Conclusion

# Experiments - Qualitative Examples

- Baseline:
  - Instance Seg + Pairwise Ordering
  - Datasets: KINS [11] (~15,000 images) & COCOA [5] (~5,000 images)

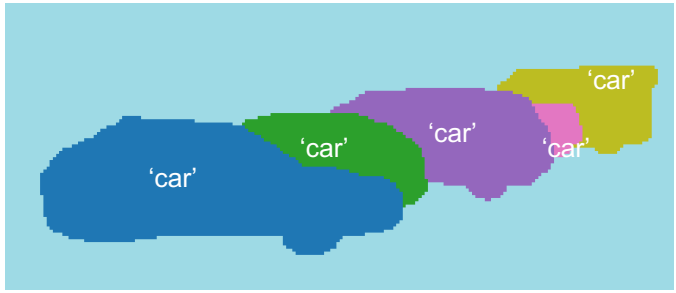


# Experiments - OOSIS Evaluation

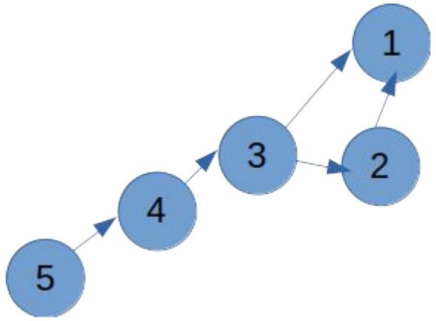
- We develop a metric for simultaneous evaluation of both outputs

Input Image


Output




**Instance masks + their semantic classes**

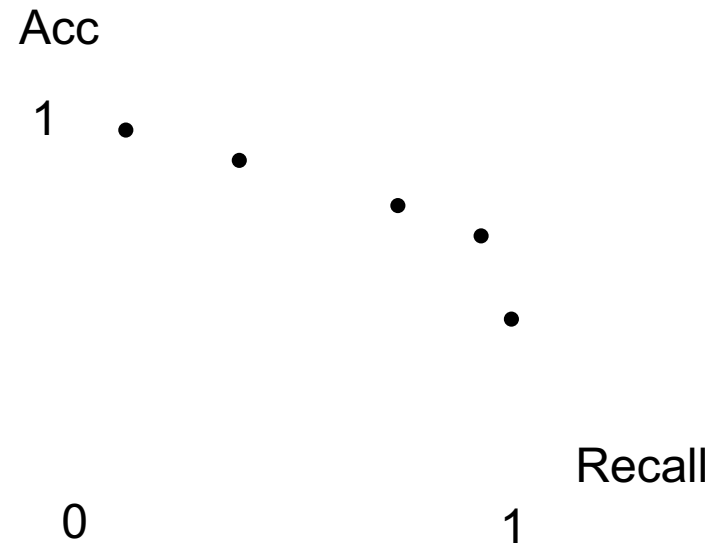



**Occlusion odering of instances**

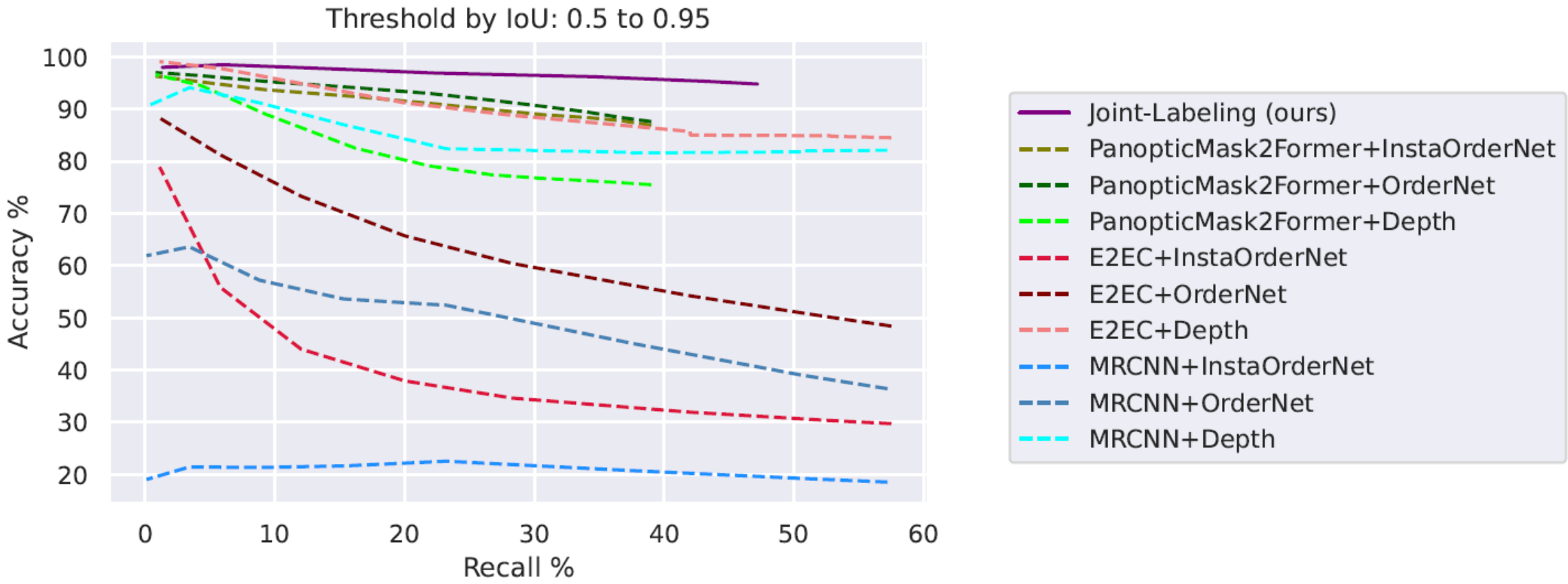
# Our Metric

- Based on Accuracy-Recall plot
- *Recall*: quality of predicted masks
- *Accuracy*: quality of occlusion ordering

- Recall can vary with altered matching criteria
  - **More / fewer** instances are considered accurate



# Experiments - OOSIS Evaluation




Our method compared to baselines on KINS.

# Experiments - Instance Masks Quality

Table 3: Instance segmentation quality by PanopticMask2Former vs ours.

| Model | Unambiguous Instance Segmentation KINS | | COCOA | |
|---|---|---|---|---|
| | $mAP_{0.5:0.95}$ ↑ | WCS ↑ | $mAP_{0.5:0.95}$ ↑ | WCS ↑ |
| PanopticMask2Former (same score) | 15.4 | 74.6 | 3.7 | 47.3 |
| PanopticMask2Former (ad-hoc score) | **21.6** | 74.6 | 7.4 | 47.3 |
| Joint-Labeling (Ours) | **21.6** | **77.4** | **9.1** | **61.2** |

# Experiments – Oriented Occ Boundary

- Using standard metrics
- Ours vs. the state-of-the-art P2ORM [12]

The performance of our oriented occlusion boundary model vs. the state-of-the-art P2ORM model on KINS dataset.

| Model | Oriented Occ Boundary | | |
|---|---|---|---|
| | ODS↑ | OIS↑ | AP↑ |
| P2ORM | 77.8 | 80.6 | 79.4 |
| Ours | **84.5** | **86** | **88.3** |

# Outline

- Introduction
- Related Work
- Our Approach
- Experiments & Results
- **Conclusion**

# Conclusion

- We address OOSIS using deep learning.
- We propose a novel joint semantic segmentation & oriented occlusion boundary model outperforming the SoTA
- Our approach can be thought of as an instance/panoptic segmentation method, but requires occlusion ground truth
- We introduce a new metric for OOSIS

# Thank You!
# Questions?

# References

[1] Yang, Y., Hallman, S., Ramanan, D., and Fowlkes, C. C. Layered object models for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(9):1731–1743, 2011.

[2] Isola, P. and Liu, C. Scene collaging: Analysis and synthesis of natural images with semantic layers. In Proceedings of the IEEE International Conference on Computer Vision, pp. 3048–3055, 2013.

[3] Tighe, J., Niethammer, M., and Lazebnik, S. Scene parsing with object instances and occlusion ordering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3748–3755, 2014.

[4] Zhang, Z., Schwing, A. G., Fidler, S., and Urtasun, R. Monocular object instance segmentation and depth ordering with cnns. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2614–2622, 2015.

[5] Zhu, Y., Tian, Y., Metaxas, D., and Dollár, P. Semantic amodal segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1464–1472, 2017.

[6] Lee, H. and Park, J. Instance-wise occlusion and depth orders in natural scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21210–21221, 2022.

[7] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid scene parsing network. In CVPR, 2017.

# References

[8] Boykov, Y., Veksler, O., and Zabih, R. Fast approximate energy minimization via graph cuts. IEEE Transactionson Pattern Analysis and Machine Intelligence, 23(11):1222–1239, November 2001.

[9] Veksler, O. Efficient Graph-Based Energy Minimization Meth. in Comp. Vis. PhD thesis, 1999.

[10] Kolmogorov, V. and Rother, C. Minimizing nonsubmodular functions with graph cuts-a review. IEEE transactions on pattern analysis and machine intelligence, 29(7):1274–1279, 2007.

[11] Qi, L., Jiang, L., Liu, S., Shen, X., and Jia, J. Amodal instance segmentation with kins dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[12] Xuchong Qiu, Yang Xiao, Chaohui Wang, and Renaud Marlet. Pixel-pair occlusion relationship map (p2orm): formulation, inference and application. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pp. 690–708. Springer, 2020.

[13] He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pp. 2961–2969, 2017a.

# References

[14] Zhang, T., Wei, S., and Ji, S. E2ec: An end-to-end contour-based method for high-quality high-speed instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4443–4452, 2022

[15] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3), 2022.

[16] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. Advances in neural information processing systems, 27, 2014.

[17] Forrester Cole, Kevin Sanik, Doug DeCarlo, Adam Finkelstein, Thomas Funkhouser, Szymon Rusinkiewicz, and Manish Singh. How well do line drawings depict shape? In ACM SIGGRAPH 2009 papers, pages 1–9. 2009.

[18] Ken Nakayama. Biological image motion processing: a review. Vision research, 25(5):625–660, 1985.

[19] Gaetano Kanizsa, Paolo Legrenzi, and Paolo Bozzi. Organization in vision: Essays on gestalt perception. (No Title), 1979.

# References

[20] Derek Hoiem, Andrew N Stein, Alexei A Efros, and Martial Hebert. Recovering occlusion boundaries from a single image. In 2007 IEEE 11th International Conference on Computer Vision, pages 1–8. IEEE, 2007.

[21] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8533–8542, 2020.

[22] Peng Wang and Alan Yuille. Doc: Deep occlusion estimation from a single image. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 545–561. Springer, 2016.

[23] Guoxia Wang, Xiaochuan Wang, Frederick WB Li, and Xiaohui Liang. Doobnet: Deep object occlusion boundary detection from an image. In Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14, pages 686–702. Springer, 2019.

[24] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3668–3678, 2015.