Robust Instance Segmentation through Reasoning about Multi-Object Occlusion

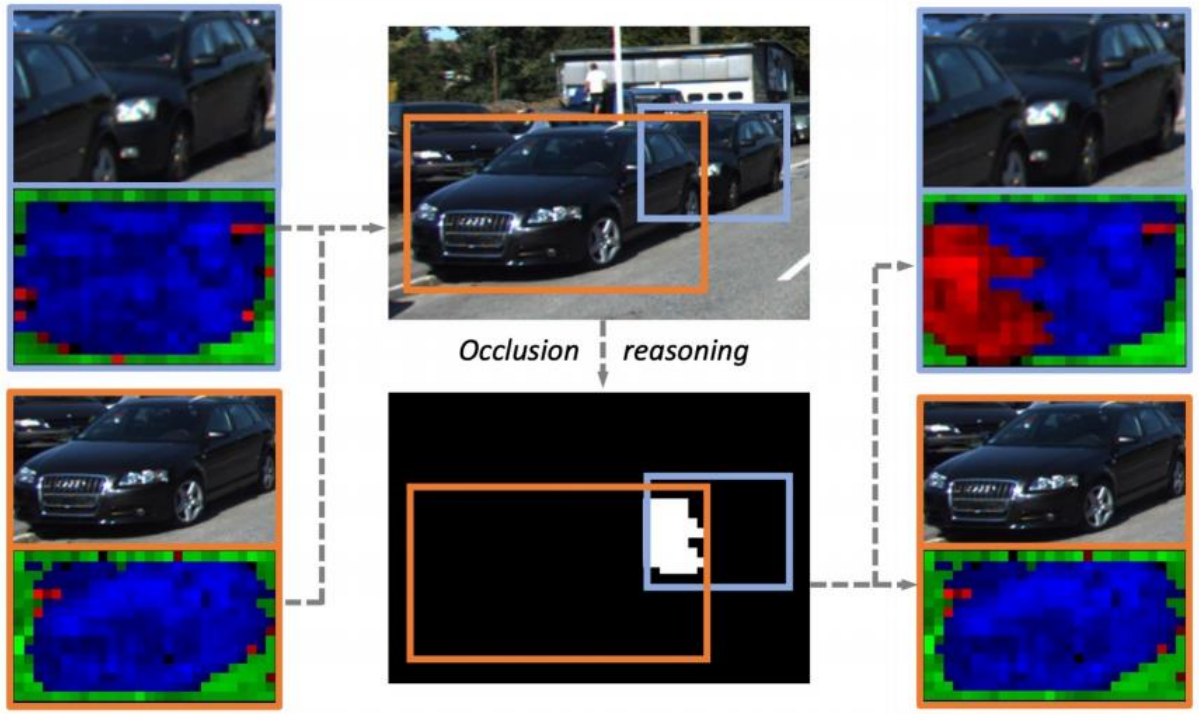
Xiaoding Yuan[1] Adam Kortylewski[2] Yihong Sun[2] Alan Yuille[2]

[1]Tongji University, [2]Johns Hopkins University

arXiv:2012.02107v1 [cs.CV] 3 Dec 2020

Speaker：Li Hao

Introduction& Motivation



当图像中包含多个物体部分遮挡时用神经网络分析复杂场景是具有挑战性，与人类相比，深度网络在识别部分遮挡物体的方面鲁棒性较差，主要困难来源于物体顺序和位置的组合变异性以及场景中可能存在已知或未知的物体。而现有的图像分析方法大多是独立处理对象，没有考虑到附近对象的相对遮挡。

Introduction& Motivation

目前解决遮挡问题的一种方法是数据增强，虽然鲁棒性增强了，但是部分遮挡对象的分类性能差。

合成深度网络（compositional deep networks or CompositionalNets）比数据增强更具有鲁棒性，但是其也是独立地对待图中对象，没有明确使用对象间的相互关系。

Contributions

本文有以下贡献：

1）提出了一种多物体实例分割的网络，其基于CompositionalNets，通过多物体遮挡推理增强了对遮挡的鲁棒性，并且只需要bounding box来进行监督。

2）提出了一个Occlusion Reasoning Module (ORM)，可以在多个物体的生成模型中进行有效的推理，能检测到错误的前馈（feed-forward）预测，并利用通过推理对象的遮挡顺序来进行更正。

3）在KINS 数据集中，实现了遮挡下实例分割的state-of-the-art。

Method — CompNets for Single Objects

the fully connected classification head is replaced with a differentiable compositional model

M： the number of mixtures of compositional models per object category

$$p(F|\Theta_y)=\sum_m \nu_m p(F|\theta_y^m),\ \nu_m \in \{0,1\},\ \sum_{m=1}^{M} \nu_m=1 \quad (1)\quad \Theta_y=\{\theta_y^m=\{\mathcal{A}_y^m,\chi_y^m,\Lambda\}|m=1,\ldots,M\}$$

$$p(F|\theta_y^m)=\prod_i p(f_i|\mathcal{A}_{i,y}^m,\chi_{i,y}^m,\Lambda) \qquad (2)$$

$$\chi_y^m=\{\chi_{i,y}^m|i\in[H,W]\} \qquad \mathcal{A}_y^m=\{\mathcal{A}_{i,y}^m|i\in[H,W]\}$$

Method — CompNets for Single Objects

feature likelihood is defined as composition of a foreground and a context likelihood

$$p(f_i|\mathcal{A}^m_{i,y}, \chi^m_{i,y}, \Lambda) = p(i|m,y)\; p(f_i|\mathcal{A}^m_{i,y}, \Lambda) \qquad (3)$$

$$+ (1 - p(i|m,y))\; p(f_i|\chi^m_{i,y}, \Lambda). \qquad (4)$$

Foreground 似然

context 似然

$$\mathcal{A}^m_{i,y} = \{\alpha^m_{i,k,y} | k = 1, \ldots, K\} \qquad \sum_{k=0}^{K} \alpha^m_{i,k,y} = 1$$

$$\Lambda = \{\lambda_k = \{\sigma_k, \mu_k\} | k = 1, \ldots, K\}$$

the parameters of von-Mises-Fisher distributions

$$p(f_i|\mathcal{A}^m_{i,y}, \Lambda) = \sum_k \alpha^m_{i,k,y} p(f_i|\lambda_k), \qquad (5)$$

$$p(f_i|\lambda_k) = \frac{e^{\sigma_k \mu_k^T f_i}}{Z(\sigma_k)}, ||f_i|| = 1, ||\mu_k|| = 1. \qquad (6)$$

Method — CompNets for Single Objects

用了一个outlier模型来增强其对部分遮挡的鲁棒性:

$$p(F|\theta_y^m,\beta)=\prod_i p(f_i|\beta,\Lambda)^{1-z_i^m}\,p(f_i|\mathcal{A}_{i,y}^m,\Lambda)^{z_i^m}. \quad (7) \qquad \mathcal{Z}^m=\{z_i^m\in\{0,1\}|i\in\mathcal{P}\}$$

outlier模型: $$p(f_i|\beta,\Lambda)=\sum_k \beta_{n,k}p(f_i|\sigma_k,\mu_k). \quad (8)$$

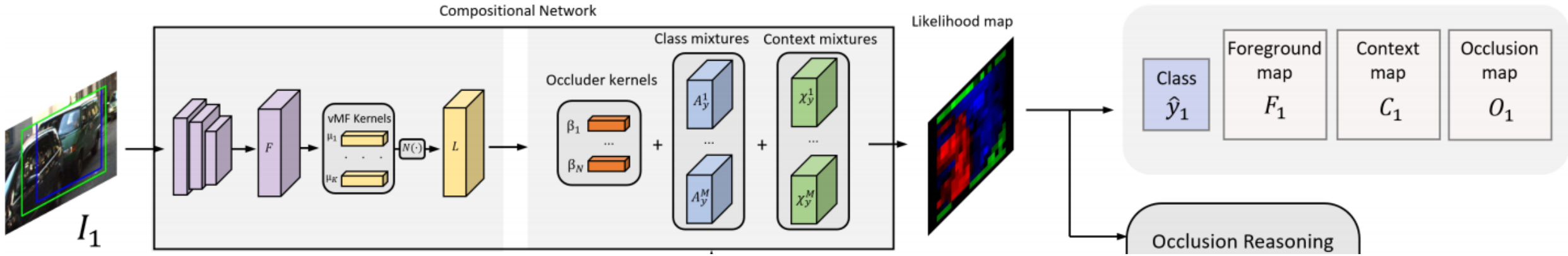
比较模型的似然项可以实现实例分割。

occlusion O $$p(f_i=\mathcal{O})=p(i|m,y)\;p(f_i|\beta,\Lambda) \quad (9)$$

foreground F $$p(f_i=\mathcal{F},y)=p(i|m,y)\;p(f_i|\mathcal{A}_{i,y}^m,\Lambda) \quad (10)$$

context C $$p(f_i=\mathcal{C},y)=(1-p(i|m,y))\;p(f_i|\chi_{i,y}^m,\Lambda) \quad (11)$$

Method — CompNets for Single Objects



Method — CompNets for Multiple Objects

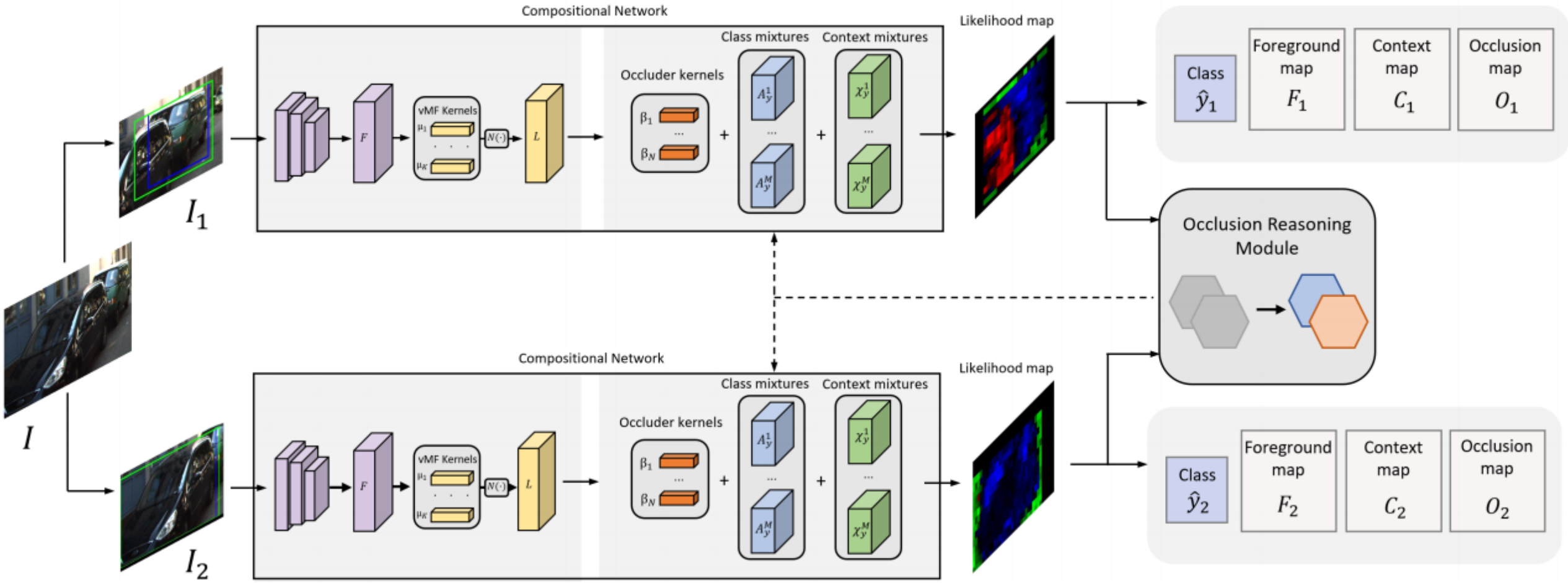
assuming independence between objects neglects the relations between them and leads to inconsistencies in the segmentation results.

$$p(F|\theta^m_{y_1},\ldots,\theta^m_{y_N},\beta)=\prod_i\prod_{n=1}^{N+1}p_n(f_i)^{z_{i,n}}, \qquad (12)$$

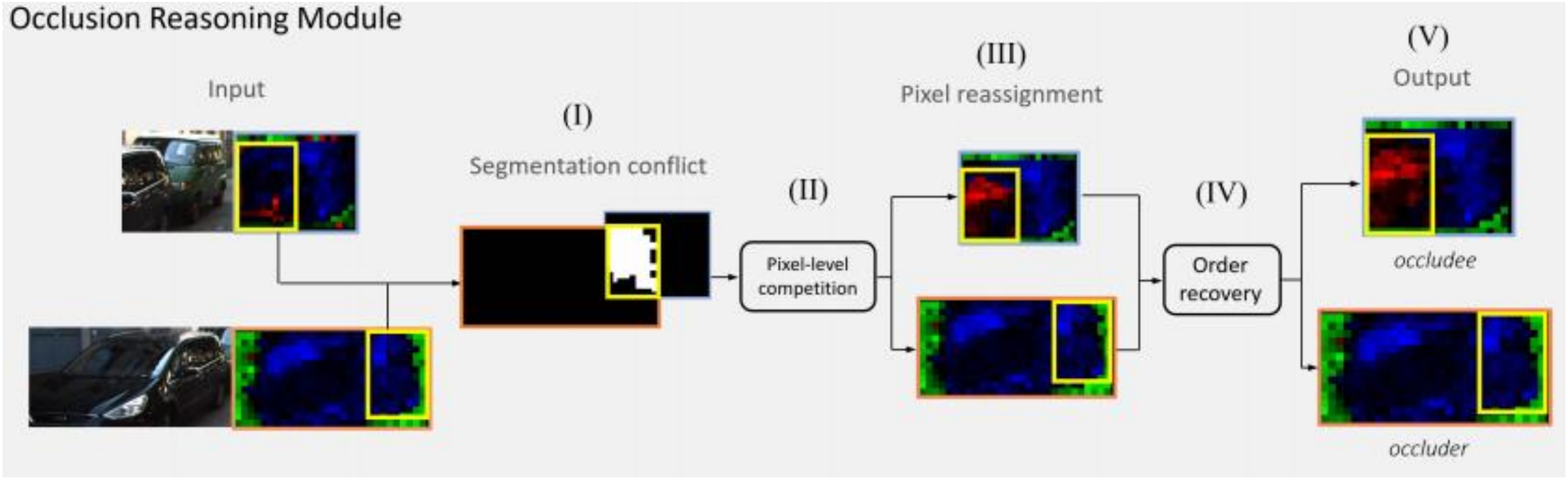
$p_n(f_i)=p(F|\theta^m_{y_n},\beta)$ $\quad p_{N+1}(f_i)=p(f_i|\beta,\Lambda)$ $\qquad \sum_n z_{i,n}=1$ and $z_{i,n}\in\{0,1\}$

Maximizing the model likelihood is difficult （because it involves multiple objects and the visibility at each pixel， depends on the visibility of the neighboring pixels）

Proposed network architecture



Pipeline — Reasoning about Multi-Object Occlusion



Experiments

Modal and amodal instance segmentation

on the KINS

	Mask	L0	L1	L2	L3	Mean
Mask R-CNN	✓	85.8	81.5	72.7	51.9	73
CompNet	✗	75.8	67.7	44.4	23.3	64.3
Ours (iter=2)	✗	**75.9**	**69.2**	**54.0**	**34.6**	**67.2**

	Mask	L0	L1	L2	L3	Mean
PCNet-M	✓	83.1	77.5	68.5	51.6	70.2
BBTP	✗	77.9	71.6	67	67.8	71.1
CompNet	✗	76.6	76.1	75.9	74.7	76.2
Ours (iter=2)	✗	**76.9**	**76.4**	**76.5**	**76.5**	**76.7**

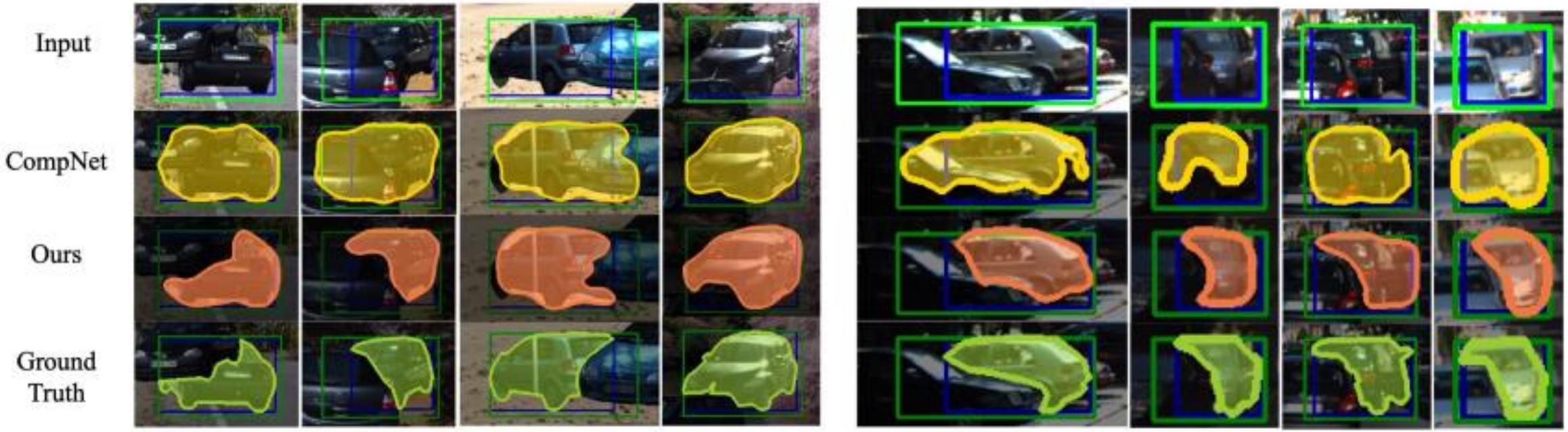
(a) 2 objects (b) 4 objects (c) 2 objects with unknown occlusion

on occlusion challenge

	2 Objects					4 Objects					2 Objects + Unknown Occlusion				
Occ Level	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean
Mask R-CNN	88.2	86.3	69.1	58.2	82.3	88.7	88	74.8	63	78.6	90.5	86.8	72.2	57.1	76.7
CompNet	77.8	67.3	51.0	26.3	66.9	**76.7**	67.1	50.2	26.1	56.0	**78.9**	72.2	57.8	36.0	63.6
Ours (iter=1)	**78.0**	**75.3**	65.4	45.6	72.9	75.2	**72.9**	61.9	43.0	65.0	77.9	**73.3**	**62.0**	**41.7**	**65.8**
Ours (iter=2)	**78.0**	**75.3**	**65.7**	**47.2**	**73.1**	75.2	**72.9**	**62.2**	**44.0**	**65.3**	78.0	**73.3**	**62.0**	**41.7**	**65.8**

	2 Objects					4 Objects					2 Objects + Unknown Occlusion				
Occ Level	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean
PCNet-M	82.4	81	69.3	47	70	87.2	79.3	63.7	41.3	67.9	-	-	-	-	-
BBTP	**80.5**	73.6	69.5	72.8	74.1	**80.5**	71.9	64	66	70.6	**83.7**	77.3	67.9	60.6	72.4
CompNet	78.0	76.6	75.0	72.1	76.7	77.3	75.4	74.1	71.4	74.8	78.4	**78.1**	76.1	71.9	76.5
Ours (iter=1)	79.9	**80.0**	79.2	77.7	**79.7**	78.6	78.9	78.1	76.6	78.2	78.6	78.0	**76.2**	**72.1**	**76.6**
Ours (iter=2)	79.9	**80.0**	**79.3**	**78.1**	**79.7**	80.0	**80.0**	**79.3**	**78.1**	**79.5**	78.5	**78.1**	**76.2**	**72.1**	**76.6**

Experiments



(a) Synthetic dataset (b) KINS

Figure 6: Qualitative results for modal segmentation on KINS and images from our occlusion challenge. The top row show the input images including bounding box annotations. Images in the second row are generated by the baseline CompNet, and the third row shows the results by our CompNet with multi-object ORM. The last row shows the ground truth.

Experiments

Ablation study（ verify the effectiveness of the order recovery ）

	2 objects		4 objects		2 + unknown	
	Modal	Amodal	Modal	Amodal	Modal	Amodal
NOD	70.5	77.8	58.5	75.2	65.0	76.5
OD	**73.1**	**79.7**	**65.3**	**79.5**	**65.8**	**76.6**

Thanks!