BLADE:

Box-Level Supervised Amodal Segmentation through Directed Expansion

Zhaochen Liu[1,2], Zhixuan Li[3], Tingting Jiang[1,4]

[1] National Engineering Research Center of Visual Technology, National Key Laboratory for Multimedia Information Processing, School of Computer Science, **Peking University**

[2] AI Innovation Center, School of Computer Science, **Peking University**

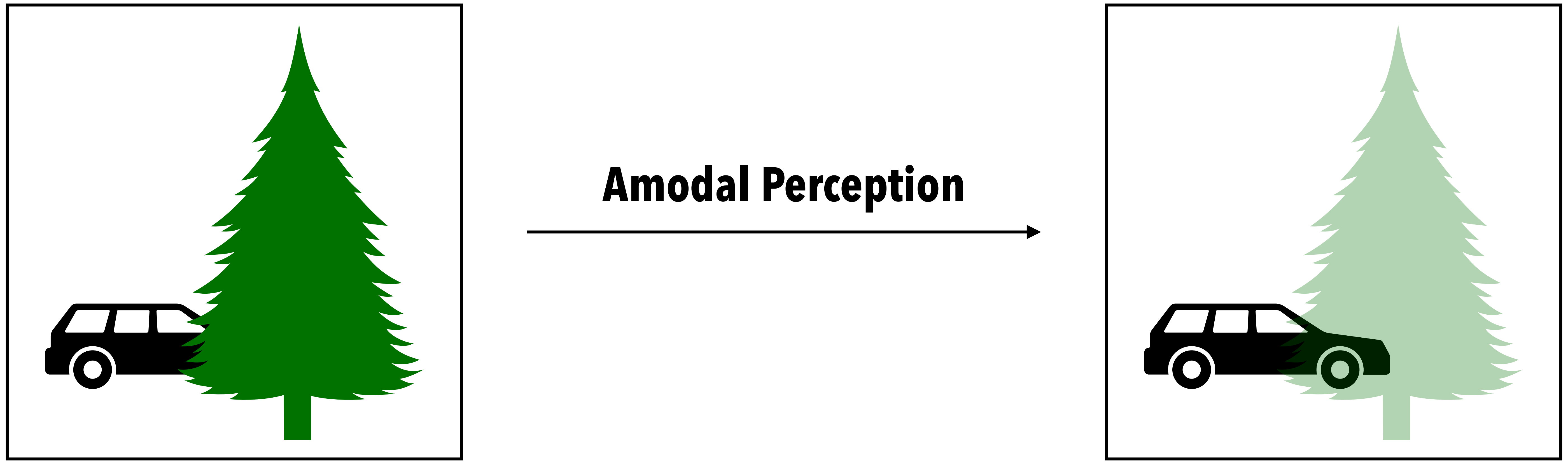
[3] School of Computer Science and Engineering, **Nanyang Technological University**

[4] National Biomedical Imaging Center, **Peking University**

The 38th Annual AAAI Conference on Artificial Intelligence

FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Amodal Perception



- Amodal perception is to **infer the complete shape of occluded objects**.
- A **vital** ability of human's cognitive system.
- Essential potential for **tremendous** real-world applications (autonomous driving, robotic gripping, novel view synthesis, …).

Related Work

- In computer vision, amodal instance segmentation has aroused **broad** concern since it was proposed, which aims to predict complete shapes of partially occluded objects.

Direct Optimization
Li et al., 2016,
Zhu et al., 2017,
Qi et al., 2019,
…

Depth Relationships
Zhang et al., 2019,
…

Compositional Models
Wang et al., 2020,
…

Shape Priors
Xiao et al., 2021,
Li et al., 2022,
…

Correlation
Follmann et al., 2019,
Ke et al., 2021,
…

Amodal Completion
Ehsani et al., 2018,
Dhamo et al., 2019,
Ling et al., 2020,
…

Challenge

- However, annotating pixel-level ground-truth amodal masks for such objects is **labor-intensive and error-prone** due to the absence of visible cues in occluded regions.

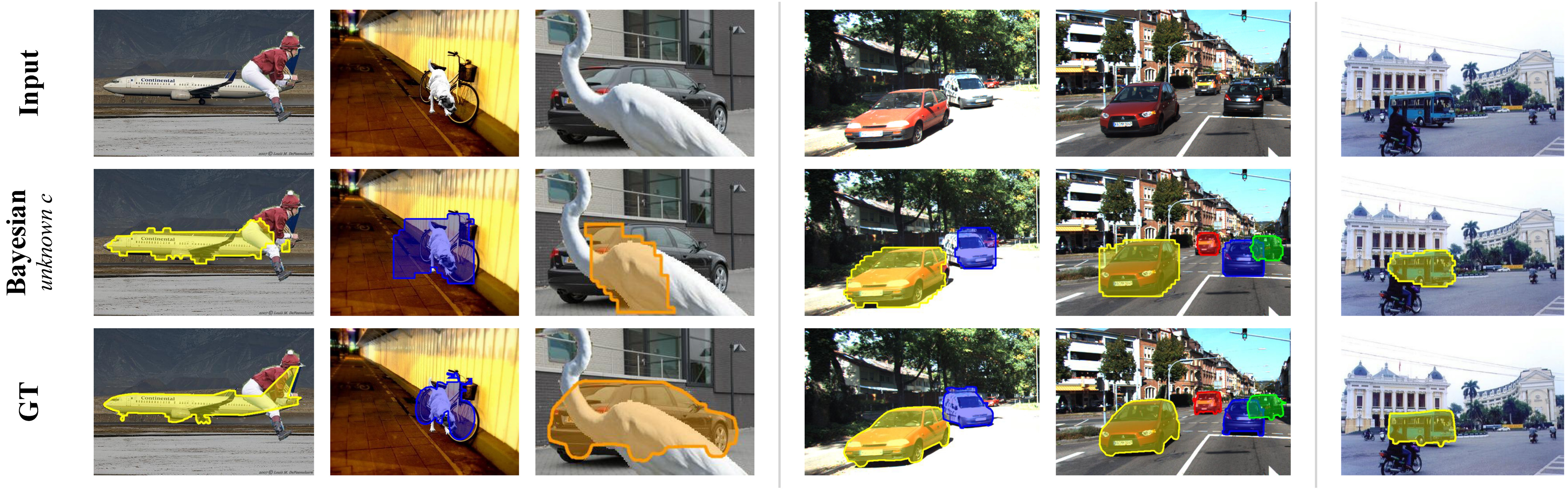
- To solve the challenges of pixel-level annotation, Bayesian-Amodal (Sun et al., 2022), a **weakly supervised** approach is proposed that utilizes **ground-truth bounding boxes** as an alternative supervision signal.

Bayesian-Amodal

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

BLADE

- Nevertheless, the amodal mask generated by the Bayesian-Amodal approach exhibits **low resolution** and **uneven boundaries**.

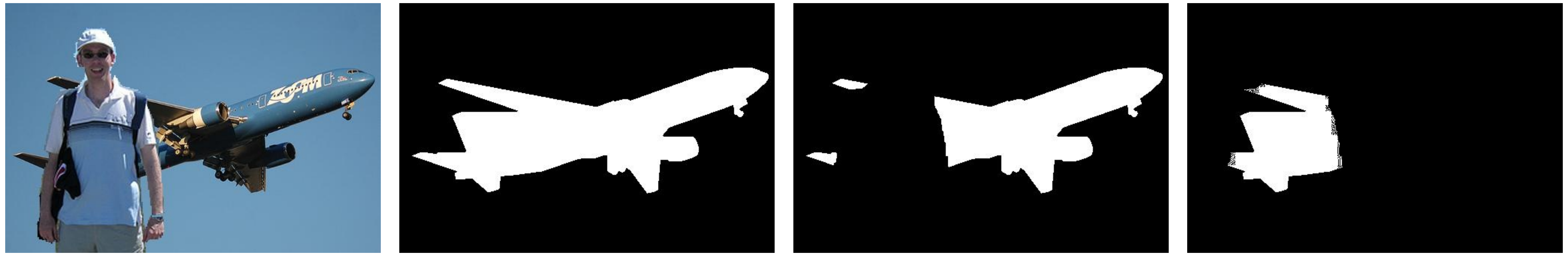


Introduction

- *How to obtain amodal masks with both* ***high-resolution*** *and* ***accurate boundaries*** *solely through* ***box-level supervision****?*
- *To deal with this challenge, we propose the* ***B****ox-****L****evel supervised* ***A****modal segmentation network through* ***D****irected* ***E****xpansion,* **BLADE***, a weakly-supervised method.*

Method

- An amodal mask $\mathbf{M}_a$ can be decomposed.
- Inspired by this, we design a hybrid structure with multiple branches.



$$\mathbf{M}_a = \mathbf{M}_v + \mathbf{M}_o$$

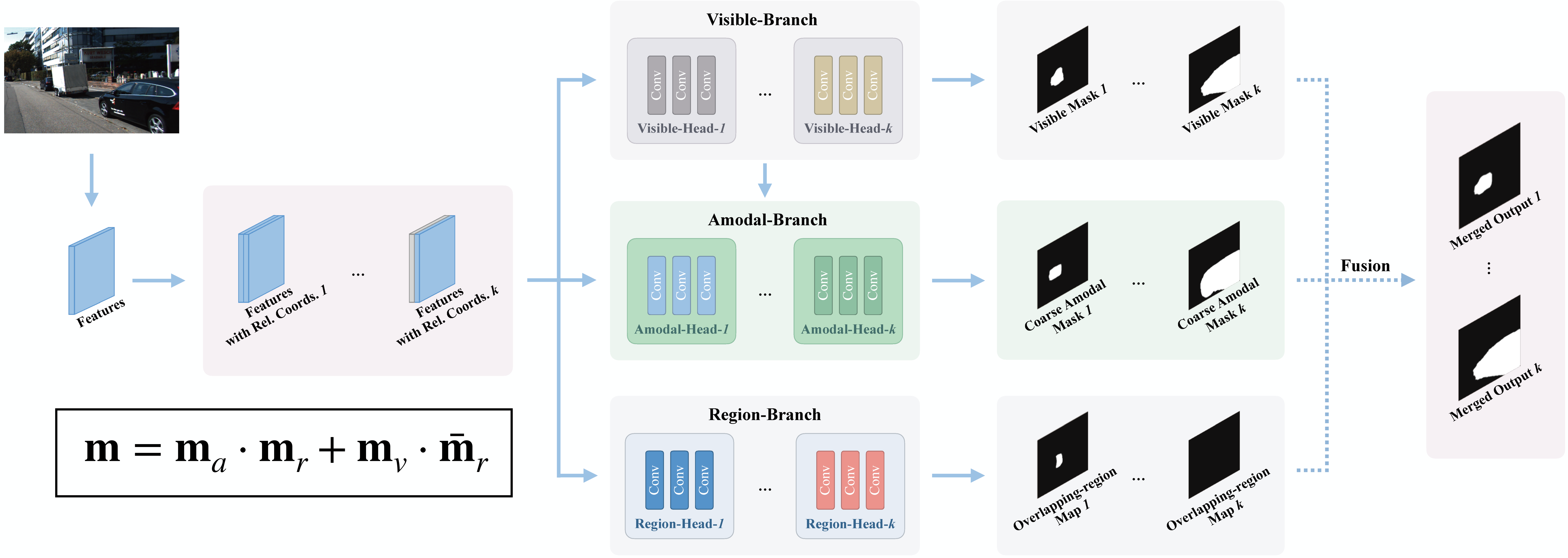
General segmentation problem

Clues

Method | Network Architecture

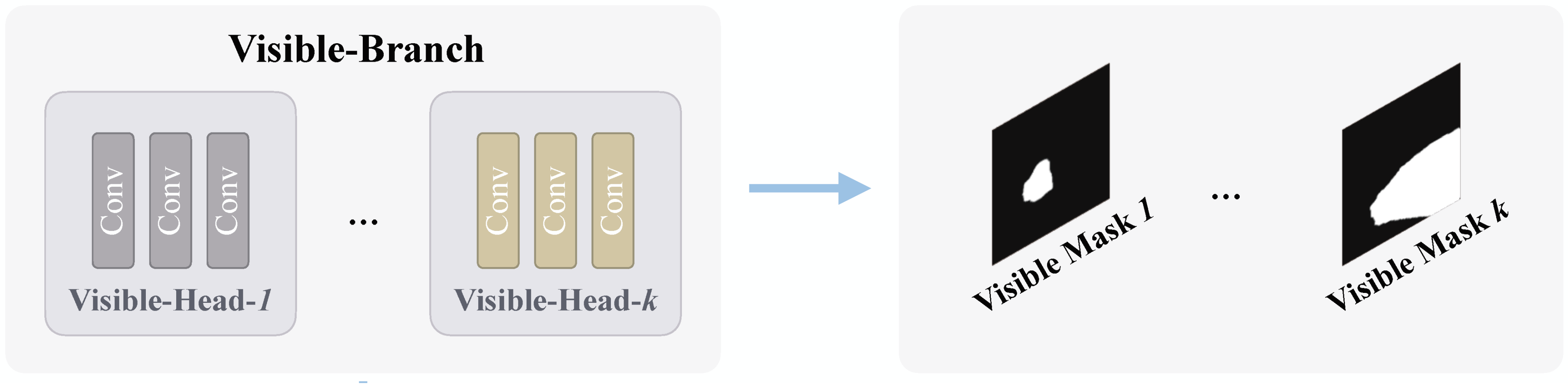
The 38th Annual AAAI Conference on Artificial Intelligence BLADE

- The three branches share the same multi-scale features extracted from the image
- The three branches all adopt dynamically-generated instance-aware mask heads containing varying instance-by-instance parameters (refer to CondInst, Tian et al., 2020).

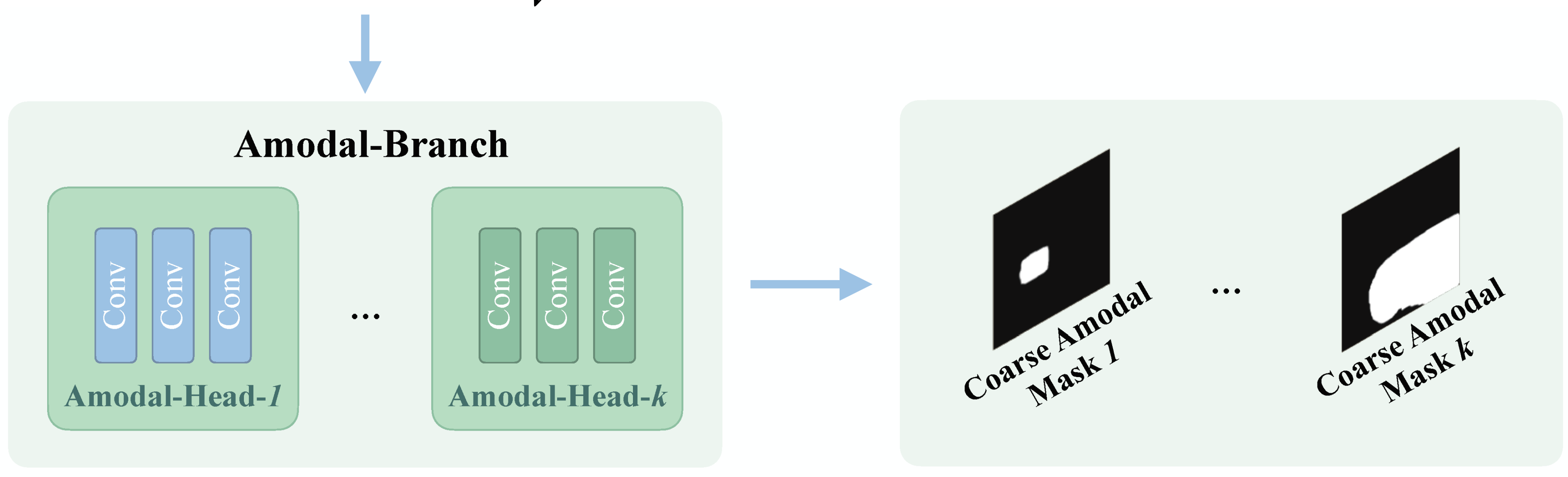


$$\mathbf{m} = \mathbf{m}_a \cdot \mathbf{m}_r + \mathbf{m}_v \cdot \bar{\mathbf{m}}_r$$

Method | Multiple Branch



Predicted $\mathbf{m}_v$



Visible-Brach

- The original mask heads with projection loss and pairwise loss in BoxInst (Tian et al., 2021) are used.
- $\mathbf{B}_v$ (the bounding box of visible portion) annotations are applied as the supervision.

Amodal-Brach

- We feed it the predicted $\mathbf{m}_v$ from visible-branch in addition to the features and relative coordinates.
- $\mathbf{B}_a$ (the bounding box of complete object) annotations are applied as the supervision.

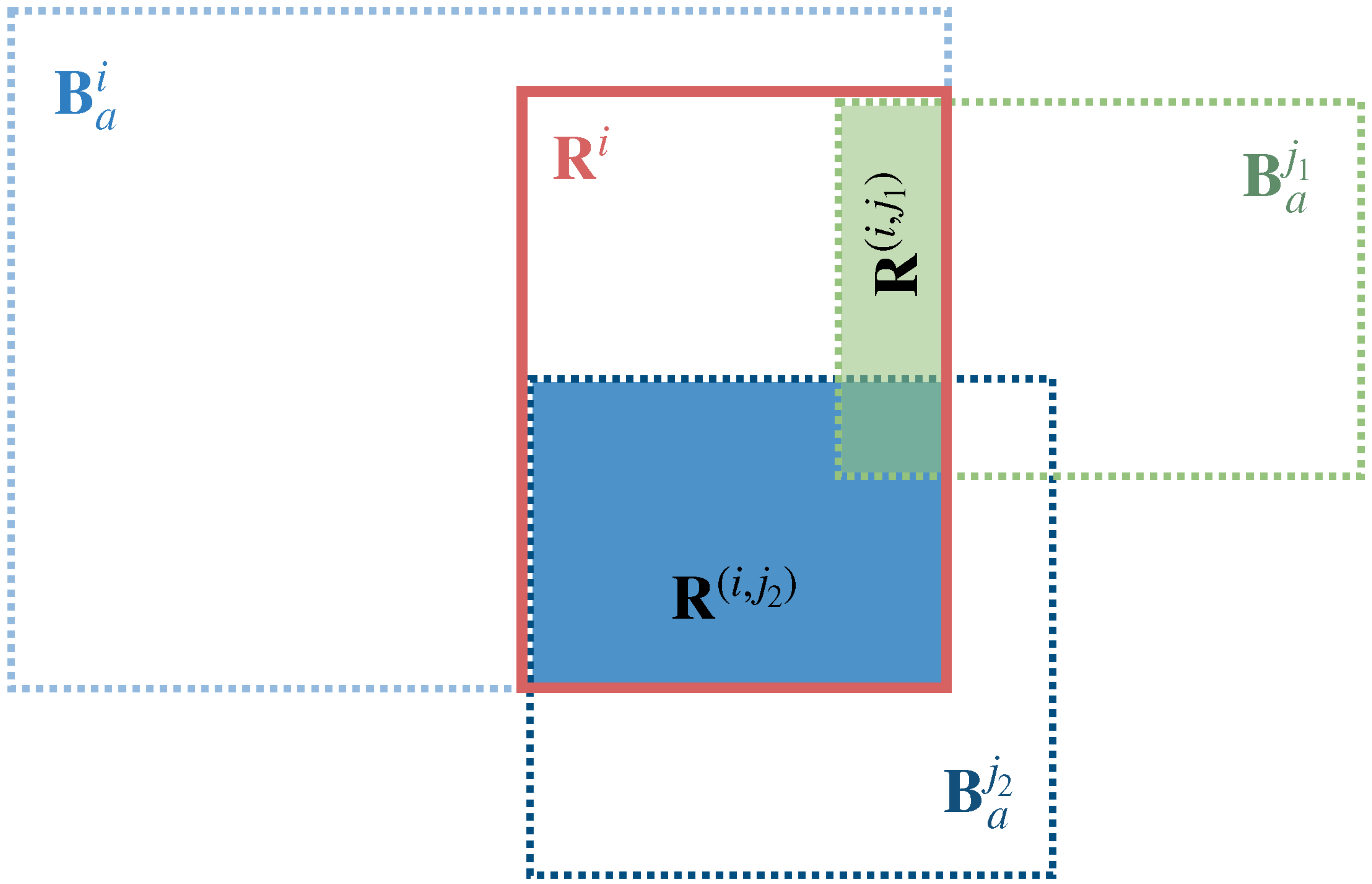
- What about the region-branch?

Method | Overlapping Region

- The **tightest** bounding box that covers **all intersecting areas** of the amodal bounding box of the object and those of other objects.
- The occluded portion of each object should be inside if exists.

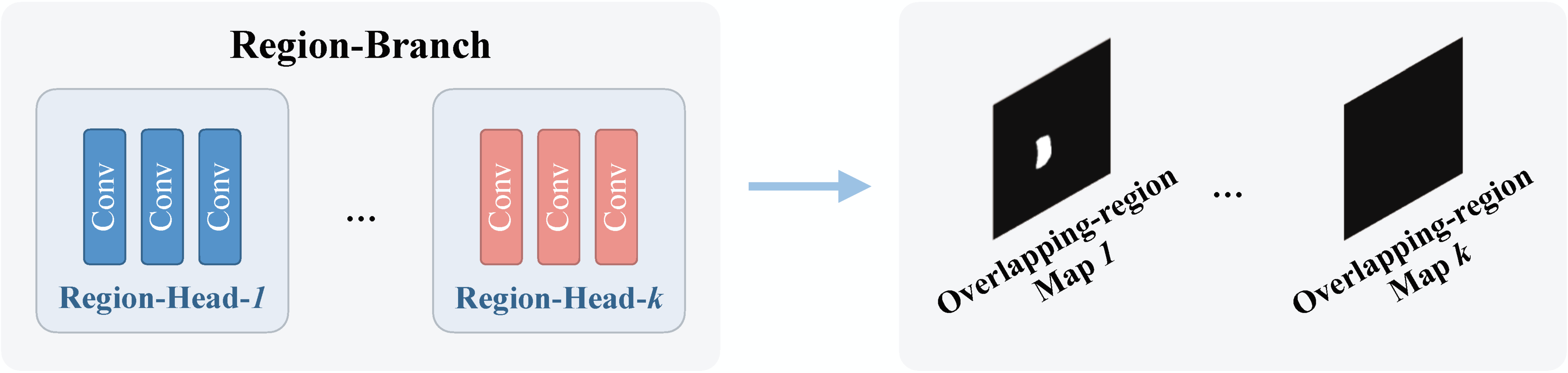
Method | Overlapping Region

- If there are multiple intersecting areas, the **envelope box** is used as the ground-truth overlapping region.
- For the example in the figure, both $\mathbf{B}_a^{j_1}$ and $\mathbf{B}_a^{j_2}$ overlaps $\mathbf{B}_a^{i}$, then the red box $\mathbf{R}^i$ is defined as the overlapping region of instance i.



Method | Region-Branch

- The prediction of the four parameters $\mathbf{R}^i = (x^i_{min}, y^i_{min}, x^i_{max}, y^i_{max})$
  -> The prediction of the corresponding **bitmask**
- A simple pixel-level BCE loss
- Better robustness



Method | Directed Expansion

- The overall loss function of amodal-branch is

$$L^a = \alpha_1^a L_{proj}^a + \alpha_2^a L_{pair}^a + \alpha_3^a L_{con}.$$

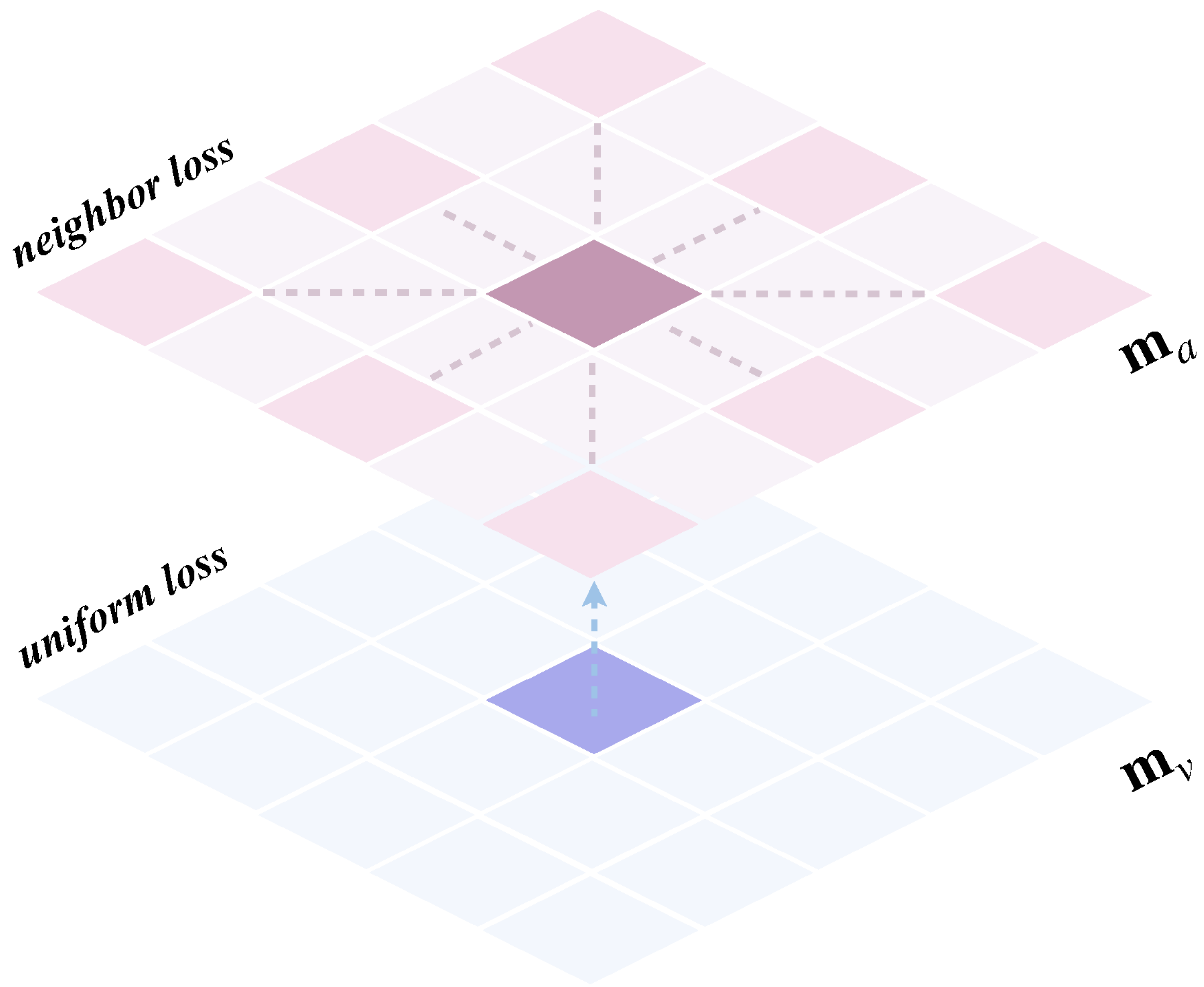
- Utilizing the input $\mathbf{m}_v$ as clues, we introduce a **connectivity loss** L_{con} in it.
- L_{con} is to direct the expansion from predicted visible mask $\mathbf{m}_v$ to predicted amodal mask $\mathbf{m}_a$.

Method | Connectivity Loss

- The connectivity loss contains two terms, namely neighbor loss and uniform loss.

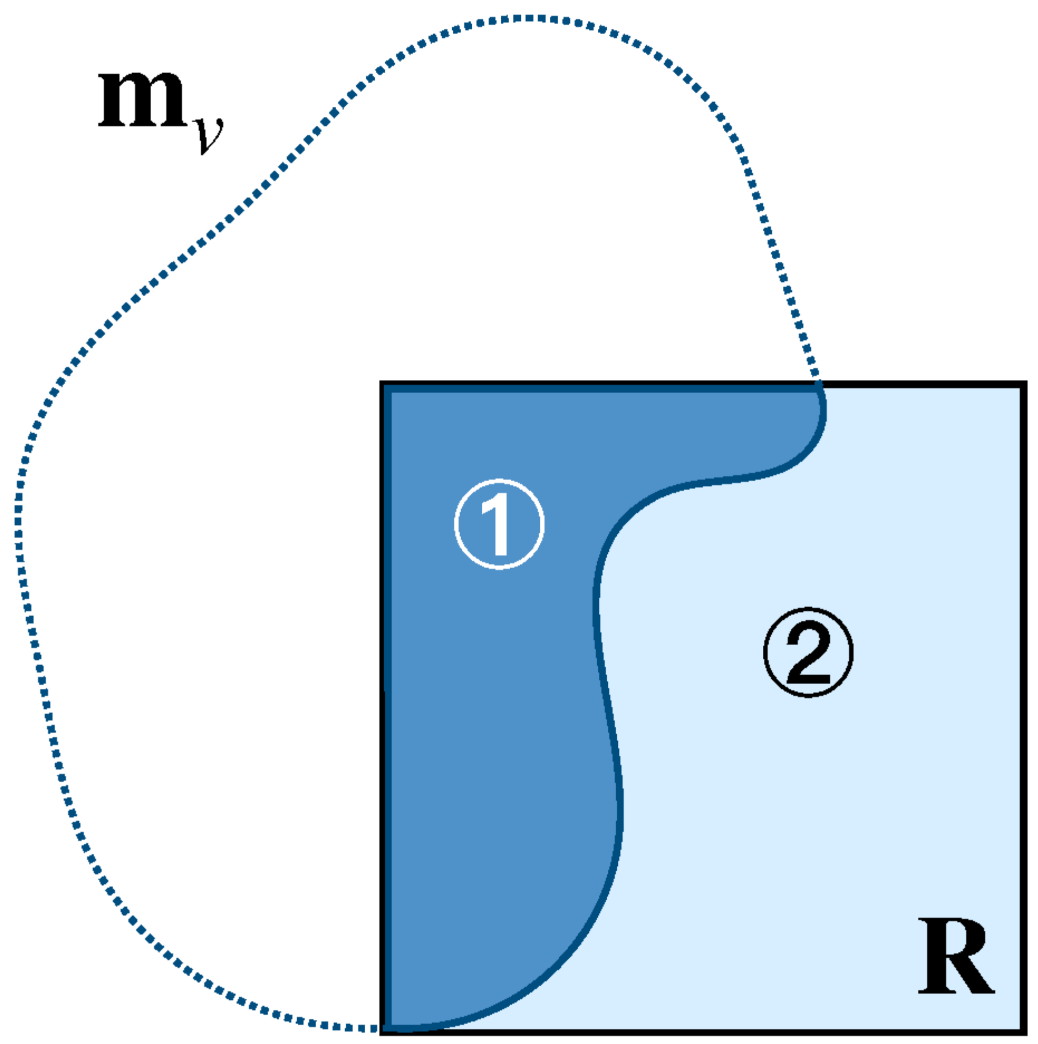
$$L_{con} = l_{ne} + l_{un}$$

- l_{ne}: The label consistency of each pixel with its neighbors in $\mathbf{m}_a$.
- l_{un}: The consistency of corresponding pixels between $\mathbf{m}_a$ and $\mathbf{m}_v$.



Method | Connectivity Loss

- l_{ne} is applied to predicted-overlapping-visible pixels (region ①).
- l_{un} is applied to the whole overlapping region $\mathbf{R}$ (region ①+②).



Method | Neighbor Loss

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA
BLADE

- Consider an undirected graph $G = (V_{pov}, E_{pov})$.
- V_{pov}: The set of predicted-overlapping-visible pixels satisfies

$$\forall (i,j) \in V_{pov}, (i,j) \in \mathbf{R} \ \wedge \ \mathbf{m}_v(i,j) > t.$$

- E_{pov}: The set of edges that connect each pixel with its **eight** neighbors and contain at least one pixel in V_{pov}.
- t: The threshold of the visible-branch.

Method | Neighbor Loss

- For an edge $e = ((i_1, j_1), (i_2, j_2)) \in E_{pov}$, the ground-truth consistency value $c_e = 1$ when the labels of its two endpoints are the same while $c_e = 0$ when the labels are different.
- The predicted consistency value $\tilde{c}_e$ can be defined as

$$\tilde{c}_e = \mathbf{m}_a(i_1, j_1) \cdot \mathbf{m}_a(i_2, j_2) + (1 - \mathbf{m}_a(i_1, j_1)) \cdot (1 - \mathbf{m}_a(i_2, j_2)).$$

- We adopt the BCE loss

$$l_{ne} = -\frac{1}{N_e} \sum_{e \in E_{pov}} c_e \log \tilde{c}_e + (1 - c_e) \log(1 - \tilde{c}_e)$$

to minimize the gap between all $\tilde{c}_e$ and corresponding c_e, where N_e is the number of edges in E_{pov}.

Method | Uniform Loss

- $\mathbf{m}_a(i,j)$: The prediction of the probability that pixel (i,j) belongs to the object.
- $\mathbf{m}_v(i,j)$: The prediction of the probability that pixel (i,j) belongs to the visible portion of the object.
- Therefore, any $\mathbf{m}_a(i,j)$ should NOT be less than $\mathbf{m}_v(i,j)$.
- Observing this, the uniform loss is defined as

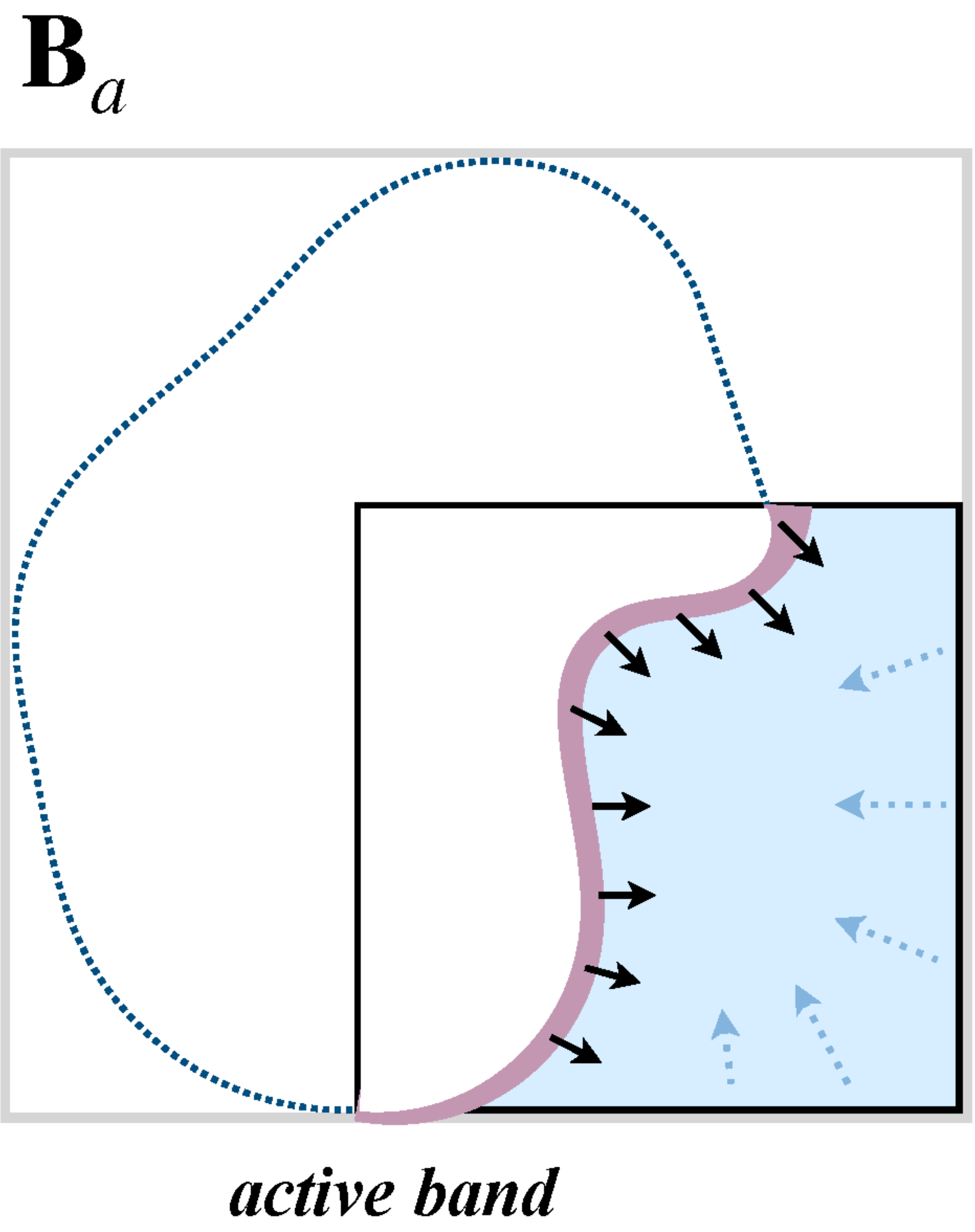
$$l_{un} = \frac{K}{N_{\mathbf{R}}} \sum_{(i,j)\in\mathbf{R}} \max(\mathbf{m}_v(i,j) - \mathbf{m}_a(i,j), 0)$$

to penalize those pixels with reduced values from $\mathbf{m}_v$ to $\mathbf{m}_a$, where $\mathbf{R}$ is the set of pixels in the overlapping region and $N_{\mathbf{R}}$ is the number of these pixels.

Method | Directed Expansion

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA
BLADE

- By introducing the connectivity loss, an active band is built as the initiation of expansion.
- Multiple losses for the amodal-branch reach a balance of encouragement and inhibition of expansion thus directing a **moderate** expansion.



Experiments

- **Datasets:** OccludedVehicles (Wang et al., 2020), KINS (Qi et al., 2019) and COCOA-cls (Follmann et al., 2019).
- **Metric:** Mean intersection-over-union (IoU).
- **Baselines:** BBTP (Hsu et al., 2019), BoxInst (Tian et al., 2021), and Bayesian-Amodal (Sun et al., 2022).

Experiments | Comparison

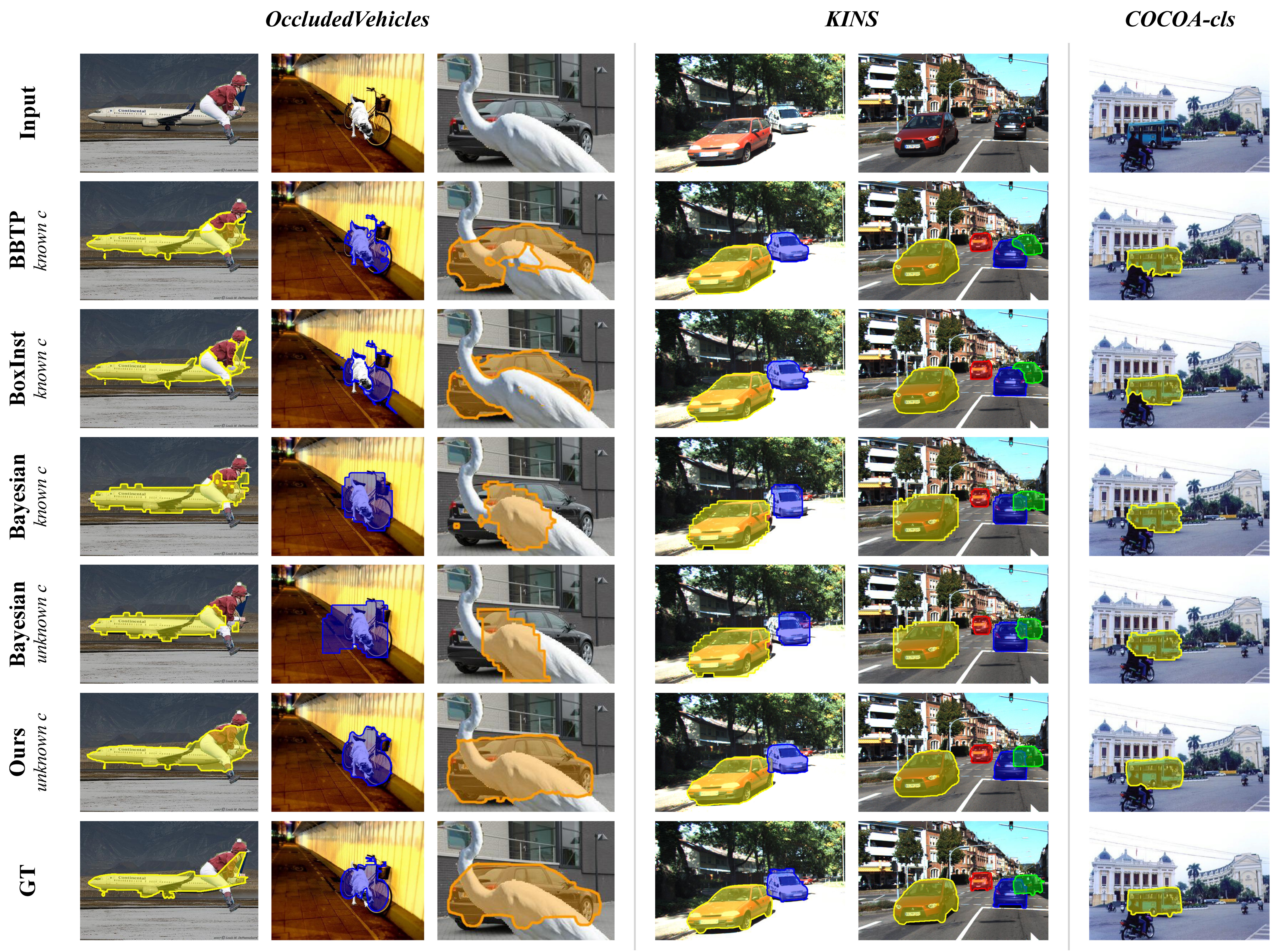
- Our proposed approach **outperforms** existing weakly-supervised methods with large margins and significantly **reduces** the performance gap with fully-supervised methods.

Method	known c	OccludedVehicles										Mean
		FG-0	FG-1			FG-2			FG-3			
		-	BG-1	BG-2	BG-3	BG-1	BG-2	BG-3	BG-1	BG-2	BG-3	
BBTP (Hsu et al. 2019)	Yes	66.5	59.7	58.4	57.9	54.4	51.0	48.9	50.4	44.7	40.2	53.2
BoxInst (Tian et al. 2021)	Yes	72.3	52.5	53.5	53.9	37.7	38.1	38.2	23.0	22.8	23.7	41.6
Bayesian-Amodal (Sun, Kortylewski, and Yuille 2022)	Yes	63.9	59.7	59.6	59.7	57.2	56.8	56.8	55.0	53.9	53.4	57.6
Bayesian-Amodal (Sun, Kortylewski, and Yuille 2022)	No	63.0	59.5	59.5	59.5	56.2	55.9	55.6	51.9	50.6	48.3	56.0
Ours	No	**73.2**	**70.5**	**69.7**	**68.9**	**69.7**	**68.1**	**66.2**	**68.2**	**64.5**	**62.8**	**68.2**

Method	Supervision	known c	KINS					COCOA-cls				
			FG-0	FG-1	FG-2	FG-3	Mean	FG-0	FG-1	FG-2	FG-3	Mean
SAM (ViT-H) (Kirillov et al. 2023)	-	Yes	86.7	75.0	50.8	39.0	62.9	82.7	74.9	59.2	42.3	64.8
VRSP (Xiao et al. 2021)	fully	-	84.7	75.8	74.5	67.1	75.5	82.1	77.7	74.5	72.9	76.8
AISFormer (Tran et al. 2022)	fully	-	85.8	76.4	75.0	69.4	76.7	80.6	76.9	70.9	62.1	72.6
BBTP (Hsu et al. 2019)	weakly	Yes	77.0	68.3	58.9	53.9	64.5	57.3	49.4	40.7	35.0	45.6
BoxInst (Tian et al. 2021)	weakly	Yes	**82.0**	73.3	56.6	43.6	63.9	76.8	67.0	57.2	34.0	58.8
Bayesian-Amodal (Sun, Kortylewski, and Yuille 2022)	weakly	Yes	72.3	69.6	66.2	58.5	66.7	65.3	65.0	64.3	**61.4**	64.0
Bayesian-Amodal (Sun, Kortylewski, and Yuille 2022)	weakly	No	69.9	68.1	63.2	47.3	62.1	58.3	59.8	58.6	53.5	57.6
Ours	weakly	No	81.6	**74.5**	**73.7**	**63.6**	**73.4**	**80.3**	**76.5**	**69.9**	57.9	**71.2**

Experiments | Comparison

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20–27, 2024 | VANCOUVER, CANADA
BLADE



Experiments | Ablation Study

- On the KINS dataset
- UN: The uniform loss

  NE: The neighbor loss

  FS: The fusion structure
- Small adjustments of the weights in L^a

  -> Certain but not dramatic performance changes
- Our currently selected weights

$$\alpha_1^a = 2.0, \alpha_2^a = 1.0, \alpha_3^a = 1.0$$

achieve good performance.

	UN	NE	FS	FG-0	FG-1	FG-2	FG-3	Mean
1	✓	✓	✓	81.6	74.5	**73.7**	**63.6**	**73.4**
2		✓	✓	81.6	**74.9**	70.2	57.3	71.0
3	✓		✓	82.8	73.2	56.8	41.6	63.6
4			✓	**82.9**	72.8	56.7	40.3	63.2
5	✓	✓		76.6	66.7	63.9	56.2	65.9
6		✓		77.3	66.9	62.9	53.2	65.1
7	✓			82.3	74.2	60.0	44.7	65.3
8				82.2	73.1	56.2	40.0	62.9

α_1^a	α_2^a	α_3^a	FG-0	FG-1	FG-2	FG-3	Mean
1.0	1.0	1.0	81.3	72.3	69.6	62.1	71.3
2.0	**1.0**	**1.0**	81.6	74.5	73.7	63.6	**73.4**
1.0	2.0	1.0	82.0	73.4	72.7	64.8	73.2
1.0	1.0	2.0	79.9	71.6	68.3	60.1	70.0

Summary

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20–27, 2024 | VANCOUVER, CANADA
BLADE

- **Problem:** Box-level supervised amodal segmentation
- **Key Idea: Directed expansion**
  - A structure of multi-branch fusion based on the overlapping region
  - Conservative strategy and expansion-encouraged strategy
  - A connectivity loss for reasonable expansion
- **Results:** Our method significantly outperforms current methods

Thanks!

BLADE:

Box-Level Supervised Amodal Segmentation through Directed Expansion

Zhaochen Liu, Zhixuan Li, Tingting Jiang

{dreamerliu, ttjiang}@pku.edu.cn, zhixuanli520@gmail.com

The 38th Annual AAAI Conference on Artificial Intelligence

FEBRUARY 20-27, 2024 | VANCOUVER, CANADA