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

Abstract

Perceiving the complete shape of occluded objects is essential for human and machine intelligence. While the amodal segmentation task is to predict the complete mask of partially occluded objects, it is time-consuming and labor-intensive to annotate the pixel-level ground truth amodal masks. Box-level supervised amodal segmentation addresses this challenge by relying solely on ground truth bounding boxes and instance classes as supervision, thereby alleviating the need for exhaustive pixel-level annotations. Nevertheless, current box-level methodologies encounter limitations in generating low-resolution masks and imprecise boundaries, failing to meet the demands of practical real-world applications. We present a novel solution to tackle this problem by introducing a directed expansion approach from visible masks to corresponding amodal masks. Our approach involves a hybrid end-to-end network based on the overlapping region - the area where different instances intersect. Diverse segmentation strategies are applied for overlapping regions and non-overlapping regions according to distinct characteristics. To guide the expansion of visible masks, we introduce an elaborately-designed connectivity loss for overlapping regions, which leverages correlations with visible masks and facilitates accurate amodal segmentation. Experiments are conducted on several challenging datasets and the results show that our proposed method can outperform existing state-of-the-art methods with large margins.

Introduction

Amodal perception is a vital ability of human's cognitive system (Nanay 2018; Kanizsa, Legrenzi, and Bozzi 1979) for inferring the complete shape of occluded objects easily. Amodal perception shows essential potential for tremendous real-world applications including autonomous driving (Qi et al. 2019; Breitenstein and Fingscheidt 2022), robotic gripping (Wada et al. 2018; Wada, Okada, and Inaba 2019) and novel view synthesis (Li et al. 2022; Gkitsas et al. 2021). For example, deducting the complete shape and range from the visible region of target objects (pedestrians or vehicles) is critical for accurate object recognition and routine planning in self-driving (Ao, Ke, and Ehinger 2023).

Figure 1: An illustration of the overlapping region. The overlapping region of an object is the tightest bounding box that covers all intersecting areas of its amodal bounding box and those of other objects, so the occluded portion of each object should be inside if exists.

In computer vision, amodal instance segmentation has aroused broad concern since it was proposed in AIS (Li and Malik 2016), which aims to predict complete shapes of partially occluded objects. 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 mitigate the challenges of pixel-level annotation, Bayesian-Amodal (Sun, Kortylewski, and Yuille 2022), a weakly supervised approach is proposed that utilizes ground-truth bounding boxes as an alternative supervision signal instead of the intricate ground-truth amodal masks. This method employs a Bayesian model to effectively address the amodal segmentation problem. Nevertheless, the amodal mask generated by the Bayesian-Amodal approach exhibits low resolution and uneven boundaries. This outcome arises from the Bayesian model's inherent coarse shape priors under box-level supervision, which inadequately align with the demands of real-world applications.

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 amodal segmen-

[*]These authors contributed equally.
[†]Corresponding author.

tation method. The key insight of BLADE is to first deduct the visible mask from the detected bounding box, and then expand it to the amodal mask with the guidance of the correlation between the two masks. The correlation reflects the resemblance and distinction between visible and occluded regions in terms of shape and appearance, thus indicating the direction and extent of the expansion. Compared with expanding in a naively unguided manner, the proposed approach can provide more explicit guidance, easing the network's learning burden and contributing to more accurate amodal mask prediction. To depict the aforementioned correlation, a new hybrid end-to-end network is proposed based on the *overlapping region*. As shown in Fig. 1, the overlapping region reveals the intersection between an instance and others, encompassing the occluded portion of the instance. And expansion-encouraged and relatively conservative strategies are designed for overlapping regions and non-overlapping regions, respectively.

Specifically, our method extends the box-level supervised instance segmentation technique introduced in BoxInst (Tian et al. 2021). While BoxInst effectively performs segmentation on non-overlapping regions, we broaden its capabilities to enable amodal segmentation. Our proposed network first predicts the visible mask and corresponding expanded coarse amodal mask, and then fuses them according to the estimated overlapping region. To achieve this, besides the original visible branch contained by BoxInst, we introduce two additional branches, namely amodal-branch and region-branch, to distinguish overlapping regions and non-overlapping regions. The three branches, visible-branch, amodal-branch, and region-branch, predict the visible mask, the coarse amodal mask, and the position of overlapping region, respectively. As for the amodal-branch, observing that within the overlapping region, the visible portion of an object maintains a significant adjacency with its occluded counterpart, we design a connectivity loss to encourage the expansion of the visible segment within the overlapping area towards its encompassing vicinity, finally reaching the coverage of occluded segments through the cooperation with other losses. As for the visible-branch, a general segmentation approach is adopted in which there is no expansion-encouraged factor. The final output amodal prediction is determined by combining the outcomes of three branches. Concretely, for the overlapping region, the segmentation result is from the amodal-branch, while in the remaining areas the segmentation result is from the visible-branch.

We have conducted experiments on three challenging datasets, including OccludedVehicles (Wang et al. 2020), KINS (Qi et al. 2019) and COCOA-cls (Follmann et al. 2019). The results show that our proposed approach outperforms existing weakly-supervised methods with large margins and significantly reduces the performance gap with fully-supervised methods. Our contributions can be summarized as follows:

- We introduce a novel hybrid end-to-end network utilizing the overlapping region. This approach enables diverse segments of an instance to employ tailored segmentation strategies while facilitating collaborative interaction.
- We propose a novel connectivity loss for the overlapping region, guiding the visible mask to expand towards the amodal mask. This approach leverages the correlation with the visible segment, facilitating the accurate prediction of occluded components.
- Our approach significantly outperforms the existing box-level supervised instance segmentation method, reaching state-of-the-art performance.

Related Work

Amodal segmentation is to predict the shape of both visible parts and occluded parts of partially occluded objects. Malik first proposed the task and provided a synthetic dataset (Li and Malik 2016). Subsequently, the KINS (Qi et al. 2019) and Amodal COCO (Zhu et al. 2017) datasets based on real-world images and human-predicted annotations were built. Many fully supervised approaches have been proposed in this field. In addition to direct optimization methods (Li and Malik 2016; Zhu et al. 2017; Qi et al. 2019), researchers have introduced and utilized depth relationships (Zhang et al. 2019), shape priors (Xiao et al. 2021; Li et al. 2022), compositional models (Wang et al. 2020) and the correlation between visible and occluded segments (Follmann et al. 2019; Ke, Tai, and Tang 2021) to improve performance. And some related work (Ehsani, Mottaghi, and Farhadi 2018; Dhamo, Navab, and Tombari 2019; Ling et al. 2020) achieves amodal segmentation based on amodal perception and completion. However, fully supervised methods face a common problem that data annotation takes considerable time and effort. Especially for real-world images, humans can only figure out the approximate shape and range of the occluded parts of an object based on experience, which may vary due to different annotators and increase system error. Noticing this problem, some researchers proposed self-supervised amodal perception and completion methods (Zhan et al. 2020; Nguyen and Todorovic 2021). Nevertheless, these methods introduce the correlation with occluders, requiring the category of occluders during training and testing, which brings great limitations to practical applications. In contrast, only bounding boxes and class annotations of instances are adopted as supervision in our approach thus avoiding the above issues.

Box-level supervised amodal segmentation was proposed in Bayesian-Amodal (Sun, Kortylewski, and Yuille 2022), which is exactly the setting adopted by our method. Box-level supervised amodal segmentation uses bounding boxes and categories of instances as supervision during training, which tackles time-consuming, labor-intensive, and error-prone pixel-level labeling and better supports larger-scale training. Based on related work (Kortylewski et al. 2020, 2021), Bayesian-Amodal also proposed a method that outperformed alternative weakly supervised methods, which replaced the fully connected classifier in neural networks with a Bayesian generative model of the neural network features. However, the results of this method have a relatively low resolution so we cannot get more detailed information. In this paper, we propose a new hybrid network with better accuracy, in which distinct and proper strategies are adopted

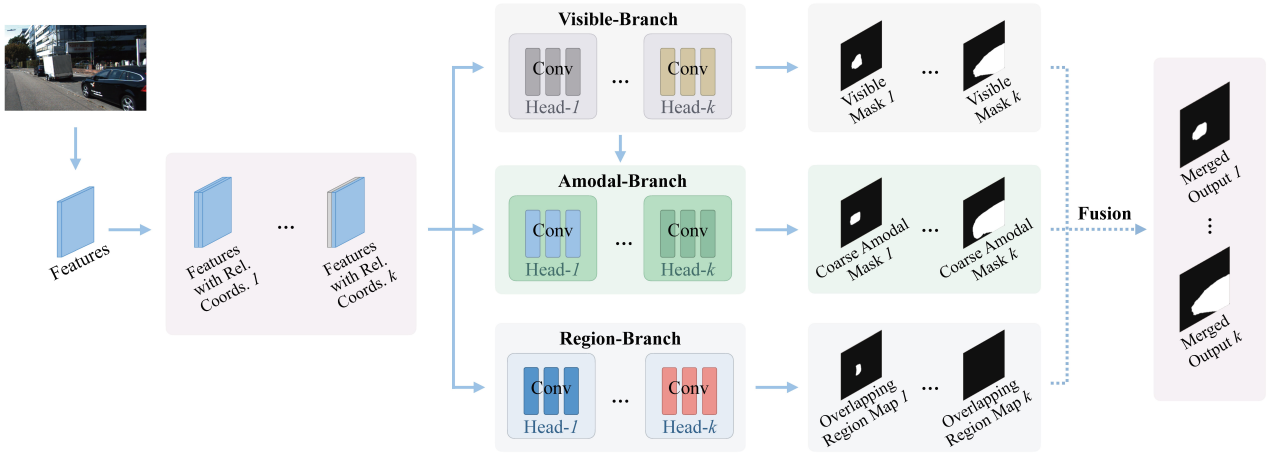

Figure 2: A schematic illustration of the proposed BLADE approach. Extracted features with relative coordinates maps are input to visible-branch, amodal-branch, and region-branch to predict the visible mask, coarse amodal mask, and overlapping region map of each instance respectively, which all adopt dynamically generated instance-aware mask heads. Exploiting the correlation, predicted visible masks are also input to the amodal branch for our proposed connectivity loss that directs the expansion from visible masks to corresponding coarse amodal masks. The final outputs use coarse amodal masks in predicted overlapping regions and visible masks in other regions.

for different regions and the correlation between visible and occluded regions is utilized.

Box-level supervised segmentation is an important and concerned field of computer vision, which comparatively focuses on the segmentation of non-occluded instances. Box-level supervision is weak, so related methods often introduce some observations and priors as assistance. SDI (Khoreva et al. 2017) utilizes the object shape priors. OSIS (Pham et al. 2018) introduces the Bayesian model for better formulation. BBTP (Hsu et al. 2019) exploits the bounding box tightness prior. WSIS (Arun, Jawahar, and Kumar 2020) builds an annotation consistency framework. In (Sun et al. 2020), a two-stage transfer learning framework employing valid generated masks from GrabCut (Rother, Kolmogorov, and Blake 2004) is designed. Recently, BoxInst (Tian et al. 2021) raises an observation that proximal pixels with similar colors tend to have the same label, and proposes a concise and effective method incorporating a projection loss and a pairwise loss. BoxInst achieves leading performance and greatly narrows the performance gap between weakly and fully supervised instance segmentation. However, these methods do not perform well if directly transferred to an amodal segmentation task. In our approach, a novel connectivity loss encourages the mask to expand from the visible region towards the occluded region is designed and helps achieve the goal of amodal segmentation.

Methodology Design

Problem Definition

Amodal segmentation aims to predict the amodal mask $\mathbf{M}_a$ for a given image $\mathbf{I}$ and some region-of-interest (Xiao et al. 2021). In this paper, we follow the task setting of box-level supervised amodal segmentation and take only bounding box annotations and class labels as the supervision signal.

Specifically, we take ground-truth $\mathbf{B}_v$ (the bounding box of visible portion), $\mathbf{B}_a$ (the bounding box of the complete object), and $\mathbf{c}$ (the class label) of each instance as the supervision for training. During the test, we take only $\mathbf{B}_{test}$, the object bounding box of the visible area, as input to determine the region of interest and match the corresponding instance. The network ultimately outputs an amodal mask $\mathbf{m}$ of the instance.

Network Architecture

Apparently, $\mathbf{M}_a$ can be decomposed as

$$\mathbf{M}_a = \mathbf{M}_v + \mathbf{M}_o, \tag{1}$$

where $\mathbf{M}_v$ and $\mathbf{M}_o$ represent the visible mask and the occluded mask respectively. The prediction of $\mathbf{M}_v$ can exactly be formulated as a less complicated general segmentation problem and provide the estimation of $\mathbf{M}_o$ with clues.

Inspired by this, we design a hybrid structure with multiple branches: as shown in Fig. 2, our network consists of visible-branch, amodal-branch, and region-branch, which are used to predict the visible mask $\mathbf{m}_v$, the coarse amodal mask $\mathbf{m}_a$, and the map of the overlapping region $\mathbf{m}_r$ respectively. The final output $\mathbf{m}$ fuses these three items, using $\mathbf{m}_a$ in the predicted overlapping region while using $\mathbf{m}_v$ in the other region. That is to say

$$\mathbf{m} = \mathbf{m}_a * \mathbf{m}_r + \mathbf{m}_v * \bar{\mathbf{m}}_r, \tag{2}$$

where $\bar{\mathbf{m}}_r$ is inverse $\mathbf{m}_r$ indicates non-overlapping region.

Our proposed network is developed on the recent box-level supervised instance segmentation method BoxInst (Tian et al. 2021). For the visible-branch, $\mathbf{B}_v$ annotations are applied as the supervision. We directly use the original mask heads with projection loss and pairwise loss in BoxInst which are set up just for our demand. For

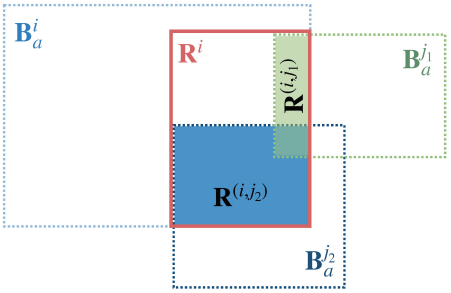

Figure 3: If there are multiple intersecting areas, the envelope box is used as the ground-truth overlapping region. For the example in the figure, both $\mathbf{B}_a^{j_1}$ and $\mathbf{B}_a^{j_2}$ overlaps $\mathbf{B}_a^i$, then the red box $\mathbf{R}^i$ is defined as the overlapping region of instance i.

the amodal-branch, $\mathbf{B}_a$ annotations are applied as the supervision. We feed the amodal-branch the predicted $\mathbf{m}_v$ from visible-branch in addition to the features and relative coordinates. Utilizing the input $\mathbf{m}_v$ as clues, we introduce a connectivity loss to direct the expansion from $\mathbf{m}_v$ to $\mathbf{m}_a$. Meanwhile, projection loss and pairwise loss are also used to limit the range of expansion. For the region-branch, we transform the prediction of the four parameters of some ground truth bounding box into the prediction of the corresponding bitmask to improve robustness while using a simple pixel-level BCE loss. The above three branches share the same multi-scale features extracted from the image and all adopt dynamically-generated instance-aware mask heads containing varying instance-by-instance parameters (Jia et al. 2016), thereby significantly increasing the flexibility and decreasing the amount of parameters. The entire network is trained together, so the multiple branches form joint supervision, in which the correlation between the visible region and the occluded region is implied and mutual assistance is built.

Overlapping Region

As shown in Fig. 1, the overlapping region is the area where different instances overlap. Overlapping and occlusion are two accompanying phenomena: if an object is occluded, it must overlap another object, which means the overlapping region contains the occluded region. Therefore, We can use the easily accessible overlapping region as an approximate estimation of the corresponding occluded region to determine where the expanded amodal mask $\mathbf{m}_a$ is applied.

Under the setting of box-level supervision, the overlapping region of each instance is recorded by one bounding box. We use the intersection between the amodal bounding box of each instance and those of other instances within the same image, obviating the necessity for introducing additional annotations. Suppose there are n instances in an image, and their amodal bounding boxes are

$$\mathbf{B}_a^i = (x_{min}^i, y_{min}^i, x_{max}^i, y_{max}^i),\ i = 1, ..., n, \tag{3}$$

then for instance i, the overlapping region with instance j is

$$\mathbf{R}^{(i,j)} = (x_{min}^{(i,j)}, y_{min}^{(i,j)}, x_{max}^{(i,j)}, y_{max}^{(i,j)}) \tag{4}$$

if exists, where

$$\begin{aligned} x_{min}^{(i,j)} &= \max(x_{min}^i, x_{min}^j),\ y_{min}^{(i,j)} = \max(y_{min}^i, y_{min}^j), \\ x_{max}^{(i,j)} &= \min(x_{max}^i, x_{max}^j),\ y_{max}^{(i,j)}) = \min(y_{max}^i, y_{max}^j). \end{aligned} \tag{5}$$

As shown in Fig. 3, we take the envelope box of all existing overlapping region $\mathbf{R}^{(i,j)}$ as the ground-truth overlapping region of instance i, which is

$$\begin{aligned} \mathbf{R}^i = (&\min_{j \in \mathbf{V}_i}\{x_{min}^{(i,j)}\}, \min_{j \in \mathbf{V}_i}\{y_{min}^{(i,j)}\}, \\ &\max_{j \in \mathbf{V}_i}\{x_{max}^{(i,j)}\}, \max_{j \in \mathbf{V}_i}\{y_{max}^{(i,j)}\}), \end{aligned} \tag{6}$$

where $\mathbf{V}_i$ is the indexes of all instances that possess the valid overlapping region with instance i. During training and testing, we actually use the corresponding bitmask instead of the four-parameter representation, in which the pixel value is 1 if within the overlapping region and is 0 if outside. A pixel-level BCE loss

$$L^r = -\frac{\alpha^r}{N} \sum_{\mathbf{m}_r} p \log \tilde{p} + (1-p) \log(1-\tilde{p}) \tag{7}$$

is applied for region-branch accordingly, where $p, \tilde{p}$ is the ground truth value and the predicted value of some pixel, N is the total number of pixels and α^r is a constant coefficient.

Connectivity Loss

As mentioned above, the overall loss function of amodal-branch can be written as

$$L^a = \alpha_1^a L_{proj}^a + \alpha_2^a L_{pair}^a + \alpha_3^a L_{con}, \tag{8}$$

where α_i^a is the constant weight of each term. As shown in Fig. 4, the newly-introduced connectivity loss consists of two parts

$$L_{con} = l_{ne} + l_{un}, \tag{9}$$

namely neighbor loss and uniform loss.

The neighbor loss is applied to predicted-overlapping-visible pixels (pixels in the overlapping region that are predicted to belong to the visible mask $\mathbf{m}_v$ in visible-branch) in the predicted coarse amodal mask $\mathbf{m}_a$, which measures the label consistency of each pixel with its neighbors. We design this loss based on an observation that in the overlapping region, the visible segment of an instance is mostly adjacent to its occluded counterpart. Consider an undirected graph $G = (V_{pov}, E_{pov})$. V_{pov} is the set of predicted-overlapping-visible pixels satisfies

$$\forall (i,j) \in V_{pov}, (i,j) \in \mathbf{R} \ \wedge\ \mathbf{m}_v(i,j) > t, \tag{10}$$

where $\mathbf{m}_a(i,j)$ is the value of predicted coarse amodal mask $\mathbf{m}_a$ at position (i,j), t is the threshold of the visible-branch, and E_{pov} is the set of edges that connect each pixel with its eight neighbors and contain at least one pixel in V_{pov}. 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

$$(i_1,j_1),(i_2,j_2) \in \mathbf{B}_a \vee (i_1,j_1),(i_2,j_2) \notin \mathbf{B}_a, \tag{11}$$

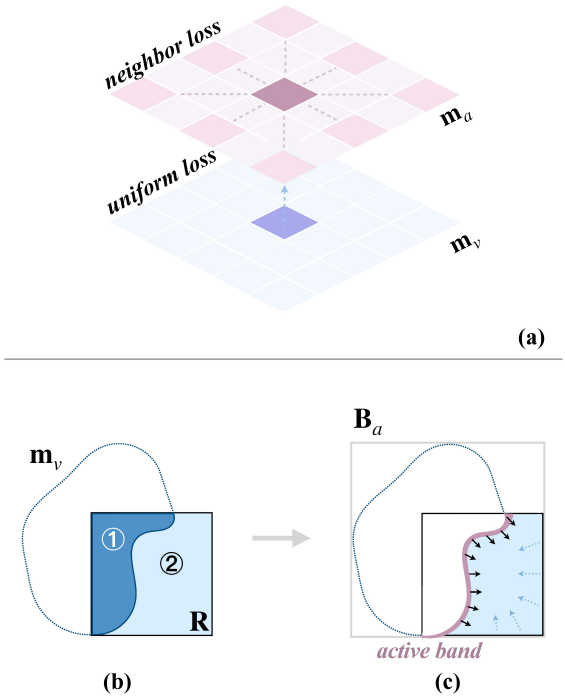

Figure 4: An illustration of the connectivity loss. (a) The connectivity loss contains two terms, namely neighbor loss and uniform loss. The neighbor loss measures the label consistency of each pixel with its neighbors in $\mathbf{m}_a$, while the uniform loss reflects the consistency of corresponding pixels between $\mathbf{m}_a$ and $\mathbf{m}_v$. (b) The neighbor loss is applied to predicted-overlapping-visible pixels (region ①), while the uniform loss is applied to the whole overlapping region $\mathbf{R}$ (region ①+②). (c) By the action of 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.

while $c_e = 0$ when the labels are different. The predicted consistency value $\tilde{c}_e$ can be defined as

$$\begin{aligned}\tilde{c_e} =&\mathbf{m}_a(i_1,j_1) \cdot \mathbf{m}_a(i_2,j_2) + \\ &(1-\mathbf{m}_a(i_1,j_1)) \cdot (1-\mathbf{m}_a(i_2,j_2)),\end{aligned} \tag{12}$$

which is positively correlated with the consistency and certainty. We adopt the frequently-used BCE loss

$$l_{ne} = -\frac{1}{N_e}\sum_{e\in E_{pov}} c_e \log \tilde{c}_e + (1-c_e)\log(1-\tilde{c}_e) \tag{13}$$

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

The uniform loss is applied to the whole overlapping region, which ensures the consistency between the predicted coarse amodal mask $\mathbf{m}_a$ and the predicted visible mask $\mathbf{m}_v$. The values $\mathbf{m}_a(i,j)$ and $\mathbf{m}_v(i,j)$ can be regarded as predictions of the probability that pixel (i,j) belongs to the object and the probability it belongs to the visible portion of the object, respectively. Therefore, it's obvious that any $\mathbf{m}_a(i,j)$ should NOT be less than $\mathbf{m}_v(i,j)$. Observing this, the uniform loss is defined as

$$l_{un} = \frac{K}{N_{\mathbf{R}}}\sum_{(i,j)\in\mathbf{R}} \max(\mathbf{m}_v(i,j) - \mathbf{m}_a(i,j), 0) \tag{14}$$

to penalize those pixels with reduced values from $\mathbf{m}_v$ to $\mathbf{m}_a$, where K is a constant coefficient used to regulate the severity of the penalty, $\mathbf{R}$ is the set of pixels in the overlapping region and $N_{\mathbf{R}}$ is the number of these pixels.

By introducing the connectivity loss, we build an active band in the predicted coarse amodal mask $\mathbf{m}_a$ containing the neighbors of predicted-overlapping-visible pixels near edges, which tends to increase the value to share the same label as pixels in $\mathbf{m}_v$. We use neighbors with a one-pixel gap to the center pixel to increase the width of the active band, which can be adjusted as needed. The active band will further expand outward within a certain range until reaching a balanced state due to the existence of pairwise loss and projection loss. Intuitively, diverse losses adopted in amodal-branch form an antagonistic effect of encouragement and inhibition of expansion, resulting in both under-expansion and over-expansion being largely avoided.

Experiments

Datasets and Metric

Datasets. Our experiments are conducted on three challenging datasets for amodal segmentation, including OccludedVehicles (Wang et al. 2020), KINS (Qi et al. 2019) and COCOA-cls (Follmann et al. 2019).

The OccludedVehicles dataset is a synthetic dataset based on PASCAL3D+ (Xiang, Mottaghi, and Savarese 2014) in which occluders are pasted randomly. It consists of 51801 objects that are evenly distributed among three foreground occlusion levels, FG-1, FG-2, FG-3, three background occlusion levels BG-1, BG-2, BG-3, and a non-occluded level FG-0. The three foreground occlusion levels correspond to 20-40%, 40-60%, and 60-80% of the object being occluded, while the three background occlusion levels correspond to 1-20%, 20-40%, and 40-60% of the context being occluded.

The KINS dataset is based on real-world images with real occlusion. We follow the experimental settings of Bayesian-Amodal (Sun, Kortylewski, and Yuille 2022) that restricts the scope of evaluation to vehicles with a height greater than 50 pixels and divides the objects into four occlusion levels FG-0, FG-1, FG-2, FG-3, which correspond to 0%, 1-30%, 30-60%, and 60-90% of the object being occluded. The subset we adopt consists of 14826 objects.

The COCOA-cls dataset is an extension of the real occlusion dataset Amodal COCO (Zhu et al. 2017) and consists of 766 objects. It's divided into four foreground occlusion levels FG-0, FG-1, FG-2, and FG-3, which correspond to 0%, 1-20%, 20-40%, and 40-70% of the object being occluded.

Metric. As for the metric of evaluation, we adopt mean Intersection-over-Union (IoU) like related work (Sun, Kortylewski, and Yuille 2022). IoU calculates the ratio of intersecting pixels between the predicted amodal mask and corresponding ground truth amodal mask to their union, therefore a larger value indicates a more accurate segmentation result.

Method	known c	OccludedVehicles										
		FG-0	FG-1			FG-2			FG-3			Mean
		-	BG-1	BG-2	BG-3	BG-1	BG-2	BG-3	BG-1	BG-2	BG-3	
BBTP	Yes	66.5	59.7	58.4	57.9	54.4	51.0	48.9	50.4	44.7	40.2	53.2
BoxInst	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	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	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)	-	Yes	86.7	75.0	50.8	39.0	62.9	82.7	74.9	59.2	42.3	64.8
VRSP	fully	-	84.7	75.8	74.5	67.1	75.5	82.1	77.7	74.5	72.9	76.8
AISFormer	fully	-	85.8	76.4	75.0	69.4	76.7	80.6	76.9	70.9	62.1	72.6
BBTP	weakly	Yes	77.0	68.3	58.9	53.9	64.5	57.3	49.4	40.7	35.0	45.6
BoxInst	weakly	Yes	**82.0**	73.3	56.6	43.6	63.9	76.8	67.0	57.2	34.0	58.8
Bayesian-Amodal	weakly	Yes	72.3	69.6	66.2	58.5	66.7	65.3	65.0	64.3	**61.4**	64.0
Bayesian-Amodal	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**

Table 1: The comparison of amodal segmentation performance on the synthetic-occlusion OccludedVehicles dataset and the real-occlusion KINS and COCOA-cls datasets.

Implementation Details

We implement our model based on BoxInst (Tian et al. 2021) and Detectron2 (Wu et al. 2019) on the PyTorch framework (Paszke et al. 2019). We use an FCOS module (Tian et al. 2019; Tian, Shen, and Chen 2020) to detect objects and a ResNet-50-FPN backbone (He et al. 2016; Lin et al. 2017) to extract multi-scale features for subsequent processes. In the training, we use data at all occlusion levels and choose $\alpha_1^a = 2.0, \alpha_2^a = 1.0, \alpha_3^a = 1.0$. We conduct 60000 iterations with a batch size of 6, during which we adopt a 3-stage learning rate of 0.01 in the first 40000 iterations, 0.001 in the 40000-54000 iterations, and 0.0001 in the 54000-60000 iterations. The training is completed on 3 NVIDIA GeForce RTX 2080Ti GPUs taking about 5 hours per time. In the testing, we evaluate objects at each occlusion level separately to obtain the mean IoU at each level, and then use their average value as the mean IoU of the entire test set.

Comparison with Existing Methods

Baselines. We benchmark our proposed BLADE method against BBTP (Hsu et al. 2019), BoxInst (Tian et al. 2021), and Bayesian-Amodal (Sun, Kortylewski, and Yuille 2022). These baselines are all under box-level supervision, among which BBTP and BoxInst are the two best-performing methods for box-level supervised general segmentation, and Bayesian-Amodal is a state-of-the-art approach for box-level supervised amodal segmentation. For the convenience of practical applications, our method only requires the bounding box $\mathbf{B}_{test}$ of the visible portion as the input, which is supported by our expansion-based path. But other methods may require inputting the bounding box of the entire object to know the object center c. Bayesian-Amodal provides multiple options, among which we choose both the end-to-end known-c model and the end-to-end unknown-c model for comparison. Designed for normal segmentation, BBTP and BoxInst can only segment the visible parts without knowing c but cannot predict the amodal masks. To compare their performance of amodal segmentation with our method, we adopt the known-c setting for them.

Synthetic Occlusion. As shown in Table 1, our proposed method significantly outperforms existing methods on the OccludedVehicles dataset. Our model performs best at all occlusion levels. Especially in levels with high occlusion ratios, our method shows remarkable advantages. Compared with Bayesian-Amodal (unknown c), our model improves by more than 12% in mean IoU.

Real Occlusion. As shown in Table 1, our method also performs well on the KINS dataset which is based on real-world images. Although our method is weakly-supervised, which is not comparable to fully-supervised methods, we provide some results in Table 1 for better evaluation including SAM (Kirillov et al. 2023), VRSP (Xiao et al. 2021), and AISFormer (Tran et al. 2022). Among all these methods, our approach achieves very competitive mean performance. Compared with Bayesian-Amodal (unknown c), our model improves by over 11% in mean IoU.

Transferability. Since the number of objects in the COCOA-cls dataset is quite small, we transfer the model trained on the OccludedVehicles dataset to the COCOA-cls dataset, and use the union of the training set and the test set of COCOA-cls for evaluation. As shown in Table 1, our model reflects better transferability than others. Compared with Bayesian-Amodal (unknown c), our model improves by more than 13% in mean IoU.

Some qualitative results are also shown in Fig. 5, which indicate that masks predicted by our method possess higher resolution and accuracy than others. Our predictions cover more complete occluded segments than BBTP and BoxInst (such as the car in the 3rd column), while also exhibiting more precise and smooth boundaries than Bayesian-Amodal (such as the aeroplane in the 1st column).

Ablation Study

We conduct the ablation study on the KINS dataset, the result of which is shown in Table 2. In these experiments, we all adopt the unknown c setting.

The Effect of Fusion Structure. To verify the effectiveness of the fusion structure, we cancel the region-branch and the structure of fusing multiple branches' results in the baselines from the 4th row to the 8th row, directly taking the prediction of the amodal-branch as output. Compared with these experimental results, the corresponding results with the fusion structure show significantly better performance.

The Effect of Neighbor Loss. To validate the importance of the neighbor loss, we conduct the experiments at the 3rd and 4th rows. Compared with the results at the 1st and 2nd rows, their performance show a notable drop, especially at levels with high occlusion ratios.

The Effect of Uniform Loss. To evaluate the effect of the uniform loss, we introduce the baselines at the 2nd and 4th rows which have no uniform loss. Consequently, a gap emerges between the performance of them and the performance of corresponding models with the uniform loss.

The Effect of Different Weights in L^a. Small adjustments of the weights in L^a will result in certain but not dramatic performance changes, and our selected weights $\alpha_1^a = 2.0, \alpha_2^a = 1.0, \alpha_3^a = 1.0$ achieve good performance.

Conclusion

In this work, we achieve box-level supervised amodal segmentation through directed expansion. Our approach introduces a fusion structure based on the overlapping region. Conservative strategy and expansion-encouraged strategy are applied to non-overlapping regions and overlapping regions, respectively. Utilizing predicted visible masks as clues, a connectivity loss is incorporated for reasonable expansion. Experimental results indicate our method significantly outperforms other state-of-the-art methods.

	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
a	1.0	1.0	1.0	81.3	72.3	69.6	62.1	71.3
b	**2.0**	**1.0**	**1.0**	81.6	74.5	73.7	63.6	**73.4**
c	1.0	2.0	1.0	82.0	73.4	72.7	64.8	73.2
d	1.0	1.0	2.0	79.9	71.6	68.3	60.1	70.0

Table 2: The ablation study on the KINS dataset. UN, NE, and FS respectively represent the uniform loss, the neighbor loss, and the fusion structure. The fusion structure indicates whether to adopt L^r and the corresponding region-branch.

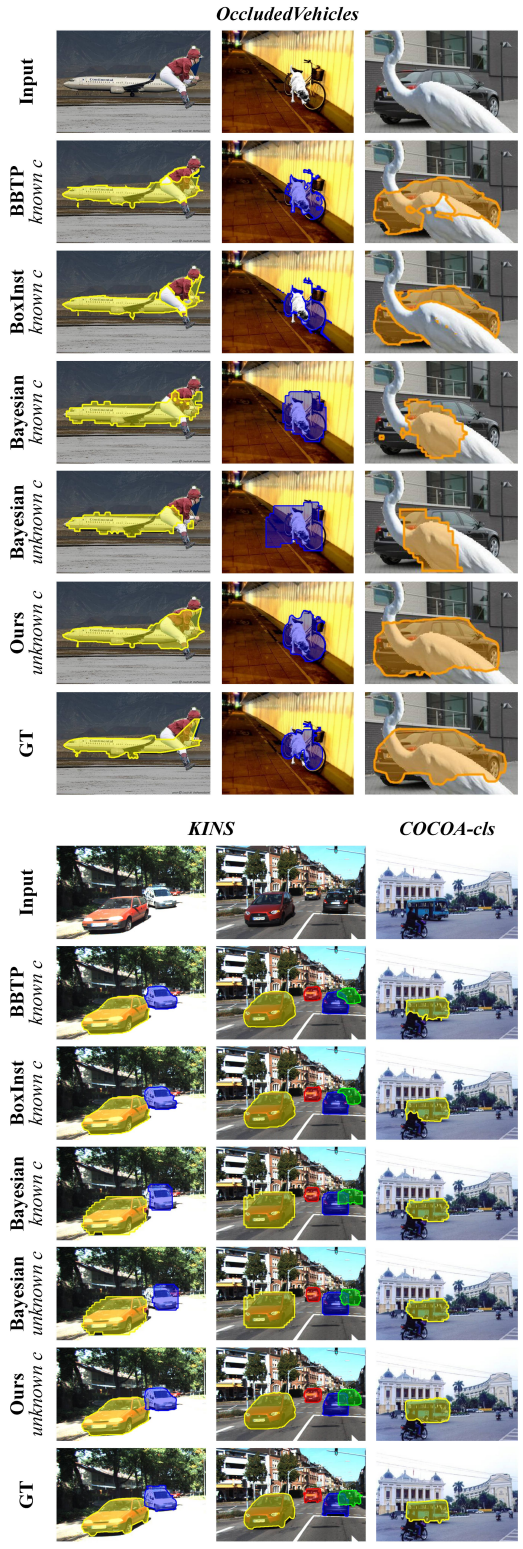

Figure 5: Qualitative examples of our approach compared with corresponding ground-truth amodal masks and estimations of BBTP, BoxInst, Bayesian-Amodal (both the known-c model and the unknown-c model). Zoom in for a better view.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China under contract 62088102. This work was also partially supported by Qualcomm. We also acknowledge High-Performance Computing Platform of Peking University for providing computational resources.

References

Ao, J.; Ke, Q.; and Ehinger, K. A. 2023. Image amodal completion: A survey. *Computer Vision and Image Understanding*, 103661.

Arun, A.; Jawahar, C.; and Kumar, M. P. 2020. Weakly supervised instance segmentation by learning annotation consistent instances. In *Proceedings of the European Conference on Computer Vision*, 254–270.

Breitenstein, J.; and Fingscheidt, T. 2022. Amodal cityscapes: a new dataset, its generation, and an amodal semantic segmentation challenge baseline. In *IEEE Intelligent Vehicles Symposium*, 1018–1025.

Dhamo, H.; Navab, N.; and Tombari, F. 2019. Object-driven multi-layer scene decomposition from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5369–5378.

Ehsani, K.; Mottaghi, R.; and Farhadi, A. 2018. SeGAN: Segmenting and generating the invisible. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6144–6153.

Follmann, P.; König, R.; Härtinger, P.; Klostermann, M.; and Böttger, T. 2019. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 1328–1336.

Gkitsas, V.; Sterzentsenko, V.; Zioulis, N.; Albanis, G.; and Zarpalas, D. 2021. Panodr: Spherical panorama diminished reality for indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3716–3726.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Hsu, C.-C.; Hsu, K.-J.; Tsai, C.-C.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in Neural Information Processing Systems*, 32: 6586–6597.

Jia, X.; De Brabandere, B.; Tuytelaars, T.; and Gool, L. V. 2016. Dynamic filter networks. *Advances in Neural Information Processing Systems*, 29: 667–675.

Kanizsa, G.; Legrenzi, P.; and Bozzi, P. 1979. Organization in vision: Essays on Gestalt perception. *Praeger Publishers*.

Ke, L.; Tai, Y.-W.; and Tang, C.-K. 2021. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4019–4028.

Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; and Schiele, B. 2017. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 876–885.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kortylewski, A.; He, J.; Liu, Q.; and Yuille, A. L. 2020. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8940–8949.

Kortylewski, A.; Liu, Q.; Wang, A.; Sun, Y.; and Yuille, A. 2021. Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *International Journal of Computer Vision*, 129: 736–760.

Li, K.; and Malik, J. 2016. Amodal instance segmentation. In *Proceedings of the European Conference on Computer Vision*, 677–693.

Li, Z.; Ye, W.; Jiang, T.; and Huang, T. 2022. 2D amodal instance segmentation guided by 3D shape prior. In *Proceedings of the European Conference on Computer Vision*, 165–181.

Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ling, H.; Acuna, D.; Kreis, K.; Kim, S. W.; and Fidler, S. 2020. Variational amodal object completion. *Advances in Neural Information Processing Systems*, 33: 16246–16257.

Nanay, B. 2018. The importance of amodal completion in everyday perception. *i-Perception*, 9(4).

Nguyen, K.; and Todorovic, S. 2021. A weakly supervised amodal segmenter with boundary uncertainty estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7396–7405.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32: 8026–8037.

Pham, T.; Do, T.-T.; Carneiro, G.; Reid, I.; et al. 2018. Bayesian semantic instance segmentation in open set world. In *Proceedings of the European Conference on Computer Vision*, 3–18.

Qi, L.; Jiang, L.; Liu, S.; Shen, X.; and Jia, J. 2019. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3014–3023.

Rother, C.; Kolmogorov, V.; and Blake, A. 2004. "GrabCut" interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3): 309–314.

Sun, Y.; Kortylewski, A.; and Yuille, A. 2022. Amodal segmentation through out-of-task and out-of-distribution generalization with a Bayesian model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1215–1224.

Sun, Y.; Liao, S.; Gao, C.; Xie, C.; Yang, F.; Zhao, Y.; and Sagata, A. 2020. Weakly supervised instance segmentation based on two-stage transfer learning. *IEEE Access*, 8: 24135–24144.

Tian, Z.; Shen, C.; and Chen, H. 2020. Conditional convolutions for instance segmentation. In *Proceedings of the European Conference on Computer Vision*, 282–298.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9627–9636.

Tian, Z.; Shen, C.; Wang, X.; and Chen, H. 2021. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5443–5452.

Tran, M.; Vo, K.; Yamazaki, K.; Fernandes, A.; Kidd, M.; and Le, N. 2022. AISFormer: Amodal instance segmentation with transformer. In *Proceedings of the British Machine Vision Conference*.

Wada, K.; Kitagawa, S.; Okada, K.; and Inaba, M. 2018. Instance segmentation of visible and occluded regions for finding and picking target from a pile of objects. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2048–2055.

Wada, K.; Okada, K.; and Inaba, M. 2019. Joint learning of instance and semantic segmentation for robotic pick-and-place with heavy occlusions in clutter. In *Proceedings of the International Conference on Robotics and Automation*, 9558–9564.

Wang, A.; Sun, Y.; Kortylewski, A.; and Yuille, A. L. 2020. Robust object detection under occlusion with context-aware compositionalnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12645–12654.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2. Accessed: 2023-03-27.

Xiang, Y.; Mottaghi, R.; and Savarese, S. 2014. Beyond pascal: A benchmark for 3D object detection in the wild. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 75–82.

Xiao, Y.; Xu, Y.; Zhong, Z.; Luo, W.; Li, J.; and Gao, S. 2021. Amodal segmentation based on visible region segmentation and shape prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2995–3003.

Zhan, X.; Pan, X.; Dai, B.; Liu, Z.; Lin, D.; and Loy, C. C. 2020. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3784–3792.

Zhang, Z.; Chen, A.; Xie, L.; Yu, J.; and Gao, S. 2019. Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2124–2132.

Zhu, Y.; Tian, Y.; Metaxas, D.; and Dollár, P. 2017. Semantic amodal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1464–1472.