Box-supervised Instance Segmentation with Level Set Evolution

Wentong Li[1], Wenyu Liu[1], Jianke Zhu[1(✉)], Miaomiao Cui[2], Xiansheng Hua[2], and Lei Zhang[3]

[1]Zhejiang University [2]Alibaba Damo Academy
[3]The Hong Kong Polytechnic University
{liwentong, liuwenyu.lwy, jkzhu}@zju.edu.cn, miaomiao.cmm@alibaba-inc.com, xshua@outlook.com, cslzhang@comp.polyu.edu.hk

Abstract. In contrast to the fully supervised methods using pixel-wise mask labels, box-supervised instance segmentation takes advantage of the simple box annotations, which has recently attracted a lot of research attentions. In this paper, we propose a novel single-shot box-supervised instance segmentation approach, which integrates the classical level set model with deep neural network delicately. Specifically, our proposed method iteratively learns a series of level sets through a continuous Chan-Vese energy-based function in an end-to-end fashion. A simple mask supervised SOLOv2 model is adapted to predict the instance-aware mask map as the level set for each instance. Both the input image and its deep features are employed as the input data to evolve the level set curves, where a box projection function is employed to obtain the initial boundary. By minimizing the fully differentiable energy function, the level set for each instance is iteratively optimized within its corresponding bounding box annotation. The experimental results on four challenging benchmarks demonstrate the leading performance of our proposed approach to robust instance segmentation in various scenarios. The code is available at: https://github.com/LiWentomng/boxlevelset.

Keywords: Instance segmentation, level set, box supervision

1 Introduction

Instance segmentation aims to obtain the pixel-wise labels of the interested object, which plays an important role in many applications, such as autonomous driving and robotic manipulation. Though having achieved promising performance, most of the existing instance segmentation approaches [9, 13, 18, 43, 48] are trained in a supervised manner, which heavily depend on the pixel-wise mask annotations and incur expensive labeling costs.

To deal with this problem, box-supervised instance segmentation takes advantage of the simple box annotation rather than the pixel-wise mask labels, which has recently attracted a lot of research attentions [16, 24–26, 44, 47]. To enable pixel-wise supervision with box annotation, some methods [26, 47] focus on generating the pseudo mask labels by an independent network, which

needs to employ extra auxiliary salient data [47] or post-processing methods like MCG [39] and CRF [23] to obtain precise pseudo labels. Due to the involved multiple separate steps, the training pipeline becomes complicated with many hyper-parameters. Several recent approaches [16, 44] suggest a unified framework using the pairwise affinity modeling, e.g., neighbouring pixel pairs [16] and colour pairs [44], enabling an end-to-end training of the instance segmentation network. The pairwise affinity relationship is defined on the set of partial or all neighbouring pixel pairs, which oversimplifies the assumption that the pixel or colour pairs are encouraged to share the same label. The noisy contexts from the objects and background with similar appearance are inevitably absorbed, leading to inferior instance segmentation performance.

In this paper, we propose a novel single-shot box-supervised instance segmentation approach to address the above limitations. Our approach integrates the classical level set model [7, 37] with deep neural network delicately. Unlike the existing box-supervised methods [16, 25, 26, 44], we iteratively learn a series of level set functions for implicit curve evolution within the annotated bounding box in an end-to-end fashion. Different from fully-supervised level set-based methods [15, 17, 49, 54], our proposed approach is able to train the level set functions in a weakly supervised manner using only the bounding box annotations, which are originally used for object detection.

Specifically, we introduce an energy function based on the classical continuous Chan-Vese energy functional [7], and make use of a simple and effective mask supervised method, i.e., SOLOv2 [48], to predict the instance-aware mask map as the level set for each instance. In addition to the input image, the deep structural features with long-range dependencies are introduced to robustly evolve the level set curves towards the object's boundary, which is initialized by a box projection function at each step. By minimizing the fully differentiable energy function, the level set for each instance is iteratively optimized within its corresponding bounding box annotation. Extensive experiments are conducted on four challenging benchmarks for instance segmentation under various scenarios, including general scene, remote sensing and medical images. The leading qualitative and quantitative results demonstrate the effectiveness of our proposed method. Especially, on remote sensing and medical images, our method outperforms the state-of-the-art methods by a large margin.

The highlights of this work are summarized as follows:

1) We propose a novel level set evolution-based approach to instance segmentation. To the best of our knowledge, this is the first deep level set-based method that tackles the problem of box-supervised instance segmentation.

2) We incorporate the deep structural features with the low-level image to achieve robust level set evolution within bounding box region, where a box projection function is employed for level set initialization.

3) Our proposed method achieves new state-of-the-arts of box-supervised instance segmentation on COCO [30] and Pascal VOC [11] datasets, remote sensing dataset iSAID [50] and medical dataset LiTS [3].

2 Related Work

2.1 Box-supervised Instance Segmentation

The existing instance segmentation methods can be roughly divided into two categories. The first group [9, 13, 22, 55] performs segmentation on the regions extracted from the detection results. Another category [4, 5, 43, 48, 51] directly segments each instance in a fully convolutional manner without resorting to the detection results. However, all these methods rely on the expensive pixel-wise mask annotations.

Box-supervised instance segmentation, which only employs the bounding box annotations to obtain pixel-level mask prediction, has recently been receiving increasing attention. Khoreva *et al.* [20] proposed to predict the mask with box annotations under the deep learning framework, which heavily depends on the region proposals generated by the unsupervised segmentation methods like GrabCut [40] and MCG [39]. Based on Mask R-CNN [13], Hsu *et al.* [16] formulated the box-supervised instance segmentation into a multiple instance learning (MIL) problem by making use of the neighbouring pixel-pairwise affinity regularization. BoxInst [44] uses the color-pairwise affinity with box constraint under an efficient RoI-free CondInst framework [43]. Despite the promising performance, the pairwise affinity relationship is built on either partial or all neighbouring pixel pairs with the oversimplified assumption that spatial pixel or color pairs are encouraged to share the same label. This inevitably introduces noises, especially from the nearby background or similar objects. Besides, the recent methods like BBAM [26] and DiscoBox [25] focus on the generation of proxy mask labels, which often require multiple training stages or networks to achieve promising performance. Unlike the above methods, our proposed level set-based approach is learned implicitly in an end-to-end manner, which is able to iteratively align the instance boundaries by optimizing the energy function within the box region.

2.2 Level Set-based Segmentation

As a classical variational approach, the level set methods [1, 37] have been widely used in image segmentation, which can be categorized into two major groups: region-based methods [7, 36, 45] and edge-based methods [6, 33]. The key idea of level set is to represent the implicit curve by an energy function in a higher dimension, which is iteratively optimized by using gradient descent. Some works [15, 17, 21, 49, 54] have been proposed to embed the level set into the deep network in an end-to-end manner and achieve promising segmentation results. Wang *et al.* [49] predicted the evolution parameters and evolved the predicted contour by incorporating the user clicks on the boundary points. The energy function is based on the edge-based level set method in [6]. Levelset R-CNN [15] performs the Chan-Vese level set evolution with the deep features based on Mask R-CNN [13], where the original image is not used in the optimization. Yuan *et al.* [54] built a piecewise-constant function to parse each constant sub-region corresponding to a different instance based on the Mumford–Shah

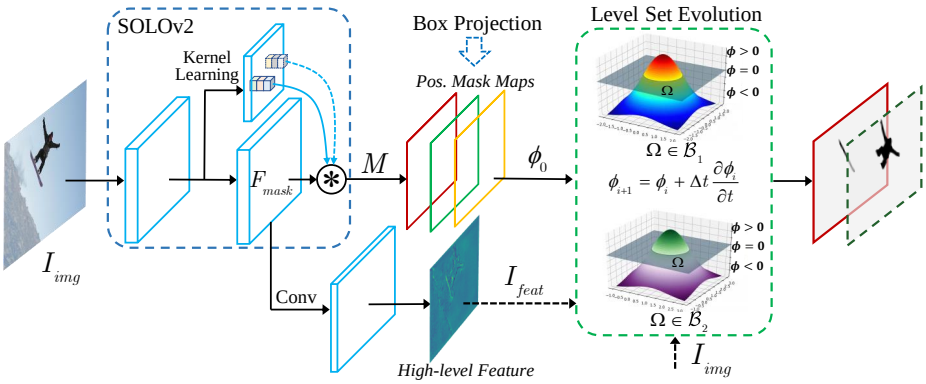

Fig. 1. Overview of our method. Our framework is designed based on SOLOv2 [48]. The positive mask maps M are obtained by level set evolution within the bounding box region. With the iterative energy minimization, the accurate instance segmentation can be obtained with box annotations only. The category branch is not shown here for simpler illustration.

model [36], which achieves instance segmentation by a fully convolutional network. The above methods perform level set evolution between deep features and ground-truth mask in a fully supervised manner, which train the network to predict different sub-regions and get object boundaries. Our proposed approach performs level set evolution only using the box-based annotations without the pixel-wise mask supervision.

Kim *et al.* [21] performed level set evolution in an unsupervised manner, which is mostly related to our proposed approach. To achieve N-class semantic segmentation, it employs the global multi-phase Mumford–Shah function [36] that only evolves on the low-level features of the input image. Our method is based on the Chan-Vese functional [7], which is constrained within the local bounding box with the enriched information from both input image and high-level deep features. Moreover, the initialization of level set is generated automatically for robust curve evolution.

3 Proposed Method

In this section, we present a novel box-supervised instance segmentation method, which incorporates the classical continuous Chan-Vese energy-based level set model [7] into deep neural network. To this end, we introduce an energy function, which enables the neural network to learn a series of level set functions evolving to the instance boundaries implicitly. In specific, we take advantage of an effective mask-supervised SOLOv2 model [48] to dynamically segment objects by locations and predict the instance-aware mask map of full-image size. To facilitate the box-supervised instance segmentation, we treat each mask map as the level set function ϕ for its corresponding object. Furthermore, we make

use of both the input image I_{img} and high-level deep features I_{feat} as the input to evolve the level set, where a box projection function is employed to encourage the network to automatically estimate an initial level set ϕ_0 at each step. The level set for each instance is iteratively optimized within its corresponding bounding box annotation. Fig. 1 gives the overview of our proposed framework.

3.1 Level Set Model in Image Segmentation

We first give a brief review of the level set methods [7, 36, 45], which formulate the image segmentation as a consecutive energy minimization problem. In the Mumford-Shah level set model [36], the segmentation of a given image I is obtained by finding a parametric contour C, which partition the image plane $\Omega \subset \mathbb{R}^2$ into N disjoint regions $\Omega_1, \cdots, \Omega_N$. The Mumford-Shah energy functional $\mathcal{F}^{MS}(u, C)$ can be defined as below:

$$\mathcal{F}^{MS}(u_1, \cdots, u_N, \Omega_1, \cdots, \Omega_N) = \sum_{i=1}^{N} (\int_{\Omega_i} (I - u_i)^2 dxdy + \mu \int_{\Omega_i} |\nabla u_i|^2 dxdy + \gamma |C_i|), \tag{1}$$

where u_i is a piecewise smooth function approximating the input I, ensuring the smoothness inside each region Ω_i. μ and γ are weighted parameters.

Chan and Vese [7] later simplified the Mumford-Shah functional as a variational level set, which has been explored aplenty [31, 34, 46, 52]. Specially, it can be derived as follows,

$$\begin{aligned} \mathcal{F}^{CV}(\phi, c_1, c_2) = &\int_{\Omega} |I(x,y) - c_1|^2 H(\phi(x,y)) dxdy \\ &+ \int_{\Omega} |I(x,y) - c_2|^2 (1 - H(\phi(x,y))) dxdy + \gamma \int_{\Omega} |\nabla H(\phi(x,y))| \, dxdy \end{aligned} \tag{2}$$

where H is the Heaviside function and $\phi(x, y)$ is the level set function, whose zero crossing contour $C = \{(x, y) : \phi(x, y) = 0\}$ divides the image space Ω into two disjoint regions, inside contour C: $\Omega_1 = \{(x, y) : \phi(x, y) > 0\}$ and outside contour C: $\Omega_2 = \{(x, y) : \phi(x, y) < 0\}$. In Eq. (2), the first two terms intend to fit the data, and the third term regularizes the zero level contour with a non-negative parameter γ. c_1 and c_2 are the mean values of input $I(x, y)$ inside C and outside C, respectively. The image segmentation is achieved by finding a level set function $\phi(x, y) = 0$ with c_1 and c_2 that minimize the energy $\mathcal{F}^{CV}$.

3.2 Box-supervised Instance Segmentation

Our proposed method exploits the level set evolution with Chan-Vese energy-based model [7] to achieve high-quality instance segmentation using the box annotations only.

Level Set Evolution within Bounding Box. Given an input image $I(x, y)$, we aim to predict the object boundary curve by evolving a level set implicitly within the region of annotated bounding box $\mathcal{B}$. The mask prediction $M \in$

$\mathbb{R}^{H\times W\times S^2}$ by SOLOv2 contains $S\times S$ potential instance maps of size $H\times W$. Each potential instance map contains only one instance whose center is at location (i,j). The mask map predicted for the location (i,j) with the category probability $p^*_{i,j}>0$ is regarded as the positive instance sample. We treat each positive mask map within box $\mathcal{B}$ as the level set $\phi(x,y)$, and its corresponding pixel space of input image $I(x,y)$ is referred as Ω, i.e., $\Omega\in\mathcal{B}$. C is the segmentation boundary with zero level $C=\{(x,y):\phi(x,y)=0\}$, which partitions the box region into two disjoint regions, i.e., foreground object and background.

To obtain the accurate boundary for each instance, we learn a series of level sets $\phi(x,y)$ by minimizing the following energy function:

$$\begin{aligned}\mathcal{F}(\phi,I,c_1,c_2,\mathcal{B}) = &\int\limits_{\Omega\in\mathcal{B}} |I^*(x,y)-c_1|^2\sigma(\phi(x,y))dxdy \\ &+\int\limits_{\Omega\in\mathcal{B}} |I^*(x,y)-c_2|^2(1-\sigma(\phi(x,y)))dxdy+\gamma\int\limits_{\Omega\in\mathcal{B}}|\nabla\sigma(\phi(x,y))|\,dxdy,\end{aligned} \tag{3}$$

where $I^*(x,y)$ denotes the normalized input image $I(x,y)$, γ is a non-negative weight, and σ denotes the *sigmoid* function that is treated as the characteristic function for level set $\phi(x,y)$. Different from the traditional Heaviside function [7], the *sigmoid* function is much smoother, which can better express the characteristics of the predicted instance and improve the convergence of level set evolution during the training process. The first two items in Eq. (3) force the predicted $\phi(x,y)$ to be uniform both inside region Ω and outside area $\bar{\Omega}$. c_1 and c_2 are the mean values of Ω and $\bar{\Omega}$, which are defined as below:

$$c_1(\phi)=\frac{\int\limits_{\Omega\in\mathcal{B}} I^*(x,y)\sigma(\phi(x,y))dxdy}{\int\limits_{\Omega\in\mathcal{B}}\sigma(\phi(x,y))dxdy},\quad c_2(\phi)=\frac{\int\limits_{\Omega\in\mathcal{B}} I^*(x,y)(1-\sigma(\phi(x,y)))dxdy}{\int\limits_{\Omega\in\mathcal{B}}(1-\sigma(\phi(x,y)))dxdy}. \tag{4}$$

The energy function $\mathcal{F}$ can be optimized with gradient back-propagation during training. With the time step $t\geqslant 0$, the derivative of energy function $\mathcal{F}$ upon ϕ can be written as follows:

$$\frac{\partial\phi}{\partial t}=-\frac{\partial\mathcal{F}}{\partial\phi}=-\nabla\sigma(\phi)[(I^*(x,y)-c_1)^2-(I^*(x,y)-c_2)^2+\gamma div\left(\frac{\nabla\phi}{|\nabla\phi|}\right)], \tag{5}$$

where ∇ and div are the spatial derivative and divergence operator, respectively. Therefore, the update of ϕ is computed by

$$\phi_i=\phi_{i-1}+\Delta t\frac{\partial\phi_{i-1}}{\partial t}. \tag{6}$$

The minimization of the above terms can be viewed as an implicit curve evolution along the descent of energy function. The optimal boundary C of the instance is obtained by minimizing the energy $\mathcal{F}$ via iteratively fitting ϕ_i as follows:

$$\inf_{\Omega\in\mathcal{B}}\{\mathcal{F}(\phi)\}\approx 0\approx\mathcal{F}(\phi_i). \tag{7}$$

Input Data Terms. The energy function in Eq. (3) encourages the curve evolution based on the uniformity of regions inside and outside the object. The input image I_u represents the essential low-level features, including shape, colour,

image intensities, etc. However, such low-level features usually vary with illumination variations, different materials and motion blur, making the level set evolution less robust.

In addition to the normalized input image, we take into account the high-level deep features I_f, which embed the image semantic information, to obtain more robust results. To this end, we make full use of the unified and high-resolution mask feature F_{mask} from all FPN levels in SOLOv2, which is further fed into a convolution layer to extract the high-level features I_f. Besides, the features I_f are enhanced by the tree filter [27, 41], which employs minimal spanning tree to model long-range dependencies and preserve the object structure. The overall energy function for level set evolution can be formulated as follows:

$$\mathcal{F}(\phi) = \lambda_1 * \mathcal{F}(\phi, I_u, c_{u_1}, c_{u_2}, \mathcal{B}) + \lambda_2 * \mathcal{F}(\phi, I_f, c_{f_1}, c_{f_2}, \mathcal{B}), \tag{8}$$

where λ_1 and λ_2 are weights to balance the two kinds of features. c_{u_1}, c_{u_2} and c_{f_1}, c_{f_2} are the mean values for input terms I_u and I_f, respectively.

Level Set Initialization. Conventional level set methods are sensitive to the initialization that is usually manually labeled. In this work, we employ a box projection function [44] to encourage the model to automatically generate a rough estimation of the initial level set ϕ_0 at each step.

In particular, we utilize the coordinate projection of ground-truth box to x-axis and y-axis and calculate the projection difference between the predicted mask map and the ground-truth box. Such a simple scheme limits the predicted initialization boundary within the bounding box, providing a good initial state for curve evolution. Let $m^b \in \{0,1\}^{H\times W}$ denote the binary region by assigning one to the locations in the ground-truth box, and zero otherwise. The mask score predictions $m^p \in (0,1)^{H\times W}$ for each instance can be regarded as the foreground probabilities. The box projection function $\mathcal{F}(\phi_0)_{box}$ is defined as below:

$$\mathcal{F}(\phi_0)_{box} = \mathcal{P}_{dice}(m^p_x, m^b_x) + \mathcal{P}_{dice}(m^p_y, m^b_y), \tag{9}$$

where m^p_x, m^b_x and m^p_y, m^b_y denote the x-axis projection and y-axis projection for mask prediction m^p and binary ground-truth region m^b, respectively. $\mathcal{P}_{dice}$ represents the projection operation measured by 1-D dice coefficient [35].

3.3 Training and Inference

Loss Function. The loss function L to train our proposed network consists of two items, including L_{cate} for category classification and L_{inst} for instance segmentation with box annotations:

$$L = L_{cate} + L_{inst}, \tag{10}$$

where L_{cate} is the Focal Loss [29]. For L_{inst}, we employ the presented differentiable level set energy as the optimization objective:

$$L_{inst} = \frac{1}{N_{pos}} \sum_k \mathbb{1}_{\{p^*_{i,j}>0\}} \{\mathcal{F}(\phi) + \alpha \mathcal{F}(\phi_0)_{box}\}, \tag{11}$$

Table 1. Performance comparisons on Pascal VOC `val` 2012. "$*$" denotes the results of GrabCut reported from BoxInst [44]. All entries are the results using *box-supervision*.

methods	backbone	AP	AP_{25}	AP_{50}	AP_{70}	AP_{75}
GrabCut* [40]	ResNet-101	19.0	-	38.8	-	17.0
SDI [20]	VGG-16	-	-	44.8	-	16.3
Liao *et al.* [28]	ResNet-101	-	-	51.3	-	22.4
Sun *et al.* [42]	ResNet-50	-	-	56.9	-	21.4
BBTP [16]	ResNet-101	23.1	-	54.1	-	17.1
BBTP w/ CRF [16]	ResNet-101	27.5	-	59.1	-	21.9
Arun *et al.* [2]	ResNet-101	-	73.1	57.7	33.5	31.2
BBAM [26]	ResNet-101	-	76.8	63.7	39.5	31.8
BoxInst [44]	ResNet-50	34.3	-	59.1	-	34.2
BoxInst [44]	ResNet-101	36.5	-	61.4	-	37.0
DiscoBox [25]	ResNet-50	-	71.4	59.8	41.7	35.5
DiscoBox [25]	ResNet-101	-	72.8	62.2	45.5	37.5
Ours	ResNet-50	36.3	76.3	64.2	43.9	35.9
Ours	ResNet-101	**38.3**	**77.9**	**66.3**	**46.4**	**38.7**

where N_{pos} indicates the number of positive samples, and $p^*_{i,j}$ denotes the category probability at target location (i, j). $\mathbb{1}$ represents the indicator function, which ensures only the positive instance mask samples perform the level set evolution. $\mathbb{1}$ is set to one if $p^*_{i,j} > 0$, and zero otherwise. α is the weight parameter, which is set to 3.0 empirically in our implementation.

Inference. It is worth noting that the level set evolution is only employed during training to generate implicit supervisions for network optimization. *The inference process is the same as the original SOLOv2 network.* Given the input image, the mask prediction is directly generated with efficient matrix non-maximum suppression (NMS). Comparing to SOLOv2, our proposed network introduces only one additional convolution layer to generate the high-level features with negligible cost.

4 Experiments

To evaluate our proposed approach, we conduct experiments on four challenging datasets, including Pascal VOC [11] and COCO [30], remote sensing dataset iSAID [50] and medical dataset LiTS [3]. On all datasets, *only box annotations are used during training.*

4.1 Datasets

Pascal VOC [11]. Pascal VOC consists of 20 categories. As in [16, 26, 44], the augmented Pascal VOC 2012 [12] dataset is used, which contains 10, 582 images for training and 1, 449 validation images for evaluation.

Table 2. Instance segmentation mask AP (%) on the COCO `test-dev`. "†" denotes the result of BBTP on the COCO `val2017` split. "∗" indicates that the BoxCaseg is trained with box and salient object supervisions.

method	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
mask-supervised:							
Mask R-CNN [13]	ResNet-101	35.7	58.0	37.8	15.5	38.1	52.4
YOLACT-700 [4]	ResNet-101	31.2	50.6	32.8	12.1	33.3	47.1
PolarMask [51]	ResNet-101	32.1	53.7	33.1	14.7	33.8	45.3
CondInst [43]	ResNet-101	39.1	60.9	42.0	21.5	41.7	50.9
SOLOv2 [48]	ResNet-101	39.7	60.7	42.9	17.3	42.9	57.4
box-supervised:							
BBTP† [16]	ResNet-101	21.1	45.5	17.2	11.2	22.0	29.8
BBAM [26]	ResNet-101	25.7	50.0	23.3	-	-	-
BoxCaseg* [47]	ResNet-101	30.9	54.3	30.8	12.1	32.8	46.3
BoxInst [44]	ResNet-101	33.2	56.5	33.6	16.2	35.3	45.1
BoxInst [44]	ResNet-101-DCN	35.0	**59.3**	35.6	**17.1**	37.2	48.9
Ours	ResNet-101	33.4	56.8	34.1	15.2	36.8	46.8
Ours	ResNet-101-DCN	**35.4**	59.1	**36.7**	16.8	**38.5**	**51.3**

Table 3. Deep variational instance segmentation methods on COCO `val`. "Sup." denotes the form of supervision, i.e., *Mask* or *Box*. *Our method is only supervised with box annotations yet achieves competitive results.*

method	backbone	Sup.	AP
DeepSnake [38]	DLA-34 [53]	*Mask*	30.5
Levelset R-CNN [15]	ResNet-50	*Mask*	34.3
DVIS-700 [54]	ResNet-50	*Mask*	32.6
DVIS-700 [54]	ResNet-101	*Mask*	**35.7**
Ours	ResNet-101	***Box***	33.0
Ours	ResNet-101-DCN	***Box***	35.0

COCO [30]. COCO has 80 general object classes. Our models are trained on `train2017` (115K images), and evaluated on `val2017` (5K images) and `test-dev` split (20K images).

iSAID [50]. It is a large-scale high-resolution remote sensing dataset for aerial instance segmentation, containing many small objects with complex backgrounds. The dataset comprises 1,411 images for training and 458 validation images for evaluation with 655,451 instance annotations.

LiTS [3]. The Liver Tumor Segmentation Challenge (LiTS) dataset[1] consists of 130 volume CT scans for training and 70 volume CT scans for testing. We randomly partition all the scans having mask labels into the training and validation dataset with the ratio of 4:1.

[1] https://competitions.codalab.org/competitions/17094

Table 4. Results of mask AP (%) on iSAID `val`. All models are trained with "1×" schedule (12 epoch) with 600×600 input size.

method	backbone	Sup.	AP	AP_{50}	AP_{75}
Mask R-CNN [13]	R-50-C4	*Mask*	28.8	51.8	27.7
PolarMask [51]	R-50-FPN	*Mask*	27.2	48.5	27.3
CondInst [43]	R-50-FPN	*Mask*	29.5	54.5	28.3
BoxInst [44]	R-50-FPN	*Box*	17.8	41.4	12.9
Ours	R-50-FPN	*Box*	**20.1**	**41.8**	**16.6**

Table 5. Instance segmentation results on LiTS `val`. All models are trained with "1×" schedule (12 epoch). Our method outperforms BoxInst by 3.8% AP.

method	backbone	Sup.	AP	AP_{50}	AP_{75}
Mask R-CNN [13]	R-50-FPN	*Mask*	64.2	81.6	71.0
BoxInst [44]	R-50-FPN	*Box*	40.7	67.8	40.2
Ours	R-50-FPN	*Box*	**44.5**	**78.6**	**45.6**

4.2 Implementation Details

The models are trained with the AdamW [32] optimizer on 8 NVIDIA V100 GPUs. The training schedules of "1×" and "3×" are the same as `mmdetection` framework [8] with 12 epochs and 36 epochs, respectively. ResNet [14] is employed as the backbone, which is initialized with the ImageNet [10] pre-training weights. For COCO, the initial learning rate is 10^{-4} with 16 images per mini-bath. For Pascal VOC, the initial learning rate is 5×10^{-5} with 8 images per mini-bath. The scale jitter is used where the shorter image side is randomly sampled from 640 to 800 pixels on COCO and Pascal VOC datasets for fair comparison. For iSAID and LiTS, all the models on each dataset are trained with the same settings. COCO-style mask AP (%) is adopted for performance evaluation. Following [25, 26], we also report the average precision (AP) at four IoU thresholds (including 0.25, 0.50, 0.70 and 0.75) for the comparison on Pascal VOC dataset. The non-negative weight γ in Eq. 3 is set to 10^{-4} by default.

4.3 Main Results

We compare our proposed method against the state-of-the-art instance segmentation approaches, including box-supervised and fully mask-supervised methods in different scenarios.

Most box-supervised methods are evaluated on the Pascal VOC dataset. Table 1 reports the comparison results. Our method outperforms BoxInst [44] by 2.0% and 1.8% AP with ResNet-50 and ResNet-101 backbones, respectively, achieving the best performance. For AP_{25} and AP_{50}, our method can obtain 77.9% and 66.3% accuracy, largely outperforming the recent DiscoBox [25] by 5.1% and 4.1%. The high IoU threshold-based AP metrics can reflect the segmentation performance with accurate boundary, which is in line with the practical

application. Our approach achieves 38.7% AP_{75} with ResNet-101, which outperforms BoxInst [44] and DiscoBox [25] by 1.7% and 1.2%, respectively.

Table 2 shows the main results on COCO `test-dev` split. Both fully mask-supervised and box-supervised methods are compared in the evaluation. Our method outperforms BBTP [16] by 12.3% AP with the same backbone. In contrast to the recent box-supervised methods, our method outperforms BBAM [26] and BoxCaseg [47] by 7.7% AP and 2.5% AP using ResNet-101. It achieves 33.4% AP and 35.4% AP, which is higher than BoxInst [44] by 0.2% and 0.4% with ResNet-101 and ResNet-101-DCN backbones, respectively. Our approach achieves 16.8% AP_S on small objects, which is slightly lower than BoxInst [44] by 0.3%. This is because small objects lack rich features for level set evolution to distinguish the foreground object and background within the bounding box. However, our method obtains the best results for large objects, largely outperforming BoxInst [44] by 2.4% AP_L using the same ResNet-101-DCN. Our method even performs better than some recent fully mask-supervised methods, such as YOLACT [4] and PolarMask [51]. This shows that our method narrows the performance gap between mask-supervised and box-supervised instance segmentation. Fig. 2 visualizes some instance segmentation results on COCO and Pascal VOC datasets.

We then compare our method with other deep variational-based instance segmentation approaches. DeepSnake [38] is based on the classical snake method [19]. Levelset R-CNN [15] and DVIS-700 [54] are also built on level set function. These methods are all fully supervised by the mask annotations. As shown in Table 3, our method achieves comparable results to the fully supervised variational-based methods, and even outperforms DeepSnake [38] and Levelset R-CNN [15].

To further validate the robust performance of our method in more complicated scenarios, we conduct experiments on remote sensing and medical image datasets. In remote sensing, the objects of the same class are densely-distributed. For medical images, the background is highly similar to the foreground. The previous pixel relationship model-based methods are built on the neighbouring pixel pairs. They are easily affected by the noisy context. Our level set-based method drives the curve to fit the object boundary under the guidance of level set minimization, which is more robust. Table 4 and Table 5 show the mask AP results on iSAID and LiTS datasets, respectively. It can be clearly seen that our approach outperforms BoxInst [44] by 2.3% AP on iSAID and 3.8% AP on LiTS. Fig. 3 and Fig. 4 show several examples of instance segmentation on iSAID and LiTS, respectively. One can see that our method is effective in various scenarios.

4.4 Ablation Experiments

The ablation study is conducted on Pascal VOC dataset to examine the effectiveness of each module in our proposed framework.

Level Set Energy. We firstly investigate the impact of level set energy functional with different settings. Table 6 gives the evaluation results. Our method achieves 19.7% AP only with the box projection function as $\mathcal{F}_{\phi_0}$ to drive the

Fig. 2. Visualization of instance segmentation results on general scene. The model is trained with only box annotations.

network to initialize the boundary during training. This indicates that the initialization for level set function is effective to generate the initial boundary. When the original image I_u is employed as the input data term in Eq. 8, our method can achieve 22.2% AP. On the other hand, our method achieves better performance with 24.7% AP when the deep high-level features I_f are employed as the extra input data. This demonstrates that both original image and high-level features can provide useful information for robust level set evolution. Besides, the above results are constrained within the bounding box $\mathcal{B}$ region for curve evolution. When the global region with the full-image size is regarded as the Ω, there is a noticeable performance drop (24.7% vs. 21.7%). This indicates that the bounding box region can make the level set evolution smoother with less noise interference.

Number of Channels for High-level Feature. Secondly, we investigate the selection of the total number of channels for the output high-level feature I_f. As shown in Table 7, our method obtains better representation with 24.7% AP performance when the number of channels C_{I_f} is set to 9. When $C_{I_f} = 10$, the

Table 6. The impact of **level set energy** with different settings. I_u and I_f denote the input image and high-level feature as the input data terms of energy, respectively. $\mathcal{B}$ and I represent the Ω space of bounding box or the full-image region for level set evolution.

$\mathcal{F}_{\phi_0}$	$\mathcal{F}_{\phi}(I_u)$	$\mathcal{F}_{\phi}(I_f)$	$\Omega \in \mathcal{B}$	$\Omega \in I$	AP	AP_{50}	AP_{75}
✓			✓		19.7	47.4	13.9
✓	✓		✓		22.2	49.5	17.4
✓	✓	✓	✓		**24.7**	**53.3**	**20.8**
✓	✓	✓		✓	21.7	48.4	17.4

Table 7. Different channel number C_{I_f} of high-level features for curve evolution.

C_{I_f}	AP	AP_{50}	AP_{75}
5	23.3	51.3	18.7
8	24.4	52.1	20.1
9	**24.7**	**53.3**	**20.8**
10	22.0	49.2	17.3
11	21.9	49.4	16.9

Table 8. Training schedules with "1×" single-scale training and "3×" multi-scale training.

sched.	AP	AP_{50}	AP_{75}
1×	24.7	53.3	20.8
3×	**34.4**	**62.2.**	**34.6**

Table 9. The effectiveness of tree filter [41] for high-level structural features in level set.

tree filter	AP	AP_{50}	AP_{75}
w/o.	34.4	62.2	34.6
w.	**36.3**	**64.2.**	**35.9**

performance drops (24.7% vs. 22.0%). This indicates that the more channels may introduce uncertain semantic information for level set evolution.

Training Schedule. We evaluate the proposed network using different training schedules. Table 8 shows the results with 12 epochs (1×) and 36 epochs (3×). It can be observed that a longer training schedule benefits the performance of our method. Due to the relatively small size of Pascal VOC compared with COCO (about 1/10), longer training schedule leads to significant improvement (24.7% vs. 34.4%). This implies that level set evolution needs more training time to achieve better convergence for instance segmentation.

Effectiveness of Deep Structural Feature. We study the impact of tree filter [41], which models long-range dependencies and preserves object structure, on obtaining deep semantic features for level set evolution. Table 9 shows the results. One can see that by applying the tree filter to high-level deep features, +1.9% AP improvement can be achieved.

5 Conclusion

This paper presented a single-shot box-supervised instance segmentation approach that iteratively learns a series of level set functions in an end-to-end fashion. An instance-aware mask map was predicted and used as the level set, and both the original image and deep high-level features were employed as the inputs to evolve the level set curves, where a box projection function was employed to obtain the initial boundary. By minimizing the fully differentiable energy function, the level set for each instance was iteratively optimized within its

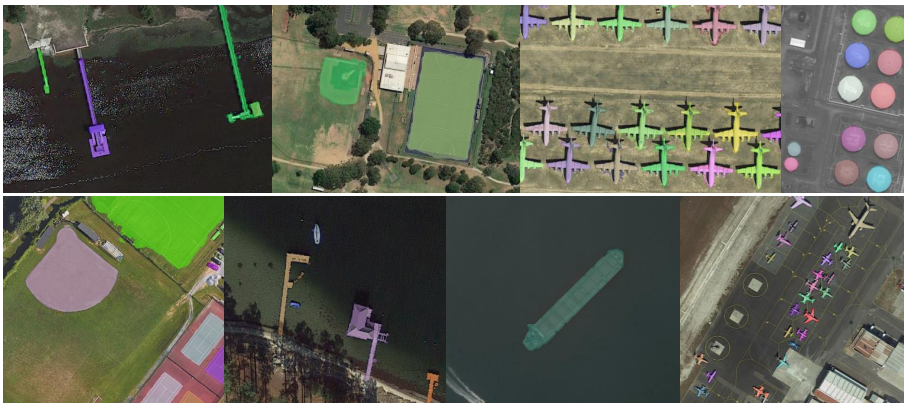

Fig. 3. Visual results of iSAID `val`. The mask predictions are obtained on the high-resolution remote sensing images only with box supervision.

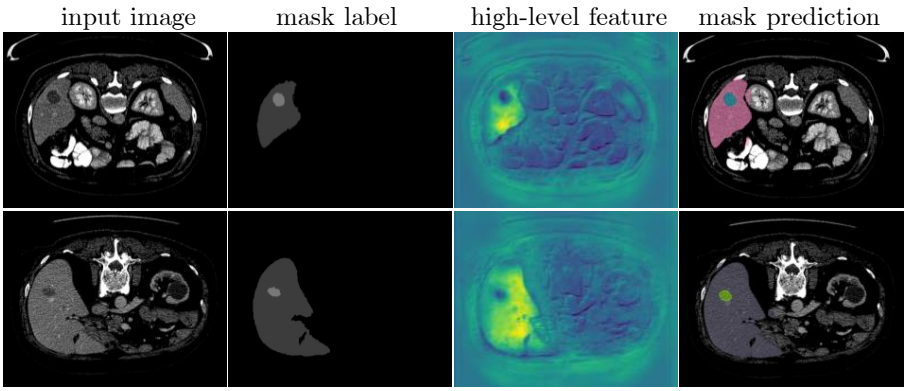

Fig. 4. Visualization examples of LiTS `val`. The high-level feature represents the input deep feature for level set evolution.

corresponding bounding box annotation. Extensive experiments were conducted on four challenging benchmarks, and our proposed approach demonstrated leading performance in various scenarios. Our work narrows the performance gap between fully mask-supervised and box-supervised instance segmentation.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants (61831015) and Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

References

1. Adalsteinsson, D., Sethian, J.A.: A fast level set method for propagating interfaces. Journal of Computational Physics **118**(2), 269–277 (1995)
2. Arun, A., Jawahar, C., Kumar, M.P.: Weakly supervised instance segmentation by learning annotation consistent instances. In: Proc. Eur. Conf. on Comp. Vis. pp. 254–270. Springer (2020)
3. Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al.: The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056 (2019)
4. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: Proc. IEEE Int. Conf. Comp. Vis. pp. 9157–9166 (2019)
5. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact++: Better real-time instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
6. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. International Journal of Computer Vision **22**(1), 61–79 (1997)
7. Chan, T., Vese, L.: Active contours without edges. IEEE Transactions on Image Processing **10**(2), 266–277 (2001)
8. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
9. Cheng, T., Wang, X., Huang, L., Liu, W.: Boundary-preserving mask r-cnn. In: Proc. Eur. Conf. on Comp. Vis. pp. 660–676. Springer (2020)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 248–255. Ieee (2009)
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)
12. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Proc. IEEE Int. Conf. Comp. Vis. pp. 991–998. IEEE (2011)
13. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: Proc. IEEE Int. Conf. Comp. Vis. pp. 2980–2988 (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 770–778 (2016)
15. Homayounfar, N., Xiong, Y., Liang, J., Ma, W.C., Urtasun, R.: Levelset r-cnn: A deep variational method for instance segmentation. In: Proc. Eur. Conf. on Comp. Vis. pp. 555–571. Springer (2020)
16. Hsu, C.C., Hsu, K.J., Tsai, C.C., Lin, Y.Y., Chuang, Y.Y.: Weakly supervised instance segmentation using the bounding box tightness prior. In: Proc. Advances in Neural Inf. Process. Syst. vol. 32, pp. 6582–6593 (2019)
17. Hu, P., Shuai, B., Liu, J., Wang, G.: Deep level sets for salient object detection. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 540–549 (2017)
18. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 6409–6418 (2019)
19. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision **1**(4), 321–331 (1988)
20. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 1665–1674 (2017)

21. Kim, B., Ye, J.C.: Mumford–shah loss functional for image segmentation with deep learning. IEEE Transactions on Image Processing **29**, 1856–1866 (2019)
22. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 9799–9808 (2020)
23. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. Proc. Advances in Neural Inf. Process. Syst. **24** (2011)
24. Kulharia, V., Chandra, S., Agrawal, A., Torr, P.H.S., Tyagi, A.: Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In: Proc. Eur. Conf. on Comp. Vis. pp. 290–308 (2020)
25. Lan, S., Yu, Z., Choy, C., Radhakrishnan, S., Liu, G., Zhu, Y., Davis, L.S., Anandkumar, A.: Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In: Proc. IEEE Int. Conf. Comp. Vis. pp. 3406–3416 (2021)
26. Lee, J., Yi, J., Shin, C., Yoon, S.: Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 2643–2652 (2021)
27. Liang, Z., Wang, T., Zhang, X., Sun, J., Shen, J.: Tree energy loss: Towards sparsely annotated semantic segmentation. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 16907–16916 (2022)
28. Liao, S., Sun, Y., Gao, C., KP, P.S., Mu, S., Shimamura, J., Sagata, A.: Weakly supervised instance segmentation using hybrid networks. In: Pro. IEEE Int. Conf. on Acous., Spee. and Signal Process. pp. 1917–1921. IEEE (2019)
29. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. IEEE Int. Conf. Comp. Vis. pp. 2980–2988 (2017)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proc. Eur. Conf. on Comp. Vis. pp. 740–755. Springer (2014)
31. Liu, S., Peng, Y.: A local region-based chan–vese model for image segmentation. Pattern Recognition **45**(7), 2769–2779 (2012)
32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
33. Malladi, R., Sethian, J.A., Vemuri, B.C.: Shape modeling with front propagation: A level set approach. IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(2), 158–175 (1995)
34. Maška, M., Daněk, O., Garasa, S., Rouzaut, A., Munoz-Barrutia, A., Ortiz-de Solorzano, C.: Segmentation and shape tracking of whole fluorescent cells based on the chan–vese model. IEEE Transactions on Medical Imaging **32**(6), 995–1006 (2013)
35. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proc. Int. Conf. on 3D Vision (3DV). pp. 565–571 (2016)
36. Mumford, D.B., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Communications on pure and applied mathematics (1989)
37. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. Journal of Computational Physics **79**(1), 12–49 (1988)
38. Peng, S., Jiang, W., Pi, H., Li, X., Bao, H., Zhou, X.: Deep snake for real-time instance segmentation. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 8533–8542 (2020)

39. Pont-Tuset, J., Arbelaez, P., T.Barron, J., Marques, F., Malik, J.: Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(1), 128–140 (2017)
40. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics (TOG) **23**(3), 309–314 (2004)
41. Song, L., Li, Y., Li, Z., Yu, G., Sun, H., Sun, J., Zheng, N.: Learnable tree filter for structure-preserving feature transform. In: Proc. Advances in Neural Inf. Process. Syst. vol. 32 (2019)
42. Sun, Y., Liao, S., Gao, C., Xie, C., Yang, F., Zhao, Y., Sagata, A.: Weakly supervised instance segmentation based on two-stage transfer learning. IEEE Access **8**, 24135–24144 (2020)
43. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Proc. Eur. Conf. on Comp. Vis. pp. 282–298 (2020)
44. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 5443–5452 (2021)
45. Vese, L.A., Chan, T.F.: A multiphase level set framework for image segmentation using the mumford and shah model. International Journal of Computer Vision **50**(3), 271–293 (2002)
46. Wang, X.F., Huang, D.S., Xu, H.: An efficient local chan–vese model for image segmentation. Pattern Recognition **43**(3), 603–618 (2010)
47. Wang, X., Feng, J., Hu, B., Ding, Q., Ran, L., Chen, X., Liu, W.: Weakly-supervised instance segmentation via class-agnostic learning with salient images. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 10225–10235 (2021)
48. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: Solov2: Dynamic and fast instance segmentation. In: Proc. Advances in Neural Inf. Process. Syst. vol. 33, pp. 17721–17732 (2020)
49. Wang, Z., Acuna, D., Ling, H., Kar, A., Fidler, S.: Object instance annotation with deep extreme level set evolution. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 7500–7508 (2019)
50. Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.S., Bai, X.: isaid: A large-scale dataset for instance segmentation in aerial images. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. Work. pp. 28–37 (2019)
51. Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., Luo, P.: Polarmask: Single shot instance segmentation with polar representation. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 12193–12202 (2020)
52. Xu, L., Lu, C., Xu, Y., Jia, J.: Image smoothing via l 0 gradient minimization. In: Pro. of the SIGGRAPH Asia Conf. pp. 1–12 (2011)
53. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 2403–2412 (2018)
54. Yuan, J., Chen, C., Li, F.: Deep variational instance segmentation. In: Proc. Advances in Neural Inf. Process. Syst. vol. 33, pp. 4811–4822 (2020)
55. Zhang, G., Lu, X., Tan, J., Li, J., Zhang, Z., Li, Q., Hu, X.: Refinemask: Towards high-quality instance segmentation with fine-grained features. In: Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. pp. 6861–6869 (2021)