ETHSeg: An Amodel Instance Segmentation Network and a Real-world Dataset for X-Ray Waste Inspection

Lingteng Qiu[1,2,3] Zhangyang Xiong[1,2] Xuhao Wang[2,3] Kenkun Liu[2] Yihan Li[2]
Guanying Chen[1,2] Xiaoguang Han[1,2*] Shuguang Cui[1,2,3]

[1]The Future Network of Intelligence Institute, CUHK-Shenzhen
[2]School of Science and Engineering, CUHK-Shenzhen [3]Shenzhen Research Institute of Big Data

Abstract

Waste inspection for packaged waste is an important step in the pipeline of waste disposal. Previous methods either rely on manual visual checking or RGB image-based inspection algorithm, requiring costly preparation procedures (e.g., open the bag and spread the waste items). Moreover, occluded items are very likely to be left out. Inspired by the fact that X-ray has a strong penetrating power to see through the bag and overlapping objects, we propose to perform waste inspection efficiently using X-ray images without the need to open the bag. We introduce a novel problem of instance-level waste segmentation in X-ray image for intelligent waste inspection, and contribute a real dataset consisting of 5,038 X-ray images (totally 30,881 waste items) with high-quality annotations (i.e., waste categories, object boxes, and instance-level masks) as a benchmark for this problem. As existing segmentation methods are mainly designed for natural images and cannot take advantage of the characteristics of X-ray waste images (e.g., heavy occlusions and penetration effect), we propose a new instance segmentation method to explicitly take these image characteristics into account. Specifically, our method adopts an easy-to-hard disassembling strategy to use high confidence predictions to guide the segmentation of highly overlapped objects, and a global structure guidance module to better capture the complex contour information caused by the penetration effect. Extensive experiments demonstrate the effectiveness of the proposed method. Our dataset is released at WIXRayNet.

1. Introduction

Nowadays, people produce an increasing amount of waste world-widely, which leads to great pressure for waste disposal. Improper disposal could bring irreversible disasters to our ecosystem, including climate warming, soil contamination, marine pollution, and so on. To reduce the harmful impact caused by the increasing amount of waste, it is urgent to develop an effective solution for proper waste disposal. In the pipeline of waste disposal, inspection for packaged garbage is a very important step as it can identify the categories and locations of waste items in the bag, providing useful information for the following processes. In the past, waste inspection is usually done by human workers manually with low efficiency. And such close contact with unknown waste increases the health risk to human workers. To improve the efficiency and reduce human contact with harmful waste, some researchers propose to use computer vision methods (*e.g.*, object detection) to identify misplaced waste items using RGB images [11, 28, 34, 40]. However, both the manual method and RGB image-based methods require garbage bags to be opened and waste items to be well-spread, bringing in costly preparation procedures. Also, it is unlikely for these methods to identify heavily occluded waste items.

Figure 1. A closed waste bag is being scanned by an X-ray machine (left), producing the corresponding X-ray image (mid). The contained waste items can be clearly seen from this image. The figure at the right is the instance-level mask prediction of our method for this X-ray image (best viewed in color)

Fortunately, we observed that X-rays have strong penetrating power that even completely occluded or buried ob-

*corresponding author, hanxiaoguang@cuhk.edu.cn

jects can be well-imaged by X-ray scanning, as shown in Fig. 1. Compared with natural images, X-ray images have distinct characteristics (see Tab. 1). First, as different materials absorb X-ray in different degrees, the colors in X-ray images indicate the thickness and the material type of the corresponding areas. Second, the edge information of all objects are well preserved even though they are occluded. These nice characteristics make it possible to inspect all waste items in packaged waste from a single X-ray image without opening the bag.

This observation motivates us to perform waste inspection efficiently using X-ray images, such that the whole pipeline of waste disposal can be simplified. In this work, we introduce a novel problem of instance-level waste segmentation in X-ray images to facilitate intelligent waste inspection. Although there are some existing methods using X-ray images for object inspection [1, 2, 27, 48, 45], they mainly focus on security inspection, where the goal is to detect prohibited items hidden in common objects. In contrast, we target predicting waste category and pixel-level mask for each waste item in X-ray image to allow a more fine-grained inspection.

As no off-the-shelf dataset is available, we contribute the first high-quality X-ray waste inspection dataset collected in campus community, named *Waste Inspection X-ray Dataset* (WIXray), for this problem. According to the current waste disposal scenario [29], we divide the common waste items into twelve categories, namely, *PlasticBottle*, *Can*, *Carton*, *GlassBottle*, *Stick*, *Tableware*, *FoodWaste*, *HeatingPad*, *Desiccant*, *MealBox*, *Battery*, and *Bulb*. Our dataset contains $5,038$ annotated X-ray images and each image contains 6.13 labeled instances on average. Sample X-ray image with annotations is shown in Fig. 2.

Since the imaging mechanism of X-ray images is largely different from natural images, directly applying existing methods designed for natural images [19] on this problem leads to decreased performance. Therefore, we propose an ***E**asy-**to**-**H**ard Instance **Seg**mentation Network* (*ETHSeg*) for X-ray inspection. First, our method adopts an *easy-to-hard disassembling* strategy that uses high-confidence instance predictions to guide the segmentation of hard instances in the highly overlapped regions. Second, a *global structure guidance* module is introduced to better capture the complex global contour information for mask prediction. Our carefully designed ETHSeg achieves much higher accuracy for X-ray waste instance segmentation.

The main contributions of this work are as follows:

- We introduce a new task of instance-level waste segmentation in X-ray images to promote the development of the intelligent waste inspection algorithm.
- We contribute an X-ray image dataset with high-quality bounding boxes and instance masks annotation as a benchmark for waste inspection. To the best of our knowledge, this is the first labeled X-ray dataset for this problem.
- We propose a new instance segmentation method that explicitly considers the occlusion and penetration effect of the X-ray image for accurate mask prediction. Extensive experiments verify the effectiveness of our method.

2. Related Work

Waste disposal and inspection Globally, millions of tons of municipal waste are generated every day, which poses a great threat to public health and the environment [12]. However, waste inspection is still mostly done manually during the last decade, which is inefficient and labor-intensive [31]. To reduce the human contact with toxic waste during the process of waste sorting, there has been some research aiming to detect waste objects using RGB cameras [11, 28, 34, 40]. However, these methods require waste objects to be detected are visible in the sight of cameras. Thus, we propose to exploit X-ray scanning to handle this problem.

X-ray image datasets X-ray has strong penetrating power, making occluded objects visible in the image. This penetrating ability has been utilized by several computer vision methods, which take X-ray images or videos as inputs [18, 49, 52, 13, 10, 6, 3, 39, 50, 2, 27]. For example, Akcay *et al*. [2] perform image classification and detection for X-ray baggage security images. Aurelia *et al*. [6] contribute a large chest x-ray dataset with multi-label annotated reports. However, most of the existing methods focus on security inspection and medical imaging analysis, and no research on X-ray waste inspection has yet been explored. This motivates us to introduce the first X-ray dataset for waste inspection.

Instance detection and segmentation Existing detection and segmentation methods are mainly designed for natural images. In the last decade, two-stage methods first become popular. Mask R-CNN [19] introduces a fully convolutional mask head to Faster R-CNN [38] detector, which is classic anchor-based two-stage detection method. This stream of detection methods have taken over the dominant position in two-stage object detection for a long period since the work of R-CNN [17, 16, 38, 7, 30, 51, 42]. QueryInst [14] proposes to do detection and segmentation by queries in a unified manner, which extends the query method of Sparse R-CNN [42]. For one-stage methods, OverFeat [41] is the first deep learning method of detection, after which many outstanding one-stage object detection methods have been proposed (*e.g*., SSD [25] and YOLO series [35, 36, 37, 4]). Many one-stage instance segmen-

tation frameworks [5, 43, 46, 47] are built on top of these one-stage detection frameworks, achieving comparable results with favorable inference speed. Recently, anchor-free approaches have attracted wide attention due to its time-efficiency and ROI-independence [21, 21, 44, 8, 43]. The representative work of detection is FCOS [44]. BlendMask [8] appends a blender module to FCOS [44] combining top-down bbox attentions and bottom-up segmentation information. CondInst [43] replaces RoI-based fixed mask branch with dynamic instance-aware networks and improves performance as well as inference time. Despite their success on natural images, they perform not as well as expected when meet X-ray images.

Amodal instance segmentation Except the common modal instance segmentation methods like [19], amodal instance segmentation is also explored to handle occlusion cases. Li and Malik [22] propose the first solution based on Iterative Bounding Box Expansion. ORCNN [15] proposes an ROI-based multi-task architecture to predict amodal mask, visible mask, and occlusion mask at the same time. Qi *et al*. [32] construct a new dataset called KINS with amodal annotations and propose Amodal Segmentation Network (ASN) with Multi-Level Coding (MLC) to improve the performance. BCNet [20] establishes a bilayer framework, in which the top GCN layer detects the occluder and the bottom GCN layer infers the occludee. Among them, BCNet is the method most related to ours, but it is designed for natural images where occluded parts of objects are invisible. In contrast, the occluded object can still be present in an X-ray image for our problem. By utilizing this characteristic, we propose our amodal instance segmentation method for X-ray image.

3. The WIXray Dataset

As there is no existing X-ray image dataset for learning waste inspection, we introduce the first X-ray image dataset with high-quality annotations to serve as a benchmark dataset for instance-level waste segmentation. This dataset is collected in our campus community with around 8 thousand residents.

3.1. Packaged Waste Collection

Waste categories According to current waste disposal scenarios, we classify the domestic wastes into four general types and twelve categories: Recyclable (*PlasticBottle*, *Can*, *Carton*, *GlassBottle*, *Stick*, and *Tableware*), Foodwaste (*FoodWaste*), Residual (*HeatingPad*, *Desiccant*, and *MealBox*), and Hazardous (*Battery* and *Bulb*). Sample X-ray images for each category are shown in Fig. 2.

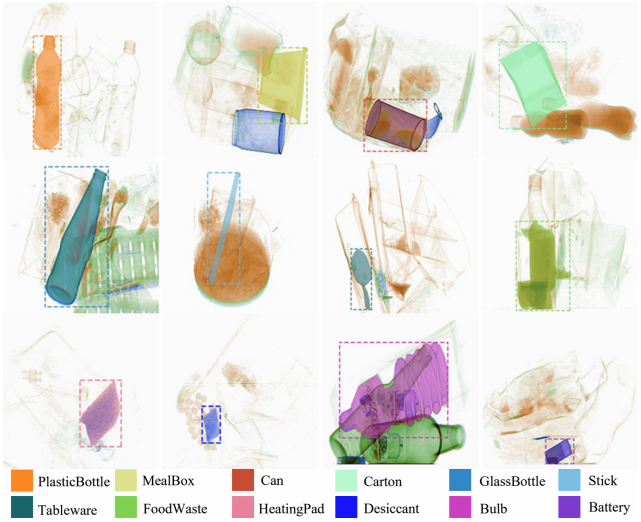

Figure 2. Examples of waste items in the WIXray dataset. Each waste category is provided with instance-level labelling.

Table 1. Characteristic of different waste categories under X-ray.

Color	Contour	Material	Waste Category
Green	Clear Outline	Glass (Dark Green)	GlassBottle
		Carboard (Light Green)	Carton
Blue	Clear Outline	Metal	Can, Battery, HeatingPad, Desiccant
Orange	Unifixed Shape Opaque	Organism	FoodWaste
	Clear Outline Transparence	Plastic or Wood	PlasticBottle, MealBox, Tableware, Stick

X-ray image capturing To increase the diversity of our dataset, we collected packaged garbage from different communal waste recycling stations and used the X-ray machine to produce the X-ray images. For some imbalanced categories in the dataset, especially the hazardous waste, we randomly put the pre-prepared specific types of waste into the waste bag to increase the number of corresponding items.

3.2. X-ray Image Annotations

X-ray image characteristics In X-ray images, different materials have different abilities to absorb X-rays, leading to varying image characteristics, as described in Tab. 1 and visualized in Fig. 2. For example, some categories of waste have robust color features while others only retain a few edge features.

Image annotation We labeled both the bounding box and the instance segmentation mask for each instance, As X-ray images have a strong penetrating effect, we can see through overlapping objects from a single view, making our dataset intrinsically different from traditional images. Instead of

Table 2. Statistics of the proposed WIXray dataset.

	Recyclable Waste						Food Waste	Residual Waste			Hazardous Waste		Total
	PlasticBottle	Can	Carton	GlassBottle	Stick	Tableware	FoodWaste	HeatingPad	Desiccant	MealBox	Battery	Bulb	
Train Set	2,900	1,298	2,024	745	3,271	492	8,200	236	438	6,004	1,093	404	27,105
Test Set	405	187	265	97	510	70	1,121	30	55	826	121	53	3,740
Total	3,305	1,485	2,289	842	3,781	562	9,321	266	493	6,830	1,214	457	30,845

only labeling unoccluded regions for an instance [24], we annotated its complete shape no matter whether they are occluded or not because each object can be completely seen in the X-ray image.

Labeling X-ray images is a challenging task as most people are not clear about X-ray image characteristics among different types of domestic waste. We recruited some environmental protection volunteers to label the collected data. In the beginning, our researchers carefully annotated the first 800 X-ray images using the *labelme* tool[1], where the waste bags were opened for visual checking. Note that these 800 images are labeled by visually compare the X-ray image and waste items with the waste bag opened. These 800 X-ray images served as references to help annotators understand the characteristics of different categories. To ensure the label quality, each labeled result was carefully reviewed by at least two inspectors.

3.3. Dataset Statistics

In total, our WIXray contains $5,038$ X-ray images and $30,881$ waste instances covering 12 common waste categories. Tab. 2 summarizes the statistics of the introduced dataset. Unlike existing X-ray datasets for security inspection [45] which only annotate a few forbidden objects, we densely labeled the common waste items in the picture. On average, our dataset contains 6.13 labeled instances per image, which is significantly larger than 2.27 instances of the HiXray dataset [45]. A larger instance number per image indicates more occlusions and contextual information, making our dataset more valuable for training and evaluation.

Images in our dataset are stored in PNG format of resolution 450×450, and split into $4,433$ for training and 605 for testing. We use this dataset as a benchmark for training and evaluating X-ray waste instance segmentation. It costs half of a year to collect and label this high-quality X-ray waste inspection dataset, and we will release this dataset to facilitate future research on this problem.

4. The Proposed Method

Existing instance segmentation methods [19] are often designed for natural images and do not consider the image characteristics of X-ray waste images, resulting in a decreased performance. In this section, we introduce a novel framework, named ***Easy-to-Hard Instance Segmentation Network*** (*ETHSeg*), to take advantage of the penetration effect and occlusion with two effective designs for instance-level waste segmentation (see Fig. 3). First, our method explicitly incorporates a *global structure guidance* module in image feature extraction to help encode the global contour context. Second, we propose an *easy-to-hard disassembling* strategy to help segment the hard examples in the occlusion regions.

4.1. Basic Segmentation for Each Instance

BCNet [20] is a state-of-the-art top-down instance segmentation method that explicitly considers object occlusion by a bilayer GCN structure. Although BCNet achieves impressive results in the natural image benchmark (*e.g.*, COCO [24]), applying it straightforwardly on our X-ray dataset leads to unsatisfactory results, due to the penetration effect and severe occlusion. Our method is built on top of BCNet, but with two substantial improvements (*i.e.*, the global structure guidance module and easy-to-hard disassembling strategy).

Bilayer convolution network (BCNet) BCNet consists of three parts: (a) a backbone with FPN [23] for image feature extraction; (b) a FCOS detector to predict object bounding boxes as instance proposals; (c) a bilayer GCN structure for instance segmentation. Given input feature $\mathbf{X} \in \mathbb{R}^{(HW)\times C}$, the GCN in the bilayer structure can be represented as:

$$\mathbf{Z} = \sigma(\mathbf{AXW}) + \mathbf{X}, \tag{1}$$

$$\mathbf{A} = softmax(F(\mathbf{X}, \mathbf{X})), \tag{2}$$

$$F(\mathbf{X}, \mathbf{X}) = \theta(\mathbf{X})\phi(\mathbf{X})^T, \tag{3}$$

where $\mathbf{Z}$ is the updated feature, $\mathbf{A} \in \mathbb{R}^{(HW)\times(HW)}$ is a self-attention map, $\mathbf{W}$ is a learnable output transformation matrix, and σ is a normalization layer with ReLU. F measures the dot-product similarity between two nodes $\mathbf{X}_i$ and $\mathbf{X}_j$, where θ and ϕ are trainable transformations implemented by 1×1 convolution.

The first GCN layer in BCNet takes the ROI feature $\mathbf{X}_{roi}$ as input to produce an updated feature $\mathbf{Z}_0$, and infers the contour and mask of the *occluder*. The updated feature is then added to the ROI feature $\mathbf{X}_f = \mathbf{X}_{roi} + \mathbf{Z}_0$ as input

[1] https://github.com/wkentaro/labelme

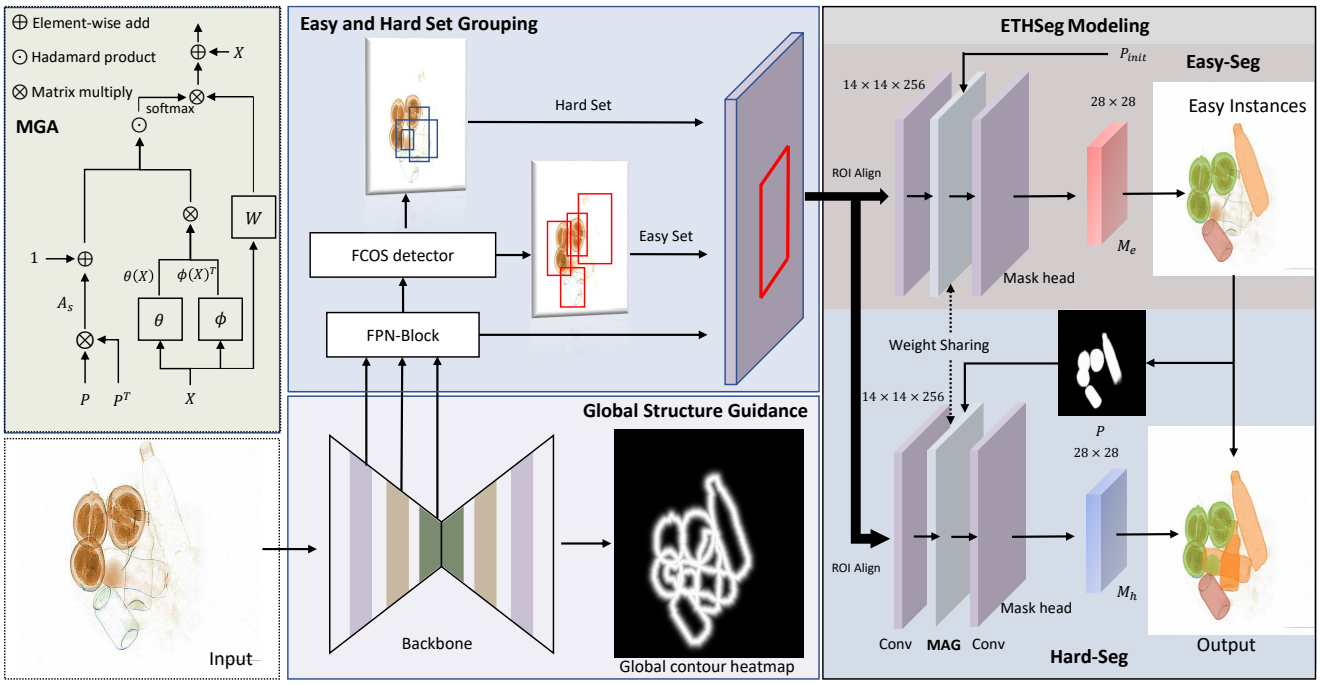

Figure 3. Overview of the proposed Easy-to-Hard Instance Segmentation Network (ETHSeg). First, we extract image features by a backbone and predict a global contour heatmap supervised by the Global Structure Guidance. Second, we employ FCOS detect head to obtain both the easy set and hard via the Easy and Hard Set Grouping. Last, our ETHSeg Modeling adopts an easy-to-hard disassembling strategy for mask prediction. The diagram at the upper left illustrates the details of the mask-guided attention (MGA).

for the second GCN layer to predict the contour and mask of the *occludee* (*i.e.*, target). More details of BCNet can be found in [20].

Global structure guidance BCNet crops the extracted feature by ROI-Align to be the input of the segmentation head to predict the mask and contour, and supervisions on the mask and contour are applied on the ROI area. However, this supervision is too local to help the network capture the complex relationship between different instances in highly overlapping regions. This is particularly true for X-ray images as the penetration effect brings more contours.

Therefore we design a global structure guidance module, which can be easily plugged into any existing top-down methods, to extract global contour context to guide the segmentation of ROI areas by multi-tasks learning. As shown in Fig. 3, we upsample the low-resolution feature map into high-resolution ones and predict multi-scale global contour heatmap $\{\hat{J}_i | i = 1, 2, 3\}$. Inspired by the human pose estimation methods [33], the ground-truth contour map J_i is represented as the heatmap with Gaussian distributions centered at the contour points with a variance of σ. As the global contour map contains overlapping information between different instances, integrating a global structure guidance module in the network is beneficial for extracting global semantic features to distinguish overlapping objects.

4.2. Easy-to-Hard Disassembling

In the X-ray image instance segmentation, the "hard" instances are generally objects overlapping with multiple other objects in the cluttered region. Inspired by the way human perform instance segmentation where "easy" (*i.e.*, easily distinguishable) instances are first identified before segmenting the hard ones, we propose a novel easy-to-hard disassembling strategy to simulate this process. Our method first splits the object bounding boxes into an easy and a hard set, and then utilizes masks extracted from easy instances to help predict highly occluded instances.

Easy and hard set grouping Given bounding box set $\mathbf{B} = \{\mathbf{B}_0, \mathbf{B}_1, \ldots, \mathbf{B}_N\} \in \mathbb{R}^{N \times 5}$, where N is the number of bounding boxes processed by NMS. $\mathbf{B}_i = [c_x, c_y, w, h, s]$ is a box detected from object detector, where (c_x, c_y) is the center coordinate, w and h are the width and height, and s is the predictive confidence. Based on the predictive score of each bounding box, we first split the box set $\mathbf{B}$ into two parts, namely, the high-confidence set $\mathbf{B}_{\text{easy}}$ with N_e boxes and the low-confidence set $\mathbf{B}_{\text{hard}}$ with (N_h)

boxes. After getting two groups of boxes, our goal is to exploit the easy set to generate useful information to guide the segmentation of the hard one.

Note that we have tried splitting the box set into three or four groups in the implementation, but found the improvement over using two groups is marginal. So we empirically divided the box set into two groups.

Easy-to-hard segmentation As the bounding boxes in the high-confidence set generally produce good segmentation results, we first predict the instance masks for all the boxes in the easy set, denoted as $\mathbf{M}_e = \{\mathbf{M}_e^i | i = 0, 1, \ldots, N_e\}$, using a segmentation head $\mathbf{G}_s$:

$$\mathbf{M}_e^i = \mathbf{G}_s(\mathbf{Z}_e^i), \tag{4}$$

$$\mathbf{Z}_e^i = \sigma(\mathbf{A}\mathbf{X}_e^i\mathbf{W}) + \mathbf{X}_e^i, \tag{5}$$

where $\mathbf{Z}_e$ is the updated feature by the GCN layer, $\mathbf{X}_e^i$ is the ROI feature of the i^{th} box in the easy set, and $\mathbf{A}$ is the self-attention map defined as in Eq. (2).

Note that the bilayer GCN structure of BCNet works on a *single* ROI feature, where the feature from the occluder branch is added back to the ROI feature to help the prediction of the occludee branch. However, in our case, features extracted from the easy and hard set belong to *different* ROIs, making direct summation of features not practical.

In this work, we propose to use the estimated mask information from the easy set to enhance the self-attention map for predicting masks of the hard set. Specifically, we first convert these predicted masks $\mathbf{M}_e$ back to the input image space according to their ROI coordinates, and merge them into a single mask $\mathbf{P}$ via an element-wise max operation. Mask $\mathbf{P}$ is a soft mask, each value indicates the probability of a pixel location contain object instances. As this mask provide strong information for the locations of easy instances, we propose a *mask-guided attention* to guide the segmentation of hard instances.

For an ROI feature of the j^{th} box in the hard set $\mathbf{X}_h^j$, we perform ROI-align on the mask $\mathbf{P}$ to get a pixel-aligned mask $\mathbf{P}^j$ for $\mathbf{X}_h^j$. The the mask-guided attention $\mathbf{A}_g$ is defined as

$$\mathbf{A}_g = softmax(F(\mathbf{X}_h^j, \mathbf{X}_h^j) \odot \mathbf{A}_s)), \tag{6}$$

$$\mathbf{A}_s = 1 + \mathbf{P}^j(\mathbf{P}^j)^T, \tag{7}$$

where $\odot$ denotes Hadamard product and $\mathbf{A}_s$ is a similarity matrix whose elements represent the probability of two nodes contain object.

Finally, the masks of the hard set $\mathbf{M}_h = \{\mathbf{M}_h^j | j = 0, 1, \ldots, N_h\}$ can be computed as

$$\mathbf{M}_h^j = \mathbf{G}_s(\mathbf{Z}_h^j), \tag{8}$$

$$\mathbf{Z}_h^j = \sigma(\mathbf{A}_g\mathbf{X}_h^j\mathbf{W}) + \mathbf{X}_h^j. \tag{9}$$

Through our mask-guided attention, the hard set can use the similarity matrix built from the easy mask to boost their attention graph, thus improving the predicted accuracy for the hard set.

4.3. Loss Function

The objective function for our ETHSeg can now be formulated. First, We employ FCOS [44] as the object detector due to its anchor-free efficiency, and the loss function for detection $\mathcal{L}_{\text{Detect}}$ is defined as [44]

$$\mathcal{L}_{\text{Detect}} = \mathcal{L}_{\text{Regression}} + \mathcal{L}_{\text{Center}} + \mathcal{L}_{\text{classes}}. \tag{10}$$

Second, the loss function for the segmentation network $\mathcal{L}_{\text{mask}}$ consists of three components:

$$\mathcal{L}_{\text{mask}} = \lambda_1\mathcal{L}_e + \lambda_2\mathcal{L}_h + \lambda_3\mathcal{L}_{\text{heatmap}}, \tag{11}$$

$$\mathcal{L}_{\text{e}} = \mathcal{L}_{\text{Occulder}}(\mathbf{M}_e) + \mathcal{L}_{\text{Occludee}}(\mathbf{M}_e), \tag{12}$$

$$\mathcal{L}_{\text{h}} = \mathcal{L}_{\text{Occulder}}(\mathbf{M}_h) + \mathcal{L}_{\text{Occludee}}(\mathbf{M}_h), \tag{13}$$

$$\mathcal{L}_{\text{heatmap}} = \sum_{i=1}^{3} \mathcal{L}_{MSE}(\hat{J}_i, J_i), \tag{14}$$

where $\mathcal{L}_e$, and $\mathcal{L}_h$ denote the segmentation losses for the easy and hard bounding box sets, and $\mathcal{L}_{\text{heatmap}}$ supervises the prediction of the global contour heatmap. The segmentation loss for the occluder ($\mathcal{L}_{\text{Occulder}}$) and occludee ($\mathcal{L}_{\text{Occludee}}$) in each ROI are the same as in BCNet [20]. λ_1, λ_2, and λ_3 are hyper-parameters to balance the loss functions, which are empirically tuned to be $\{0.5, 1.0, 0.5\}$ using the training set.

Finally, the whole instance segmentation framework can be trained in an end-to-end manner defined by a multi-task loss function $\mathcal{L}$:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{detect}} + \mathcal{L}_{\text{mask}}, \tag{15}$$

where $\lambda = 1.0$ is the loss weight.

5. Experiments

In this section, we compare our methods with existing instance segmentation approaches on our benchmark dataset.

5.1. Implementation Details

Global Structure guidance module Given an image with a resolution of 800, the backbone network and FPN will extract five different feature maps $\{P3, P4, P5, P6, P7\}$ with height and width of $\{100, 50, 25, 13, 7\}$. See [44] for more details. We first step by step upsample P_i to have the dimension as P_{i+1}, and concatenate them as the updated $\tilde{P}_{i+1}$, for i ranged from 3 to 5. Next, we utilize the updated $\tilde{P}_{i+1}$ to obtain the corresponding global contour map $\hat{J}_i$. The spatial size of the global contour map is a quarter of the input image size.

Table 3. Detection and instance segmentation results on the proposed WIXray.

Methods	Backbone	Detection AP			Segmentation AP		
		overall	AP_{50}	AP_{75}	overall	AP_{50}	AP_{75}
Faster R-CNN [38]	ResNet-101-FPN	43.46	62.40	48.17	-	-	-
Cascade R-CNN [7]	ResNet-101-FPN	46.30	63.84	50.55	-	-	-
Sparse R-CNN [42]	ResNet-101-FPN	48.85	64.88	54.42	-	-	-
Mask RCNN [19]	ResNet-101-FPN	45.32	63.87	50.03	42.86	59.71	47.13
Cascade Mask R-CNN [7]	ResNet-101-FPN	46.86	64.18	52.49	43.97	60.44	47.64
ORCNN [15]	ResNet-101-FPN	42.32	57.53	47.89	37.70	52.51	42.74
QueryInst [14]	ResNet-101-FPN	48.23	64.48	53.26	44.34	61.03	49.05
SSD [25]	VGG-16	36.48	58.84	40.46	-	-	-
YOLOv3 [37]	DarkNet-53	39.57	60.80	44.76	-	-	-
FCOS* [44]	ResNet-101-FPN	48.39	66.80	52.20	-	-	-
SOLOv2 [47]	ResNet-101-FPN	-	-	-	44.39	61.32	48.83
YOLACT [5]	ResNet-101-FPN	37.65	59.87	40.03	36.18	55.12	38.26
BlendMask [8]	ResNet-101-FPN	47.38	63.72	51.77	43.61	59.62	46.55
CondInst [43]	ResNet-101-FPN	47.72	64.42	51.93	43.77	60.10	47.68
BCNet [20]	ResNet-101-FPN	48.45	65.63	52.05	45.11	61.32	49.20
ETHSeg (ours)	ResNet-101-FPN	48.73	66.68	53.32	**46.85 (+1.74)**	**63.22 (+1.90)**	**50.95 (+1.91)**

* indicates that FCOS [44] was trained with the setting of BCNet (https://github.com/lkeab/BCNet).

Training In terms of the object detector, we follow the training strategies suggested in FCOS [44]. For training our segmentation head, we choose both the ground-truth boxes and object proposals whose predicted scores and IOU with ground-truth are larger than 0.05 and 0.3 as our proposals. The threshold used for easy-to-hard grouping is set to 0.65.

Both the detector and the segmentation network could be end-to-end trained as typical top-down methods. SGD with momentum is employed for training $15K$ iterations with $1K$ warm-up iterations. We set the batch size to 16 and the initial learning rate to 0.01. The learning rate is decayed by a factor set of 0.1 in $7K$ and $12K$ iterations. The variance σ used to generate the ground-truth global contour map is set to 8.

Inference During inference, we keep at most 50 proposal boxes generated by FCOS whose predicted scores are larger than 0.3 with a 0.6 NMS threshold. Next, according to our easy-to-hard disassembling strategy, we first predict the masks for the easy set, and then use these masks to guide the mask prediction for the hard set.

5.2. Results and Comparisons

We employed the MMDetection toolkit [9] to implement existing instance segmentation methods for comparisons. To ensure a fair comparison, all the compared methods used ResNet-101-FPN as the backbone and were initialized from the COCO pre-trained models. We trained these methods on the training set of our dataset using SGD and AdamW [26].

We also attempted to evaluate existing amodal instance segmentation methods (*e.g.* ORCNN [15]) on our dataset. Note that our dataset is not perfectly suitable for amodal segmentation, as overlapping waste items might be penetrated by the X-ray and no apparent occlusion orders can be inferred. We modified our dataset with a simple assumption that the smaller objects occlude the larger objects.

Table 4. Instance segmentation results of different variant models using the same detection results.

Model	AP	AP_{50}	AP_{75}
BCNet + Detection from ETHSeg	45.53	62.38	49.95
BCNet + Global Structure Guidance	45.98(+0.45)	62.65	49.96
BCNet + Easy-to-Hard Disassembling	46.12(+0.59)	62.00	**51.02**
ETHSeg	**46.85(+1.32)**	**63.22**	50.95

As shown in Tab. 3, we compared our method with those state-of-the-art object detectors on the WIXray Dataset. We can find that Our ETHSeg performs better than both existing one-stage and two-stage methods in all evaluation metrics. Specifically, our method illustrated its effectiveness by outperforming Cascaded Mask R-CNN[19] and QueryInst[14] by 2.98 and 2.51 segmentation AP respectively. Compared to the one-stage instance segmentation methods, our method, with the same detector, exceeds BCNet by 1.74 segmentation AP.

The visualization results are shown in Fig. 4. It is obvious that our method is able to detect occlusion objects more accurately and estimated contours are closer to the ground truth thanks to our ingenious design.

5.3. Ablation Study

We conduct a series of ablation studies to verify the effectiveness of our global structure guidance module and the easy-to-hard disassembling strategy in our framework.

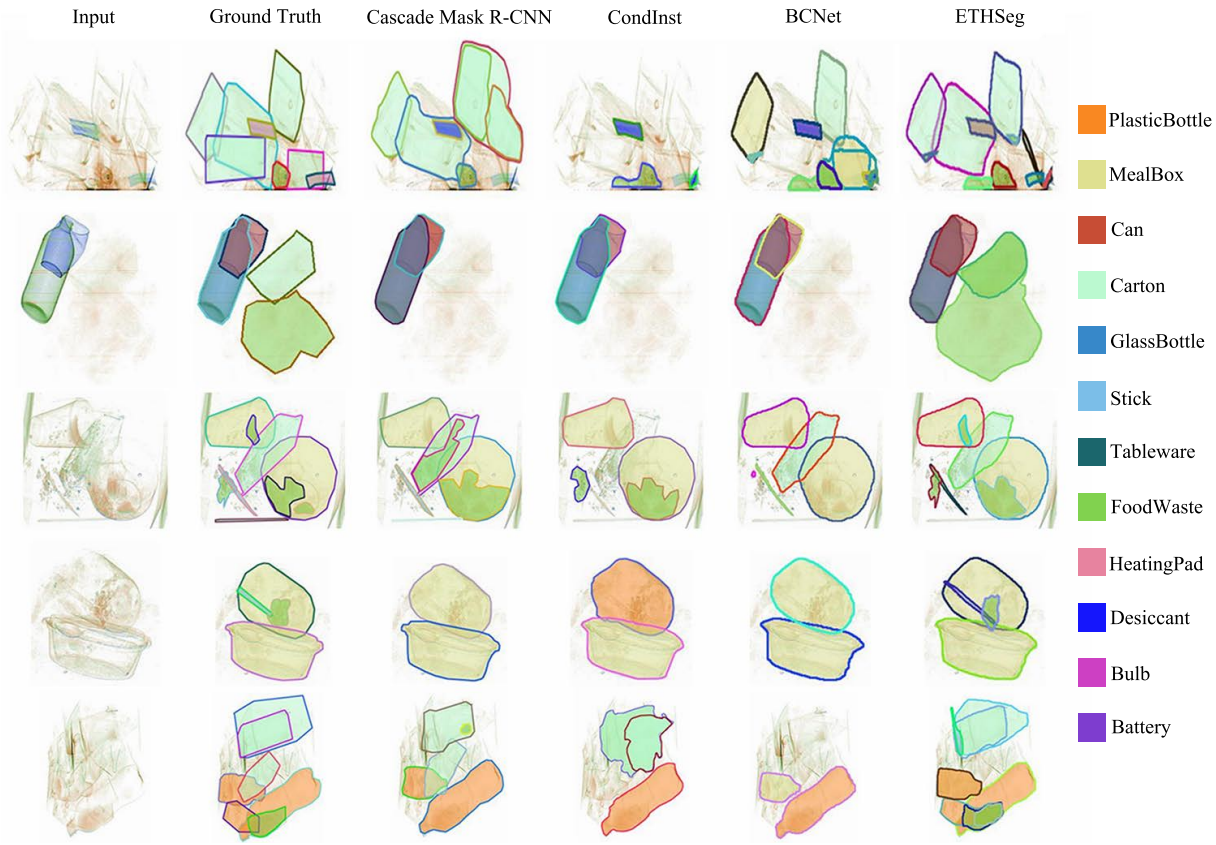

Figure 4. Qualitative comparison of instance segmentation on the proposed WIXray dataset. The mask color indicates the waste category and the boundary line is merely for identifying the instance contour.

Tab. 4 tabulates the quantitative comparison of four variant models using the *same* detection results: 1) BCNet; 2) with global structure guidance module; 3) using easy-to-hard disassembling strategy; 4) our whole framework.

Compared to the original BCNet, adding the global structure guidance module helps to improve the bounding box detection and mask prediction. This module improves the mask AP on the test set from 45.53 to 45.98. Furthermore, integrating our easy-to-hard disassembling strategy in BCNet effectively improves the mask prediction of the low-confidential proposals, increasing the mask AP to 46.12. Last, including both the easy-to-hard disassembling strategy and global structure guidance module in BCNet produces more accurate results. Compared with the baseline (*i.e.*, BCNet), our ETHSeg achieves a significant improvement of 1.32 AP for instance segmentation.

6. Conclusions

We have introduced a novel problem of instance-level waste segmentation in X-ray images, which enables accurate waste inspection without opening the waste bags. Then we created an X-ray image dataset with high-quality annotations as a benchmark for learning instance-level waste segmentation. As existing methods for natural image instance segmentation cannot well handle the penetration effect and severe occlusions existing in the X-ray image, we proposed a new method, called ETHSeg, that explicitly considers these image characteristics to achieve better performance. Experimental results on our benchmark dataset clearly demonstrate the effectiveness of our method.

Despite promising results have been shown for X-ray waste inspection, our work has the following limitations. First, we rely on the penetration effect of the X-ray for waste inspection. However, objects with low-density materials appear to be low-contrast or transparent in the X-ray image, making it difficult to inspect those objects. Second, our method still has difficulty in segmenting small objects.

Acknowledgment The work was supported in part by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, and by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104. Thanks to Chaoyue Duan et al. for their contributions in dataset collection and the ITSO in CUHKSZ for their High-Performance Computing Services.

References

[1] S. Akcay and T. P. Breckon. An evaluation of region based object detection strategies within x-ray baggage security imagery. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1337–1341. IEEE, 2017.

[2] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon. Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. *IEEE transactions on information forensics and security*, 13(9):2203–2215, 2018.

[3] N. Andriyanov, A. K. Volkov, A. K. Volkov, A. Gladkikh, and S. Danilov. Automatic x-ray image analysis for aviation security within limited computing resources. In *IOP Conference Series: Materials Science and Engineering*, volume 862, page 052009. IOP Publishing, 2020.

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[5] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9157–9166, 2019.

[6] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

[7] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[8] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8573–8581, 2020.

[9] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[10] X. Chen, J. Li, Y. Zhang, Y. Lu, and S. Liu. Automatic feature extraction in x-ray image based on deep learning approach for determination of bone age. *Future Generation Computer Systems*, 110:795–801, 2020.

[11] Y. Chu, C. Huang, X. Xie, B. Tan, S. Kamal, and X. Xiong. Multilayer hybrid deep-learning method for waste classification and recycling. *Computational Intelligence and Neuroscience*, 2018, 2018.

[12] J. L. Domingo and M. Nadal. Domestic waste composting facilities: a review of human health risks. *Environment international*, 35(2):382–389, 2009.

[13] W. Du, H. Shen, J. Fu, G. Zhang, and Q. He. Approaches for improvement of the x-ray image defect detection of automobile casting aluminum parts based on deep learning. *NDT & E International*, 107:102144, 2019.

[14] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu. Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6910–6919, 2021.

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

[16] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[18] S. Hassantabar, M. Ahmadi, and A. Sharifi. Diagnosis and detection of infected tissue of covid-19 patients based on lung x-ray image using convolutional neural network approaches. *Chaos, Solitons & Fractals*, 140:110170, 2020.

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

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

[21] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[22] K. Li and J. Malik. Amodal instance segmentation. In *European Conference on Computer Vision*, pages 677–693. Springer, 2016.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[26] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[27] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2019.

[28] G. Mittal, K. B. Yagnik, M. Garg, and N. C. Krishnan. Spotgarbage: smartphone app to detect garbage using deep learning. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 940–945, 2016.

[29] A. G. Mukherjee, U. R. Wanjari, R. Chakraborty, K. Renu, B. Vellingiri, A. George, S. R. CR, and A. V. Gopalakrishnan. A review on modern and smart technologies for efficient waste disposal and management. *Journal of Environmental Management*, 297:113347, 2021.

[30] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2019.

[31] D.-U. Park, S.-H. Ryu, S.-B. Kim, and C.-S. Yoon. An assessment of dust, endotoxin, and microorganism exposure during waste collection and sorting. *Journal of the Air & Waste Management Association*, 61(4):461–468, 2011.

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

[33] L. Qiu, X. Zhang, Y. Li, G. Li, X. Wu, Z. Xiong, X. Han, and S. Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *European Conference on Computer Vision*, pages 488–504. Springer, 2020.

[34] S. L. Rabano, M. K. Cabatuan, E. Sybingco, E. P. Dadios, and E. J. Calilung. Common garbage classification using mobilenet. In *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, pages 1–4. IEEE, 2018.

[35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[36] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[37] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[38] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[39] T. W. Rogers, N. Jaccard, E. J. Morton, and L. D. Griffin. Automated x-ray image analysis for cargo security: Critical review and future promise. *Journal of X-ray science and technology*, 25(1):33–56, 2017.

[40] V. Ruiz, Á. Sánchez, J. F. Vélez, and B. Raducanu. Automatic image-based waste classification. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 422–431. Springer, 2019.

[41] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[42] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021.

[43] Z. Tian, C. Shen, and H. Chen. Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer, 2020.

[44] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[45] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu. Towards real-world prohibited item detection: A large-scale x-ray benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5412–5421, 2021.

[46] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li. Solo: Segmenting objects by locations. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020.

[47] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen. Solov2: Dynamic and fast instance segmentation. *arXiv preprint arXiv:2003.10152*, 2020.

[48] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 138–146, 2020.

[49] X. Wu, L. Qiu, X. Gu, and Z. Long. Deep learning-based generic automatic surface defect inspection (asdi) with pixelwise segmentation. *IEEE Transactions on Instrumentation and Measurement*, 70:1–10, 2020.

[50] M. Xu, H. Zhang, and J. Yang. Prohibited item detection in airport x-ray security images via attention mechanism based cnn. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 429–439. Springer, 2018.

[51] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen. Dynamic r-cnn: Towards high quality object detection via dynamic training. In *European Conference on Computer Vision*, pages 260–275. Springer, 2020.

[52] Y. Zhang, S. Miao, T. Mansi, and R. Liao. Unsupervised x-ray image segmentation with task driven generative adversarial networks. *Medical image analysis*, 62:101664, 2020.