A Multi-Level Approach to Waste Object Segmentation

Tao Wang [1,2,3], Yuanzheng Cai [1,2,*], Lingyu Liang [4] and Dongyi Ye [2]

[1] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350108, China; twang@mju.edu.cn

[2] College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China; yiedy@fzu.edu.cn

[3] NetDragon Inc., Fuzhou 350001, China

[4] School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China; eelyliang@scut.edu.cn

* Correspondence: yuanzheng_cai@mju.edu.cn

check for updates

Abstract: We address the problem of localizing waste objects from a color image and an optional depth image, which is a key perception component for robotic interaction with such objects. Specifically, our method integrates the intensity and depth information at multiple levels of spatial granularity. Firstly, a scene-level deep network produces an initial coarse segmentation, based on which we select a few potential object regions to zoom in and perform fine segmentation. The results of the above steps are further integrated into a densely connected conditional random field that learns to respect the appearance, depth, and spatial affinities with pixel-level accuracy. In addition, we create a new RGBD waste object segmentation dataset, MJU-Waste, that is made public to facilitate future research in this area. The efficacy of our method is validated on both MJU-Waste and the Trash Annotation in Context (TACO) dataset.

1. Introduction

Waste objects are commonly found in both indoor and outdoor environments such as household, office or road scenes. As such, it is important for a vision-based intelligent robot to localize and interact with them. However, detecting and segmenting waste objects are much more challenging than most other objects. For example, waste objects could either be incomplete or damaged, or both. In many cases, their presence could only be inferred from scene-level contexts, e.g., via reasoning about their contrast to the background and judging by their intended utilities. On the other hand, one key challenge to accurately localizing waste objects is the extreme scale variation resulting from the variable physical sizes and the dynamic perspectives, as shown in Figure 1. Due to the large number of small objects, it is difficult even for most humans to accurately delineate waste object boundaries without zooming in to see the appearance details clearly. For the human vision system, however, attention can either be shifted to cover a wide area of the visual field, or narrowed to a tiny region as when we scrutinize a small area for details (e.g., [1–4]). Presented with an image, we can immediately recognize the meaning of the scene and the global structure, which allow us to easily spot objects of interest. We can consequently attend to those object regions to perform fine-grained delineation. Inspired by how the human vision system works, we solve the waste object segmentation problem in a similar manner by integrating visual cues from multiple levels of spatial granularity.

Figure 1. Example images from MJU-Waste and TACO [5] datasets and their zoomed-in object regions. Detecting and localizing waste objects require both scene level and object level reasoning. See text for details.

The general idea of exploiting objectness has long been proven effective for a wide range of vision-based applications [6–9]. In particular, several works have already demonstrated that objectness reasoning can positively impact semantic segmentation [10–13]. However, in this work we propose a simple yet effective strategy for waste object proposal that neither require pretrained objectness models nor additional object or part annotations. Our primary goal is to address the extreme scale variation which is much less common in generic objects. In order to obtain accurate and coherent segmentation results, our method performs joint inference at three levels. Firstly, we obtain a coarse segmentation at the scene-level to capture the global context and to propose potential object regions. We note that our simple object region proposal strategy captures the objectness priors reasonably well in practice. This is followed by an object-level segmentation to recover fine structural details for each object region proposal. In particular, adopting two separate models at the scene and object levels respectively allows us to disentangle the learning of the global image contexts from the learning of the fine object boundary details. Finally, we perform joint inference to integrate results from both the scene and object levels, as well as making pixel-level refinements based on color, depth, and spatial affinities. The main steps are summarized and illustrated in Figure 2. We obtain significantly superior results with our method, greatly surpassing a number of strong semantic segmentation baselines.

Recent years witnessed a huge success of deep learning in a wide spectrum of vision-based perception tasks [14–17]. In this work, we would also like to harness the powerful learning capabilities of convolutional neural network (CNN) models to address the waste object segmentation problem. Most state-of-the-art CNN-based segmentation models exploit the spatial-preserving properties of fully convolutional networks [17] to directly learn feature representations that could translate into class probability maps either at the scene level (i.e., semantic segmentation) or the object level (i.e., instance segmentation). One of the key limitations when it comes to applying these general-purpose models directly for waste object segmentation is that they are unable to handle the extreme object scale variation due to the delicate contention between global semantics and accurate localization under a fixed feature resolution, and the resulting segmentation can be inaccurate for the abundant small objects with complex shape details. Based upon this observation, we propose to learn a multi-level model that allows us to adaptively zoom into object regions to recover fine structural details, while retaining a

scene-level model to capture the long-range context and to provide object proposals. Furthermore, such a layered model can be jointly reasoned with pixel-level refinements under a unified Conditional Random Field (CRF) [18] model.
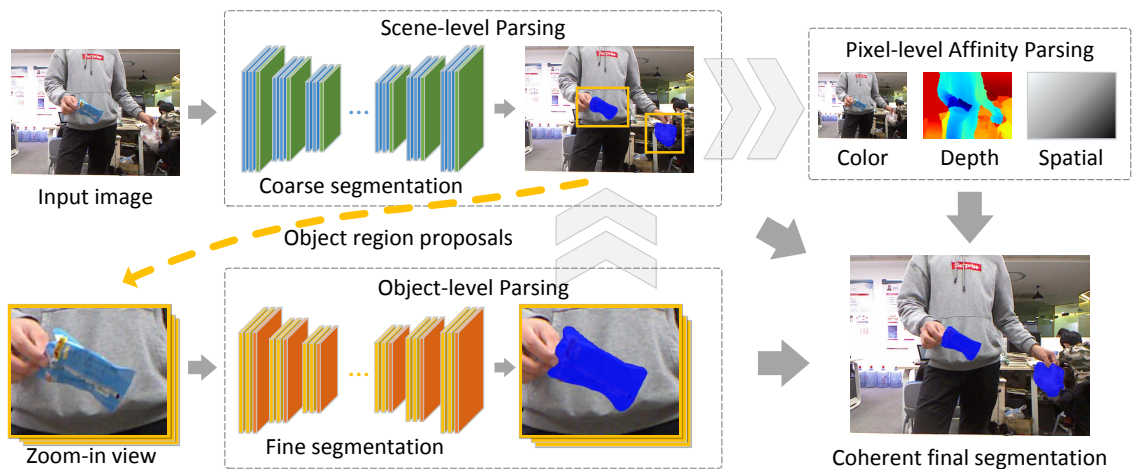
Figure 2. Overview of the proposed method. Given an input image, we approach the waste object segmentation problem at three levels: (i) scene-level parsing for an initial coarse segmentation, (ii) object-level parsing to recover fine details for each object region proposal, and (iii) pixel-level refinement based on color, depth, and spatial affinities. Together, joint inference at all these levels produces coherent final segmentation results.

The main contributions of our work are three-fold. Firstly, we propose a deep-learning based waste object segmentation framework that integrates scene-level and object-level reasoning. In particular, our method does not require additional object-level annotations. By virtue of a simple object region proposal method, we are able to learn separate scene-level and object-level segmentation models that allow us to achieve accurate localization while preserving the strong global contextual semantics. Secondly, we develop a strategy based on densely connected CRF [19] to perform joint inference at the scene, object, and pixel levels to produce a highly accurate and coherent final segmentation. In addition to the scene and object level parsing, our CRF model further refines the segmentation results with appearance, depth, and spatial affinity pairwise terms. Importantly, this CRF model is also amenable to a filtering-based efficient inference. Finally, we collected and annotated a new RGBD [20] dataset, MJU-Waste, for waste object segmentation. We believe our dataset is the first public RGBD dataset for this task. Furthermore, we evaluate our method on the TACO dataset [5], which is another public waste object segmentation benchmark. To the best of our knowledge, our work is among the first in the literature to address waste object segmentation on public datasets. Experiments on both datasets verify that our method can be used as a general framework to improve the performance of a wide range of deep models such as FCN [17], PSPNet [21], CCNet [22] and DeepLab [23].

We note that the focus of this work is to obtain accurate waste object boundary delineation. Another closely related and also very challenging task is waste object detection and classification. Ultimately, we would like to solve for waste instance segmentation with fine-grained class information. However, existing datasets do not provide a large number of object classes with sufficient training data. In addition, differentiating waste instances under a single class label is also challenging. For example, the best Average Precision (AP) obtained in [5] are in the 20s for the TACO-1 classless litter detection task where the goal is to detect and segment litter items with a single class label. Therefore, in this paper we adopt a research methodology under which we gradually move toward richer models while maintaining a high level of performance. In this regard, we formulate our problem as a two-class (waste vs. background) semantic segmentation one. This allows us to obtain high quality segmentation results as we demonstrate with our experiments.

In the remainder of this paper, Section 2 briefly reviews the literature on waste object segmentation and related tasks, as well as recent progress in semantic segmentation. We then describe details of our method in Section 3. Afterwards, Section 4 presents findings from our experimental evaluation, followed by closing remarks in Section 5.

2. Related Work

2.1. Waste Object Segmentation and Related Tasks

The ability to automatically detect, localize and classify waste objects is of wide interest in the computer and robotic vision community. However, there are relatively limited works in the literature that address the specific task of waste object segmentation. We believe this is partially due to the poor availability of public waste segmentation datasets until very recently. Therefore, in this paper we propose the MJU-Waste dataset with 2475 RGBD images each annotated with a pixelwise waste object mask. To facilitate future research, we make our dataset publicly available. To the best of our knowledge, this is the only public dataset of this kind in addition to the TACO dataset [5] of 1500 color images. Below we briefly review some recent works on waste object classification, detection and segmentation which are closely related ours.

Yang and Thung [24] addressed the waste classification problem and compared the performance of shallow and deep models. In their work, they collected the TrashNet dataset of 2500 images of single pieces of waste. Based on their data, Bircanoğlu et al. [25] and Aral et al. [26] performed detailed comparisons among various deep architectures. Additionally, Awe et al. [27] created a synthetic dataset for waste object detection based on Faster RCNN [16]. Similarly, Chu et al. [28] proposed a hybrid CNN approach for waste classification with a dataset of 5000 waste objects. Vo et al. [29] created another dataset VN-trash with 5904 images for deep transfer learning. Furthermore, Ramalingam et al. [30] presented a debris classification model for floor-cleaning robots with a cascade CNN and an SVM. Yin et al. [31] proposed a lightweight CNN for food litter detection in table cleaning tasks. Rad et al. [32] presented an approach similar to OverFeat [33] for litter object detection. Another similar approach based on Faster RCNN [16] is presented by Wang and Zhang [34]. Contrary to the above works, we address the waste object segmentation problem that requires accurate delineation of object boundaries.

In terms of methods that involves a segmentation component, Bai et al. [35] designed a robot for picking up garbage on the grass with a two-stage perception approach. Firstly, they used SegNet [36] for ground segmentation to allow the robot to move toward waste objects. After a close-range image is acquired, ResNet [14] is used for object classification. Here the segmentation module is used for background modeling of the grassland only, hence no object segmentation is performed. Deepa et al. [37] presented a garbage coverage segmentation method in water terrain based on color transformation and K-means. In addition, Mittal et al. [38] proposed an approach based on the Fully Convolutional Network (FCN) [17] for coarse garbage segmentation. Their method is based on extracting image patches and combining their predictions, and therefore cannot capture the finer object boundary details. Zeng et al. [39] proposed a multi-scale CNN based garbage detection method from airborne hyperspectral data. In their method, a binary segmentation map is generated as the input to selective search [7] for the purpose of obtaining bounding box-based region proposals. All these above works do not address the specific task of waste object segmentation for robotic interaction.

Perhaps being the closest to our work, Zhang et al. [40] proposed an object segmentation method for waste disposal lines based on RGBD sensors. Their method begins with background subtraction on the 3D point cloud, and then attempts to find an optimal projection plane for subsequent object segmentation. Another work [41] from the same group proposed a relabeling method for ambiguous regions after the background is subtracted. Unlike their methods, we take a data-driven approach to address the problem in a much more challenging scenario. Specifically, our method does not

assume a particular background model and is able to segment waste objects in both hand-held and in-the-wild scenarios.

Another work that is conceptually similar to ours is from Grard et al. [42]. They explored an interesting interactive setting for object segmentation from a cluttered background. A user is asked to click on an object to extract, and their model uses a dual-objective FCN trained on synthetic depth images to produce the object mask. In our work, however, we aim at a more challenging scenario that does not require human interaction.

2.2. Semantic Segmentation

The task of assigning a class label to every pixel in an image is a long-established fundamental problem in computer vision [43–51]. Since the pioneering work of Long, Shelhamer and Darrell [17], researchers proposed a large number of architectures and techniques to improve upon the FCN to capture multi-scale context [52–58] or to improve results at object boundaries [59–66]. Other recent works have explored encoder-decoder structures [36,67,68] and contextual dependencies based on the self-attention mechanism [22,69–71]. For example, some of the recent influential works include PSPNet [21] which proposed the pyramid pooling module to combine representations from multiple scales and RefineNet [72] in which a multi-path refinement network is proposed for high-resolution semantic segmentation. Furthermore, Gated SCNN [73] proposed a two-stream structure that explicitly enhances shape prediction. We note that the performance of semantic segmentation algorithms is clearly related to the backbone architecture being used. For example, the recently proposed ResNeSt [15] model based on Split-Attention blocks provided large performance improvements to a number of vision tasks including semantic segmentation.

We note that our work is similar to DeepLab [74] whose main contributions include the atrous spatial pyramid pooling (ASPP) module to capture the multi-scale context and the use of dense CRF [19] to improve results at object boundaries. In their follow up work [23,75], they improved the ASPP module by adding image pooling and a decoder structure. Our method differs from the DeepLab series in two important aspects. Firstly, our method is tailored to the task of waste object segmentation and introduces layered deep models that perform scene-level parsing and object-level parsing respectively. Secondly, the dense CRF model in our work integrates information from both the scene and object level parsing results, as well as pixel-level affinities which can additionally encode local geometric information via input depth images. In fact, we demonstrate through our experiments that the proposed method is a general framework which can be applied in conjunction with DeepLab and other strong semantic segmentation baselines such as PSPNet [21] and CCNet [22] to improve their results by a clear margin on the waste object segmentation task.

There have been a few works that explored semantic segmentation with RGBD data [76–79]. For example, FuseNet [80] proposed a two-stream encoder that extracts features from both color and depth images in an encoder-decoder type of network. Qi et al. [81] constructed a graph neural network based on spatial affinities inferred from depth. Contrary to existing works, we use depth affinities as a means to refining waste object segmentation results. Our use of depth information is flexible in that the model can cope with situations where the depth modality is present or absent without re-training.

Lastly, there have been a few works that tackle objectness aware semantic segmentation [10,11]. Apart from addressing the problem in the novel waste object segmentation domain, our method uses a simple yet effective strategy for object region proposal that does not require any additional object or part annotations.

3. Our Approach

In this section, let us formally introduce the waste object segmentation problem and the proposed approach. We begin with the definition of the problem and notations. Given an input color image and optionally an additional depth image, our model outputs a pixelwise labeling map, as shown in Figure 2. Mathematically, denote the input color image as $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the optional depth image as

$\mathbf{D} \in \mathbb{R}^{H \times W}$, and the semantic label set as $\mathcal{C} = \{1, 2, \ldots, C\}$, where C is the number of classes. Our goal is to produce a structured semantic labeling $\mathbf{x} \in \mathcal{C}^{H \times W}$. We note that in deep models, the labeling of $\mathbf{x}$ at image coordinate (i, j), x_{ij}, is usually obtained via multi-class softmax scores on a spatial-preserving convolutional feature map $\mathbf{C} \in \mathbb{R}^{H \times W \times C}$, i.e., $x_{ij} = \underset{k'}{\operatorname{argmax}} \frac{\exp(\mathbf{C}^{i,j,k'})}{\sum_{k=1}^{C} \exp(\mathbf{C}^{i,j,k})}$. In practice, it is common that the convolutional feature map $\mathbf{C}$ is downsampled w.r.t. the original image resolution, but we can always assume that the resolution can be restored with interpolation.

3.1. Layered Deep Models

In this work, we apply deep models at both the scene and the object levels. For this purpose, let us define a number of image regions in which we obtain deeply trained feature representations. Firstly, let $\mathcal{R}_0 = \{(i, j)_{i \in \{1 \ldots H\}, j \in \{1 \ldots W\}}\}$ be the set of all spatial coordinates on the image plane, or the entire image region. This is the region in which we perform scene-level parsing (i.e., coarse segmentation). In addition, we perform object-level parsing (i.e., fine segmentation) on a set of non-overlapping object region proposals. We denote each of these additional regions as $\mathcal{R}_l = \{(i, j)_{i \in \{1 \ldots H_l\}, j \in \{1 \ldots W_l\}}\}, 1 \leq l \leq L$. Details on generating these regions are discussed in Section 3.3. We apply our coarse segmentation feature embedding network $\mathcal{F}_c$ and the fine segmentation feature embedding network $\mathcal{F}_f$ to the appropriate image regions as follows:

$$
\begin{aligned}
\mathbf{C}_0 &= \mathcal{F}_c(\mathcal{R}_0 \Rightarrow \mathbf{I}), \\
\mathbf{C}_l &= \mathcal{F}_f(\mathcal{R}_l \Rightarrow \mathbf{I}), 1 \leq l \leq L
\end{aligned}
\tag{1}
$$

where $\mathcal{R}_l \Rightarrow \mathbf{I}$ denotes cropping the region $\mathcal{R}_l$ from image $\mathbf{I}$. Here $\mathbf{C}_0 \in \mathbb{R}^{H \times W \times C}$ and $\mathbf{C}_l \in \mathbb{R}^{H_l \times W_l \times C}$, and we note that these feature maps are upsampled where necessary. In addition, the spatial dimension may be image and region specific for both $\mathcal{R}_0$ and $\mathcal{R}_l$, which poses a practical problem for batch-based training. To address this issue, during CNN training we resize all image regions so that they have a common shorter side length, followed by randomly cropping a fixed-sized patch as part of the data augmentation procedure. We refer the readers to Section 4.2 for details. In Figure 3, the processes shown in blue and yellow illustrate the steps described in this section.
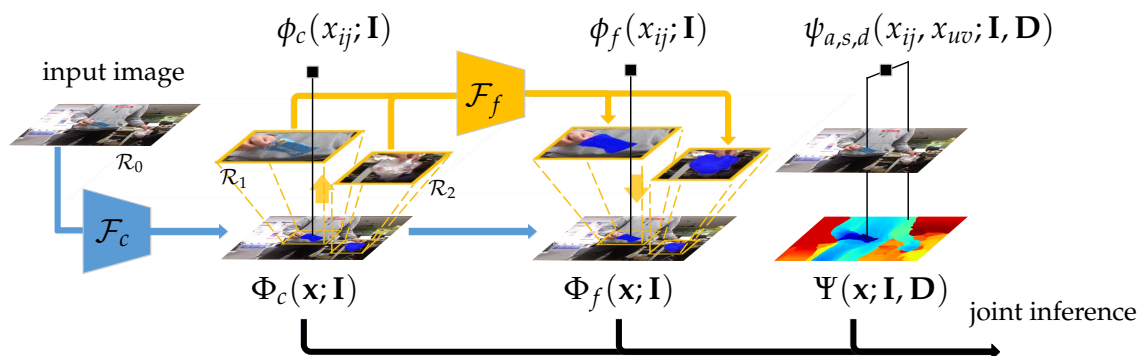
Figure 3. The graphical representation of our CRF. $\mathcal{F}_c$ and $\mathcal{F}_f$ represent the feature embedding functions for the coarse and the fine segmentation networks. Our model consists of the scene-level unary term $\Phi_c(\mathbf{x}; \mathbf{I})$, the object-level unary term $\Phi_f(\mathbf{x}; \mathbf{I})$, and the pixel-level pairwise term $\Psi(\mathbf{x}; \mathbf{I}, \mathbf{D})$.

3.2. Coherent Segmentation with CRF

Given the layered deep models, we now introduce our graphical model for predicting coherent waste object segmentation results. Specifically, the overall energy function of our CRF model consists of three main components:

$$
E(\mathbf{x}, \mathbf{I}, \mathbf{D}) = \Phi_c(\mathbf{x}; \mathbf{I}) + \alpha \cdot \Phi_f(\mathbf{x}; \mathbf{I}) + \Psi(\mathbf{x}; \mathbf{I}, \mathbf{D})
\tag{2}
$$

where $\Phi_c(\mathbf{x}; \mathbf{I})$ represents the scene-level coarse segmentation potentials, $\Phi_f(\mathbf{x}; \mathbf{I})$ denotes the object-level fine segmentation potentials, and $\Psi(\mathbf{x}; \mathbf{I}, \mathbf{D})$ is the pairwise potentials that respect the color, depth, and spatial affinities in the input images. α is the weight for the relative importance among the two unary terms. The graphical representation of our CRF model is shown in Figure 3. We describe the details of these three terms below.

Scene-level unary term. The scene-level coarse segmentation unary term is given by $\Phi_c(\mathbf{x}; \mathbf{I}) = \sum_{(i,j) \in \mathcal{R}_0} \phi_c(x_{ij}; \mathbf{I}) = \sum_{(i,j) \in \mathcal{R}_0} -\log(P_c(x_{ij}; \mathbf{I}))$ where $P_c(x_{ij}; \mathbf{I})$ is a pixelwise softmax on the feature map $\mathbf{C}_0$ as follows:

$$P_c(x_{ij}; \mathbf{I}) = \frac{\exp(\mathbf{C}_0^{i,j,k'})}{\sum_{k=1}^{C} \exp(\mathbf{C}_0^{i,j,k})} [\![x_{ij} = k']\!] \tag{3}$$

where $[\![\cdot]\!]$ denotes the indicator function. This term produces a coarse segmentation map based on the long-range contexts from the input image. Importantly, we use the output of this term to generate our object region proposals, as discussed in Section 3.3.

Object-level unary term. The object-level fine segmentation unary term is given by $\Phi_f(\mathbf{x}; \mathbf{I}) = \sum_{(i,j) \in \mathcal{R}_0} \phi_f(x_{ij}; \mathbf{I})$ where $\phi_f(x_{ij}; \mathbf{I})$ is defined as:

$$\phi_f(x_{ij}; I) = \begin{cases} -\log(P_f(x_{ij}, l; \mathbf{I})), & \text{if } (i,j) \in \mathcal{R}_l, 1 \le l \le L, \\ \phi_c(x_{ij}; \mathbf{I}), & \text{otherwise} \end{cases} \tag{4}$$

The formulation above states that if a pixel location (i, j) belongs to one of the L object region proposals, a negative log-probability obtained via fine segmentation is adopted. Otherwise, the object-level unary term falls back to the scene-level unary potentials. Here the probability $P_f(x_{ij}, l; \mathbf{I})$ given by the fine segmentation model is obtained as follows:

$$P_f(x_{ij}, l; \mathbf{I}) = \frac{\exp(\mathbf{C}_l^{\mathcal{T}_l^1(i), \mathcal{T}_l^2(j), k'})}{\sum_{k=1}^{C} \exp(\mathbf{C}_l^{\mathcal{T}_l^1(i), \mathcal{T}_l^2(j), k})} [\![x_{ij} = k']\!] \tag{5}$$

where $\mathcal{T}_l^1(\cdot)$ and $\mathcal{T}_l^2(\cdot)$ are translation functions that map the image coordinates to that of the l-th object proposal region, and $\mathbf{C}_l$ is the output feature embedding from the fine segmentation model for the l-th object proposal region. We note that the object-level unary potentials typically recover more fine details along object boundaries, as opposed to the scene-level unary potentials. In general, it would become too computationally expensive to compute scene-level potentials at a comparable resolution for the entire image. In most cases, computing the object-level unary term on less than 3 object region proposals are sufficient, see Section 3.3 for details. Additionally, our object-level potentials are obtained via a separate deep model that allows us to decouple the learning of long-range contexts from the learning of fine structural details.

Pixel-level pairwise term. Although the object-level unary potentials provide finer segmentation details, accurate boundary details could still be lost for some irregularly shaped waste objects. This poses a practical challenge for detail-preserving global inference.

Following [19], we address this challenge by introducing a pairwise term that is a linear combination of Gaussian kernels in a joint feature space that includes color, depth, and spatial coordinates. This allows us to produce coherent object segmentation results that respect the appearance, depth, and spatial affinities in the original image resolution. More importantly, this form of the pairwise

term allows for efficient global inference [19]. Specifically, our pairwise term $\Psi(\mathbf{x}; \mathbf{I}, \mathbf{D})$ includes an appearance term $\psi_a(x_{ij}, x_{uv}; \mathbf{I})$, a spatial smoothing term $\psi_s(x_{ij}, x_{uv})$ and a depth term $\psi_d(x_{ij}, x_{uv}; \mathbf{D})$:

$$
\Psi(\mathbf{x}; \mathbf{I}, \mathbf{D}) = \sum_{\substack{(i,j)\in\mathcal{R}_0 \\ (u,v)\in\mathcal{R}_0}} [\![x_{ij} \neq x_{uv}]\!] \Big[\psi_a(x_{ij}, x_{uv}; \mathbf{I})
$$

$$
+ \psi_s(x_{ij}, x_{uv}) + \psi_d(x_{ij}, x_{uv}, \mathbf{D}) \Big] \tag{6}
$$

where $[\![x_{ij} \neq x_{uv}]\!]$ is the Potts label compatibility function. The appearance term and the smoothing term follow [19] and take the following form:

$$
\psi_a(x_{ij}, x_{uv}; \mathbf{I}) = w^{(a)} \exp\left(-\frac{|p_{ij} - p_{uv}|^2}{2\theta_\alpha^2} - \frac{|I_{ij} - I_{uv}|^2}{2\theta_\beta^2} \right) \tag{7}
$$

$$
\psi_s(x_{ij}, x_{uv}) = w^{(s)} \exp\left(-\frac{|p_{ij} - p_{uv}|^2}{2\theta_\gamma^2} \right) \tag{8}
$$

where I_{ij} and p_{ij} are the image appearance and position features at the pixel location (i, j). In addition, when a input depth image $\mathbf{D}$ is available, we are able to enforce an additional pairwise term induced by geometric affinities:

$$
\psi_d(x_{ij}, x_{uv}; \mathbf{D}) = w^{(d)} \exp\left(-\frac{|p_{ij} - p_{uv}|^2}{2\theta_\delta^2} - \frac{|D_{ij} - D_{uv}|^2}{2\theta_\epsilon^2} \right) \tag{9}
$$

where D_{ij} is the depth reading at the pixel location (i, j). We note that in practice, any missing values in $\mathbf{D}$ are filled in with a median filter [82] beforehand, see Section 4.1 for details. In addition, Equation (9) can be conveniently added or removed depending on the depth data availability. We simply discard Equation (9) when training models for the TACO dataset which only contains color images.

3.3. Generating Object Region Proposals

In this work, we follow a simple strategy to generate object region proposals $\mathcal{R}_l = \{(i,j)_{i \in \{1...H_l\}, j \in \{1...W_l\}}\}, 1 \leq l \leq L$. In particular, the output from the scene-level coarse segmentation model is a good indication of the waste object locations. See Figure 3 for an example. We begin with extracting the connected components in the foreground class labelings of $\mathbf{x} \in \mathcal{C}^{H \times W}$ from the maximum a posterior (MAP) estimate of the scene-level unary term $\Phi_c(\mathbf{x}; \mathbf{I})$. For each connected component, a tight bounding box $\mathcal{R}_l^t$ is extracted. This is followed by extending $\mathcal{R}_l^t$ by 30% in four directions (i.e., N,S,W,E), subject to the image boundary truncation. Finally, we merge overlapping regions and remove those below or above certain size thresholds (details in Section 4.2) to obtain a concise set of final object region proposals $\mathcal{R}_l, 1 \leq l \leq L$. Example object region proposals obtained using this procedure are shown in Figure 4, and we note that any similar implementation should also work satisfactorily.

Most images from the MJU-Waste dataset contain only one hand-held waste object per image. For DeepLabv3 with a ResNet-50 backbone, for example, only 2.4% of all images from MJU-Waste produce 2 or more object proposals. For the TACO dataset, 24.0% of all images produce 2 or more object proposals. However, only 8.1% and 0.6% of all images produce more than 3 and 5 object proposals, respectively.
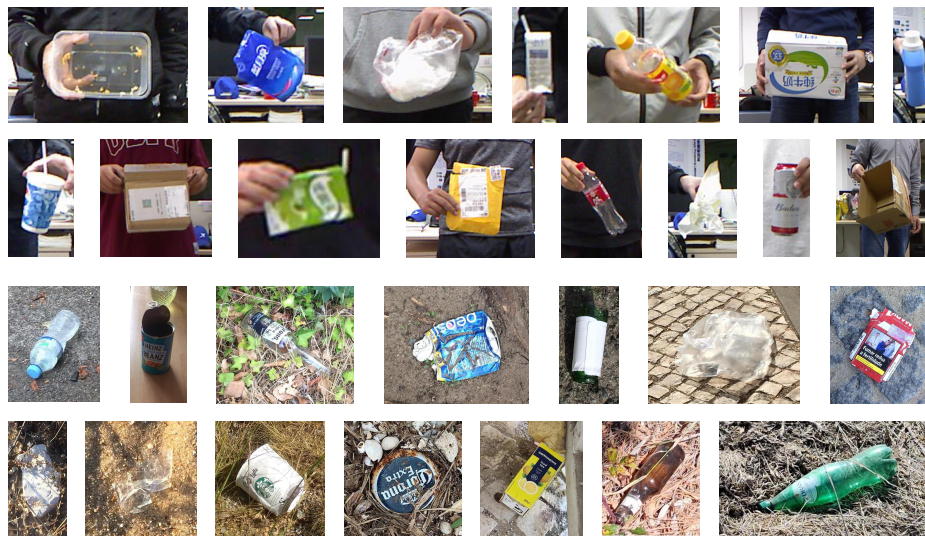
Figure 4. Example object region proposals. The first two rows show the object region proposals from the MJU-Waste dataset. The remaining two rows show the object region proposals from the TACO dataset.

3.4. Model Inference

Following [19], we use the mean field approximation of the joint probability distribution $P(\mathbf{x})$ that computes a factorized distribution $Q(\mathbf{x})$ which minimizes the KL-divergence [83,84] $KL(Q||P)$. For our model, this yields the following message passing-based iterative update equation:

$$
\begin{aligned}
Q_{ij}(x_{ij} = k) = &\frac{1}{Z_{ij}} \exp \Big\{ - \phi_c(x_{ij}) - \alpha \cdot \phi_f(x_{ij}) \\
&- \sum_{k' \in \mathcal{C}} [\![k, k']\!] \sum_{uv \neq ij} \big[\psi_a(x_{ij}, x_{uv}) + \psi_s(x_{ij}, x_{uv}) + \psi_d(x_{ij}, x_{uv}) \big] Q_{uv}(k') \Big\}
\end{aligned}
\tag{10}
$$

where the input color image $\mathbf{I}$ and the depth image $\mathbf{D}$ are omitted for notation simplicity. In practice, we use the efficient message passing algorithm proposed in [19]. The number of iterations is set to 10 in all experiments.

3.5. Model Learning

Let us now move on to discuss details pertaining to the learning of our model. Specifically, we learn the parameters of our model by piecewise training. First, the coarse segmentation feature embedding network $\mathcal{F}_c$ is trained with standard cross-entropy (CE) loss on the predicted coarse segmentation. Based on the coarse segmentation for the training images, we extract object region proposals with the method discussed in Section 3.3. This allows us to then train the fine segmentation feature embedding network $\mathcal{F}_f$ using the cropped object regions in a similar manner. Next, we learn the weight and the kernel parameters of our CRF model. We initialize them to the default values used in [19] and then use grid search to finetune their values on a held-out validation set. We note that our model is not too sensitive to most of the parameters. On each dataset, we use fixed values of these parameters for all CNN architectures. See Section 4.2 for details.

4. Experimental Evaluation

In this section, we compare the proposed method with state-of-the-art semantic segmentation baselines. We focus on two challenging scenarios for waste object localization: the hand-held setting (for applications such as service robot interactions or smart trash bins) and waste objects "in the wild".

In our experiments, we found that one of the common challenges for both scenarios is the extreme scale variation causing standard segmentation algorithms to underperform. Our proposed method, however, greatly improves the segmentation performance in these adverse scenarios. Specifically, we evaluate our method on the following two datasets:

- **MJU-Waste Dataset.** In this work, we created a new benchmark for waste object segmentation. The dataset is available from https://github.com/realwecan/mju-waste/. To the best of our knowledge, MJU-Waste is the largest public benchmark available for waste object segmentation, with 1485 images for training, 248 for validation and 742 for testing. For each color image, we provide the co-registered depth image captured using an RGBD camera. We manually labeled each of the image. More details about our dataset are presented in Section 4.1.
- **TACO Dataset.** The Trash Annotations in COntext (TACO) dataset [5] is another public benchmark for waste object segmentation. Images are collected from mainly outdoor environments such as woods, roads and beaches. The dataset is available from http://tacodataset.org/. Individual images in this dataset are either under the CC BY 4.0 license or the ODBL (c) OpenLitterMap & Contributors license. See http://tacodataset.org/ for details. The current version of the dataset contains 1500 images, and a split with 1200 images for training, 150 for validation and 150 for testing is available from the authors. In all experiments that follow, we use this split from the authors.

We summarize the key statistics of the two datasets in Table 1. Once again, we emphasize that one of the key characteristics of waste objects is that the number of objects per class can be highly imbalanced (e.g., in the case of TACO [5]). In order to obtain sufficient data to train a strong segmentation algorithm, we use a single class label for all waste objects, and our problem is therefore defined as a binary pixelwise prediction one (i.e., waste vs. background). For the quantitative evaluation that follows, we report the performance of baseline methods and the proposed method by four criteria: Intersection over Union (IoU) for the waste object class, mean IoU (mIoU), pixel Precision (Prec) for the waste object class, and Mean pixel precision (Mean). Let TP, FP and FN denote the total number of true positive, false positive and false negative pixels, respectively. The four criteria used are defined as follows:

- Intersection over Union (IoU) for the c-th class is the intersection of the prediction and ground-truth regions of the c-th class over the union of them, defined as:

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c} \tag{11}$$

- mean IoU (mIoU) is the average IoU of all C classes:

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c} \tag{12}$$

- Pixel Precision (Prec) for the c-th class is the percentage of correctly classified pixels of all predictions of the c-th class:

$$\text{Prec}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \tag{13}$$

- Mean pixel precision (Mean) is the average class-wise pixel precision:

$$\text{Mean} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \tag{14}$$

We note that the image labelings are typically dominated by the background class, therefore IoU and Prec reported on the waste objects only are more sensitive than mIoU and Mean which consider both waste objects and the background.

Table 1. Key statistics of the two datasets used in our experimental evaluation. Data splits are the number of training + validation + test images. The image size varies in TACO so we report the average image size here. Currently, MJU-Waste uses a single class label for all waste objects (in addition to the background class). For TACO, there are 60 categories which belong to 28 super (top) categories.

	Modalities	Images	Data Split	Image Size	Objects Per Image	Classes
MJU-Waste	RGBD	2475	1485 + 248 + 742	640 × 480	1.02	single
TACO	RGB only	1500	1200 + 150 + 150	≈3223 × 2825	3.19	60 (28)

4.1. The MJU-Waste Dataset

Before we move on to report our findings from the experiments, let us more formally introduce the MJU-Waste dataset. We created this dataset by collecting waste items from a university campus, bringing them back to a lab, and then take pictures of people holding waste items in their hands. All images in the dataset are captured using a Microsoft Kinect RGBD camera [20]. The current version of our dataset, MJU-Waste V1, contains 2475 co-registered RGB and depth image pairs. Specifically, we randomly split the images into a training set, a validation set and a test set of 1485, 248 and 742 images, respectively.

Due to sensor limitations, the depth frames contain missing values at reflective surfaces, occlusion boundaries, and distant regions. We use a median filter [82] to fill in the missing values in order to obtain high quality depth images. Each image in MJU-Waste is annotated with a pixelwise mask of waste objects. Example color frames, ground-truth annotations, and depth frames are shown in Figure 5. In addition to semantic segmentation ground-truths, object instance masks are also available.
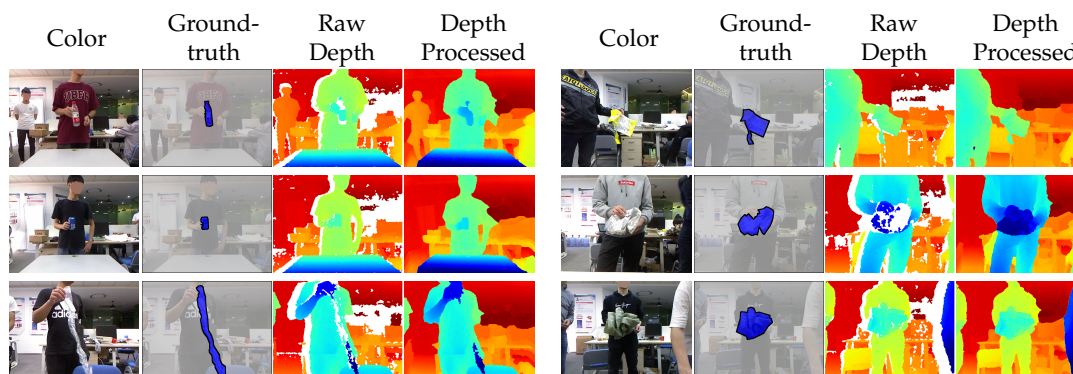
Figure 5. Example color frames, ground-truth annotations, and depth frames from the MJU-Waste dataset. Ground-truth masks are shown in blue. Missing values in the raw depth frames are shown in white. These values are filled in with a median filter following [82].

4.2. Implementation Details

Here we report the key implementation details of our experiments, as follows:

- **Segmentation networks $\mathcal{F}_c$ and $\mathcal{F}_f$.** Following [21,74], we use the polynomial learning rate policy with the initial learning rate set to 0.001 and the power factor set to 0.9. The total number of iterations are set to 50 epochs on both datasets with a batch size of 4 images. In all experiments, we use the ImageNet-pretrained backbones [85] and a standard SGD optimizer with momentum and weight decay factors set to 0.9 and 0.0001, respectively. To avoid overfitting, standard data augmentation techniques including random mirroring, resizing (with a resize factor between 0.5

and 2), cropping and random Gaussian blur [21] are used. The base (and cropped) image sizes for $\mathcal{F}_c$ and $\mathcal{F}_f$ are set to 520(480) and 260(240) pixels during training, respectively.

- **Object region proposals.** To maintain a concise set of object region proposals, we empirically set the minimum and maximum number of pixels $N_{\min}$ and $N_{\max}$ in an object region proposal. For MJU-Waste, $N_{\min}$ and $N_{\max}$ are set to 900 and 40,000, respectively. For TACO, $N_{\min}$ and $N_{\max}$ are set to 25,000 and 250,000, due to the larger image sizes. Object region proposals that are either too small or too big will simply be discarded.

- **CRF parameters.** We initialize the CRF parameters with the default values in [19] and follow a simple grid search strategy to find the optimal values of CRF parameters in each term. For reference, the CRF parameters used in our experiments are listed in Table 2. We note that our model is somewhat robust to the exact values of these parameters, and for each dataset we use the same parameters for all segmentation models.

- **Training details and codes.** In this work, we use a publicly available implementation to train the segmentation networks $\mathcal{F}_c$ and $\mathcal{F}_f$. The CNN training codes are available from: https://github.com/Tramac/awesome-semantic-segmentation-pytorch/. We use the default training settings unless otherwise specified earlier in this section. For CRF inference we use another public implementation. The CRF inference codes are available from: https://github.com/lucasb-eyer/pydensecrf/. The complete set of CRF parameters are summarized in Table 2.

Table 2. CRF parameters used in our experiments. Depth terms are not applicable to the TACO dataset.

	α	$w^{(a)}$	$w^{(s)}$	$w^{(d)}$	θ_α	θ_β	θ_γ	θ_δ	θ_ϵ
MJU-Waste	1	3	1	1	20	20	1	10	20
TACO	1	3	1	-	100	20	10	-	-

4.3. Results on the MJU-Waste Dataset

The quantitative performance evaluation results we obtained on the test set of MJU-Waste are summarized in Table 3. Methods using our proposed multi-level model have "ML" in their names. For this dataset, we report the performance of the following baseline methods:

- **FCN-8s [17].** FCN is a seminal work in CNN-based semantic segmentation. In particular, FCN proposes to transform fully connected layers into convolutional layers that enables a classification net to output a probabilistic heatmap of object layouts. In our experiments, we use the network architecture as proposed in [17], which adopts a VGG16 [86] backbone. In terms of the skip connections, we choose the FCN-8s variant as it retains more precise location information by fusing features from the early *pool*3 and *pool*4 layers.

- **PSPNet [21].** PSPNet proposes the pyramid pooling module for multi-scale context aggregation. Specifically, we choose the ResNet-101 [14] backbone variant for a good tradeoff between model complexity and performance. The pyramid pooling module concatenates the features from the last layer of the *conv*4 block with the same features applied with $1 \times 1, 2 \times 2, 3 \times 3$ and 6×6 average pooling and upsampling to harvest multi-scale contexts.

- **CCNet [22].** CCNet presents an attention-based context aggregation method for semantic segmentation. We also choose the ResNet-101 backbone for this method. Therefore, the overall architecture is similar to PSPNet except that we use the Recurrent Criss Cross Attention (RCCA) module for context modeling. Specifically, given the *conv*4 features, the RCCA module obtains a self-attention map to aggregate the context information in horizontal and vertical directions. Similarly, the resultant features are concatenated with the *conv*4 features for downstream segmentation.

- **DeepLabv3 [23].** DeepLabv3 proposes the Atrous Spatial Pyramid Pooling (ASPP) module for capturing the long-range contexts. Specifically, ASPP proposes the parallel dilated convolutions with varying atrous rates to encode features from different sized receptive fields. The atrous rates

used in our experiments are 12, 24 and 36. In addition, we experimented with both ResNet-50 and ResNet-101 backbones on the MJU-Waste dataset to explore the performance impact of different backbone architectures.

Table 3. Performance comparisons on the test set of MJU-Waste. For each method, we report the IoU for waste objects (IoU), mean IoU (mIoU), pixel Precision for waste objects (Prec) and Mean pixel precision (Mean). See Section 4.3 for details.

Dataset: MJU-Waste (Test)	Backbone	IoU	mIoU	Prec	Mean
Baseline Approaches					
FCN-8s [17]	VGG-16	75.28	87.35	85.95	92.83
PSPNet [21]	ResNet-101	78.62	89.06	86.42	93.11
CCNet [22]	ResNet-101	83.44	91.54	**92.92**	**96.35**
DeepLabv3 [23]	ResNet-50	79.92	89.73	86.30	93.06
DeepLabv3 [23]	ResNet-101	**84.11**	**91.88**	89.69	94.77
Proposed Multi-Level (ML) Model					
FCN-8s-ML	VGG-16	82.29 (+7.01)	90.95 (+3.60)	91.75 (+5.80)	95.76 (+2.93)
PSPNet-ML	ResNet-101	81.81 (+3.19)	90.70 (+1.64)	89.65 (+3.23)	94.73 (+1.62)
CCNet-ML	ResNet-101	86.63 (+3.19)	93.17 (+1.63)	**96.05** (+3.13)	**97.92** (+1.57)
DeepLabv3-ML	ResNet-50	84.35 (+4.43)	92.00 (+2.27)	91.73 (+5.43)	95.78 (+2.72)
DeepLabv3-ML	ResNet-101	**87.84** (+3.73)	**93.79** (+1.91)	94.43 (+4.74)	97.14 (+2.37)

We refer interested readers to the public implementation discussed in Section 4.2 for the network details of the above baselines. For each baseline method, we additionally implement our proposed multi-level modules and then present a direct performance comparison in terms of IoU, mIoU, Prec and Mean improvements. We show that our method provides a general framework under which a number of strong semantic segmentation baselines could be further improved. For example, FCN-8s benefits the most from a multi-level approach (i.e., +7.01 points of IoU improvement), partially due to the relatively low baseline performance. Even for the best-performing baseline, DeepLabv3 with a ResNet-101 backbone, our multi-level model further improves its performance by +3.73 IoU points. We note that such a large quantitative improvement can also be visually significant. In Figure 6, we present qualitative comparisons between FCN-8s, DeepLabv3 and their multi-level counterparts. It is clear that our approach helps to remove false positives in some non-object regions. More importantly, it is evident that multi-level models more precisely follow object boundaries.

In Table 4, we additionally perform ablation studies on the validation set of MJU-Waste. Specifically, we compare the performance of the following variants of our method:

- **Baseline.** DeepLabv3 baseline with a ResNet-50 backbone.
- **Object only.** The above baseline with additional object-level reasoning. This method is implemented by retaining only the two unary terms of Equation (2). All pixel-level pairwise terms are turned off. This will test if the object-level reasoning will contribute to the baseline performance.
- **Object and appearance.** The baseline with object-level reasoning plus the appearance and the spatial smoothing pairwise terms. The depth pairwise terms are turned off. This will test if the additional pixel affinity information (without depth, however) is useful. It also verifies the efficacy of the depth pairwise terms.
- **Appearance and depth.** The baseline with all pixel-level pairwise terms but without the object-level unary term. This will test if an object-level fine segmentation network is necessary, as well as the performance contribution of the pixel-level pairwise terms alone.

- **Full model.** Our full model with all components proposed in Section 3.
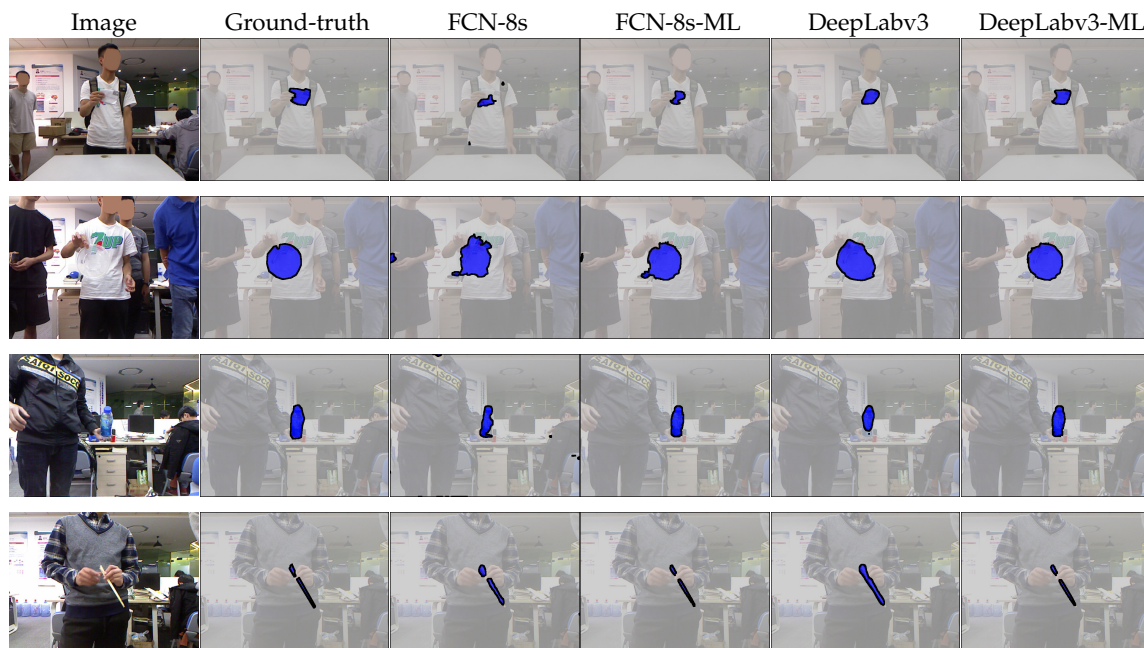
Figure 6. Example segmentation results on the MJU-Waste test set. Input images and ground-truth annotations are shown in the first two columns. Baseline methods are FCN-8s (VGG-16) and DeepLabv3 (ResNet-50). Our proposed methods (FCN-8s-ML and DeepLabv3-ML) more accurately recover object boundaries. Best viewed electronically, zoomed in.

Table 4. Results from our ablation studies carried out on the validation set of MJU-Waste. The baseline method is DeepLabv3 with a ResNet-50 backbone. We add different components proposed in Section 3 individually to test their performance impact. See Section 4.3 for details.

Dataset: MJU-Waste (val)	Object?	Appearance?	Depth?	IoU	mIoU	Prec	Mean
Baseline	✗	✗	✗	80.86	90.24	87.49	93.67
+ components	✓	✗	✗	81.43	90.53	88.06	93.96
	✓	✓	✗	85.44	92.58	91.79	95.83
	✗	✓	✓	83.45	91.57	91.84	95.83
Full model	✓	✓	✓	**86.07**	**92.90**	**92.77**	**96.32**

Results are clear that the full model performs the best, producing superior performance by all four criteria. This validates that the various components proposed in our method all positively impact the final results.

In terms of the computational efficiency, we report a breakdown of the average per-image inference time in Table 5. The baseline method corresponds to the scene-level inference only; additional object and pixel level inference incurs extra computational costs. These runtime statistics are obtained with an i9 desktop CPU and a single RTX 2080Ti GPU. Our full model with DeepLabv3 and ResNet-50 runs at approximately 0.8 second per image. Specifically, the computational costs for object-level inference are mainly a result from the object region proposals and the forward pass of the object region CNN. The pixel-level inference time, on the other hand, is mostly the result from the iterative mean-field approximation. It should be noted that the inference times reported here are obtained based on public implementations as mentioned in Section 4.2, without any specific optimization.

More example results obtained on the test set of MJU-Waste with our full model are shown in Figure 7. Although the images in MJU-Waste are captured indoors so that the illumination variations

are less significant, there are large variations in the clothing colors and, in some cases, the color contrast between the waste objects and the clothes is small. In addition, the orientation of the objects also exhibits large variations. For example, the objects can be held with either one or both hands. During the data collection, we simply ask the participants to hold objects however they like. Despite these challenges, our model is able to reliably recover the fine boundary details in most cases.

Table 5. Average per-image inference time on MJU-Waste. The baseline method is DeepLabv3 with a ResNet-50 backbone, which corresponds to the scene-level inference time. Additional object and pixel level inference incurs extra computational costs. System specs: i9-9900KS CPU, 64GB DDR4 RAM, RTX 2080Ti GPU. Test batch size set to 1 with FP32 precision. See Section 4.3 for details.

| MJU-Waste (val)	Scene-Level	Object-Level	Pixel-Level	Total
inference time (ms)	52	352	398	802
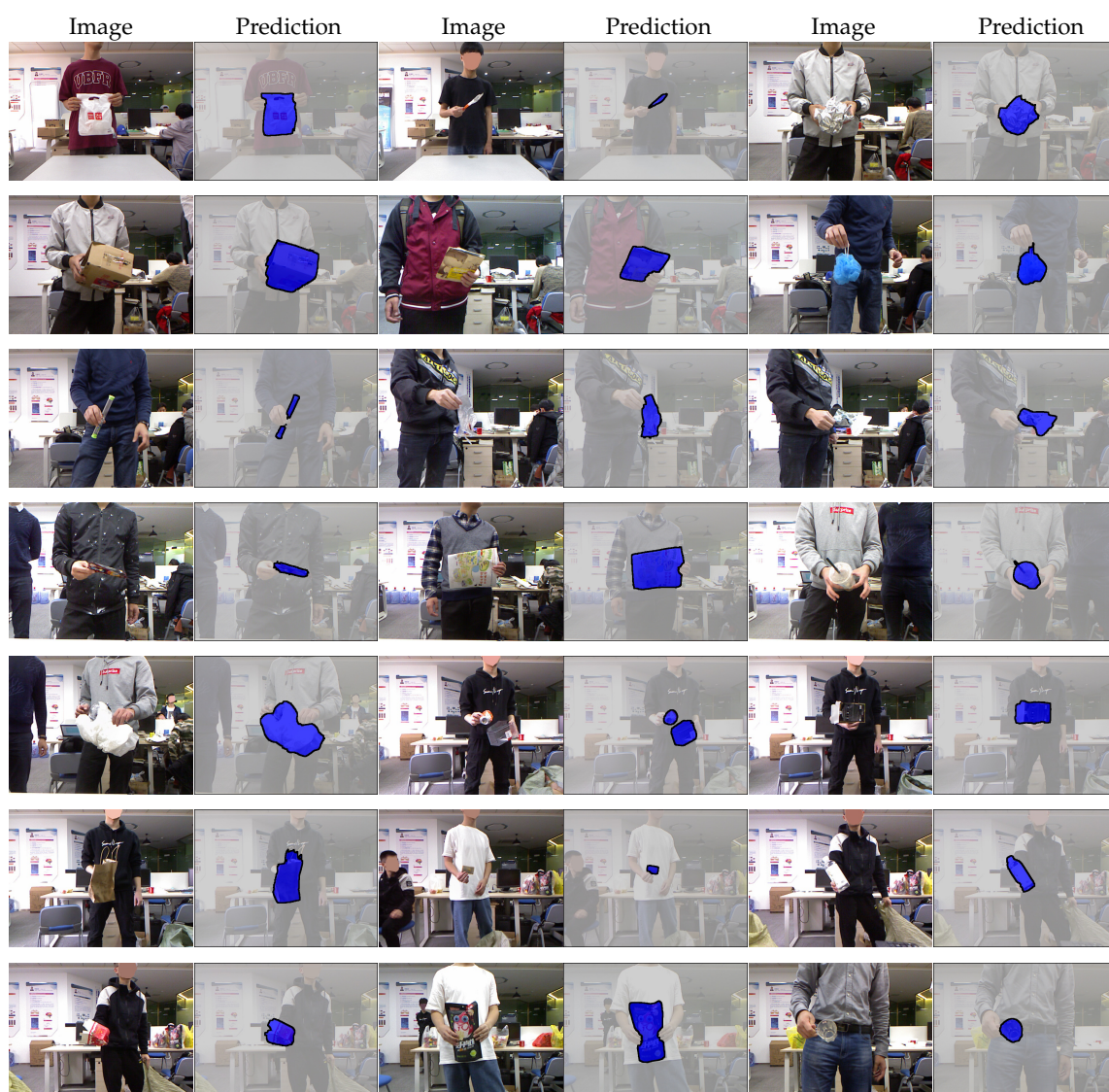
Figure 7. Segmentation results on MJU-Waste (test). Method is DeepLabv3-ML (ResNet-50).

4.4. Results on the TACO Dataset

We additionally evaluate the performance of our method on the TACO dataset. TACO contains color images only, so we exclude Equation (9) for training and evaluating models on this dataset.

This dataset presents a unique challenge for localizing waste objects "in-the-wild". In general, TACO is different to MJU-Waste in two important aspects. Firstly, multiple waste objects with extreme scale variation are more common (see Figures 8 and 9 for examples). Secondly, unlike MJU-Waste the backgrounds are diverse, such as in road, grassland and beach scenes. Quantitative results obtained on the TACO test set are summarized in Table 6. Specifically, we compare our multi-level model against two baselines: FCN-8s [17] and DeepLabv3 [23]. Again, in both cases our multi-level model is able to improve the baseline performance by a clear margin. Qualitative comparisons of the segmentation results are presented in Figure 8. It is clear that our multi-level method is able to more closely follow object boundaries. More example segmentation results are presented in Figure 9. We note that the changes in illumination and orientation are generally greater on TACO than on MJU-Waste, due to the fact that there are many outdoor images. Particularly, in some beach images it is very challenging to spot waste objects due to the poor illumination and the weak color contrast. Furthermore, object scale and orientation vary greatly as a result of different camera perspectives. Again, our model is able to detect and segment waste objects with high accuracy in most images, demonstrating the efficacy of the proposed method.
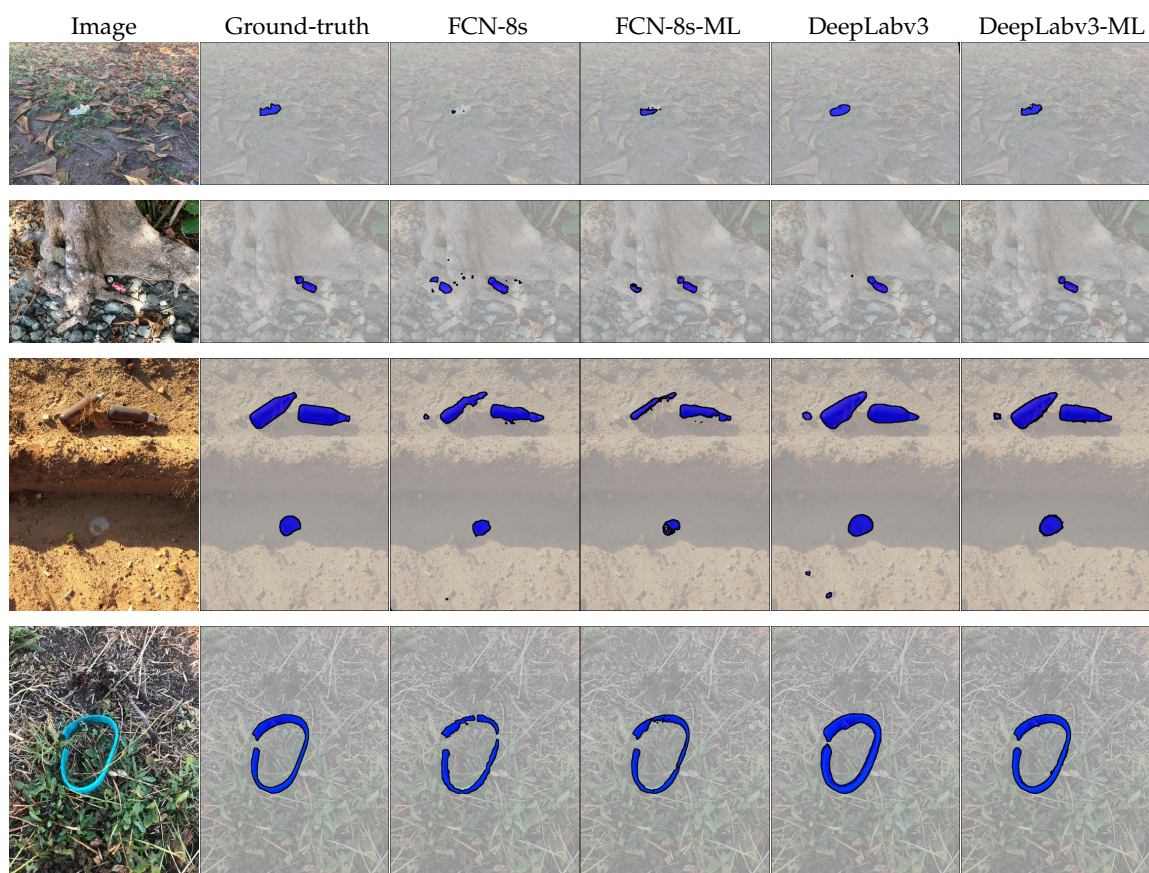
Figure 8. Example segmentation results on the TACO test set. Input images and ground-truth annotations are shown in the first two columns. Baseline methods are FCN-8s (VGG-16) and DeepLabv3 (ResNet-101). Our proposed methods (FCN-8s-ML and DeepLabv3-ML) more accurately recover object boundaries. Best viewed electronically, zoomed in.

Image	Prediction	Image	Prediction	Image	Prediction

Figure 9. Segmentation results on TACO (test). Method is DeepLabv3-ML (ResNet-101).

Table 6. Performance comparisons on the test set of TACO. For each method, we report the IoU for waste objects (IoU), mean IoU (mIoU), pixel Precision for waste objects (Prec) and mean pixel precision (Mean). See Section 4.4 for details.

Dataset: TACO (Test)	Backbone	IoU	mIoU	Prec	Mean
Baseline Approaches					
FCN-8s [17]	VGG-16	70.43	84.31	85.50	92.21
DeepLabv3 [23]	ResNet-101	**83.02**	**90.99**	**88.37**	**94.00**
Proposed Multi-Level (ML) Model					
FCN-8s-ML	VGG-16	74.21 (+3.78)	86.35 (+2.04)	90.36 (+4.86)	94.65 (+2.44)
DeepLabv3-ML	ResNet-101	**86.58** (+3.56)	**92.90** (+1.91)	**92.52** (+4.15)	**96.07** (+2.07)

5. Conclusions

We presented a multi-level approach to waste object localization. Specifically, our method integrates the appearance and the depth information from three levels of spatial granularity: (1) A scene-level segmentation network captures the long-range spatial contexts and produces an initial coarse segmentation. (2) Based on the coarse segmentation, we select a few potential object regions and then perform object-level segmentation. (3) The scene and object level results are then integrated into a pixel-level fully connected conditional random field to produce a coherent final localization. The superiority of our method is validated on two public datasets for waste object segmentation. As part of our work, we collected the MJU-Waste dataset that is made publicly available to facilitate future research in this area. We hope that our method could serve as a modest attempt to induce further exploration into vision-based perception of waste objects in complex real-world scenarios. For example, possible future work may explore the training of robust segmentation models that work on multiple datasets with large object appearance and camera perspective variations.

Author Contributions: Conceptualization, T.W.; Formal analysis, Y.C.; Funding acquisition, Y.C.; Methodology, T.W. and L.L.; Writing—original draft, T.W.; Writing—review & editing, D.Y. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eriksen, C.W.; James, J.D.S. Visual attention within and around the field of focal attention: A zoom lens model. *Percept. Psychophys.* **1986**, *40*, 225–240. [CrossRef] [PubMed]
2. Shulman, G.L.; Wilson, J. Spatial frequency and selective attention to local and global information. *Perception* **1987**, *16*, 89–101. [CrossRef] [PubMed]
3. Pashler, H.E. *The Psychology of Attention*; MIT Press: Cambridge, MA, USA, 1999.
4. Palmer, S.E. *Vision Science: Photons to Phenomenology*; MIT Press: Cambridge, MA, USA, 1999.
5. Proença, P.F.; Simões, P. TACO: Trash Annotations in Context for Litter Detection. *arXiv* **2020**, arXiv:2003.06975.

6. Alexe, B.; Deselaers, T.; Ferrari, V. Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2189–2202. [CrossRef]

7. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]

8. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*; Springer: Zurich, Switzerland, 2014; pp. 391–405.

9. Cheng, M.M.; Zhang, Z.; Lin, W.Y.; Torr, P. BING: Binarized normed gradients for objectness estimation at 300fps. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3286–3293.

10. Wang, Y.; Liu, J.; Li, Y.; Yan, J.; Lu, H. Objectness-aware semantic segmentation. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 307–311.

11. Xia, F.; Wang, P.; Chen, L.C.; Yuille, A.L. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *European Conference on Computer Vision*; Springer: Amsterdam, The Netherlands, 2016; pp. 648–663.

12. Alexe, B.; Deselaers, T.; Ferrari, V. Classcut for unsupervised class segmentation. In *European Conference on Computer Vision*; Springer: Crete, Greece, 2010; pp. 380–393.

13. Vezhnevets, A.; Ferrari, V.; Buhmann, J.M. Weakly supervised semantic segmentation with a multi-image model. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 643–650.

14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

15. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Zhang, Z.; Lin, H.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. ResNeSt: Split-Attention Networks. *arXiv* **2020**, arXiv:2004.08955.

16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; NeurIPS Foundation: Montreal, QC, Canada, 2015; pp. 91–99.

17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

18. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; MIT Press: Cambridge, MA, USA, 2009.

19. Krähenbühl, P.; Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*; NeurIPS Foundation: Granada, Spain, 2011; pp. 109–117.

20. Han, J.; Shao, L.; Xu, D.; Shotton, J. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Trans. Cybern.* **2013**, *43*, 1318–1334.

21. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.

22. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the 2019 International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 603–612.

23. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.

24. Yang, M.; Thung, G. *Classification of Trash for Recyclability Status*; CS229 Project Report; Stanford University: Stanford, CA, USA, 2016.

25. Bircanoğlu, C.; Atay, M.; Beşer, F.; Genç, Ö.; Kızrak, M.A. RecycleNet: Intelligent waste sorting using deep neural networks. In Proceedings of the 2018 Innovations in Intelligent Systems and Applications (INISTA), Thessaloniki, Greece, 3–5 July 2018; pp. 1–7.

26. Aral, R.A.; Keskin, Ş.R.; Kaya, M.; Hacıömeroğlu, M. Classification of trashnet dataset based on deep learning models. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 2058–2062.

27. Awe, O.; Mengistu, R.; Sreedhar, V. Smart trash net: Waste localization and classification. CS229 Project Report; Stanford University: Stanford, CA, USA, 2017.

28. Chu, Y.; Huang, C.; Xie, X.; Tan, B.; Kamal, S.; Xiong, X. Multilayer hybrid deep-learning method for waste classification and recycling. *Comput. Intell. Neurosci.* **2018**, *2018*, 5060857. [CrossRef]

29. Vo, A.H.; Son, L.H.; Vo, M.T.; Le, T. A Novel Framework for Trash Classification Using Deep Transfer Learning. *IEEE Access* **2019**, *7*, 178631–178639. [CrossRef]

30. Ramalingam, B.; Lakshmanan, A.K.; Ilyas, M.; Le, A.V.; Elara, M.R. Cascaded machine-learning technique for debris classification in floor-cleaning robot application. *Appl. Sci.* **2018**, *8*, 2649. [CrossRef]

31. Yin, J.; Apuroop, K.G.S.; Tamilselvam, Y.K.; Mohan, R.E.; Ramalingam, B.; Le, A.V. Table Cleaning Task by Human Support Robot Using Deep Learning Technique. *Sensors* **2020**, *20*, 1698. [CrossRef] [PubMed]

32. Rad, M.S.; von Kaenel, A.; Droux, A.; Tieche, F.; Ouerhani, N.; Ekenel, H.K.; Thiran, J.P. A computer vision system to localize and classify wastes on the streets. In *International Conference on Computer Vision Systems*; Springer: Shenzhen, China, 2017; pp. 195–204.

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

34. Wang, Y.; Zhang, X. Autonomous garbage detection for intelligent urban management. *MATEC Web Conf.* **2018**, *232*, 01056. [CrossRef]

35. Bai, J.; Lian, S.; Liu, Z.; Wang, K.; Liu, D. Deep learning based robot for automatically picking up garbage on the grass. *IEEE Trans. Consum. Electron.* **2018**, *64*, 382–389. [CrossRef]

36. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]

37. Deepa, T.; Roka, S. Estimation of garbage coverage area in water terrain. In Proceedings of the 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), Bengaluru, India, 17–19 August 2017; pp. 347–352.

38. Mittal, G.; Yagnik, K.B.; Garg, M.; Krishnan, N.C. Spotgarbage: Smartphone app to detect garbage using deep learning. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, 12–16 September 2016; pp. 940–945.

39. Zeng, D.; Zhang, S.; Chen, F.; Wang, Y. Multi-scale CNN based garbage detection of airborne hyperspectral data. *IEEE Access* **2019**, *7*, 104514–104527. [CrossRef]

40. Qiu, Y.; Chen, J.; Guo, J.; Zhang, J.; Liu, S.; Chen, S. Three dimensional object segmentation based on spatial adaptive projection for solid waste. In *CCF Chinese Conference on Computer Vision*; Springer: Tianjin, China, 2017; pp. 453–464.

41. Wang, C.; Liu, S.; Zhang, J.; Feng, Y.; Chen, S. RGB-D Based Object Segmentation in Severe Color Degraded Environment. In *CCF Chinese Conference on Computer Vision*; Springer: Tianjin, China, 2017; pp. 465–476.

42. Grard, M.; Brégier, R.; Sella, F.; Dellandréa, E.; Chen, L. Object segmentation in depth maps with one user click and a synthetically trained fully convolutional network. In *Human Friendly Robotics*; Springer: Reggio Emilia, Italy, 2019; pp. 207–221.

43. He, X.; Zemel, R.S.; Carreira-Perpiñán, M.Á. Multiscale conditional random fields for image labeling. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004; Volume 2, p. II.

44. Shotton, J.; Winn, J.; Rother, C.; Criminisi, A. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision*; Springer: Graz, Austria, 2006; pp. 1–15.

45. Ladický, L.; Russell, C.; Kohli, P.; Torr, P.H. Associative hierarchical crfs for object class image segmentation. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 739–746.

46. Gould, S.; Fulton, R.; Koller, D. Decomposing a scene into geometric and semantically consistent regions. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009; pp. 1–8.

47. Kumar, M.P.; Koller, D. Efficiently selecting regions for scene understanding. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3217–3224.

48. Munoz, D.; Bagnell, J.A.; Hebert, M. Stacked hierarchical labeling. In *European Conference on Computer Vision*; Springer: Crete, Greece, 2010; pp. 57–70.

49. Tighe, J.; Lazebnik, S. Superparsing: scalable nonparametric image parsing with superpixels. In *European Conference on Computer Vision*; Springer: Crete, Greece, 2010; pp. 352–365.

50. Liu, C.; Yuen, J.; Torralba, A. Nonparametric scene parsing via label transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2368–2382. [CrossRef]

51. Lempitsky, V.; Vedaldi, A.; Zisserman, A. Pylon model for semantic segmentation. In *Advances in Neural Information Processing Systems*; NeuraIPS Foundation: Granada, Spain, 2011; pp. 1485–1493.

52. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.

53. Mostajabi, M.; Yadollahpour, P.; Shakhnarovich, G. Feedforward semantic segmentation with zoom-out features. In Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3376–3385.

54. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.

55. Lin, G.; Shen, C.; Van Den Hengel, A.; Reid, I. Efficient piecewise training of deep structured models for semantic segmentation. In Proceedings of the 2016 Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3194–3203.

56. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.

57. Ding, H.; Jiang, X.; Shuai, B.; Qun Liu, A.; Wang, G. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2393–2402.

58. Lin, D.; Ji, Y.; Lischinski, D.; Cohen-Or, D.; Huang, H. Multi-scale context intertwining for semantic segmentation. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 603–619.

59. Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; Torr, P.H. Conditional random fields as recurrent neural networks. In Proceedings of the 2015 International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1529–1537.

60. Schwing, A.G.; Urtasun, R. Fully connected deep structured networks. *arXiv* **2015**, arXiv:1503.02351.

61. Liu, Z.; Li, X.; Luo, P.; Loy, C.C.; Tang, X. Semantic image segmentation via deep parsing network. In Proceedings of the 2015 International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1377–1385.

62. Jampani, V.; Kiefel, M.; Gehler, P.V. Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks. In Proceedings of the 2016 Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4452–4461.

63. Chandra, S.; Kokkinos, I. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *European Conference on Computer Vision*; Springer: Amsterdam, The Netherlands, 2016; pp. 402–418.

64. Pohlen, T.; Hermans, A.; Mathias, M.; Leibe, B. Full-resolution residual networks for semantic segmentation in street scenes. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4151–4160.

65. Gadde, R.; Jampani, V.; Kiefel, M.; Kappler, D.; Gehler, P.V. Superpixel convolutional networks using bilateral inceptions. In *European Conference on Computer Vision*; Springer: Amsterdam, The Netherlands, 2016; pp. 597–613.

66. Li, X.; Jie, Z.; Wang, W.; Liu, C.; Yang, J.; Shen, X.; Lin, Z.; Chen, Q.; Yan, S.; Feng, J. Foveanet: Perspective-aware urban scene parsing. In Proceedings of the 2017 International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 784–792.

67. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the 2015 International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1520–1528.

68. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing And Computer-Assisted Intervention*; Springer: Munich, Germany, 2015; pp. 234–241.

69. Yuan, Y.; Wang, J. Ocnet: Object context network for scene parsing. *arXiv* **2018**, arXiv:1809.00916.

70. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Change Loy, C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 267–283.

71. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.

72. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.

73. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5229–5238.

74. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]

75. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 801–818.

76. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision*; Springer: Zurich, Switzerland, 2014; pp. 345–360.

77. Li, Z.; Gan, Y.; Liang, X.; Yu, Y.; Cheng, H.; Lin, L. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In *European Conference on Computer Vision*; Springer: Amsterdam, The Netherlands, 2016; pp. 541–557.

78. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2650–2658.

79. Hu, X.; Yang, K.; Fei, L.; Wang, K. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1440–1444.

80. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*; Springer: Taipei, Taiwan, 2016; pp. 213–228.

81. Qi, X.; Liao, R.; Jia, J.; Fidler, S.; Urtasun, R. 3D graph neural networks for rgbd semantic segmentation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5199–5208.

82. Lai, K.; Bo, L.; Ren, X.; Fox, D. A large-scale hierarchical multi-view rgb-d object dataset. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 1817–1824.

83. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]

84. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.

85. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; NeurIPS Foundation: Lake Tahoe, CA, USA, 2012; pp. 1097–1105.

86. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.