Uncertainty Aware Proposal Segmentation for Unknown Object Detection

Yimeng Li and Jana Košecká
George Mason University, Fairfax, VA, USA
{yli44, kosecka}@gmu.edu

Abstract

Recent efforts in deploying Deep Neural Networks for object detection in real world applications, such as autonomous driving, assume that all relevant object classes have been observed during training. Quantifying the performance of these models in settings when the test data is not represented in the training set has mostly focused on pixel-level uncertainty estimation techniques of models trained for semantic segmentation. This paper proposes to exploit additional predictions of semantic segmentation models and quantifying its confidences, followed by classification of object hypotheses as known vs. unknown, out of distribution objects. We use object proposals generated by Region Proposal Network (RPN) and adapt distance aware uncertainty estimation of semantic segmentation using Radial Basis Functions Networks (RBFN) for class agnostic object mask prediction. The augmented object proposals are then used to train a classifier for known vs. unknown objects categories. Experimental results demonstrate that the proposed method achieves parallel performance to state of the art methods for unknown object detection and can also be used effectively for reducing object detectors' false positive rate. Our method is well suited for applications where prediction of non-object background categories obtained by semantic segmentation is reliable.

1. Introduction

The last decade marked big progress in the design of Deep Network models for object detection and semantic segmentation. Highly accurate pixel-level classification of known object and background categories has been achieved by training state-of-the-art models on large fully annotated datasets [7, 19]. When applying these models in real-world settings it is often the case that objects that are not represented in the training set appear at test time. This is of particular importance in autonomous driving, where unknown objects can appear on the road or close to the road and become a potential threat to safety. One class of methods approach this problem using methodologies for uncertainty estimation of deep networks models, such as Dropout [17] or ensemble methods [3]. In driving scenarios the estimation of uncertainties of semantic segmentation using Dropout often does not coincide with novel objects, making it difficult to generate reliable novel object hypotheses (see Figure. 2). Many highly uncertain regions correspond to correctly classified background classes or boundaries between different semantic categories.

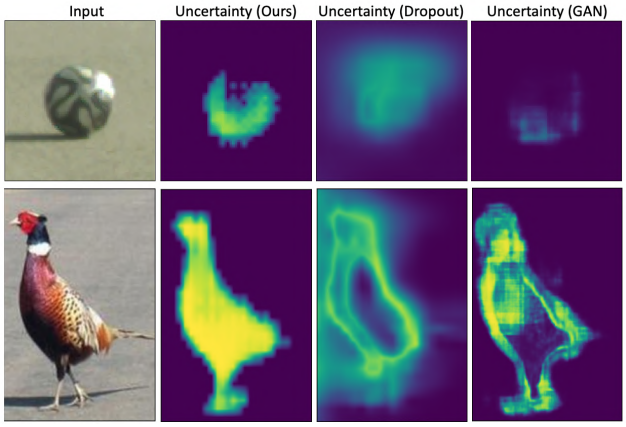

Figure 1. Detecting the out of distribution (OOD) objects from object proposals. Our approach predicts high uncertainty on the OOD object while Dropout [17] method is distracted by the background and GAN [20] method neglects the object (football).

In this paper, we propose a new approach for detection of out of distribution objects by leveraging pixel level predictions obtained by semantic segmentation and their associated distance aware uncertainty estimates using Radial Basis Functions Networks (RBFN). Instead of making final predictions at the pixel level as done by methods that rely only on semantic segmentation, we use class agnostic object proposals generated by region proposal network (RPN), that are both segmented and further classified as known or unknown objects. The premise of our approach is that pixels belonging to background classes (*e.g.* road, vegetation, building) can be classified with high confidence by semantic segmentation, while all object's pixels (known and unknown) will have higher uncertainty. Once we make sure a

pixel belong to an object, we can further decide it belongs to an unknown object if it has high uncertainty being classified into all known object categories. This assumption is suitable for embodied agents which operate over extended periods in their native environments where non-object background classes are less likely to change. Our contributions can be summarized as follows: (i) A new Radial Basis Function Network and novel regularization terms for segmentation of object proposals; (ii) Distance aware uncertainty estimation for object mask prediction; (iii) Detailed ablation study showing the effects of object detection and semantic segmentation features and evaluation of the proposed method on the available datasets [25, 5, 20] showing parallel performance to state-of-the-art. (iv) Improved the false positive rate of the modern object detector [12] and semantic segmentation models [6], by identifying detections with large uncertainty.

2. Related Work

Uncertainty estimation. Several methods proposed for uncertainty estimation have been introduced in the context of image classification tasks. In deep learning setting early approach by [4] proposed an extreme value parameter redistribution method to tackle the open-set recognition problem. A similar problem to open-set recognition is to detect out-of-distribution (OOD) examples as samples from a different dataset. Authors in [13] use simple statistics derived from softmax distributions to determine whether an example is misclassified or from a different distribution from the training data. To overcome the difficulties of stochastic methods such as dropout [17] that require multiple network passes, in [26] propose deterministic uncertainty estimation method. They learn a feature space using RBF-kernels and suggest that feature distance to the nearest center well quantifies the uncertainty of prediction. Authors in [21] presented a normalization method for deep neural networks to maintain the feature distance in the intermediate layers, in [18] they learn a representative in-distribution data embedding by taking additional background data as adversarial examples.

Semantic Segmentation. Several approaches for detecting unknown objects using pixel level predictions in semantic segmentation framework have been introduced in the autonomous driving setting along with the datasets to evaluate them. Authors in [5] proposed a dataset for this task by synthesizing unexpected objects into CityScapes images. They also implemented existing uncertainty estimation methods [10, 22] for semantic segmentation models with evaluations on the proposed dataset. The methods did not work well for detecting the unknown objects, making it difficult to aggregate pixel level uncertainty predictions to object hypotheses. This is an inherent limitation of methods based on semantic segmentation. Authors in [28] identified the outliers from Lidar-projected point cloud with prototype models used for few shot learning and reasoned about embeddings of unknown categories as distances from the prototypes. Instead of relying on uncertainty estimation techniques, Lis et al. [20] used GANs for synthesizing images from predicted semantic segmentation and designed a discrepancy network for identifying the differences between the original input image and a synthesized image leading the performance on available benchmarks. Biase et al. [8] improve Lis's re-synthesis method by using uncertainty maps to guide the discrepancy network to focus on high-uncertain areas. In [16], authors propose to detect the unexpected objects through standardized max logits. This method does not require additional training, only requiring further processing on the segmentation results. In [23], authors use normalizing flow to learn the data density embedding space. In the test stage pixels with low density are considered as foreground objects hypotheses. The method is evaluated on indoors scenes only.

Figure 2. Uncertainty estimation result with DeeplabV3+ and RBFN. Note that high uncertainty sometimes show up on background regions(top right corner).

Panoptic segmentation. The recently proposed panoptic segmentation approaches are also relevant to our work. Semantic segmentation model can provide pixel-level features for classification of the known object and background categories, while object detection models make predictions over bounding boxes that aggregate features over larger regions. Authors in [30] suggest to use both semantic segmentation features and object mask features to determine whether the pixel belongs to object or background, while authors in [27] present an instance center prediction head to assign pixels from the same class to different object instances. In [9, 14] authors show that features obtained by training object detector model help predicting higher-quality object masks compared to semantic segmentation features. In [15], authors proposed a open world object detection framework that used contrastive learning in the feature space to discover unknown object classes. In presented work we also explore the efficacy of using both semantic segmentation features and object detection features for unknown object detection.

3. Approach

We propose a novel approach for unknown object detection that starts from object proposals using RPN [12] and Edge Boxes [32] and associated feature maps from state-of-the-art object detection and semantic segmentation models. In the first stage we train a proposal segmentation model with pixel-level uncertainty for object mask prediction (see Figure 1). In the second stage, we pool the features from object mask region into one feature vector for object class prediction. If a proposal has high uncertainty being classified into all known object categories, then it is labeled as an unknown object.

3.1. Proposal Segmentation

The proposal segmentation model takes an object proposal o_i and its associated features f_i extracted from Mask-RCNN model [12] and DeeplabV3+ semantic segmentation model [6][1]. To label each pixel of the proposal into one of the semantic categories and estimate its uncertainty, we process the initial feature maps by adding additional convolutional layers before passing it to RBF Network. Details of the architectures are described in section 3.3.

RBF Network. For uncertainty estimation, we adapt Radial Basis Functions Network (RBFN) [1, 26] and its feature-distance uncertainty estimation framework. This deterministic uncertainty quantification approach forgoes the disadvantages of dropout or ensemble methods that require multiple passes through network. The predictions of RBFN are made by computing a kernel function and a distance function, between the feature vector computed by deep model and the centroids. The uncertainty of the prediction is measured as the distance between the model output and the closest centroid. Data points with feature vectors that are far away from centroids do not belong to any class and can be considered out of distribution.

The feature extraction module $g(f(u,v))$ consists of four convolutional layers, taking features at spatial location $f(u,v)$ in the feature map, followed by RBFN-classifier that has two learnable parts: K centers $\mu_{c,k}$ for each class and a weight component $w_{c,k}$ for each center. We apply the radial basis function (gaussian) to the feature output $g(f(u,v))$ and the class centers as a measure of distance between them:

$$h_c(g(f(u,v)),\mu_c) = \sum_{k=1}^{K} w_{c,k} \exp(-\frac{\|g(f(u,v)) - \mu_{c,k}\|^2}{2\sigma^2}) \tag{1}$$

where σ is the scale term for Gaussian kernel. Class c with the minimum distance (i.e. maximum h_c) is the final prediction. Uncertainty τ is computed as the difference between one (upper bound of h_c) and the distance to the predicted class:

$$\tau = 1 - \max h_c(g(f(u,v)),\mu_c) \tag{2}$$

The adoption of this model and uncertainty estimation framework for semantic segmentation comes with its challenges. We describe next how to tackle them using novel regularization term.

Boundary Regularization. In practical settings it has been observed that RBF networks are difficult to optimize and can frequently map the out of distribution features to in distribution features, also referred to as *features collapse* problem. Figure 4 shows an example where an out-of-distribution (OOD) object is confidently classified into the 'ground' class. This has been observed in [26], where authors suggested adding gradient penalty to the loss function. In the context of semantic segmentation task gradient penalty is computed per pixel and causes loss explosion during training. Further conflict with batch normalization, causes the gradient penalty to reduce model's overall performance.

We propose a regularization method better suited for the segmentation task, where pixel level predictions are sought. We observed that the boundary pixels between background and object usually have high uncertainty because their receptive field includes features from both object and background pixels. If we consider these pixels as outliers, we can confine the computed embedding to be either object or background pixels. In other words, enforcing a uniform distribution for the boundary pixels D_{bd} and maximizing the classification performance on the remaining in distribution pixels D_{in}. This is captured by the following loss function:

$$L(g,w;D_{in},D_{bd}) = L_{in}(g,w;D_{in}) + L_{bd}(g,w;D_{bd}) \tag{3}$$

where

$$L_{in}(g(u,v),y) = -\sum_{c} y_c \log(h_c) + (1-y_c) log(1-h_c) \tag{4}$$

is the standard cross-entropy loss for in distribution pixels between each class distance h_c and a one-hot encoding of the label y. L_{bd} is the same loss for boundary pixels where the label encoding is fully zeros.

Toy Example in 2D Dimension. We use a toy example (Figure 5) to explain the proposed regularization method. During training stage, the in-distribution data is represented by two Gaussians, the red one for background features and the green one for object feature vectors (see Figure 5(a)). During testing, we add out of distribution features, here denoted by uniformly distributed blue points. We train RBF network with different regularization terms, to quantify their ability to classify out of distribution data points. Figure 5(b) shows the estimated uncertainty of applying

[1]For MASK-RCNN we take 14x14 ROI aligned feature map being passed to mask branch and for DeeplabV3 we take the feature map for all pixels in the proposal followed by ROI align stage to yield another set of channels with spatial support of 14x14.

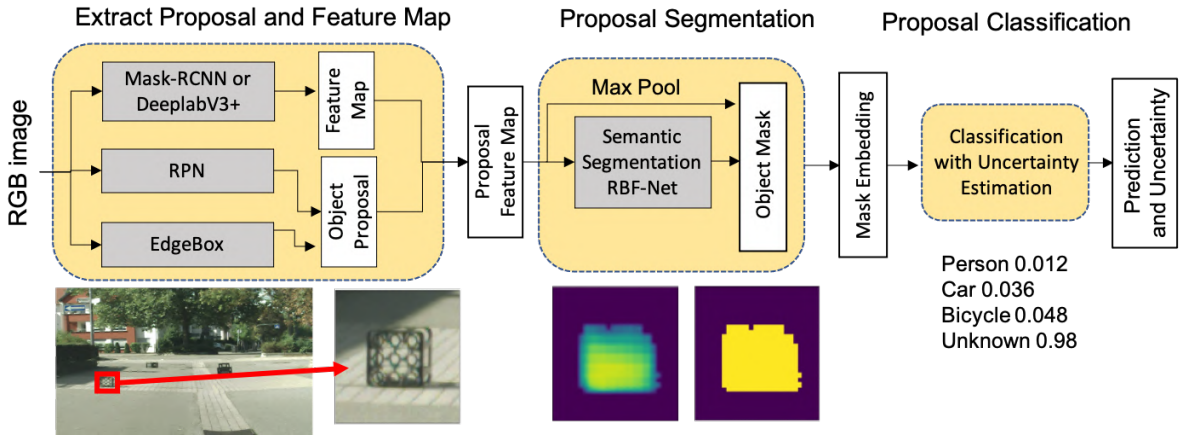

Figure 3. Proposed approach pipeline. Given an input image, we extract object proposals and proposal feature maps. Proposal segmentation module predicts the foreground object mask from the feature map. Proposal classification module pools the feature vector from the feature map given the object mask, followed by classification of object as unknown or in distribution objects and **it's uncertainty estimation**. Object classification with high uncertainty is decided as an outlier.

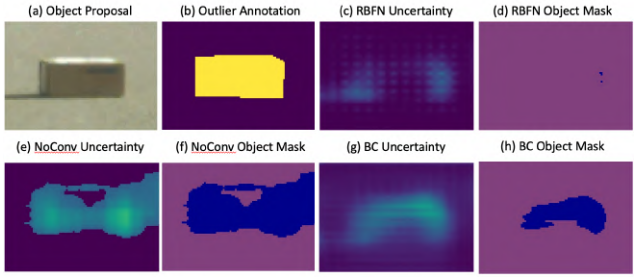

Figure 4. (a)(b) show one example object proposal of out of distribution object (OOD) from Lost & Found dataset and its annotation. (c)(d), (e)(f), (g)(h) show the proposal segmentation result using RBFN, RBFN-NoConv and RBFN with boundary constraint models. Note that enforcing the boundary constraint helps the detection of OOD object.

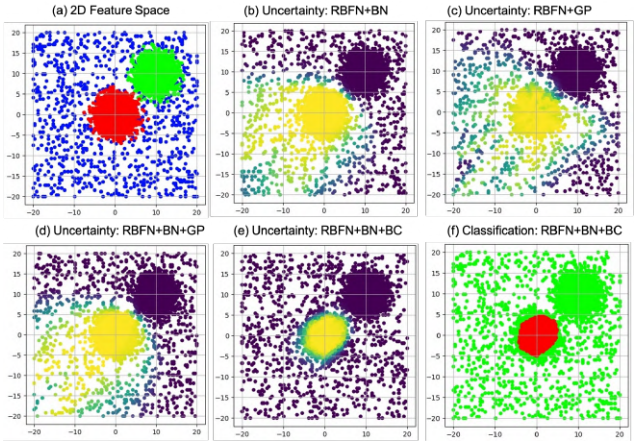

Figure 5. Visualization of point classification on 2d space. (a) shows the whole feature space including two Gaussian (red Gaussian blob represents in-dist background features and green blob represents in-dist object features) and the blue points follows a uniform distribution (OOD data). (b)(c)(d)(e) shows the uncertainty estimation results (brighter color means lower uncertainty) of classifying all the points into the red Gaussian blob by using RBFN combined with different regularization methods. (f) is the final classification result by thresholding the uncertainty value and separating the red Gaussian blob from the OOD points.

RBF-Net method, where some OOD points also have low uncertainty and got miss-classified into the in distribution classes. This is the feature collapse phenomenon. Figure 5(c) shows that gradient penalty reduces the uncertainty distribution problem to some extent. Figure 5(d) shows that gradient penalty contradicts with batch normalization [11] causing the number of OOD points with low uncertainty increase. The last two plots of Figure 5(e) and (f) show that the boundary points work as a strong constraint to the point embeddings with only points in the center of the blob having high confidence (bright color). This enables us to separate the (background) points belonging to the center Gaussian from the other OOD points given the estimated uncertainty in Figure 5(f).

The demonstration of these effects on the proposal segmentation is in Figures 1 and 4. The details of implementation, trade-offs between generalization ability of RBF models on known classes and ability to reliably classify out of distribution objects and the effect of different regularization terms can be found in Section 4.2.

3.2. Proposal Classification

The previous sections described an approach for semantic segmentation of object proposals using uncertainty associated with pixel level predictions. By thresholding pix-

Figure 6. Visualization of object proposal segmentation on Cityscapes val. Three examples are presented, including a car, a pedestrian and one OOD object. From left to right, we have the RGB input, semantic segmentation result and per pixel uncertainty estimation result. Uncertainty values on the background pixels are ignored. Note that the uncertainty values vary substantially inside the object: not only pixels inside an out-of-dist object have high uncertainty, but also pixels close to the boundary of known objects have higher uncertainty.

els with low uncertainty of background classes we obtain a binary object mask. We apply max-pooling on the features associated with the mask passing the resulting feature to RBFN model for classification with uncertainty detailed in Section 3.3. We auto-label the object proposals from training data by computing the IoU between proposals and ground-truth object bounding boxes. If IoU is larger than a threshold, the proposal is labeled as the ground-truth object category. We use such labeled object proposals as training data to train the proposal classification model.

3.3. Implementation Details

Object Proposals. We take the top 1000 object proposals predicted by Mask-RCNN [29] trained on Cityscapes. Usually 500 lower ranked proposals have large portion of background. We keep them in the training set so our model is able to distinguish between background categories. During testing, we also use object proposals generated by using EdgeBox method [32].

Network Architecture Details. The architecture of the segmentation model is inspired by the mask head in Mask-RCNN [12]. The input feature map (14x14) is passed through 4 convolutional layers to learn the representation and one deconvolutional layer to resize the feature map from 14x14 to 28x28, before passing it to the RBFN classification module for final prediction. We use the same hyper-parameters for all the RBF classification layers: 512 centers for each class and feature dimension of 256-dim. The scale σ in the Gaussian kernel is set to be 0.1.

Model Training. Both proposal segmentation model and classification model is trained using SGD with initial learning rate 0.1, momentum 0.9 and batch size 64. The learning rate is divided by 10 every 10 epochs. For the RBF classification layer, we update the centers through exponential moving average method [24] with momentum 0.999.

4. Experiments

We perform three main experiments: evaluation of the proposed model on outdoor scenes (Sec. 4.1), ablation study of the proposed model (Sec. 4.2) and evaluation of the proposed model in indoor scenes (Sec. 4.3). To evaluate our approach on out of distribution (OOD) object detection, we compare all test methods uncertainty estimation output with the ground-truth OOD annotation and compute metrics associated with a binary classification task. We use AUROC[2], to evaluate proposal segmentation performance. For evaluation on the whole image, we also compute average precision (AP) to deal with in-distribution and OOD data unbalanced situation.

Since the number of false positives is also relevant for safety-critical applications, we also compute the false positive rate (FPR_{95}) at 95% true positive rate (TPR), which is also used in [5]. For the classification of in distribution object classes, we simply use classification accuracy (Acc).

4.1. OOD Object Detection in Outdoor Scenes

We train the proposed model and the baseline methods on Cityscapes [7] and evaluate on the following three datasets containing OOD objects not covered by Cityscapes.

FS Lost & Found (L&F) [25, 5]. This dataset contains 100 real images captured with the same camera setup as Cityscapes. Pixel-level annotations are available to distinguish between two classes, OOD objects (e.g. cargo boxes and toy cars) and classes present in Cityscapes. We select 62 images containing objects with sufficient spatial support and object proposal size during evaluation. The unexpected objects in the rest images are neglected by EdgeBox [32] because of unnoticeable size.

Fishyscapes Static (FS) [5]. This dataset contains 30 images with unknown objects super-imposed synthetically through image compositing techniques. Objects not covered by Cityscapes (including aeroplane, bird, cat, cow, dog, horse, sheep) are randomly resized and positioned onto Cityscapes validation images. Postprocessing techniques like lightning and shadow adaptation are applied to make the images more genuine.

Road Anomaly (RA) [20]. This dataset contains 60 real images collected from Internet. These include OOD objects located on or near the road to mimic the traffic scenes. Various OOD objects including animals, rocks, lost tires and construction equipment are present. Note that most images from this dataset have a very different background setting than Cityscapes. We evaluate on this dataset to compare generalization of different methods to other outdoor scenes.

Baselines. We switch DeeplabV3+ last classification layer with RBFN [26] as the first baseline and denote it as DeeplabV3+-RBFN. Pixel-level uncertainty is computed as

[2] Area Under ROC Curve, denoted as AC in the results Table.

per pixel's feature distance to the closest class centers.

We use the GAN method [20] as the second baseline. It uses CycleGAN to generate a synthetic image from the semantic segmentation of the input image. The synthesis on the OOD object regions is expected to be poor as these outliers are not covered in the training data. We take the pixel discrepancy image between input image and synthesized image as the uncertainty estimation result for comparison since the discrepancy value is in the range $[0, 1]$.

We also compare to the state of the art method [8] that came out a few months ago and denote it as Resynthesis++. This method is built on th GAN method[20] while the uncertainty maps are also considered in the final inference. We take the trained model and evaluate it for the Whole Image Segmentation task only.

Proposal Segmentation. We compare our proposal segmentation method with the baseline methods on proposal segmentation task. Object proposals are chosen if they overlap with any OOD objects. We treat proposal segmentation on OOD objects as a binary segmentation task, where one pixel's uncertainty of classification indicates the probability of it belonging to an OOD object. Proposal segmentation results for DeeplabV3+-RBFN and GAN are cut out from the whole image uncertainty map results. The proposal segmentation model uses semantic segmentation (SSeg) features from pre-trained DeepLabV3+ model as input. Table 1 presents the result. Our method performs particularly well on L&F dataset. This is because images in L&F have similar background to the training data. On FS, our model performs slightly worse than the GAN method. We hypothesise that since synthetic OOD objects in FS are blended into the background, they have a small distance to background in the feature space, which is detrimental to the feature distance based methods.

Whole Image Segmentation. Here we compare the performance of our method with the baselines on the entire image. We first rank object proposals by their objectness score, removing the ones with large IoU. Proposal's feature map is then passed through the proposal segmentation (Prop-Seg) and proposal classification (Prop-Cls) model to compute per pixel uncertainty u_{seg} and proposal's overall uncertainty u_{cls}. Proposal segmentation uses SSeg features as input while Prop-Cls uses object detection (ObjDet) features obtained from training Mask-RCNN on Cityscapes. If u_{cls} is below a threshold (in practice, we use 0.3), then this proposal's result is discarded, otherwise per pixel uncertainty is computed as $u_{seg} \cdot u_{cls}$. We accumulate the results from all the remaining proposals and embed them to an empty uncertainty image as the final result. Table 2 presents the results. Our method achieves parallel performance to the recent Resynthesis++ method across all the datasets. It performs quite well on AP compared to baselines as it is less affected by the uncertainty computed on the background regions. It performs slightly worse on Road Anomaly dataset as its images are collected from Internet and the background is different from images in Cityscapes. This results in our method mistakenly recognizing some background proposals as OOD objects. Figure 7 shows some qualitative results of the proposed method and baselines.

4.2. Ablation Study

In this section we attempt to get a better understanding of how certain components of our model contribute to the overall performance. We evaluate four aspects of the proposal segmentation (Prop-Seg) model: input representation, uncertainty estimation method, training data and regularization method. The proposal classification (Prop-Cls) model is affected by two aspects: input representation and training data. All the evaluations are done over object proposals overlapping with the OOD object labels.

Feature Representation: ObjDet vs. SSeg We vary the input visual representation with remaining aspects fixed. Results for proposal segmentation model (Prop-Seg) are presented in lines 1, 3 of Table 3, and lines 1, 3 of Table 4 for proposal classification (Prop-Cls). For the proposal segmentation task, using SSeg feature performs better than ObjDet feature particularly with the FPR_{95} metric under all three datasets. This is because the SSeg feature is optimized over the background pixels while ObjDet feature is mainly trained with the object pixels. This endows SSeg features with a better description power in distinguishing the foreground objects from the background. For the proposal classification task, ObjDet features beat SSeg features by over 30% AP on the OOD datasets. This shows that ObjDet features are more suitable even for the OOD objects.

Regularization We compare the regularization methods for Prop-Seg. Results are presented in line 7 (no convolutional (no-conv) layers so no need for regularization), line 3 (original model), line 2 with gradient penalty (GP) and line 8 with our proposed boundary pixel constraint of Table 3. Note that the no-conv model performs well on FS but not so much on L&F and RA. This is because in FS the OOD objects are synthesized onto the image while in L&F and RA we have real OOD objects. Without having the convolutional layers to further fine-tune the input data for the RBFN layer, the no-conv model generalizes poorly to new data. This is denoted by the high FPR_{95} on L&F.

Comparing line 1 with line 2 of Table 3, we observe that adding gradient penalty (GP) during training hinders the model's performance. We don't show the model using SSeg feature as input and having GP for regularization because we encountered loss explosion when training the model. On the other hand, the model trained with the boundary pixel constraint performs slightly better on AC and FPR_{95}. This means that with the help of the boundary constraint, our model learns a more robust representation and is able to de-

Method	L&F (AC↑/AP↑/FPR_{95} ↓)	RA (AC↑/AP↑/FPR_{95} ↓)	FS (AC↑/AP↑/FPR_{95} ↓)
DeeplabV3+-RBFN [26]	73.8 / 40.3 / 55.5	60.1 / 39.9 / 72.3	82.7 / **64.3** / 42.8
GAN [20]	85.8 / 58.5 / 33.1	70.6 / 54.0 / 55.7	**84.0** / 63.9 / **40.0**
Ours	**92.1** / **70.3** / **23.4**	**76.2** / **56.5** / **47.3**	82.8 / 56.3 / 43.0

Table 1. Comparison of Proposal Segmentation Performance.

Method	L&F (AC↑/AP↑/FPR_{95} ↓)	RA (AC↑/AP↑/FPR_{95} ↓)	FS (AC↑/AP↑/FPR_{95} ↓)
DeeplabV3+-RBFN [26]	68.9 / 3.3 / 54.7	73.2 / 20.0 / 54.1	78.2 / 14.7 / 44.9
GAN [20]	84.2 / 10.1 / 28.9	**86.1** / 42.3 / *32.2*	82.6 / 16.1 / *40.2*
Resynthesis++ [8]	**95.2** / **53.8** / **13.8**	84.6 / 41.5 / **45.6**	**92.7** / **56.3** / **26.5**
Ours	*90.7* / *32.8* / *24.9*	78.7 / **45.3** / 37.3	*88.0* / *34.3* / 41.5

Table 2. Whole Image Segmentation Performance. Bold numbers denote the top performance and italic numbers denote the second performance.

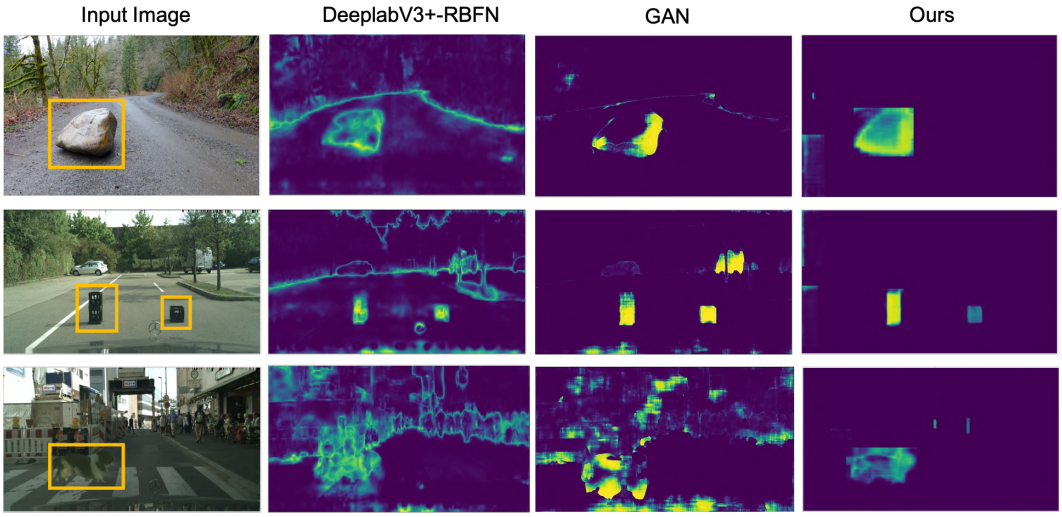

Figure 7. Whole Image OOD Object Segmentation Results. Orange box denotes the OOD object. Each row corresponds to a different input image. From top to bottom, the input image is selected from Road Anomaly, Lost & Found and Fishyscapes datasets. From left to right, we have the input image, Deeplab-RBFN [26] result, GAN [20] result and our method result.

tect hard OOD object examples as showed in Figure 4. This is important for safety-critical applications.

Uncertainty Estimation Methods Here we compare different uncertainty estimation methods. We experiment with three techniques: Entropy, Dropout and RBFN. The entropy method is implemented by using a linear layer in the end of the model for classification. Uncertainty is computed as the entropy of the output probabilities. The Dropout method is implemented by adding dropout layers after all convolutional layers. Uncertainty is estimated by performing multiple forward passes through the model with dropout enabled, and computing the entropy of the averaged predicted probability vector. Line 3, 4 and 5 of Table 3 present the results. RBFN outperforms the Entropy and Dropout by a large gap on L&F and RA. However, none of the three methods perform well on FS. This indicates that current uncertainty estimation methods are not sensitive to synthesized outliers.

Training Data Here we evaluate how the training data affects the Prop-Seg and Prop-Cls model's performance. Marshal [23] suggests to perform density estimation with background features only for the background segmentation task. We followed their idea and trained a Prop-Seg model using only the background pixels (Line 6 of Table 3). Comparing with line 3 where all class labels are being used, model trained with only background pixels performs much worse across all datasets. This shows the effectiveness of having negative examples (object pixels) during training even for feature distance based models.

r	Input Rep	Uncertainty Estimation	Trained Classes	Regularization	L&F (AC↑/AP↑/FPR$_{95}$ ↓)	RA (AC↑/AP↑/FPR$_{95}$ ↓)	FS (AC↑/AP↑/FPR$_{95}$ ↓)
1	ObjDet	RBFN	All	-	84.8 / 59.2 / 37.9	76.1 / **60.7** / 49.5	74.4 / 48.6 / 48.9
2	ObjDet	RBFN	All	Gradient Penalty	84.0 / 61.5 / 43.1	73.3 / 57.5 / 55.4	68.1 / 43.9 / 61.0
3	SSeg	RBFN	All	-	90.4 / **71.7** / 30.9	74.0 / 59.5 / 48.0	81.0 / 61.3 / 44.9
4	SSeg	Dropout	All	-	81.6 / 52.7 / 45.6	70.6 / 50.0 / 62.8	82.4 / 59.8 / 42.7
5	SSeg	Entropy	All	-	79.3 / 49.4 / 49.7	71.9 / 53.8 / 61.4	77.6 / 53.4 / 51.3
6	SSeg	RBFN	bg only	-	74.1 / 43.9 / 21.3	68.0 / 48.6 / 62.8	69.7 / 49.3 / 62.6
7	SSeg	RBFN	All	No Conv	82.4 / 56.2 / 43.4	74.7 / 60.4 / 52.0	**83.3** / **62.0** / **40.2**
8	SSeg	RBFN	All	Boundary Constraint	**92.1** / 70.3 / **23.4**	**76.2** / 56.5 / **47.3**	82.8 / 56.3 / 43.0

Table 3. Ablation Study on our Proposal Segmentation Model, illustrating the performance of models trained with different visual input, different uncertainty estimation techniques and without the regularization methods.

Method	Cityscapes (Acc)	L&F (AC/AP)	RA (AC/AP)
ObjDet (9 classes)	**97.8**	**95.5 / 40.5**	**95.4 / 67.1**
ObjDet (4 classes)	97.1	93.3 / 25.3	88.3 / 44.5
SSeg (9 classes)	93.1	80.3 / 8.7	78.5 / 28.2
SSeg (4 classes)	89.2	86.1 / 11.5	68.4 / 17.1

Table 4. Proposal Classification on Outdoor Scenes.

4.3. OOD Object Detection in Indoor Scenes

We train a Mask-RCNN and DeeplabV3+ on ADE20K [31] and extract proposals and ObjDet/SSeg feature maps. Two object classes, 'vase' and 'lamp' are ignored during training. During testing, we select object proposals containing these two classes and evaluate our approach on ADE20K and AVD [2] datasets. The results with different input representation are in Table 5. The model using ObjDet features performs slightly better. In indoor scenes proposals have fewer background pixels, taking out the advantage of SSeg features for background representation. Figure 8 shows proposal segmentation results.

Method	ADE20K (AC↑/AP↑/FPR$_{95}$ ↓)	AVD (AC↑/AP↑/FPR$_{95}$ ↓)
ObjDet	**92.3** / **92.8** / **26.8**	**94.9** / **97.0** / 17.9
SSeg	90.9 / 89.4 / 31.3	94.3 / 96.4 /**16.8**

Table 5. Proposal Segmentation on Indoor Scenes: ADE20K and AVD.

5. Conclusion

We proposed a two-step proposal segmentation and classification method using RBFN for unknown object detection. We examine the performance of the proposal segmentation model using different backbone features and a variety of regularization methods. The proposed regularization through boundary pixel constraint proved to be most useful for finding hard out-of-distribution examples. We present comprehensive comparison of the model to alternative approaches in the literature. The proposed method can be used to flag false positives made by modern object detectors. In the experiment, we also demonstrate the method's generalization to indoor scenes. In the future we plan to integrate the RBFN prototype model into a region proposal network to detect general objects more effectively. We are also interested in seeing if the proposed method can detect adversarial attacks on modern semantic segmentation and object detection models.

Figure 8. Top: segmentation of novel ADE20K and AVD instances. Bottom: uncertainty based segmentation, for proposals with more than one object. Note that the uncertainty on the table and wall is apparently lower than the foreground unknown objects.

References

[1] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, NY, USA:, 2012.

[2] Phil Ammirato, Patrick Poirson, Eunbyung Park, Jana Košecká, and Alexander C Berg. A dataset for developing and benchmarking active vision. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*,

pages 1378–1385. IEEE, 2017.

[3] A. Pritzel B. Lakshminarayanan and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *In Advances in Neural Information Processing Systems (NIPS)*, pages 3–18, 2017.

[4] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *arXiv preprint arXiv:1904.03215*, 2019.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[8] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16918–16927, 2021.

[9] Cheng-Yang Fu, Tamara L Berg, and Alexander C Berg. Imp: Instance mask projection for high accuracy semantic segmentation of things. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5178–5187, 2019.

[10] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In *Advances in neural information processing systems*, pages 3581–3590, 2017.

[11] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] Dan Hendrycks and Kevin Gimpel. A baseline for detecting missclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[14] Rui Hou, Jie Li, Arjun Bhargava, Allan Raventos, Vitor Guizilini, Chao Fang, Jerome Lynch, and Adrien Gaidon. Real-time panoptic segmentation from dense detections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8523–8532, 2020.

[15] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. *arXiv preprint arXiv:2103.02603*, 2021.

[16] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15425–15434, 2021.

[17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

[18] Yi Li and Nuno Vasconcelos. Background data resampling for outlier-aware classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13218–13227, 2020.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2152–2161, 2019.

[21] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33, 2020.

[22] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv: 1802.10501*, 2018.

[23] Nicolas Marchal, Charlotte Moraldo, Hermann Blum, Roland Siegwart, Cesar Cadena, and Abel Gawel. Learning densities in feature space for reliable segmentation of indoor scenes. *IEEE Robotics and Automation Letters*, 5(2):1032–1038, 2020.

[24] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

[25] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1099–1106. IEEE, 2016.

[26] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, 2020.

[27] Mark Weber, Jonathon Luiten, and Bastian Leibe. Single-shot panoptic segmentation. *arXiv preprint arXiv:1911.00764*, 2019.

[28] Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying unknown instances for autonomous driving. In *Conference on Robot Learning*, pages 384–393. PMLR, 2020.

[29] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[30] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019.

[31] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[32] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 391–405. Springer, 2014.