Individual topical study report

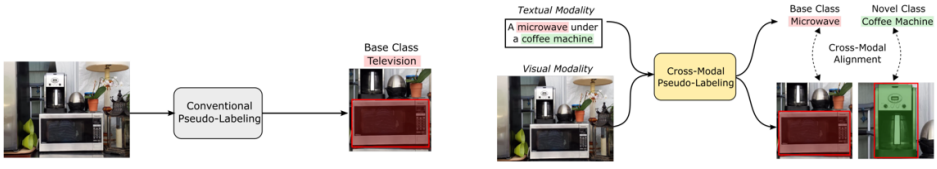
Introduction

In this report, I would like to explore the field of unknown-aware object detection with the assistance of semantic segmentation, which is one of the main topic in this course.
Detecting objects that are not seen from the previous training stage has become a trend in object detection. However, there are more and more findings suggest that the information semantic segmentation provides is extremely helpful for detecting those unknown objects.
As a results, I will provide clear overviews for the following three papers, “Learning to Detect Every Thing in an Open World”, “Residual Pattern Learning for Pixel-wise Out-of-Distribution Detection in Semantic Segmentation”, “Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling”. All of them are recently proposed and accepted in top conferences.

Methods

Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling (CVPR 2022)

https://arxiv.org/abs/2111.12698 (https://arxiv.org/abs/2111.12698)



Open-Vocabulary

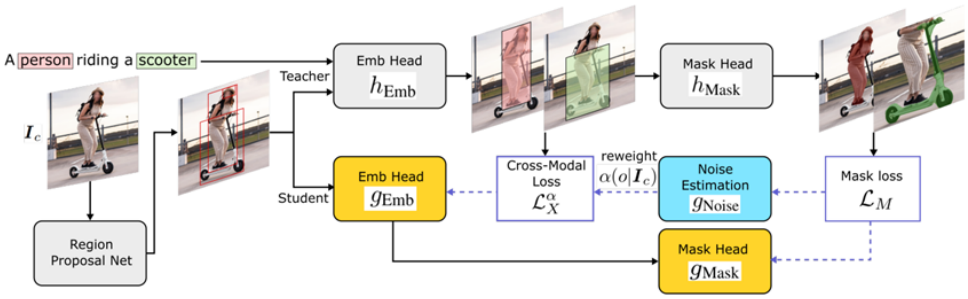
The concept of open-vocabulary in detection and segmentation is to leverage large-scale pretrained image-text models to provide additional information (textual modality) for the model. In this work, a pretrained Bert is introduced to convert object nouns appeared in the caption of the image to word embeddings.

Moreover, the original fully connected layer for classification in Mask-RCNN will be replaced by an embedding layer that projects visual embeddings of region proposals to word embeddings. As such, we can calculate the cosine similarity between the word embedding of a region proposal and the corresponding word embedding provided by the pretrain Bert.

Open-Vocabulary Instance Segmentation

The goal of open-vocabulary instance segmentation aims not only segmenting known classes with mask annotations but also unknown classes without annotations. With the aid of open-vocabulary as stated above, it is now possible to generate pseudo masks via word semantics and visual features. To move further, this paper proposed a mechanism to selectively distills mask knowledge from the teacher model to the student model by estimating the mask noise levels. In such fashion, the student model can surpass the upper-bound of the noisy teacher model, while leveraging the pseudo mask annotations generated by the teacher model.

Overall architecture



Teacher model

The goal of the teacher model is to use caption-image pairs to generate pseudo labels and learn a joint embedding space between word semantic and visual features.
As mentioned earilier, the teacher model is based on Mask-RCNN and replace the last fully connected layer with an embedding layer. The score of each region is based on

the cosine similarity with the Bert word embedding for each class.
The main issue here is that the teahcer model along does not encode enough informaion when it comes to unknown objects. High-level information obtained from pretrained Bert cannot guarantee good ability in segmenation.

Cross-modal pseudo labeling

The purpose of cross-modal pseudo labeling is to distill information from word embeddings and aligned object regions into the student embedding head.
In this stage, the teacher model has not entered mask head, and the aligned object regions for each object noun in the caption is obtained. Student model is taught to predict the correct aligned object region for any given object noun.

$$\mathcal{L}_X(\mathcal{Y}_c|\boldsymbol{I}_c;g) = -\sum_{o\in\mathcal{O}_c}\log\frac{\mathrm{e}^{\boldsymbol{v}_o^\top g_{\mathrm{Emb}}(\boldsymbol{f}_{\boldsymbol{b}_o})}}{\sum_{w\in\mathcal{V}_C}\mathrm{e}^{\boldsymbol{v}_w^\top g_{\mathrm{Emb}}(\boldsymbol{f}_{\boldsymbol{b}_o})}}$$

Pseudo-mask noises

To alleviate the noises in the pseudo mask generated by the teacher model, we need to estimate the noise level. This paper made a assumption that each pixel in pseudo mask is corrupted by a Gaussian noise. Thus, we add randomly obtained Gaussian noises into every pixel and compute the binary cross entropy loss between the teacher mask and the corrupted student mask. Noted that although the Gaussian noises are randomly generated, the variance of it is a learnable parameter. To be more precise, the noisier the teacher mask is, the higher the variance is, the higher the noise level is.

$$\mathcal{L}_M(\mathcal{Y}_c|\boldsymbol{I}_c,g) = \sum_{o\in\mathcal{O}_c}\sum_{x,y}\mathcal{L}_{\mathrm{BCE}}\left(\boldsymbol{M}_o^{xy}|g_{\mathrm{Mask}}^{xy}(\boldsymbol{f}_{\boldsymbol{b}_o})+\epsilon_o^{xy}\right)$$
$$\epsilon_o^{xy}\sim\mathcal{N}\left(0,g_{\mathrm{Noise}}^{xy}(\boldsymbol{f}_{\boldsymbol{b}_o})\right),$$

Robust student model

To sum up, the higher the noise level we estimated from mask loss, the less reliable the aligned object region is. Therfore, the loss of cross-modal pseudo labeling should be weighted by the inverse of the variance.

$$\alpha(o|\boldsymbol{I}_c) = \frac{\eta}{\sum_{x,y} g^{xy}_{\text{Noise}}(\boldsymbol{f}_{\boldsymbol{b}_o})/|\boldsymbol{b}_o|} \quad \forall o \in \mathcal{O}_c$$

$$\alpha(o|\boldsymbol{I}_c) \times \log \frac{e^{\boldsymbol{v}_o^\top g_{\text{Emb}}(\boldsymbol{f}_{\boldsymbol{b}_o})}}{\sum_{w \in \mathcal{V}_C} e^{\boldsymbol{v}_w^\top g_{\text{Emb}}(\boldsymbol{f}_{\boldsymbol{b}_o})}}$$

Residual Pattern Learning for Pixel-wise Out-of-Distribution Detection in Semantic Segmentation (preprint: 2022/11)

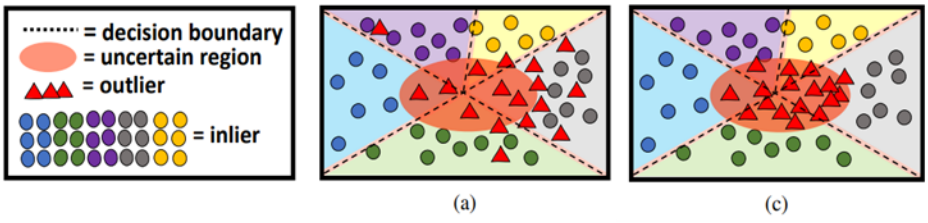
https://arxiv.org/abs/2211.14512 (https://arxiv.org/abs/2211.14512)

Out-of-Distribution Detection(OOD detection)

OOD detection is a subset of unknown-aware object detection. Unlike open-vocabulary object detection directly classifies unknown object to its corresponding class, OOD detection classifies all unknowns into the same class, unknown class. This is fairly reasonable because OOD detection does not rely on large pretrained models. A common way to detect OOD object is to use the energy formula.

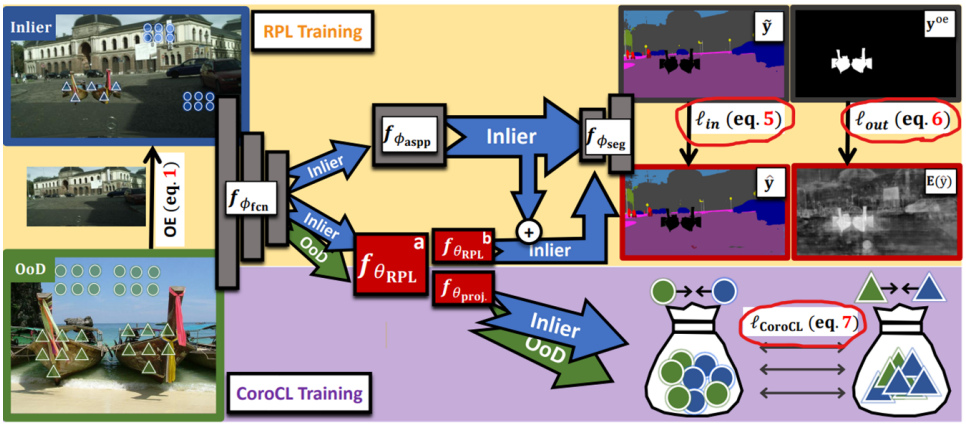
$$E(\mathbf{x}; f) = -T \cdot \log \sum_{i}^{K} e^{f_i(\mathbf{x})/T}$$

where K indicates K known classes. fi(x) represents the logit of input x. T is a hyperparameter usually set to 1. The red regions below are the uncertain regions where energy score are large.



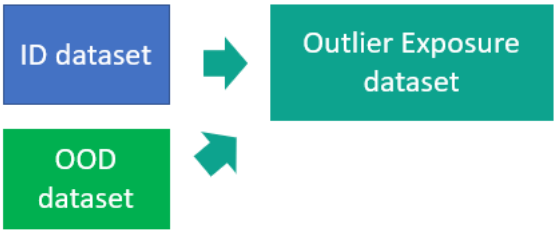
Overall architecture

The architecture includes a pretrained Mask-RCNN model that maintains good performance in known classes and a Residual Pattern Learning (RPL) head that boosts the performance in unknown classes. This setting allows the model to increase the accuracy in unknown classes, while preserving the accuracy in known classes.



Datasets

To construct a dataset with mask annotations of OOD objects, this paper utlizes a method called Outlier Exposure (ICLR 2019). Imagine that you cut and paste the OOD mask annotations from other datasets in your original dataset without overlapping with the known object annotations. Here, we cut OOD objects from COCO dataset and paste them in CityScapes dataset.



For validation, this paper adopts common OOD datasets called Fishyscape and Segment-Me-If-You-Can (SMIYC).

Residual pattern learning (RPL)

RPL is trained to approximate the freezed segmentation model to preserve the performance of known classes. However, RPL loss does not exclusively focus on known classes (In-distribution).

$$\ell_{\text{RPL}}(\mathcal{D}^{oe}, \theta_{\text{rpl}}) = \ell_{in}(\mathcal{D}^{oe}, \theta_{\text{rpl}}) + \alpha \times \ell_{out}(\mathcal{D}^{oe}, \theta_{\text{rpl}})$$

The positive energy loss (lout) specificly target the OOD objects and push their negative energy score up by maximizing them. (mi(w) is a indicator function, set to 1 when pixel w is OOD)

$$\ell_{out}(\mathcal{D}^{oe}, \theta_{\text{rpl}}) = \sum_{(\mathbf{x}_i^{oe}, \mathbf{y}_i^{oe}, \mathbf{m}_i) \in \mathcal{D}^{oe}} \sum_{\omega \in \Omega} \max(-\mathbf{m}_i(\omega) E(\hat{\mathbf{y}}_i(\omega)), 0)$$

Context-robust contrastive learning (CoroCL)

In addition to the original final layer of RPL, the paper adds another projection layer to generate embeddings of known and unknown objects. Then, apply instance-level contrastive learning to make each class objects more compact.

Learning to Detect Every Thing in an Open World (ECCV 2022)

https://arxiv.org/abs/2112.01698 (https://arxiv.org/abs/2112.01698)

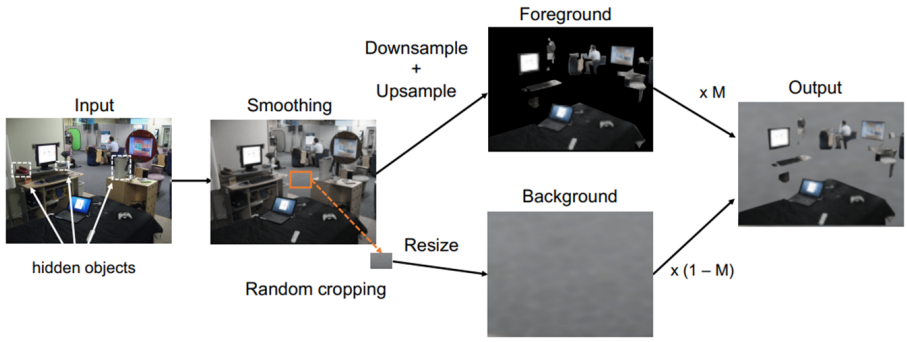
Hidden object issue in object detection

Most of the existing object detection works ignore the fact that there are unannotated objects. These objects are explicitly learned as background in the training stage. The influence might be mild if only known classes appear in the testing stage. However, if we considered unknown objects during inference, the impact will become visible.

Overall architecture

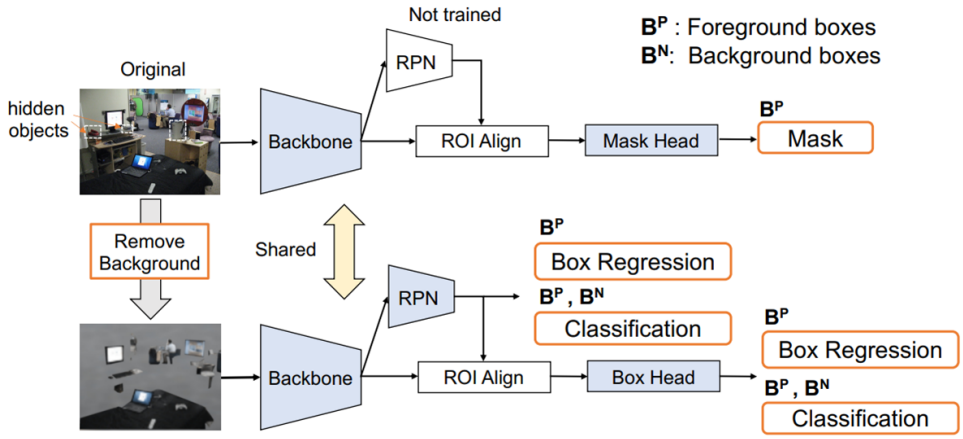
To completely resolved the issue, an intuitive way is to create an customized background without any object and paste foreground objects in the customized background.

Background Erasing (BackErase)



Background Erasing is actually very simple. Randomly sample a small region in the image and resize it to the size of the image. In such way, it is unlikely the sampled region contains foreground objects and even if it contains, after resizing, for example, 64 times, the object will be distroted and look dissimilar to an object.
After that, we paste the mask annotations in the customized background and create our own training images.

Decoupled Multi-Domain Training

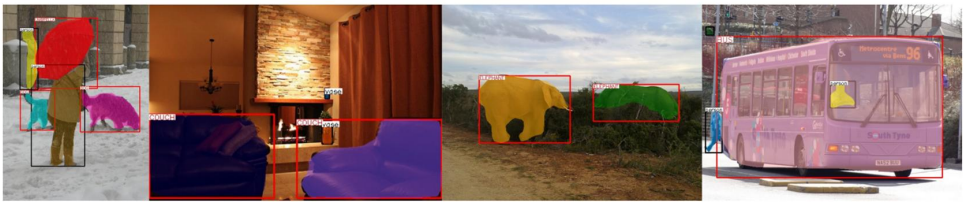


In the previous step, we make sure that the background is not contaminated by hiddem objects. Nonetheless, new problem occurs. The images we trained on are in different a domain comparing to the original images. To elaborate, the frequency of the customized background is close to zero, so **the model might rely on the frequency information to detect objects**, rather than the actual visual information.
Thus, the paper tackles the domain transfer issue by training a mask head on the original images. The function of mask head is to differntiate foreground and background, while the function of box regression is also differentiate foreground and background. As a result, the

mask head trained on original images can help alliviate the domain transfer issue. Noticed that the mask head share the same background with the box head for customized images.

Experiments

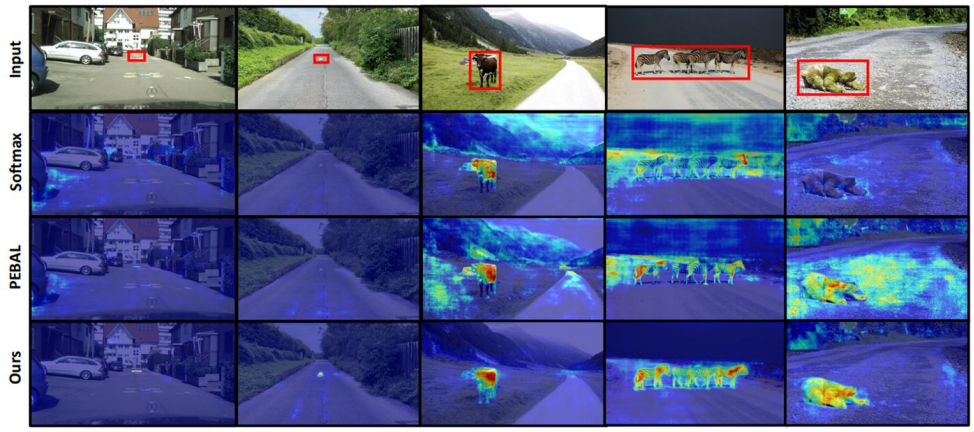
Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling (CVPR 2022)



The proposed method, XPM, obtain good mAP comparing to other caption-image and pseudo-labeling methods. However, **I don't think this is a fair comparison with other pseudo-labeling methods**, since they do not utlize captions in any of their images.

Method	Bounding Box Supervision					Instance Mask Supervision				
	Constrained		Generalized			Constrained		Generalized		
	Base	Target	Base	Target	All	Base	Target	Base	Target	All
Zero-Shot Training										
SB* [24]	29.7	0.7	29.2	0.3	24.9	-	-	-	-	-
BA-RPN* [27]	-	11.4	46.5	4.8	35.6	-	-	-	-	-
Caption Pretraining with [28]										
OVR [28]	46.8	27.5	46.0	22.8	39.9	47.2	25.9	**46.7**	20.7	39.9
SB [24]	46.9	26.9	46.3	21.2	39.7	45.9	25.7	45.3	19.6	38.6
BA-RPN [27]	46.8	26.0	46.2	20.7	39.5	46.0	25.0	45.5	19.3	38.7
OVR+OMP [19]	-	-	-	-	-	34.1	16.9	33.2	10.0	27.1
Pseudo-Labeling										
Soft-Teacher [47]	47.4	18.8	47.1	12.4	38.0	46.6	16.0	46.2	10.4	36.8
Unbiased-Teacher [48]	**47.5**	20.5	**47.2**	13.8	38.4	46.6	16.8	46.1	10.8	36.9
Cap2Det* [97]	-	-	20.1	20.3	20.1	-	-	-	-	-
XPM (Ours)	46.8	**29.9**$^{+2.4}$	46.3	**27.0**$^{+4.2}$	**41.2**	**47.3**	**33.2**$^{+7.3}$	46.3	**29.9**$^{+9.2}$	**42.0**

Residual Pattern Learning for Pixel-wise Out-of-Distribution Detection in Semantic Segmentation (preprint: 2022/11)



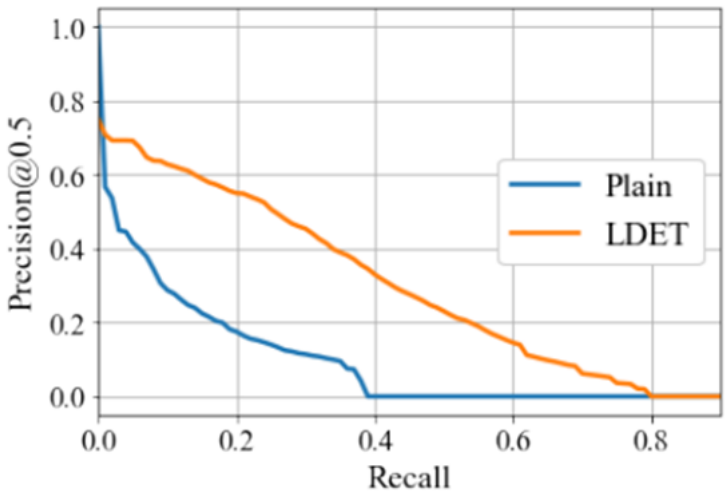
As the picture shown above,the proposed method is able to detect small OOD objects, while maintaining low false positives. **It beats state-of-the-art methods with great margin.** The FPR is extremely low in most cases and AuPRC remains high.

Methods	Fishyscapes (test)				SMIYC (test)						Overall	
	Static		L&F		Anomaly		Obstacle		L&F			
	FPR ↓	AuPRC ↑	FPR ↓	AuPRC ↑	FPR ↓	AuPRC ↑	FPR ↓	AuPRC ↑	FPR ↓	AuPRC ↑	FPR ↓	AuPRC ↑
Resynthesis [25][ICCV'19]	27.13	29.6	48.05	5.70	25.93	52.28	4.70	37.71	8.82	57.08	22.93	36.47
Embedding [1][ICCV'19]	20.25	44.03	30.02	3.55	70.76	37.52	46.38	0.82	10.36	61.70	35.55	29.52
Synboost [8][CVPR'19]	18.75	72.59	15.79	43.22	61.86	56.44	3.15	71.34	4.64	81.71	20.84	65.06
SML [17][ICCV'21]	19.64	53.11	21.52	31.05	-	-	-	-	-	-	-	-
Meta-OoD [3][ICCV'21]	8.55	86.55	35.14	29.96	15.00	**85.47**	0.75	85.07	9.70	77.90	13.83	72.99
DenseHybrid [10][ECCV'22]	5.51	72.27	6.18	43.90	62.25	42.05	6.02	80.79	-	-	-	-
PEBAL [34][ECCV'22]	1.73	92.38	7.58	44.17	40.82	49.14	12.68	4.98	8.63	75.64	14.29	53.26
RPL+CoroCL [Ours]	**0.52**	**95.96**	**2.27**	**53.99**	**11.68**	83.49	**0.58**	**85.93**	**3.22**	**82.29**	**3.65**	**80.33**

Hidden object issue in object detection

As the image and the table show, the proposed method, LDET, significantly outperfroms Mask-RCNN on COCO dataset. It detects more objects than vanilla Mask-RCNN. **LDET is capable of detecting all foreground objects, including OOD objects.**

Method	Non-VOC						All			
	Box			Mask			Box		Mask	
	AP	AR_{10}	AR_{100}	AP	AR_{10}	AR_{100}	AR_{10}	AR_{100}	AR_{10}	AR_{100}
Mask RCNN [20]	1.5	8.8	10.9	0.7	7.2	9.1	19.3	23.1	16.7	19.9
Mask $RCNN^P$	1.1	8.7	10.7	0.6	7.2	8.9	19.1	23.0	16.5	19.8
Mask $RCNN^S$	3.4	13.2	18.0	2.2	11.3	15.8	21.7	27.4	19.2	24.4
LDET	**5.0**	**18.2**	**30.8**	**4.7**	**16.3**	**27.4**	**24.4**	**36.8**	**22.4**	**33.1**



Discussions

Unknown-aware object detection is a large field to explore. The works in this field varies greatly. In this report, I introduce three representative papers trying to solve this task via the assistance of semantic segmentation. I think this line of works is meaningful and expect more future works in the field of unknown-aware object detection.