Exemplar-Based Open-Set Panoptic Segmentation Network

Jaedong Hwang[1*]   Seoung Wug Oh[2]   Joon-Young Lee[2]   Bohyung Han[1]

[1]ECE and ASRI, Seoul National University

[2]Adobe Research

[1]{jd730, bhhan}@snu.ac.kr, [2]{seoh, jolee}@adobe.com

Abstract

We extend panoptic segmentation to the open-world and introduce an open-set panoptic segmentation (OPS) task. This task requires performing panoptic segmentation for not only known classes but also unknown ones that have not been acknowledged during training. We investigate the practical challenges of the task and construct a benchmark on top of an existing dataset, COCO. In addition, we propose a novel exemplar-based open-set panoptic segmentation network (EOPSN) inspired by exemplar theory. Our approach identifies a new class based on exemplars, which are identified by clustering and employed as pseudo-ground-truths. The size of each class increases by mining new exemplars based on the similarities to the existing ones associated with the class. We evaluate EOPSN on the proposed benchmark and demonstrate the effectiveness of our proposals. The primary goal of our work is to draw the attention of the community to the recognition in the open-world scenarios. The implementation of our algorithm is available on the project webpage[1].

1. Introduction

For deeper visual understanding, researchers have constructed large-scale image-based benchmarks [6, 18, 24, 33, 43] and have studied various tasks such as image classification [14, 47], object detection [10, 23, 41, 42], semantic segmentation [4, 27, 34], instance segmentation [3, 13], and many others. Recently, researchers are getting more interested in finding the location and shape of instances and estimating semantic labels from natural scenes since they are critical for high-level understanding of visual content.

Since Kirillov *et al.* [17] recently formulated a panoptic segmentation task, which is a combination of instance segmentation and semantic segmentation, a number of studies [5, 9, 16, 20, 22, 25, 38, 51, 48] have been proposed

---

*This work was done during an internship at Adobe Research.

[1]https://cv.snu.ac.kr/research/EOPSN

Input image | Ground-truth for closed-set

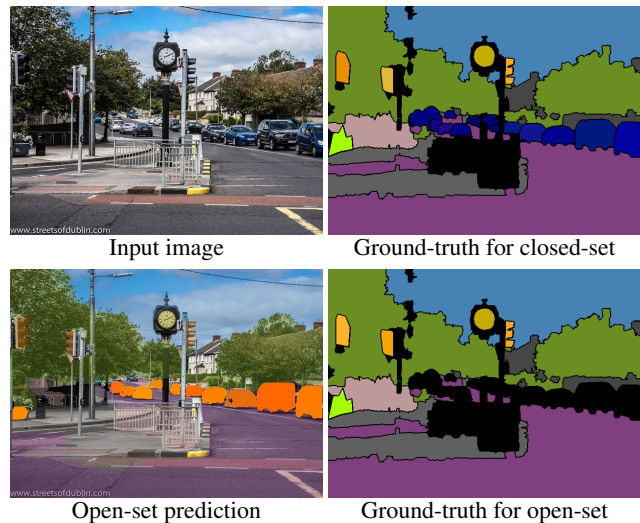Open-set prediction | Ground-truth for open-set

Figure 1: For a given image, we show the *ground-truth* for closed-set and open-set panoptic segmentation, where unlabeled regions are in black. In open-set panoptic segmentation, a model needs to find the *unknown* class instances, which are not annotated in training data. We assume that "car" is an *unknown* class in the open-set visualization.

and achieved substantial performance improvements. However, one drawback of the panoptic segmentation task is the excessive cost of dataset construction. The annotation of pixel-level panoptic segmentation labels requires significant human efforts, and is even harder than semantic segmentation. It is challenging to create a large-scale dataset containing such comprehensive annotations. There are two potential ways to tackle this problem. The first approach is to develop a weakly supervised panoptic segmentation model [21] that is free from pixel-level annotations. The other is to build a generalized model that identifies segments of not only trained classes but also unseen ones. The second approach is closely related to open-set recognition [44] that admits the existence of *unknown* class in testing.

In this paper, we extend panoptic segmentation to the open-world. Unlike the closed-set counterpart, the new

task has to deal with examples in *unknown* classes that are not supervised and acknowledged during training. Therefore, the goal of this task is to discover panoptic segmentation labels for the examples in both *known* and *unknown* classes. We call this new task *open-set panoptic segmentation* (OPS). We argue that the task is an important milestone for advancing comprehensive visual understanding and it also makes substantial impacts on a wide range of practical applications such as dataset construction, scene analysis, and image editing. Figure 1 illustrates an example of OPS.

We address the following critical challenges in the journey to OPS and make the task tractable using reasonable assumptions[2]. The first challenge comes from the definition of *object*. Since a certain object may consist of many components that belong to another objects, a coarse-grained object class can be split into multiple sub-categories. For example, a "bus" consists of various parts such as windows, side mirrors, wheels, etc. On the other hand, it is difficult to handle the *unknown* categories that belong to stuff, *e.g.*, "grass" and "water", since they are often defined as a region and their individual instances are often ill-defined. Therefore, an exhaustive panoptic segmentation for all kinds of objects and backgrounds is almost impossible. The second issue is that, unlike the open-set image classification that discriminates images in *unknown* classes, the open-set panoptic segmentation requires finding *unknown* class objects from a scene even when the *unknown* class instances, unfortunately, are labeled as background.

We propose an exemplar-based open-set panoptic segmentation network (EOPSN) as a strong baseline model for this new task. On top of Panoptic FPN [16], one of the popular closed-set panoptic segmentation approaches, we integrate an extra component that searches for *unknown* class labels inspired by the exemplar theory [30, 35] in psychology. We first find a new *unknown* class and its exemplars by clustering object features and then discover more based on the similarity to existing exemplars during training. Note that the proposed method is generic and applicable to any top-down panoptic segmentation method. Experimental results support the effectiveness of the proposed model; EOPSN outperforms the simple baselines, which are variants of [16]. The contribution of this paper is three-fold as follows:

- We define the open-set panoptic segmentation (OPS) task and make it feasible using reasonable assumptions through in-depth analysis of its inherent challenges.

- We construct a brand-new OPS benchmark by reformatting COCO [24] and present performance of several baselines, which are variants of Panoptic FPN.

- We propose a novel framework for open-set panoptic segmentation, EOPSN, based on the exemplar theory, and demonstrate its effectiveness in detecting and segmenting examples in *unknown* classes.

2. Related Works

2.1. Panoptic Segmentation

Panoptic segmentation, a joint problem of semantic segmentation and instance segmentation, has received a lot of attention since Kirillov *et al.* [17] introduce the task. There exist a large number of works for this problem and they are categorized into two groups: top-down and bottom-up approaches. The top-down techniques [9, 16, 20, 22, 25, 38, 51] typically generate object proposals and segment the proposals before combining semantic segmentation results. Mask-RCNN [13] is often deployed for instance segmentation and an encoder-decoder architecture is utilized for semantic segmentation. On top of that, AUNet [22] leverages mask-level attention to transfer knowledge from the instance segmentation head to the semantic segmentation head. UPSNet [51] proposes a parameter-free panoptic head to resolve conflicts in *thing* and *stuff* predictions. On the other hand, the bottom-up (proposal-free) methods obtain semantic segmentation outputs and then perform instance partitioning [5, 48, 52]. The bottom-up approaches are free from the instance-instance or instance-background overlap issues, but they generally achieve lower accuracy than the top-down methods. Our method is categorized to a top-down approach, which processes instances and semantic backgrounds separately.

2.2. Open-Set Learning

Open-set recognition receives a spotlight in the computer vision community recent years. The goal of open-set image classification is to classify *known* class images observed during training while recognizing examples in *unknown* classes [44]. In an open-set classification algorithm [45], *unknown* class is further divided into two subgroups depending on whether it has been exposed as a negative class during training (*seen-unknown* class) or it has never appeared (*unseen-unknown* class). On the other hand, similar to open-set object detection [8], such a strategy is not applicable to open-set panoptic segmentation since there is no information about whether unlabeled objects are seen or unseen in the training dataset.

Most open-set image classification methods [2, 26, 32, 36] make predictions to *unknown* classes at test time if the probabilities of all *known* classes are below a certain threshold. OpenMax [2] leverages $(C + 1)$-way classifier for C-class classification with background and trains the classifier using the Weibull distribution. Generated images [36] and counterfactual images [32] are employed to estimate ro-

---

[2]Section 3.3 discusses the details.

bust decision boundaries for *known* classes and identify *unknown* classes effectively. Liu *et al.* [26] propose a memory- and clustering-based model for open-set image classification when the distribution of *known* classes is long-tailed.

Recently, the open-set scenario is spread to other tasks such as object detection [8, 31] and semantic instance segmentation [37]. Miller *et al.* [31] first address open-set object detection using dropout sampling. The task is further investigated and formalized by Dhamija *et al.* [8]. Pham *et al.* [37] introduce a Bayesian optimization framework that considers both object boundaries and masks for open-set instance segmentation. This work formulates the open-set panoptic segmentation task on top of the closed-set counterpart.

2.3. Exemplar-Based Learning

Unlike prototype-based techniques [46, 50] that utilize cluster centroids as the representatives of the individual clusters, exemplar-based methods directly employ samples stored in memory. One of the well-known exemplar-based methods in machine learning is k-nearest neighbor algorithm [7], which assigns labels of new data using the training examples in neighborhood. Wu *et al.* [49] perform representation learning without manual annotations by providing each example with a separate class label. Incremental learning frameworks [12, 39, 40] often adopt exemplar-based learning to prevent catastrophic forgetting [29]. They store a small number of exemplars that represent the whole dataset or classes in the previously tasks effectively. MemAE [11] employs exemplar-based learning for anomaly detection.

Existing exemplar-based learning approaches assume that ground-truth labels for exemplars are available or make the information less critical. On the other hand, our problem is much more challenging because it requires to identify exemplars for diverse *unknown* classes automatically and learn their representations properly without labels.

3. Open-Set Panoptic Segmentation (OPS)

This section defines the open-set panoptic segmentation with its evaluation metrics. Then, we discuss the inherent challenges of the task and how we make the problem tractable via reasonable assumptions.

3.1. Definition of Label and Task

The open-set panoptic segmentation is similar to the closed-set method. The key difference is the existence of *unknown* classes, which are not available for training but appear in testing while the *known* classes are always available. Another criterion to distinguish class types is how an object is formed physically; the *thing* class, denoted by $\mathcal{C}^{\text{Th}}$, consists of the objects with concrete shape and structure while the *stuff* class, $\mathcal{C}^{\text{St}}$, involves amorphous background regions
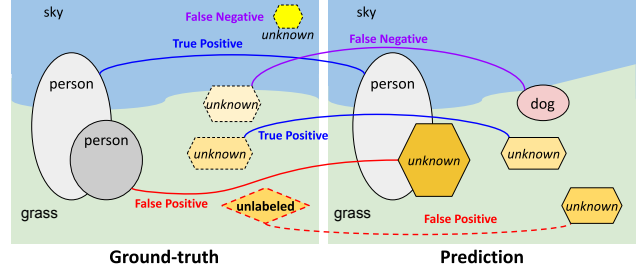

Figure 2: Toy example illustration of ground-truth and predicted panoptic segmentation of an image. Objects with black-dotted border denote *unknown* classes and an object with red-dotted border in the ground-truth is an unlabeled object. The connections across objects in the two different images indicate matching results with their attributes. Although the model correctly finds an *unknown* instance in blue, it is considered as a false positive since the instance is not labeled in the ground-truth.

(*e.g.*, "sky" or "sand") or unstructured objects (*e.g.*, "tree" or "grass"). In addition, we introduce another kind of semantic label, called *void*, which is not annotated in training data and corresponds to ambiguous or out-of-class pixels.

Given a predefined set of C semantic classes encoded by $\mathcal{C} := \{0, ..., C - 1\}$ and a set of *unknown* class codes denoted by $\mathcal{U}$, the ith pixel of an image is labeled by a tuple, $(l_i, z_i) \in (\mathcal{C} \cup \mathcal{U}) \times \mathbb{N}$, where l_i and z_i indicate a semantic class label and its instance identifier, respectively. In our problem definition, $\mathcal{U}$ actually has a single element that represents the entire *unknown* classes. The group of pixels with the same instance identifier (z_i) constitute a segment that belongs to class $l_i \in (\mathcal{C} \cup \mathcal{U})$. In summary, the open-set panoptic segmentation aims to find all segments with *known* labels or *unknown* flags in a given image.

3.2. Evaluation Metric

We utilize standard panoptic segmentation metrics [17], which include panoptic quality (PQ), segmentation quality (SQ) and recognition quality (RQ) for both *known* and *unknown* classes. The metrics are defined as

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g) \in TP} \text{IoU}(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \cdot \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}}, \quad (1)$$

where IoU means Intersection-over-Union of two regions, and *TP*, *FP*, and *FN* denote true positive, false positive, and false negative, respectively, as illustrated in Figure 2. If a predicted segment and a ground-truth region with the same semantic label are highly overlapped, *i.e.* IoU > 0.5, the identified segment is classified as the true positive set, *TP*.
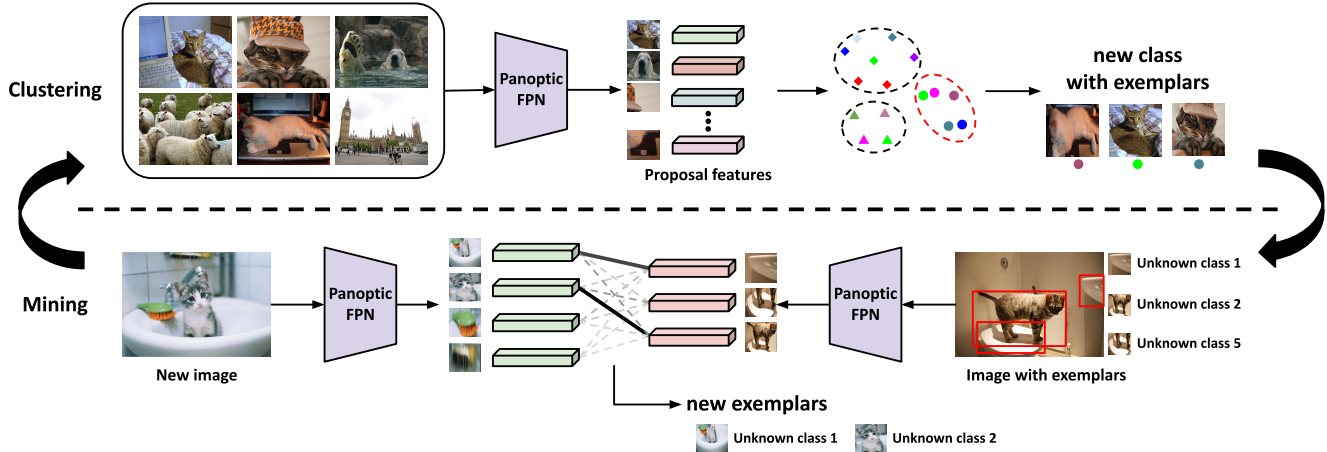
Figure 3: Illustration of the proposed framework, EOPSN. The model first identifies *unknown* classes with the corresponding exemplars using k-means clustering [28] of the detected bounding boxes in *void* areas from a small subset of images (*clustering* stage). Each element in a cluster denotes a proposal feature while each color indicates a source image. The cluster in red-dotted ellipse corresponds to one of the new *unknown* classes and each image patch in the cluster becomes an exemplar. Identified *unknown* classes and their exemplars are used to discover more exemplars by comparing with object proposals of an input image in subsequent iterations (*mining* stage). The two stages alternate to identify and augment *unknown* classes but clustering is performed sparsely to reduce computational complexity.

3.3. Challenges

Finding unknown instances   Unlike the open-set image classification, which simply recognizes unknown class images, an open-set panoptic segmentation model needs to find objects in *unknown* class from each image. It is a very difficult problem since there is no explicit semantic knowledge about the objects and the *unknown* objects might have been labeled as background during training.

Ground-truth labeling   Panoptic segmentation requires pixel-level annotations for each instance and semantic background region. Annotators are supposed to delineate all instances and background stuffs thoroughly. However, unfortunately, a certain object may consist of multiple components that correspond to other object classes, making comprehensive annotations challenging. For example, "car" is composed of several parts such as windows, side mirrors, lights, tires, etc. Also, the regions in some *stuff* labels (*e.g.*, "tree", "gravel") can be divided into several instances. Without a concrete guidline, annotations will be inconsistent and unreliable. It hinders training open-set panoptic segmentation models and measuring their performance.

Evaluation   Current metrics for evaluating panoptic segmentation performance assume that each image has a complete annotation. However, the universe set of class labels is unbounded in practice, which makes the evaluation inconsistent with true quality of panoptic segmentation. In other words, if an open-set panoptic segmentation algorithm identifies *unknown* instances very accurately but the labels of the detected objects are missing as illustrated in Figure 2, the false positive rate will be large, which leads to low RQ and PQ values. On the other hand, SQ does not suffer from such an issue but it is not straightforward to represent the overall performance since it only considers true positives.

3.4. Tractable Problem Definition

Since we face significant conceptual and practical challenges in OPS as discussed in Section 3.3, we propose a tractable version of problem definition as follows.

Assumption   We have the following three assumptions regarding *unknown* classes observed in training data. First, all *unknown* classes belong to the *thing* category. This is because potential *unknown* classes in the *stuff* category are often ill-defined. The second assumption is that parts of *known* classes cannot be *unknown* classes. For example, assuming that "car" and "person" are *known*, "tire" and "head" cannot be *unknown* instances since they are parts of the aforementioned *known* classes, respectively. However, if "tire" exists by itself, not as a part of another, it can be an *unknown* class. The final assumption is that *unknown* class objects only appear in the *void* regions during training. The purpose of the assumptions is to prevent confusion between *known* and *unknown* class regions.

Dataset   Perfect annotation for the open-set panoptic segmentation is not realistic. Hence, a reasonable way to measure the OPS quality is training on one dataset and testing on another with *unknown* classes, which is similar to

multi-domain semantic segmentation [19]. However, there is a domain gap between datasets and the definition of an *object* may be different in each dataset. Therefore, we create a dataset for OPS using an existing dataset. There are several public benchmarks for panoptic segmentation, which include COCO [24], Cityscape [6], Mapillary [33] and ADE20k [55]. However, all of them except COCO are specialized to specific environments such as driving or indoor scene. Hence, we adopt COCO, which contains natural photographs in everyday life. We generate several splits with different numbers of *unknown* classes sampled from the *thing* classes. However, our benchmark cannot evaluate unlabeled *unknown* classes properly as discussed earlier because the label information is completely missing throughout training and evaluation.

4. Method

We explain the baseline models and the proposed approach for open-set panoptic segmentation in this section.

4.1. Motivation and Overview

The core technique in OPS is how to find instances in *unknown* classes. One possible way is to leverage a class-agnostic model such as RPN [42], which predicts an objectness score. We employ this approach as our baseline. However, it has a critical limitation; the score is optimized to the *known* instances. Hence, the performance of the model depends heavily on the semantic similarity between *known* and *unknown* classes, which means that *unknown* classes are unlikely to be recognized by the model if they are not semantically related to any of *known* classes.

To tackle this drawback, we propose an exemplar-based learning framework on top of the baseline to find *unknown* classes from training data more effectively. Note that an exemplar means a bounding box corresponding to an example identified as *unknown* class. Our model first finds *unknown* classes with the associated exemplars using k-means clustering [28], which is applied to the detected bounding boxes within *void* area. A tightly coupled cluster with a high objectness score is considered as an *unknown* class, and each image patch in the cluster becomes an exemplar. The exemplars are used to find new ones by comparing the similarities between the existing exemplars and the object proposals from the images contained in the subsequent mini-batches. They are also employed as the pseudo-ground-truth bounding boxes for the future iterations. The clustering and mining procedures alternate to detect new *unknown* classes and collect examples with the identified *unknown* labels while the backbone network is concurrently optimized with both types of class labels. Figure 3 illustrates the procedure of the proposed approach.

4.2. Baseline

Since only *thing* classes can be *unknown* according to our problem definition, it is reasonable to adopt top-down panoptic segmentation methods with RPN as baseline models. We choose Panoptic FPN [16], which is composed of an instance segmentation head [13] and a semantic segmentation head on top of a shared feature extractor. The instance segmentation head is exactly the same as Mask R-CNN [13] and the semantic segmentation head is identical to the FPN-based decoder.

We modify the bounding box regressor and the mask predictor in the instance segmentation head to make them class-agnostic for handling *unknown* classes. This baseline model first predicts bounding boxes pertaining to *known* classes and finds *unknown* instances based on the objectness scores of the candidates from RPN, where the score threshold is 0.5. We also introduce the *void* class in the classification branch within the instance segmentation head to identify the bounding boxes sampled from the *void* regions. For the supervision of the *void* classes, we use the bounding boxes that more than a half of the region is inside the *void* area. Note that, since the boxes from *void* regions do not necessarily correspond to objects, they are not employed to train the RPN.

4.3. Exemplar-Based Open-Set Learning

The main goal of our exemplar-based learning is to identify bounding boxes with coherent features that belong to the same *unknown* classes. Our exemplar-based learning consists of two stages: clustering candidate proposals in *unknown* classes obtained from a subset of training images, which aims to find new categories and their exemplars, and mining new exemplars via similarity matching with the existing ones. These clustering and mining stages alternate throughout the training procedure.

For the clustering and mining stages, we extract the features from *unknown* candidate bounding boxes in *void* regions for *unknown* classes. To this end, we first reduce duplicate detections by applying Non-Maximum Suppression (NMS) with the IoU threshold 1×10^{-7}, and then sample candidate proposals with the weights based on the objectness scores given by RPN [42]. After that, 1024-dimensional features are obtained from the proposals using a backbone network, and we perform a series of operations including RoI-Align, GAP, and feature computation using two fc layers as in [13].

Clustering To find *unknown* classes, we perform k-means clustering based on the consine distance using the extracted features from candidate bounding boxes. Note that the clustering is performed at every 200 iterations using the features computed for all the examples presented in the last 200 mini-batches. We generate a large number of clusters (over-

Table 1: Open-set panoptic segmentation results on the COCO *val* set ($K = 20\%$) of the baseline approaches with different utilization of *void* regions: used as backgrounds (*Void-background*), ignored (*Void-ignorance*), supervised by Eq (3) (*Void-suppression*) and trained as a new label (*Void-train*) during training.

Utilization of *void* regions	Known									Unknown		
	PQ	SQ	RQ	PQ$^{\text{Th}}$	SQ$^{\text{Th}}$	RQ$^{\text{Th}}$	PQ$^{\text{St}}$	SQ$^{\text{St}}$	RQ$^{\text{St}}$	PQ	SQ	RQ
Void-ackground	**37.7**	76.3	**46.6**	44.8	79.3	54.1	**29.2**	72.8	**37.5**	4.0	71.1	5.7
Void-ignorance	37.2	76.3	45.9	43.9	79.0	53.1	29.1	**73.0**	37.3	3.7	71.8	5.2
Void-uppression	37.5	75.9	46.1	**45.1**	**80.6**	**54.5**	28.2	70.2	36.1	7.2	**75.3**	9.6
Void-train	36.9	**76.4**	45.5	44.0	80.3	53.3	28.2	71.7	36.0	**7.8**	73.4	**10.7**

clustering) and take a subset of clusters only that clearly correspond to true *unknown* classes. Such a cluster should have a high average objectness score but a small average cosine distance between the centroid and all elements since the instances in a true *unknown* class are prone to be clustered tightly while non-object proposals have loose connections and low objectness scores. We identify exemplars from the high-quality clusters and store them in the subsequent mining stage. Note that we maintain exemplars in multiple *unknown* classes during training but collapse all of them and make a single *unknown* class for evaluation.

Mining exemplars Our approach mines additional exemplars with the *unknown* concepts detected in the past, from the images in the incoming mini-batch. This can be done easily by comparing features of the stored exemplars and the features of the object proposals from the images in a new mini-batch. Note that, since the feature extraction network is updated over time, we need to recompute the features of the stored exemplars. We accept the proposals generated by RPN as new exemplars if their cosine similarities to any of existing exemplars are higher than a threshold.

Loss We utilize the almost same classification loss (including the regression loss), mask loss, and semantic segmentation loss adopted in Panoptic FPN [16]. The only difference is classification loss on the instance segmentation head. We utilize the cross-entropy loss over *known* classes, a background class (bg), and *unknown* classes as follows:

$$\mathcal{L}_{\text{ce}} = \sum_{c \in (\mathcal{C}^{\text{Th}} \cup \{\text{bg}\} \cup \mathcal{U})} -y_c \log p_c, \qquad (2)$$

where y_c is a (pseudo-)ground-truth label and p_c is the softmax score of class c. Additionally, we give negative supervision to the object proposals in the *void* regions so that they are not classified as *known* classes, which is given by

$$\mathcal{L}_{\text{void}} = \sum_{c \in \mathcal{C}^{\text{Th}}} -\log(1 - p_c). \qquad (3)$$

Then, the total classification loss is given by

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{ce}} + \mathbb{I}_{\text{void}} \mathcal{L}_{\text{void}}, \qquad (4)$$

where $\mathbb{I}_{\text{void}}$ is an indicator function for a box in *void* regions.

Inference Our model predicts instance segmentation and semantic segmentation results and then combine them to make a panoptic segmentation output. We first generate panoptic segmentation map using a same inference mechanism of the standard panoptic segmentation model [16] and then add predicted *unknown* instances additionally.

5. Experiments

We describe our experimental setting and present the results. We also analyze various aspects of the proposed framework. Refer to the supplementary document for implementation details and more experimental results.

5.1. Dataset and Evaluation

All exepriments are conducted on COCO [24]. We utilize 2017 panoptic segmentation train and validation splits, which contain 118K and 5K images, respectively, with annotations of 80 *thing* classes and 53 *stuff* classes. To construct an open-set setup, we remove annotations for a subset of *known* thing classes in the training dataset and consider them as *unknown* classes. We construct three different splits with varying numbers of *unknown* classes (5%, 10%, 20%)[3]. The removed classes in the the splits are shown below, where the classes are removed cumulatively.

- 5%: "car", "cow", "pizza", "toilet"
- 10%: "boat", "tie", "zebra", "stop sign"
- 20%: "dining table", "banana", "bicycle", "cake", "sink", "cat", "keyboard", "bear"

We employ the standard panoptic segmentation metrics, *i.e.*, PQ, SQ, and RQ, to evaluate performance. We report performance of *known* class and *unknown* class separately.

5.2. Quantitative Results

Table 1 presents open-set panoptic segmentation performance of the baseline models relying on Panoptic FPN by varying the utilization of *void* regions on COCO *val* set with 20% *unknown* classes. There are four different usages of

---

[3]We sorted the class labels based on their frequencies and sampled a subset of labels regularly for removal to simulate *unknown* classes.

Table 2: Open-set panoptic segmentation results on the COCO *val* set with several different *known-unknown* splits K denotes the ratio of *unknown* classes to all classes. The numbers in bold denote higher scores than the opponents.

$K(\%)$	Model	Known									Unknown		
		PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}	PQ	SQ	RQ
	Supervised	39.4	77.7	48.4	45.8	80.7	55.4	29.7	73.1	38.0	-	-	-
5	Baseline (*Void-train*)	37.7	76.7	46.4	44.2	80.4	53.5	28.3	71.3	36.2	10.0	73.8	13.5
	EOPSN	**38.0**	**76.9**	**46.8**	**44.8**	**80.5**	**54.2**	28.3	**71.9**	36.2	**23.1**	**74.7**	**30.9**
10	Baseline (*Void-train*)	36.9	75.4	45.5	43.2	79.0	52.4	28.3	70.4	36.2	8.5	73.2	11.6
	EOPSN	**37.7**	**76.8**	**46.3**	**44.5**	**80.6**	**53.8**	**28.4**	**71.8**	36.2	**17.9**	**76.8**	**23.3**
20	Baseline (*Void-train*)	36.9	**76.4**	45.5	44.0	80.3	53.3	28.2	**71.7**	36.0	7.8	73.4	10.7
	EOPSN	**37.4**	76.2	**46.2**	**45.0**	80.3	**54.5**	28.2	71.2	**36.2**	**11.3**	**73.8**	**15.3**

the boxes: training as backgrounds (*Void-background*), ignoring the boxes (*Void-ignorance*), providing the negative supervision for *known* classes using (3) to prevent the boxes from being classified as *known* class (*Void-suppression*), and training as a new class with standard cross-entropy loss (*Void-train*). All variations have similar performance for *known* classes but *unknown* classes have different characteristics. *Void-ignorance* model has the worst performance since the instances in *unknown* class are classified as *known* classes. *Void-suppression* and *Void-train* models achieve comparable accuracy and outperform the rest two options since bounding boxes in *void* area are not trained as backgrounds in RPN and do not have low objectness scores.

Table 2 summarizes the experimental results on the COCO *val* set with different *known-unknown* splits. The supervised model denotes Panoptic FPN trained on all classes without the *unknown* ones. The baseline is Panoptic FPN with the *Void-train* option, which is one of the variants presented in Table 1. EOPSN outperforms the baseline method in all aspects for *unknown* classes with large margins while still achieving competitive performance for *known* classes. Overall, both PQ and RQ in *unknown* classes are much smaller than *known* cases while SQ's are similar in both *known* and *unknown* classes. This is because SQ is computed based only on true positives.

5.3. Qualitative Results

Figure 4 illustrates the exemplars in a detected *unknown* class after the first clustering with $K = 20\%$ in EOPSN. Most exemplars contain "car" while there exist objects in other *unknown* classes such as "cake", "cow", and "bicycle" as well as instances in *known* class, "giraffe".

Figure 5 presents the comparison of the open-set panoptic segmentation methods. between the baseline model (the third row) and EOPSN (the forth row) on the COCO *val* set with $K = 20\%$. The *unknown* classes in the figures are "stop sign", "car", "keyboard", "sink", and "toilet". The second row shows ground-truths, where *unknown* classes are in orange. We observe that EOPSN successively finds
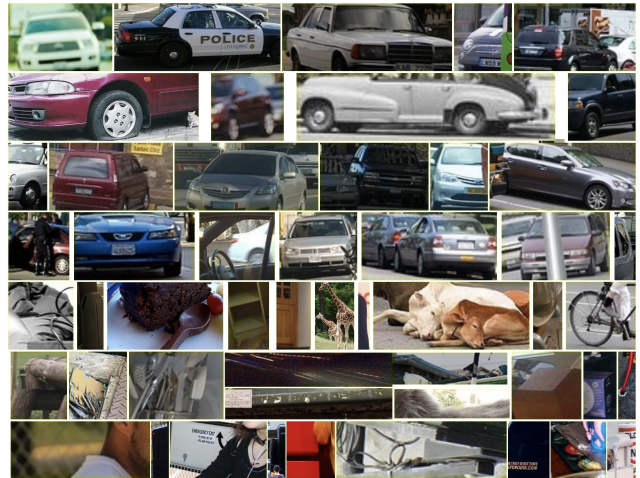

Figure 4: Visualization of the exemplars in an identified *unknown* class by the first clustering in EOPSN.

several *unknown* instances missed by the baseline model. Interestingly, EOPSN discovers the keys inside the keyboard and the bathtub unit, which are not included in the COCO classes.

6. Discussion

In addition to the challenges discussed in Section 3.3, several critical issues still remain. First of all, EOPSN mainly focuses on the classification branch in the instance segmentation head while we have not explored potential of the semantic segmentation head sufficiently. We believe that combining a bottom-up approach to form *unknown* segments would improve performance.

Second, EOPSN is based on clustering with training data to find new *unknown* classes. If an *unknown* class does not appear in the training dataset, its performance will be degraded considerably. However, as we collect more data, the number of *seen-unknown* classes will grow rapidly and the benefit of the proposed model will become salient; the tech-
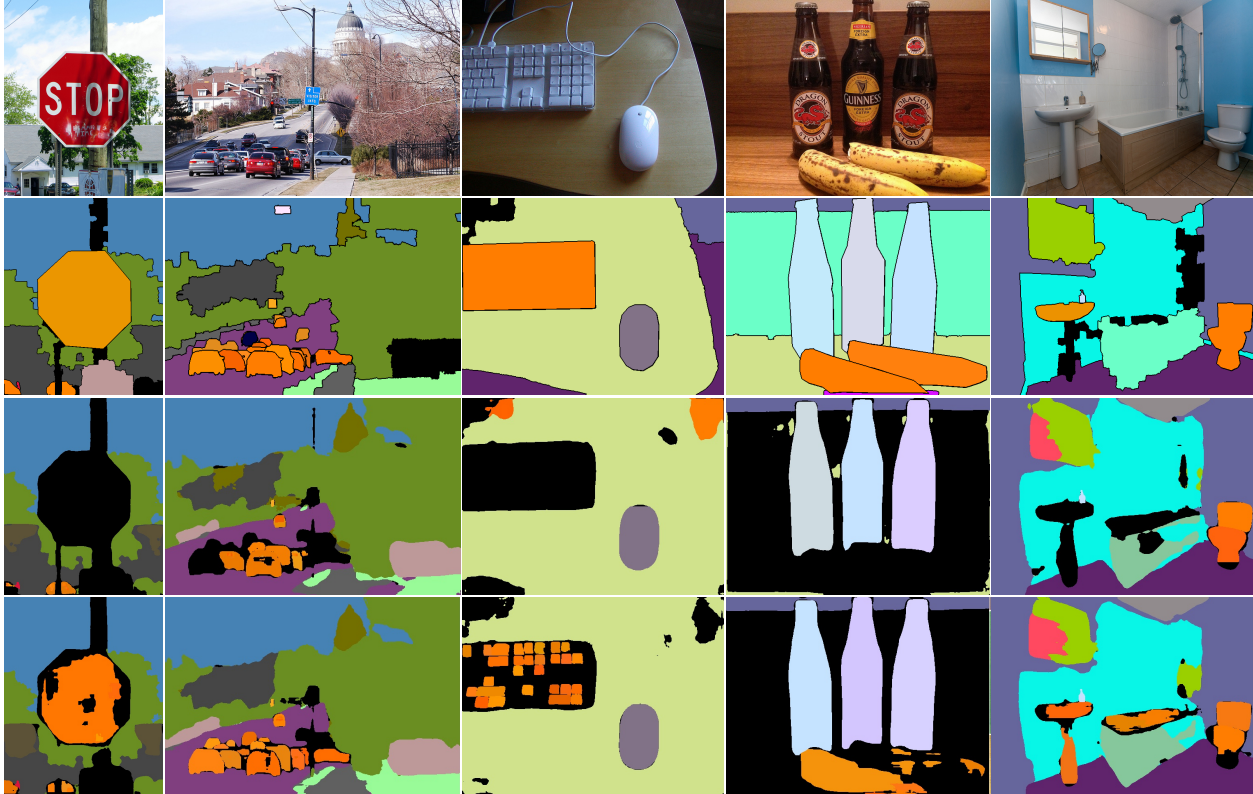
Figure 5: Qualitative results on the COCO *val* set with $K = 20\%$. The first row presents input images and the subsequent rows illustrate ground-truths, results of the baseline (Void-train), and those of EOPSN. Instances in the *unknown* class are denoted by orange color.

niques based on the class-agnostic objectness scores, *i.e.*, the baseline methods based on RPN, would have less merits since they only employ the ground-truth information of *known* classes.

Finally, a new metric should be defined for OPS. If a trained model finds more *unknown* instances in testing, which are legitimate but unlabeled, the inconsistency between true and measured accuracy will be aggravated. A straightforward solution is to employ human evaluation, but it is expensive and may incur another critical issues related to consistency. This issue may be alleviated by introducing a new metric by revising SQ. For example, by properly considering false positives of *unknown* classes, we can represent overall segmentation quality more accurately.

7. Conclusion

We introduced a novel task referred to as open-set panoptic segmentation (OPS), which involves *unknown* classes that appear in testing while not considered during training. The goal of this task is to obtain a correct panoptic segmentation map for the union of *known* and *unknown* classes. We provided a new benchmark dataset on top of COCO [24].

To tackle the challenging task, we presented EOPSN, an exemplar-based open-set panoptic segmentation network as a solid baseline model. Our framework first identifies new classes with the associated exemplars by clustering the proposals sampled from a small subset of images, and then discovers new exemplars progressively using the rest of the images during training. EOPSN outperforms the baselines based on variants of Panoptic FPN [16]. Since EOPSN is a generic approach that can be incorporated into any top-down panoptic segmentation model, the integration of the state-of-the-art networks would lead to better performance.

We hope that this work draws the attention of the computer vision community to open-set problems beyond simple recognition tasks. OPS would facilitate large-scale dataset collection that requires dense labeling and allow us to tackle more realistic tasks in challenging scenarios.

References

[1] Kelsey R Allen, Evan Shelhamer, Hanul Shin, and Joshua B Tenenbaum. Infinite mixture prototypes for few-shot learning. *ICML*, 2019.

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, 2016.

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *CVPR*, 2019.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018.

[5] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020.

[6] 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 *CVPR*, 2016.

[7] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *TIT*, 13(1):21–27, 1967.

[8] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *WACV*, 2020.

[9] Cheng-Yang Fu, Tamara L Berg, and Alexander C Berg. Imp: Instance mask projection for high accuracy semantic segmentation of things. In *ICCV*, 2019.

[10] Ross Girshick. Fast r-cnn. In *CVPR*, 2015.

[11] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, 2019.

[12] Lei Guo, Gang Xie, Xinying Xu, and Jinchang Ren. Exemplar-supported representation for effective class-incremental learning. *IEEE Access*, 8:51276–51284, 2020.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[15] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018.

[16] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019.

[17] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[19] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *CVPR*, 2020.

[20] Justin Lazarow, Kwonjoon Lee, Kunyu Shi, and Zhuowen Tu. Learning instance occlusion for panoptic segmentation. In *CVPR*, 2020.

[21] Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly-and semi-supervised panoptic segmentation. In *ECCV*, 2018.

[22] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *CVPR*, 2019.

[23] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[24] 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 *ECCV*, 2014.

[25] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In *CVPR*, 2019.

[26] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[28] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[29] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[30] Douglas L Medin and Marguerite M Schaffer. Context theory of classification learning. *Psychological review*, 85(3):207, 1978.

[31] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, 2018.

[32] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *ECCV*, 2018.

[33] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017.

[34] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[35] Robert M Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39, 1986.

[36] Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *CVPR*, 2020.

[37] Trung Pham, Vijay BG Kumar, Thanh-Toan Do, Gustavo Carneiro, and Ian Reid. Bayesian semantic instance segmentation in open set world. In *ECCV*, 2018.

[38] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *CVPR*, 2019.

[39] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *ECCV*, 2020.

[40] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.

[41] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017.

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

[44] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *TPAMI*, 35(7):1757–1772, 2012.

[45] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *TPAMI*, 36(11):2317–2324, 2014.

[46] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[48] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, 2020.

[49] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[50] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *ICML*, 2018.

[51] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, 2019.

[52] Tien-Ju Yang, Maxwell D Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *arXiv preprint arXiv:1902.05093*, 2019.

[53] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019.

[54] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019.

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.