Towards Open-World Segmentation of Parts

Tai-Yu Pan[1,*] Qing Liu[2] Wei-Lun Chao[1] Brian Price[2]
[1]The Ohio State University [2]Adobe Research
{pan.667, chao.209}@osu.edu {qingl, bprice}@adobe.com

Abstract

Segmenting object parts such as cup handles and animal bodies is important in many real-world applications but requires more annotation effort. The largest dataset nowadays contains merely two hundred object categories, implying the difficulty to scale up part segmentation to an unconstrained setting. To address this, we propose to explore a seemingly simplified but empirically useful and scalable task, class-agnostic part segmentation. *In this problem, we disregard the part class labels in training and instead treat all of them as a single part class. We argue and demonstrate that models trained without part classes can better localize parts and segment them on objects unseen in training. We then present two further improvements. First, we propose to make the model object-aware, leveraging the fact that parts are "compositions", whose extents are bounded by the corresponding objects and whose appearances are by nature not independent but bundled. Second, we introduce a novel approach to improve part segmentation on unseen objects, inspired by an interesting finding — for unseen objects, the pixel-wise features extracted by the model often reveal high-quality part segments. To this end, we propose a novel self-supervised procedure that iterates between pixel clustering and supervised contrastive learning that pulls pixels closer or pushes them away. Via extensive experiments on PartImageNet and Pascal-Part, we show notable and consistent gains by our approach, essentially a critical step towards open-world part segmentation.*

1. Introduction

Segmenting "objects" from images, such as cup, bird, vehicle, etc., is a fundamental task in computer vision and has experienced a series of breakthroughs in recent years thanks to deep learning [6, 15, 18] and large-scale data [12, 22, 30]. In many real-world applications like object grasping, behavior analysis, and image editing, however, there is often a need to go beyond "objects" and dive deeper into their compositions, *i.e.*, "parts"; for example, to segment cup handle, bird wing, vehicle wheel, etc.

Arguably, the most straightforward way to tackle this problem is to perform part "instance" segmentation, *treating each object part as a separate class; each appearance as a separate instance.* A model then must localize parts, classify them, and demarcate their boundaries. In object-level instance segmentation [15], these three sub-tasks are usually approached simultaneously, or at least, share a model backbone. Such a multi-task nature enables the model to benefit from the complementary cues among sub-tasks to attain higher accuracy. For instance, the shapes of segments often entail the class labels and vice versa.

Segmenting parts in this way, however, limits their scope to the closed world. That is, the learned model may not, or by default should not[1], generalize to object categories (and their corresponding parts) that are unseen during training. Although the largest dataset nowadays for part segmentation, PartImageNet [14], has covered more than a hundred object categories, a scale similar to representative object-level datasets like MSCOCO [30] and OpenImages [22], it is arguable not enough to cover the need in the wild.

To equip the model with the open-world capability — the ability to segment parts for unseen objects — we propose to chop off the "classification" function from the model, as it is simply not applicable to unseen parts. Namely, we remove the pre-defined fences among different object parts and instead assign a single part class to them (*i.e.*, class-agnostic). At first glance, this design choice may seem like a purely simplified version of the original problem or an unavoidable compromise. *However, we argue that it indeed helps improve the model's open-world generalizability.*

Concretely, in training a model to correctly classify *seen* object parts, we implicitly force the model to classify future unseen object parts into the background, suppressing their chances to be detected and segmented. By treating all the seen object parts as a single class, we remove the competition tension among them and in turn encourage the model

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

[1]The need to assign a "seen-class" label to every detected segment discourages the model from detecting segments that correspond to "unseen-class" classes in the first place.

to pay more attention to differentiating "parts" and "non-parts". As a result, unseen parts appear to be more like the test data in conventional supervised learning; the model can more likely detect them. Besides, removing the competition tension also encourages the model to learn the general patterns of parts, which can potentially improve the segmentation quality on unseen parts. In Sec. 4, we empirically demonstrate the effectiveness of class-agnostic training in segmenting parts from unseen objects.

We propose two further improvements towards open-world *class-agnostic* part segmentation. First, **we incorporate into the model a unique semantic cue of parts.** Compared to objects which are usually considered as "entities", *i.e*., things that can exist and appear distinctively and independently, object parts are "compositions", located within an object and often appearing together in a functionally meaningful way. We hypothesize that by making models aware of this object-part relationship, the resulting segmentation quality can be improved. To this end, we propose to introduce *class-agnostic object masks* (*e.g*., extracted by an off-the-shelf segmentation model) as an additional channel to the model. While extremely simple, we found this approach highly effective, leading to notable gains, especially on unseen objects. Moreover, it is model-agnostic and can easily be incorporated into any network architecture.

Second, **we propose a novel way to fine-tune the model using unlabeled data**, *e.g*., data it sees in its deployed environment, which may include unseen objects. We found that on unseen objects, pixel-wise features the model internally extracts often reveal high-quality segment boundaries. To take advantage of this, we propose a self-supervised approach to adapt the model backbone, which iterates between online pixel clustering (*e.g*., using k-means) and supervised contrastive learning using the cluster assignment. Concretely, we update the model backbone to pull pixels of the same clusters closer; push pixels between different clusters farther away. As will be demonstrated in Sec. 4, this approach leads to a consistent gain on unseen objects and can be further improved via a combination with self-training [2, 23]. Please see Fig. 1 for an illustration.

We validate our proposed approach, which we name Open Part Segmenter (OPS), on two part segmentation datasets, PartImageNet [14] and Pascal-Part [5]. We train the model on PartImageNet, and evaluate it on a PartImageNet out-of-distribution set and Pascal-Part: to our knowledge, we are the first to conduct a cross-dataset study for part segmentation. Data in these two sets contain a variety of unseen objects, and we use class-agnostic Average Precision (AP) as the metric. We show that OPS achieves significant and consistent gains against the baselines. On PartImageNet, we improve the AP from 38.21 to 42.61; on Pascal-Part, we improve from 9.48 to 23.02, almost a 142.8% relative gain. Importantly, all our proposed components — *class-agnostic segmentation, object mask channel, and self-supervised fine-tuning* — contribute to the gain. Moreover, if given ground-truth object masks (*e.g*., form a user in an interactive setting), OPS can encouragingly improve the AP to 85.12 on PartImageNet and 25.26 on Pascal-Part, making it a highly flexible approach. Our analyses further reveal cases that OPS can segment even finer-grained parts than the ground truths, essentially a critical step towards open-world part segmentation.

2. Related Work

Image segmentation aims to divide pixels into semantically meaningful groups. Semantic segmentation [4, 18, 47, 49] tackles this problem by classifying each pixel into a pre-defined semantic class, ignoring the fact that pixels of the same classes may belong to different instances. Instance segmentation [3, 7, 15, 24], in contrast, aims to group pixels by semantic instances (*e.g*., different bird individuals), and at the same time assigns each instance a class label. Panoptic segmentation [21, 29] further combines instance segmentation of "things" (*e.g*., cars, birds) and semantic segmentation of "stuffs" (*e.g*., sky, grass) into a unified problem.

Most image segmentation works focus on "objects" as the basic class labels [15]. Relatively fewer works treat "parts" as the basic class labels [5, 14, 31, 41, 43, 44]. In these works, part segmentation is mostly solved as a semantic segmentation problem, despite the fact that an object may contain multiple part instances of the same class (*e.g*., wheels of a truck). In applications like image editing and object grasping, it is often more important to localize part instances that users/robots can directly interact with. In this paper, we therefore focus on part instance segmentation.

Since parts are essentially "compositions" of objects, there exists a natural (hierarchical) relationship between them. Several works have attempted to leverage this information, mostly in solving hierarchical segmentation over objects and parts together and requiring specifically developed model architectures [8, 27]. In our work, we focus on parts, assuming that we can obtain object-level masks from an off-the-shelf detector. Without developing a new model architecture, we propose two fairly simple, model-agnostic ways to leverage the object-part relationship. Our approach is particularly suitable for an interactive task like image editing since the model can directly take input from users to generate "targeted" part segmentation results.

Open-world recognition. The visual recognition community has gradually moved from recognizing a pre-defined set of classes to tackling classes that are not seen in the training phases (*e.g*., rare animals, newly produced products, etc.), *e.g*., zero-shot learning [13, 39, 45]. Many recent works in an open-world setting aim to "classify" these unseen classes, for example, by exploiting external knowledge

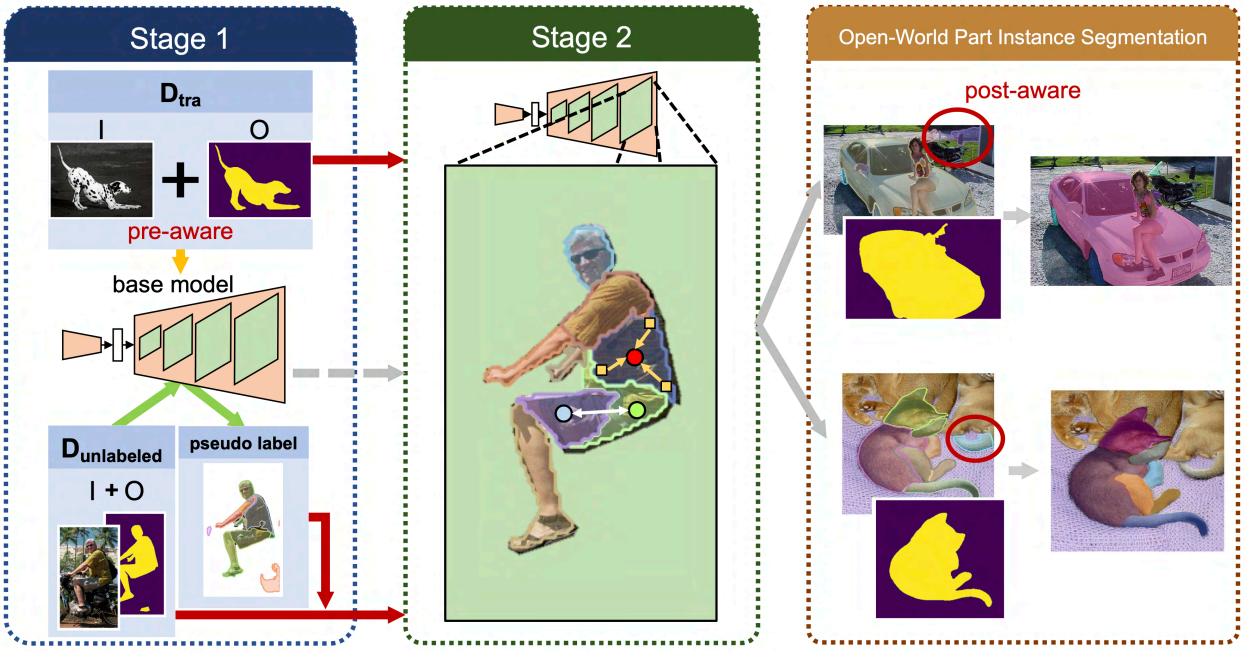

Figure 1. **Illustration of our pipeline.** We claim that class-agnostic training is the key to the open-world part instance segmentation. In stage 1, we train the model with labeled data treating all part categories as a single class. We then learn with the unlabeled data in stage 2 with our novel fine-tuning approach: (1) self-training (ST): we use pseudo labels generated for unlabeled data by the base model and jointly learn with (2) self-supervised (SS): we discover pseudo parts on the feature map and learn the affinity within a part and contrast between different parts. We also propose object-aware learning including pre-aware and post-aware, which improves the segmentation quality for parts, especially of unseen objects.

such as word vectors [34], CLIP embedding [38], etc.

Another approach, which we follow, is to "localize" these unseen items while ignoring the need to classify them [9, 26, 37]. This approach is particularly useful in image editing, manipulation, etc. The closest work to ours is [37], but there are distinct differences. First, they focused on objects while we focus on parts. Second, we propose unique improvements for part segmentation, building upon the insights that parts are "compositions", not "entities", while they focus on object "entities". Further, we propose a novel way to fine-tune our model with unlabeled data.

Semi-supervised learning, domain adaptation, and test-time training. Our fine-tuning approach on unseen objects is reminiscent of the tasks of semi-supervised learning [17, 20, 35, 36], domain adaptation [25, 28, 32, 33, 46, 48, 51], and test-time training [19, 50]. However, our fundamental problem setup is different: they mostly assume the unlabeled and labeled data share the same label space; here, we consider the case that our unlabeled data may contain unseen objects or parts. Nevertheless, we demonstrate that self-training [2, 23, 51], a strong approach in semi-supervised learning and domain adaptation performs favorably in our problem. On top of it, we further propose a novel method taking insights from the detailed investigation of intermediate features within the model for self-supervision on unlabeled data.

3. Approach

3.1. Class-Agnostic Part Segmentation

We consider part instance segmentation. We assume that every image contains one salient object and the goal is to segment that object into parts. That is, for each image, $I \in \mathbb{R}^{H \times W \times 3}$, we want to identify and segment the parts of the salient object, $\{(y_n, m_n)\}_{n=1}^{N}$, where $m \in \{0,1\}^{H \times W}$ is the binary mask of a part and y is its corresponding part class, assuming the object contains N parts (some of whom may share the same class label). Let us denote by $\mathcal{S}$ the set of object classes seen in the training data D_{tr}.

In a close-world setting, images in the test data D_{te} are expected to only contain objects that belong to $\mathcal{S}$. In the open-world setting we target, this constraint is removed, allowing the test data to contain objects unseen in D_{tr}. We denote the set of those unseen object classes by $\mathcal{U}$.

Since the test data D_{te} may contain unseen objects, their corresponding part labels are, by default, not fully covered by D_{tr}. To address this problem, we define the open-world part segmentation problem in a class-agnostic way. That is, we treat all the parts in the test data D_{te} as a single class $y_n = 1$. The goal is therefore to localize and segment object parts without assigning them class labels.

Class-aware v.s. class-agnostic training. Even though we define the evaluation protocol to be class-agnostic, it may

still be beneficial to incorporate part classes during training. To investigate the underlying effect, we compare training a model in a class-agnostic way (*i.e.*, $y_n = 1$) and a class-specific way (*i.e.*, using the original y_n in D_{tr}). In evaluation, we simply replace the part class labels predicted by the class-specific model with 1.

We leave the detailed experimental setup and results in Sec. 4. In short, in an open-world setting, we observe a consistent improvement by training the model in a class-agnostic way. Specifically, class-agnostic training enables the model to localize unseen parts with a higher recall and segment them more accurately.

In the rest of this paper, we will then employ class-agnostic training to learn the part segmentation model.

3.2. Object-Aware Learning and Inference

Object parts, by definition, are the "compositions" of the corresponding object. That is, the extent of them should not go beyond the extent of the object; object parts belonging to the same object instance should appear closely and be detected within the same object mask. While a conventional instance segmentation model may learn such prior knowledge from the training data, we argue that directly incorporating it into the model's prediction is a more straightforward way to take advantage of such a strong cue.

In this paper, we propose two simple yet highly effective ways to incorporate object masks into part segmentation.

Post-aware. The first way is post-processing. That is, after the model already outputs part segments, we directly remove all the portions outside the given object mask.

Pre-aware. We argue that the object-awareness should also be proactively included before making predictions. Instead of post-processing afterward, we hypothesize that the model can learn to incorporate the object-part relationship to directly generate higher-quality part segments. More specifically, we incorporate the class-agnostic object mask as an additional channel to the input image I, *i.e.* appending I with $O \in \{0,1\}^{H \times W}$ to become $I' \in \mathbb{R}^{H \times W \times 4}$.

Another potential benefit from this approach is that it makes predicting the background also object-aware, which could potentially improve the recall of part segmentation, especially for unseen objects. Normally, a model predicts background by $P(y = BG|I)$. When the training data are not labeled comprehensively and contain unlabeled objects, the model is forced to predict them as background, making its prediction more conservative. By incorporating the object mask as an input, the model is actually learning $P(y = BG|I, O)$ for each pixel. That is, it can rely on the input O to determine the background region and focuses more on how to segment parts within O.

Sources of object masks. There are multiple ways to obtain object masks in practice. Thanks to the large-scale datasets for object-level instance segmentation, such as MSCOCO [30], LVIS [12], and OpenImages [22], we can train an object instance segmentation model to detect more classes than existing part segmentation datasets contain.

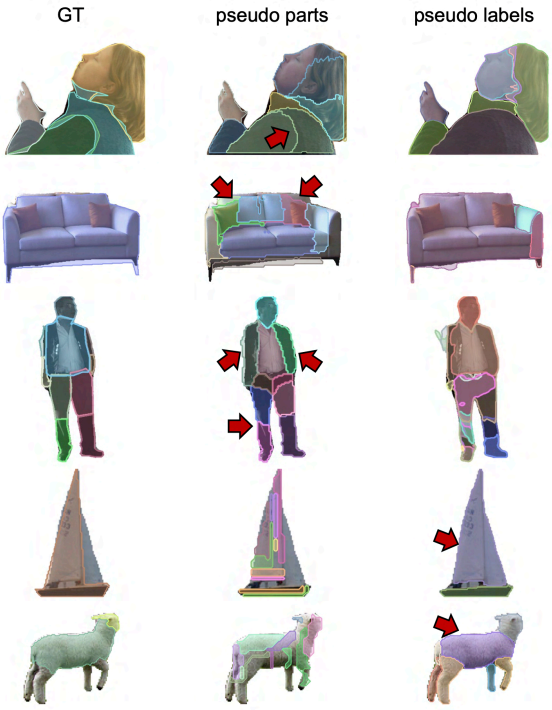

Figure 2. **Comparison between pseudo parts for self-supervised (SS) and pseudo labels for self-training (ST) of our fine-tuning approach.** In this figure we use red arrow to mark better parts. When pseudo labels are bad (row 3), we find that pseudo parts on the feature map can actually reveal good ones. They are complementary to each other in OPS.

Moreover, considering the application of image editing in an *interactive* environment, users can always define the object region of interest and even refine its boundary. All of the above suggests that the object-awareness is the feasible solution for part segmentation.

In this paper, we investigate both scenarios. We call the object masks obtained by an object detector "imperfect" masks. We assume that the users are able to provide accurate (almost "perfect") object masks and simulate such a scenario by using the "ground-truth" masks provided by the dataset. In short summary, we obtain huge improvement with both. We leave more details in Sec. 4.

3.3. Learning with Unlabeled Data

To achieve the goal of part instance segmentation in the wild, learning with unlabeled data is a more promising route than waiting for more data to be human-annotated. In this section, we investigate and present approaches to improve the part segmentation model on unlabeled data. In practice, these unlabeled data may come from existing object-level datasets (but without the part annotations), from the web,

or even from the deployed environment of the model. In our experiments, we focus on the last scenario. We assume that we have access to the images of the test data D_{te} but without any labels.

The first way we investigate is self-training (ST). That is, we generate the pseudo labels on D_{te}, using the model trained on the labeled training set D_{tr}. We then fine-tune the model using the pseudo-labeled data from D_{te} and labeled data from D_{tr} jointly. In our experiments, we observe notable gains with this approach, especially on unseen objects. However, we also find that the pseudo labels in some cases are already inferior, for example, the human in the third row of Fig. 2. Fine-tuning with such pseudo-labels can hardly improve these cases or sometimes even have negative effects on the model.

Pseudo parts by pixel-level clustering. We dig deeper into those inferior cases. We found that the majority of errors come from either under-segmentation, *e.g.*, the model segments the whole object, or boundary mis-localization. We surmise that the model itself just cannot recognize the true segment boundaries for these unseen objects.

To verify this hypothesis, we perform k-means clustering on the feature map. Specifically, we look at the feature map right before the model's prediction, which has the pixel-level resolution, and treat the feature vector on each pixel location as a data point.

To our surprise, for those inferior cases, k-means often clusters pixels into object parts more accurately than the model's prediction. In some cases, the discovered parts are even finer-grained than the ground truths, for example, the sofa and boat in Fig. 2. In other words, by comparing pixels to pixels, the features already capture the *relative* discriminative information among pixels to group them into parts. Such information, however, may not be strong enough in an *absolute* sense to trigger the model to produce parts.

With this evidence, we propose a novel self-supervised (SS) fine-tuning approach to strengthen the discriminative information among pixels. Given an unlabeled image, our approach starts with k-means on top of the feature map, followed by a supervised-contrastive-style loss that pulls features of the same cluster closer and pushes features among different clusters farther away.

Contrast between parts. Intuitively we want the features from different clusters to be dissimilar. We use a centroid to represent each cluster and maximize the pair-wise distance between centroids. Formally,

$$\mathcal{L}_c = \frac{1}{K}\sum_{i,j}^{K} \exp\left(\frac{-||c^i - c^j||_2^2}{\tau_c}\right), \tag{1}$$

where c^i and c^j is the i-th and j-th centroid from the K clusters; τ_c is the temperature.

Affinity within parts. In addition, we want the features within the same cluster to be similar with the following affinitative loss,

$$\mathcal{L}_a = \frac{1}{N}\frac{1}{K}\sum_{i}^{K}\sum_{m}^{N} \exp\left(\frac{||c^i - p_m^i||_2^2}{\tau_a}\right), \tag{2}$$

where p_m^i is the m-th pixel feature within i-th cluster; τ_a is the temperature.

Overall loss. The overall loss for self-supervised fine-tuning is a combination of contrastive and affinitative loss: $\mathcal{L}_{SS} = \lambda_c\mathcal{L}_c + \lambda_a\mathcal{L}_a$, where λ_c and λ_a are the weights for balancing two losses.

3.4. Open Part Segmenter

Finally, we combine our proposed methods in Sec. 3.1, Sec. 3.2, and Sec. 3.3 into a single pipeline named Open Part Segmenter (OPS), aiming to achieve open-world part instance segmentation (cf. Figure 1). We first train the base part segmentation model in a class-agnostic way on labeled data D_{tr}, with object masks. We then perform inference with the base model on the unlabeled data D_{te} to obtain the pseudo labels (predicted by the model directly) and pseudo parts (via clustering on the features). We then fine-tune the base model with (i) the contrastive and affinitative losses and (ii) the supervised loss using the (pseudo) labels on both D_{tr} and D_{te}. The proposed pipeline is simple but effective, showing dramatic improvements and strong generalizability to unseen objects and parts. See Sec. 4 for details.

We note that in theory we can perform more rounds of self-training to improve the results and we investigate such ideas in the supplementary material. We see consistent improvement but eventually, we see a diminishing return.

4. Experiments

4.1. Setup

This paper aims to achieve open-world part instance segmentation. To validate the concept of our problem and the effectiveness of our approaches, we use the following datasets. We assume that we have a base labeled dataset and access to unlabeled data in the model's deployed environment for training.

PartImageNet [14]. This dataset contains 16540 / 2957 / 4598 images and 109 / 19 / 30 object categories in train / val / test split which provide high-quality part segments. The dataset already considers the out-of-distribution (OOD) setting so the object categories in these three subsets are non-overlapped. We also hold out a subset of data from train with a similar size as val to tune the hyper-parameters.

Pascal-Part [5]. To further extend our study to a more realistic open-world setting for part segmentation, we also use Pascal-Part for cross-dataset evaluation. This dataset is a multi-object multi-part benchmark that contains 4998 /

5105 images for train / val, covering 20 object categories. Following Singh *et al*. [43], we parse the annotations into Pascal-Part-58/108, which the number here indicates the number of part categories. To serve our purpose, we convert the labels to instance annotations and crop the objects from scene images to become single-object multi-part as PartImageNet [14].

Evaluation. We adopt standard evaluation metrics for instance segmentation, average precision AP and AP_{50} [30]. To measure the segmentation quality of predicted parts on OOD and different datasets, we evaluate them in a class-agnostic way. In other words, we treat all the parts as a single class during evaluation regardless of whether the model was trained with or without part classes.

4.2. Implementation

Part Segmentation. We apply Mask2Former [6] for our part instance segmentation model. The backbone is ResNet-50 [16] pre-trained on ImageNet-1K [40]. The input image size is 512 and by default, it applies large-scale jittering [11] with the minimum and maximum scale of 0.1 and 2.0 respectively. In testing time, the shortest edge of the image is resized to $[400, 666]$. We train the base model on PartImageNet [14] with batch size 16 for 90K iterations. The learning rate starts at 0.0001 and is decreased by 10 at 60K and 80K. All the experiments are conducted on 8 Nvidia A100.

We compare the class-aware and class-agnostic training. The former uses the part classes as annotated while the latter treats all part classes as a single class. Both models are evaluated in a class-agnostic way for a fair comparison.

Object mask for object-awareness. We demonstrate that the part predictions should be object-aware and investigate two approaches: post-aware and pre-aware. This first approach is a post-processing method that removes all predictions that are outside the object masks. In the second approach, we concatenate the object mask as an additional channel to the input image and hypothesize the model can learn such a relationship between parts and objects.

As mentioned in Sec. 3.2, we have multiple ways to access the object masks. In the following experiments, we obtain the imperfect object masks by taking the trained model on MSCOCO [30] from [6] and performing inference on our data. We also show upper-bound results of using perfect object masks.

Learning with unlabeled data. We fine-tune the base model trained on PartImageNet [14] train data with our novel self-supervised (SS) + self-training (ST) approach on unlabeled data. We consider two settings: (1) fine-tuning on PartImageNet val and evaluate on test and (2) fine-tuning on Pascal-Part [5] train and evaluate on Pascal-Part val. Both measure the generalizability to the OOD data and the second is even cross-set.

Table 1. **Comparison between class-agnostic and class-aware training.** We first study the impact of two on PartImageNet [14] val set to validate our problem setting and approach. Class-agnostic consistently performs better than class-aware, with (imperfect, perfect) or without (none) object masks.

	none		imperf.		perf.	
class	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
agnostic	40.01	70.38	41.94	73.17	85.88	96.08
aware	39.71	69.99	41.45	72.97	84.74	95.50

Each training batch of 32 is an even mix of labeled/pseudo-labeled data and unlabeled pseudo parts data. For pseudo labels in ST, we use predictions from the base model that have confidence scores larger than 0.1. For pseudo parts in SS, we use $K = 10$ for online K-Means on normalized features $\mathcal{F}'$. We use 1 for τ_c and τ_a, 10 for λ_c and 0.5 for λ_a. We fine-tune the model for 30K and 10K iterations with learning rate $1e{-}6$ with imperfect and perfect object masks respectively.

4.3. Main Results on PartImageNet

In this section, we show our results on PartImageNet [14] val and test set. Both sets are OOD from the train set. Please see more information in Sec. 4.1. We follow the rationale of our proposed methods step by step to provide a comprehensive study with the empirical results on the open-world part instance segmentation problem.

Class-aware v.s. class-agnostic. The former trains the base model with part classes while the latter treats all the part classes as a single class. In order to investigate the underlying impact of class labels on the quality of part instance segmentation, we evaluate both approaches in a class-agnostic way. At first glance, we may think the model can leverage more information from part categories in class-aware training to refine the predicted part masks accordingly. However, we find that the class-agnostic training obtains comparable or slightly better performance. As shown in Tab. 1, 40.01 from class-agnostic is already better than 39.71 from class-aware in the plain setting without any other proposed method. The observation still holds when we further include the object masks. This suggests that learning the context alone can already achieve high-quality part segmentation. We hypothesize the model learns more general representation about parts and thus performs well on OOD data. It encourages our proposed method toward the open world. In all of the following experiments, we will directly use class-agnostic training unless stated otherwise.

Object-aware part segmentation. We propose post-aware and pre-aware for object-awareness. While the former uses object masks for post-processing to filter out unreasonable part predictions, the latter includes them as a cue for training. In Tab. 2, we show that both approaches outperform the base model that has no object-awareness (*i.e*. AP 40.01).

Table 2. **Results on post-aware and pre-aware object masks on PartImageNet [14] val and Pascal-Part-58 [5, 43].** Both approaches outperform the base models that have no object-awareness (none).

	PartImageNet val		Pascal-Part-58	
object mask	AP	AP_{50}	AP	AP_{50}
none	40.01	70.38	9.48	19.90
+ post imperf.	42.48	71.03	12.44	24.40
+ post perf.	47.10	75.25	13.06	24.97
pre imperf.	41.94	73.17	20.27	44.24
+ post imperf.	45.40	74.45	23.02	47.21
pre perf.	85.88	96.08	25.24	45.62
+ post perf.	87.61	96.19	25.26	45.67

It proves the effectiveness of our proposed methods. Furthermore, the two approaches are complementary to each other. Simply combining them obtains the further gain, *i.e.* AP 41.94 to 45.40 with imperfect masks and 85.88 to 87.61 with perfect masks.

Here we also see a big jump from not using object masks (AP 40.01) to using perfect ones (AP 87.61) in Tab. 2. This is an essential finding: considering image editing in an interactive environment, a user can always refine the object mask until it is satisfying. This encourages the application of our work to the real world.

Besides the above improvements, our proposed methods are simple and straightforward. It can be easily plugged into most existing algorithms and architectures.

Learning with unlabeled data. In this section, we demonstrate the effectiveness of our novel fine-tuning approaches, namely self-training (ST) and self-supervised (SS) learning. We fine-tune the base model with PartImageNet [14] val and evaluate on the test set. As both val and test are OOD from train, we aim to investigate the generalizability to unseen objects and parts in terms of segmentation quality.

In Tab. 3, the performance of the base model on val set is AP 41.94 with imperfect masks. Fine-tuning on val set with SS and ST alone can improve to 42.78 and 43.12 respectively. It shows the effectiveness of each individual component. By combining them, our proposed OPS model can get even higher performance.

Fine-tuning and improving on val set assumes we have access to the unlabeled data without annotations and we can leverage them in an unsupervised learning way. Here we still have another test set OOD from both train and val which is not included in the fine-tuning. In Tab. 3, we also see notable gains with all approaches (AP 40.17, 40.38, and 40.43) compared to the base model (AP 38.96). It explains that the model learns more generalizability to unseen parts and objects with fine-tuning only on val set. The proposed fine-tuning approach is an important step toward open-world part instance segmentation.

Table 3. **Results on PartImageNet [14]** We fine-tune the base model on OOD val set with proposed self-supervised (SS) and self-training (ST), with imperfect and perfect object masks. Both outperform the base model.

			Val		Test	
method	SS	ST	AP	AP_{50}	AP	AP_{50}
imperf.						
base			41.94	73.17	38.96	69.07
	✓		42.78	74.62	40.17	70.70
		✓	43.12	75.03	40.38	71.10
OPS	✓	✓	43.16	74.96	40.43	71.18
perf.						
base			85.88	96.08	83.52	94.66
	✓		86.09	96.35	83.81	94.94
		✓	86.28	96.37	83.97	95.12
OPS	✓	✓	86.19	96.43	83.86	95.05

4.4. Main Results on Pascal-Part

In Sec. 4.3, we show notable improvements on PartImageNet [14]. Here, we further extend our study to a cross-dataset setting. To this end, we use the same base model trained on PartImageNet train set, and we measure the generalizability of the model to the new test data in Pascal-Part [5]. We evaluate on two sets of ground-truth annotations, Pascal-Part-58 and Pascal-Part-108. We parse the annotations following [43] (see Sec. 4.1 for more information). To the best of our knowledge, we are the first to investigate cross-dataset scenarios in part instance segmentation.

Object-aware part segmentation. We conduct the same empirical study as in Sec. 4.3 on Pascal-Part-58. As shown in Tab. 2, the base model performs merely AP 9.48 and improves to 13.06 even after proposed post-processing with perfect mask (base + post perf.). With pre-aware imperfect masks, the performance (AP 20.27) is more than twice as the base model and also outperforms base + post perf. by a large margin. We observe similar results with perfect object masks. Furthermore, the gain is greater than what we have seen in Sec. 4.3. This suggests the effectiveness of the pre-aware approach, which can recognize more high-quality part segments, especially for data with a larger domain gap.

Learning with unlabeled data. In this section, we follow the same route as in Sec. 4.3 but fine-tune the base model using unlabeled data in the Pascal-Part train set. As shown in Tab. 4, on Pascal-Part-58, SS improves the base model from AP 20.27 and 25.24 to 20.53 and 27.13 with imperfect and perfect object masks respectively. With ST, the APs are further boosted to 24.02 and 27.69. These relative improvements of 18.4% and 9.7% are much greater than 2.9% and 0.3% in Tab. 3 on PartImageNet.

On the more fine-grained and challenging Pascal-Part-108, we also observe consistent improvements, *i.e.* relative improvements of 11.43% and 22.38% to the base model.

Figure 3. **Visualizations of part segmentation on Pascal-Part-58 [5, 43].** The first row is the ground-truth and the second is predictions from our proposed OPS model. Our model only uses part segmentation ground-truth from PartImageNet [14]. These high-quality part predictions demonstrate the feasibility of part segmentation in the wild and the effectiveness of our proposed approach.

Table 4. **Results on Pascal-Part-58 and Pascal-Part-108 [5, 43].** We fine-tune the base model on train set with the proposed self-supervised (SS) and self-training (ST), without using any additional ground-truth annotations. We show the performance with imperfect and perfect object masks. All the components in OPS consistently outperform the base model.

			Part-58 val		Part-108 val	
method	SS	ST	AP	AP_{50}	AP	AP_{50}
imperf.						
base			20.27	44.24	16.36	30.16
	✓		20.53	44.91	17.95	33.03
		✓	23.25	48.81	17.75	32.64
OPS	✓	✓	24.02	50.10	18.23	33.72
perf.						
base			25.24	45.62	13.40	29.32
	✓		27.13	49.08	13.75	29.76
		✓	27.23	48.58	15.85	33.66
OPS	✓	✓	27.69	49.75	16.40	34.60

Qualitative results. In Fig. 3, we show visualizations of part segmentation using our proposed OPS model on Pascal-Part-58 [5, 43]. Though our model is only trained using ground-truth annotations from PartImageNet [14] in a class-agnostic way, it generates high-quality part segments which align very well with GT.

In short, we demonstrate that our approaches are applicable to not only PartImageNet, OOD but closer domain, but also even more challenging cross-dataset setting. With our approach of learning with unlabeled data, OPS is able to learn better representation for more general parts, which results in superior performance on unseen objects and parts from different data domains.

Table 5. **Comparison to baselines on PartImageNet test.** We adopt standard metrics for semantic segmentation to compare with other baselines. OPS outperforms them by a large margin.

	imperfect obj. mask			perfect obj. mask		
	mIOU	fwIoU	mACC	mIOU	fwIoU	mACC
SLIC	38.15	65.10	75.61	41.19	68.09	78.12
NC	40.80	76.87	54.99	43.11	77.44	55.17
Fel.	47.97	83.82	67.70	59.73	88.90	76.10
OPS	**64.71**	**89.41**	**82.61**	**91.89**	**97.97**	**95.81**

4.5. Comparison to Other Baselines

Evaluation metric. We adopt the standard evaluation metric AP for part "instance" segmentation in previous sections. In this section, we evaluate with mIOU, fwIoU, and mACC, which are the standard metrics for semantic segmentation, to compare to other baselines. However, directly applying them without classifying parts will simply merge all the parts and cannot reflect the segmentation quality. To resolve this, we consider an oracle scenario, assigning each segmented part the class label from its highest-overlapped "ground-truth" part.

Results. We compare OPS to normalized cut (NC) [42], SLIC [1], and Felzenszwalb (Fel.) [10] as shown in Tab. 5. OPS outperforms them by a large margin in all metrics on PartImageNet test. Our method using imperfect masks even outperforms them when they use perfect masks. We argue OPS successfully captures semantic cues to attain high quality. See the supplementary material for more.

5. Conclusion

In this work, we present OPS, a method for part segmentation in an open-world setting. To be robust to unseen parts, we propose *class-agnostic* and *object-aware* learning. Combining with self-training and clustering on unlabeled data, we achieve state-of-the-art on unseen categories.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34(11):2274–2282, 2012. 8

[2] Hong-You Chen and Wei-Lun Chao. Gradual domain adaptation without indexed intermediate domains. *NeurIPS*, 34:8201–8214, 2021. 2, 3

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 2

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 2, 5, 6, 7, 8

[6] Bowen Cheng, Ishan Misra, Alexander Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1, 6

[7] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. Sparse instance activation for real-time instance segmentation. In *CVPR*, 2022. 2

[8] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *CVPR*, 2021. 2

[9] Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *CVPR*, 2019. 3

[10] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59:167–181, 2004. 8

[11] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 6

[12] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1, 4

[13] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *CVPR*, 2021. 2

[14] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. *arXiv*, 2021. 1, 2, 5, 6, 7, 8

[15] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask rcnn. In *CVPR*, 2017. 1, 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[17] Xinyue Huo, Lingxi Xie, Jianzhong He, Zijie Yang, Wengang Zhou, Houqiang Li, and Qi Tian. Atso: Asynchronous teacher-student optimization for semi-supervised image segmentation. In *CVPR*, 2021. 3

[18] Trevor Darrell Jonathan Long*, Evan Shelhamer*. Fully convolutional models for semantic segmentation. In *CVPR*, 2015. 1, 2

[19] Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907, 2021. 3

[20] Hoel Kervadec, Jose Dolz, Éric Granger, and Ismail Ben Ayed. Curriculum semi-supervised segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019. 3

[21] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 2

[22] Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Dhyanesh Narayanan, Chen Sun, Gal Chechik, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2016. 1, 4

[23] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013. 2, 3

[24] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *CVPR*, 2020. 2

[25] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *ECCV*. Springer, 2020. 3

[26] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *CVPR*, 2017. 3

[27] Xiangtai Li, Shilin Xu, Yibo Yang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Panoptic-partformer: Learning a unified model for panoptic part segmentation. In *ECCV*, 2022. 2

[28] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019. 3

[29] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *CVPR*, 2021. 2

[30] 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*. Springer, 2014. 1, 4, 6

[31] Qing Liu, Adam Kortylewski, Zhishuai Zhang, Zizhang Li, Mengqi Guo, Qihao Liu, Xiaoding Yuan, Jiteng Mu, Weichao Qiu, and Alan Yuille. Cgpart: A part segmentation dataset based on 3d computer graphics models. *arXiv*, 2021. 2

[32] Qing Liu, Adam Kortylewski, Zhishuai Zhang, Zizhang Li, Mengqi Guo, Qihao Liu, Xiaoding Yuan, Jiteng Mu, Weichao Qiu, and Alan Yuille. Learning part segmentation

through unsupervised domain adaptation from synthetic vehicles. In *CVPR*, 2022. 3

[33] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, 2021. 3

[34] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv*, 2013. 3

[35] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020. 3

[36] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015. 3

[37] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Zhe Lin, Philip Torr, and Jiaya Jia. Open-world entity segmentation. *arXiv*, 2021. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 3

[39] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*. PMLR, 2015. 2

[40] 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. 6

[41] Oindrila Saha, Zezhou Cheng, and Subhransu Maji. Improving few-shot part segmentation using coarse supervision. In *ECCV*, 2022. 2

[42] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *TPAMI*, 22(8):888–905, 2000. 8

[43] Rishubh Singh, Pranav Gupta, Pradeep Shenoy, and Ravi Sarvadevabhatla. Float: Factorized learning of object attributes for improved multi-object multi-part scene parsing. In *CVPR*, 2022. 2, 6, 7, 8

[44] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *CVPR*, 2021. 2

[45] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 2

[46] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *NeurIPS*, 32, 2019. 3

[47] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[48] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. *NeurIPS*, 32, 2019. 3

[49] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *CVPR*, 2022. 2

[50] Wentao Zhu, Yufang Huang, Daguang Xu, Zhen Qian, Wei Fan, and Xiaohui Xie. Test-time training for deformable multi-scale image registration. In *ICRA*. IEEE, 2021. 3

[51] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 3