OV-PARTS: Towards Open-Vocabulary Part Segmentation
(Supplementary Material)

Coauthor
Affiliation
Address
email

The supplementary material is organized as follows:

- Implementation Details.(Sec. A)
- Details of Benchmark Datasets: Pascal-Part-116 and ADE20K-Part-234 (Sec. B).
- Qualitative Results of Three Benchmark Tasks (Sec. C).
- Future Works and Negative societal Impacts (Sec. D).

A Implementation Details

Two-Stage Baselines. Except for the Object Mask Prompt and Compositional Prompt Tuning designs, we follow the default architecture in the original ZSseg paper. The number of part queries is set to 50. All the two-stage baselines are trained with AdamW optimizer with the initial learning rate of 1e-4 and weight decay of 1e-4. A poly learning rate policy with a power of 0.9 is adopted. The total batch size is 16 and the total training iteration is 20k. For the training of Compositional Prompt Tuning, a SGD optimizer with the initial learning rate of 2e-2 and weight decay of 5e-4 is used. And we adopt a warm-up cosine learning rate policy with 100 warm-up iterations. The total batch size is 32 and the total training iteration is 3k. We sample 128 training samples for each object part class. The length of the learnable object and part prompt tokens are 4. The object tokens are initialized from the text embedding of the template “a photo of”. The initial value of the learnable fusion weight is 0.5.

One-Stage Baselines. We adopt the original architecture of both CATSeg and CLIPSeg as described in their respective papers. For finetuning CATSeg, we utilize their pretrained model with a ResNet-101 backbone. However, while CATSeg achieves the best performance by finetuning the attention layers of CLIP’s visual encoder in open vocabulary object segmentation, we observe poor performance with the same finetuning strategy in OV-PARTS. In our experiments, we only finetune the backbone with a backbone multiplier of 0.1 and the swin transformer based decoder. We employ an AdamW optimizer with an initial learning rate of 2e-4, weight decay of 1e-4, and a cosine learning rate policy. The total batch size is 8, and the training iterations amount to 40k. For CLIPSeg, we utilize the same optimizer settings and learning rate policy as CATSeg. The training iterations are set to 20k for the zero-shot/cross-dataset part segmentation setting and 3k for the few-shot part segmentation setting. *More technical details about CLIPSeg.* CLIPSeg adds a parameter-efficient three-layered transformer decoder to the original CLIP model for segmentation. It integrates visual features from the final layer of the visual encoder and text features of all object part prompts from the text encoder through the FiLM module, forming cross-modal input token embeddings for the decoder. Furthermore, the features extracted from the 3rd, 6th, and 9th layers of CLIP’s visual encoder are projected and added to the intermediate features of the corresponding decoder layers. It’s worth noting that, the visual

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

Table A.1: Model complexity analysis on Pascal-Part-116. ZSseg+ 1/2st is first/second stage.

Model	Backbone	Image Size	Trainable Params (M)	FLOPs (G)
ZSseg+ 1st	ResNet-101c	512 $\times$ 704	61.1	103.9
ZSseg+ 2st	ViT-B/16	384 $\times$ 384	0.004	58.9
CATSeg	ResNet-101 & ViT-B/16	384 $\times$ 384	30.9	139.0
CLIPSeg	ViT-B/16	352 $\times$ 352	1.5	97.9

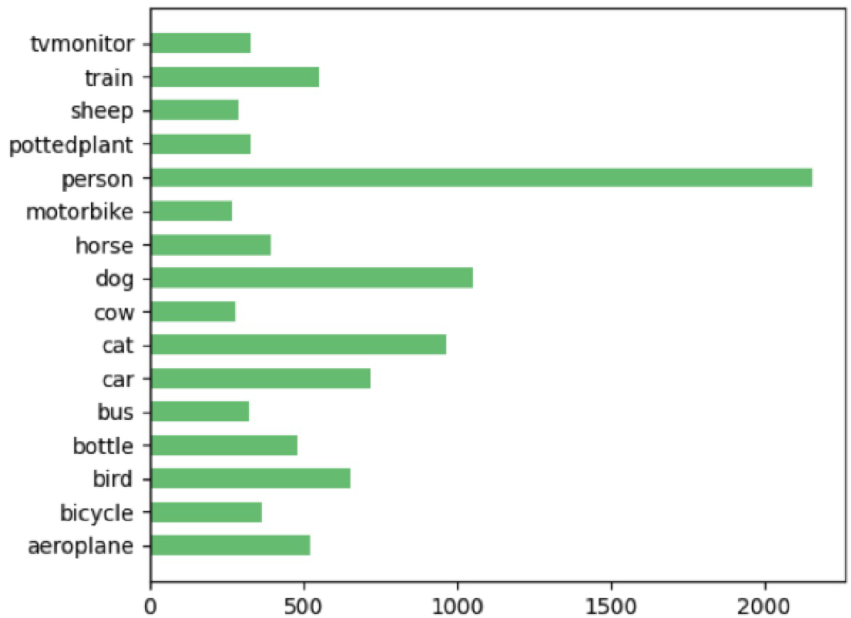

Figure A.1: The number of object masks that have corresponding part masks in Pascal-Part-116.

features extracted from the frozen CLIP visual encoder first pass through the added visual adapter, which consists of a two-layered MLP, before reaching the decoder. Finally, we replace the original binary segmentation head with a multi-class one to output the semantic segmentation map.

Model Complexity. We analyze the computational complexity of these two types of baselines and summarize the number of trainable parameters and FLOPs in Table A.1. The complexity is evaluated on Pascal-Part-116. It's evident that the one-stage CLIPSeg, which solely refines a lightweight transformer decoder and employs parameter-efficient modules, showcases the fewest trainable parameters and the lowest FLOPs. In contrast, the two-stage ZSseg+ approach, involving the training of a complete maskformer with a larger resolution, leads to the highest count of trainable parameters and FLOPs.

B Benchmark Datasets Details

B.1 Pascal-Part-116

Pascal-Part-116 contains 8431 training images and 850 testing images. Compared to the original version of Pascal-Part, we discard the directional indicator for some part classes and merge them to avoid overly intricate part definitions. The category vocabulary and merging details are as follows:

`aeroplane [body, stern, left/right wing, tail, engine, wheel]`

`bicycle [front/back wheel, saddle, handlebar, chainwheel, headlight]`

`bird [left/right wing, tail, head, left/right eye, beak, torso, neck, left/right leg, left/right foot]`

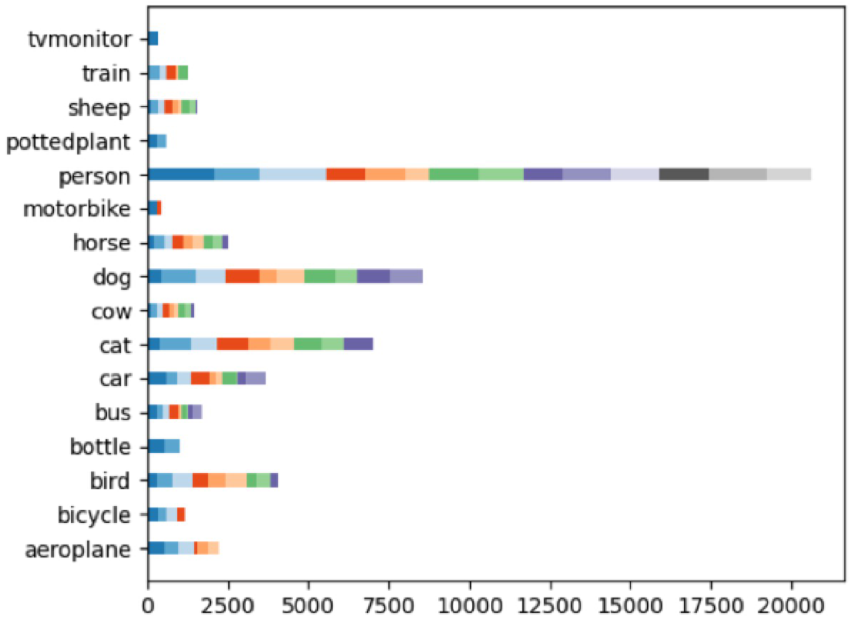

Figure B.2: The number of part masks for each object class in Pascal-Part-116. Each horizontal bar is color-coded to represent a specific part class belonging to the object. The colors of the bars are ordered from left to right based on the part sequence in the list of objects with parts.

bottle [body, cap]

bus [wheel, headlight, front, left/right side, back, roof, left/right mirror, front/back license plate, door, window]

car [wheel, headlight, front, left/right side, back, roof, left/right mirror, front/back license plate, door, window]

cat [tail, head, left/right eye, torso, neck, left-front/right-front/left-back/right-back leg, nose, left-front/right-front/left-back/right-back paw, left/right ear]

cow [tail, head, left/right eye, torso, neck, left-front-upper/left-front-lower/right-front-upper/right-front-lower/left-back-upper/left-back-lower/right-back-upper/right-back-lower leg, left/right ear, muzzle, left/right horn]

dog [tail, head, left/right eye, torso, neck, left-front/right-front/left-back/right-back leg, nose, left-front/right-front/left-back/right-back paw, left/right ear, muzzle]

horse [tail, head, left/right eye, torso, neck, left-front-upper/left-front-lower/right-front-upper/right-front-lower/left-back-upper/left-back-lower/right-back-upper/right-back-lower leg, left/right ear, muzzle, left-front/right-front/left-back/right-back hoof]

motorbike [front/back wheel, saddle, handlebar, headlight]

person [head, left/right eye, torso, neck, left-lower/right-lower/left-upper/right-upper leg, foot, nose, left/right ear, left/right eyebrow, mouth, hair, left/right lower arm, left/right upper arm, left/right hand]

pottedplant [pot, plant]

sheep [tail, head, left/right eye, torso, neck, left-front-upper/left-front-lower/right-front-upper/right-front-lower/left-back-upper/left-back-lower/right-back-upper/right-back-lower leg, left/right ear, muzzle, left/right horn]

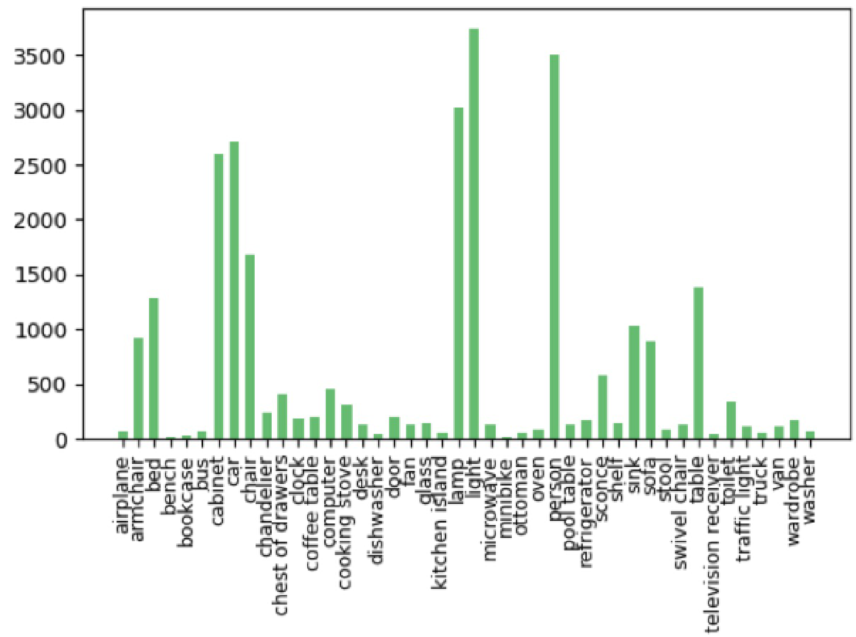

Figure B.3: The statistics for the number of object masks with part masks on ADE20K-Part-234.

`train [headlight, head, front, left/right side, back, roof, coach]`

`tvmonitor [screen]`

The unseen objects are colored blue and the removed terms are colored purple.

B.2 ADE20K-Part-234

The original subset of SceneParse150 comprises 20,210 training images and 2,000 validation images. After filtering out less frequent parts, the subset is reduced to 7,347 training images and 1,016 validation images. In ADE20K, most object parts have sparse mask annotations, and only a subset of object instances have part annotations. Hence, ADE20K-Part-234 provides the instance-level object mask annotations along with their part mask annotations. To maximize the use of labeled data and ensure authentic evaluations, different data splits are designed for the three task settings. (1) Generalized Zero-Shot Part Segmentation: Models are trained on the seen object instances from the 7,347 training images. Testing is performed on both unseen object instances from the same 7,347 training images and all object instances from the 1,016 validation images. (2) Few-Shot Part Segmentation: For each object class, 16 training images are sampled following the approach in Pascal-Part-116. we adapt the validation set from the generalized zero-shot part segmentation setting by removing the images that occur in the sampled 16-shot training set. (3) Cross-Dataset Part Segmentation: The original data split (7347/1016 training/validation images) is used since we mainly test on the Pascal-Part-116 dataset. The annotated objects with their parts are listed as follows:

```
person [arm, back, foot, gaze, hand, head, leg, neck, torso]
door [door frame, handle, knob, panel]
clock [face, frame]
toilet [bowl, cistern, lid]
cabinet [door, drawer, front, shelf, side, skirt, top]
sink [bowl, faucet, pedestal, tap, top]
lamp [arm, base, canopy, column, cord, highlight,light source, shade,tube]
sconce [arm, backplate, highlight, light source, shade]
chair [apron, arm, back, base, leg, seat, seat cushion, skirt, stretcher]
chest of drawers [apron, door, drawer, front, leg]
chandelier [arm, bulb, canopy, chain, cord, highlight, light source, shade]
```

bed [footboard, headboard, leg, side rail]
table [apron, drawer, leg, shelf, top, wheel]
armchair [apron, arm, back, back pillow, leg, seat, seat base,seat cushion]
ottoman [back, leg, seat]
shelf [door, drawer, front, shelf]
swivel chair [back, base, seat, wheel]
fan [blade, canopy, tube]
coffee table [leg, top]
stool [leg, seat]
sofa [arm, back, back pillow, leg, seat base, seat cushion, skirt]
computer [computer case, keyboard, monitor, mouse]
desk [apron, door, drawer, leg, shelf, top]
wardrobe [door, drawer, front, leg, mirror, top]
car [bumper, door, headlight, hood, license plate, logo, mirror, wheel, window, wiper]
bus [bumper, door, headlight, license plate, logo, mirror, wheel, window, wiper]
oven [button panel, door, drawer, top]
cooking stove [burner, button panel, door, drawer, oven, stove]
microwave [button panel, door, front, side, top, window]
refrigerator [button panel, door, drawer, side]
kitchen island [door, drawer, front, side, top]
dishwasher [button panel, handle, skirt]
bookcase [door, drawer, front, side]
television receiver [base, buttons, frame, keys, screen, speaker]
glass [base, bowl, opening, stem]
pool table [bed, leg, pocket]
van [bumper, door, headlight, license plate, logo, mirror, taillight, wheel, window, wiper]
airplane [door, fuselage, landing gear, propeller, stabilizer, turbine engine, wing]
truck [bumper, door, headlight, license plate, logo, mirror, wheel, windshield]
minibike [license plate, mirror, seat, wheel]
washer [button panel, door, front, side]
bench [arm, back, leg, seat]
traffic light [housing, pole]
light [aperture, canopy, diffusor, highlight, light source, shade]

B.3 Data Statistics Analysis.

We report the statistics for the number of object masks that have part annotations in Pascal-Part-116 (see Figure A.1) and ADE20K-Part-234 (see Figure B.3). The total number of part masks for each object and the proportion of each part are shown in Figure B.2 (Pascal-Part-116) and Figure B.4 (ADE20K-Part-234). In Figure B.2, the color sequence from left to right corresponds to the part word sequence as listed in Section B.1. In Figure B.4, the color sequence from bottom to up corresponds to the part word sequence as listed in Section B.2. Additionally, we report the scale distribution for the part masks of each object as shown in Figure B.8.

C Qualitative Results

The qualitative results on the comparison among **ZSseg+**, **CATSeg** and **CLIPSeg** for the challenging case "bird" are shown in Figure B.5.

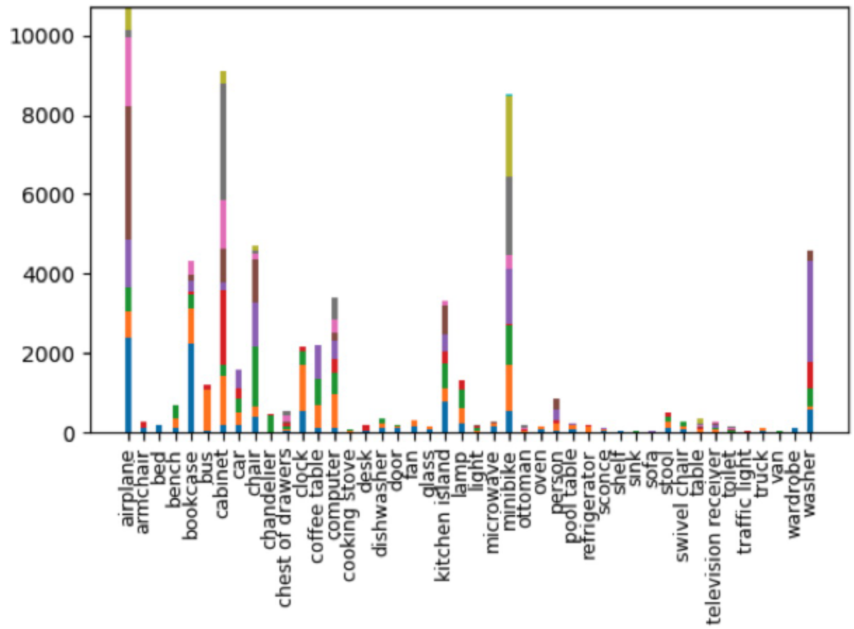

Figure B.4: The number of part masks for each object class in ADE20K-Part-234. Each horizontal bar is color-coded to represent a specific part class belonging to the object. The colors of each bar are ordered from bottom to top according to the part sequence in the list of objects with parts.

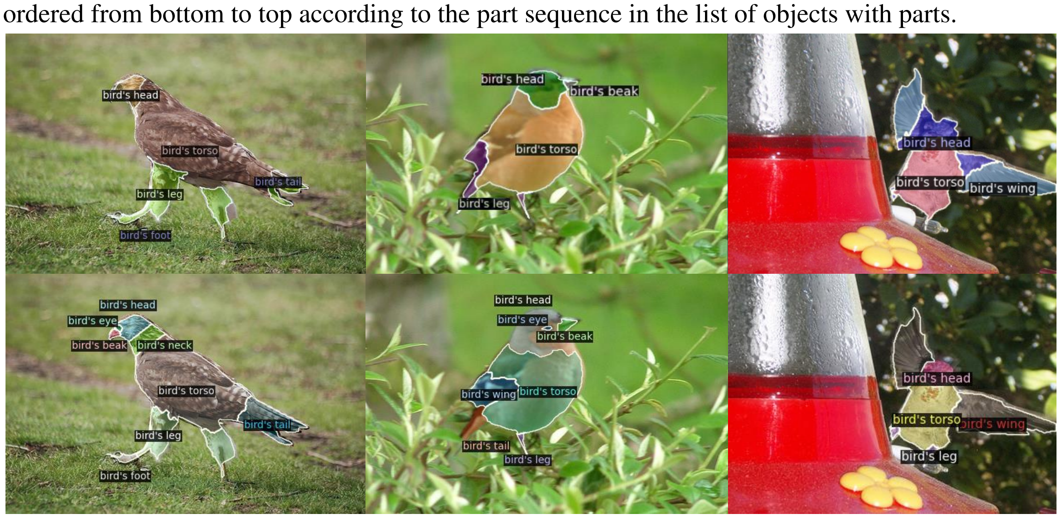

Figure B.5: Qualitaive results on **ZSseg+**, **CATSeg** and **CLIPSeg** concerning the challenging unseen "bird" class in Pascal-Part-116, as shown in the first row. The second row shows the corresponding ground truth. We can observe that CATSeg and CLIPSeg can generalize to the more novel parts: "Bird's Beak" and "Bird's Wing"

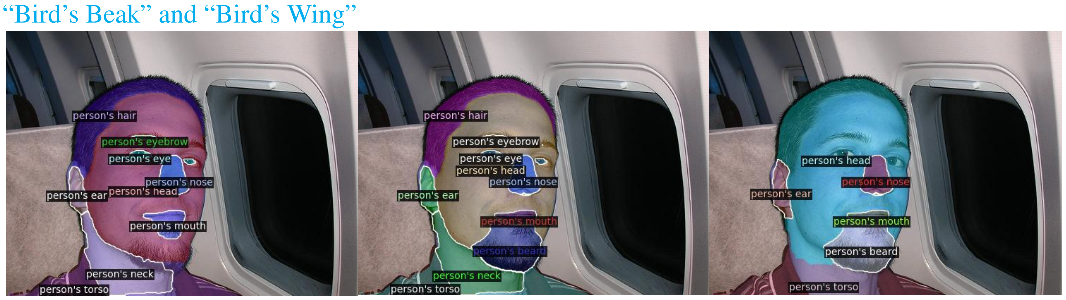

Figure B.6: Qualitative results on CATSeg's multi-granular generalization ability. From the left to the middle image, the model generalizes from "head" to the more fine-grained "beard". From the middle to the right image, the model generalizes from ["hair", "eyebrow", "eye"] to the coarse-grained "head" and also from "neck" to "torso".

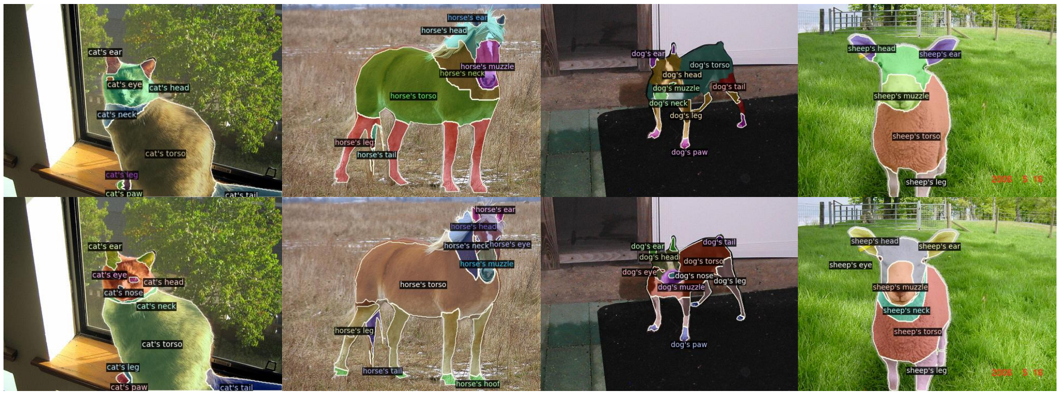

Figure B.7: More qualitative results of generalized zero-shot part segmentation on Pascal-Part-116 are in the first row. The ground truth is in the second row. The seen classes are "cat" and "horse" while the unseen classes are "dog" and "sheep".

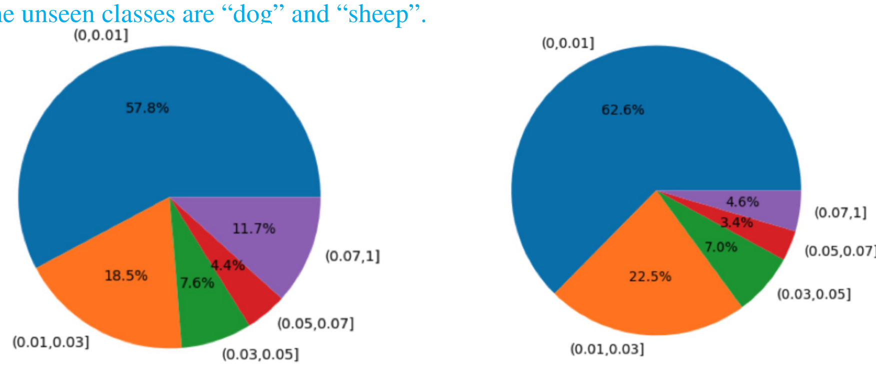

Figure B.8: The scale ratio (number of pixels in the object part mask out of all pixels in an image.) distribution of all part masks of Pascal-Part-116 (Left) and ADE20K-Part-234 (Right).

Figure B.6 shows the multi-granular generalization ability of the one-stage baselines. The adopted model is CATSeg. The visualization sample is from the "person" class in Pascal-Part-116.

We give more qualitative results on Pascal-Part-116 and ADE20K-Part-234 on the three proposed task settings. The adopted model is CLIPSeg with finetuning (VA+L+F+D). The visualization results for the **Generalized Zero-Shot Part Segmentation** on Pascal-Part-116 and ADE20K-Part-234 are shown in Figure B.7 and Figure B.9 respectively. We report the qualitative results for the **Few-Shot Part Segmentation** on Pascal-Part-116 in Figure B.10 and on ADE20K-Part-234 in Figure B.11. And the results for the **Cross-Dataset Part Segmentation** on Pascal-Part-116 are shown in Figure B.12. Furthermore, we present the qualitative results for models trained on Pascal-Part-116 and then tested on ADE20K-Part-234 are shown in Figure B.13.

D Future Works and Negative Societal Impacts

Although part-level OVSS indeed presents more challenges compared to object-level OVSS, the OV-PARTS benchmark datasets have lower quality than existing object-level OVSS benchmark datasets. The original version of Pascal-Part and ADE20K-Part are annotated without considering the open vocabulary scenario especially the analogical reasoning ability and open granularity ability that we care about in a part-level OVSS model. The benchmark datasets need to be continuously expanded and improved to encompass more diverse and complex object-part annotations.

There may be potential negative societal impacts associated with the OV-PART benchmark. The deployment of fine-grained part segmentation models in various real-world applications may lead to unintended consequences. We must ensure that the predictions be reliable and accurate in critical applications, such as medical diagnosis or autonomous vehicles. Also, there is a possibility of

Figure B.9: Qualitative results of generalized zero-shot part segmentation on ADE20K-Part-234. The first and second rows show the generalize from the seen classes [chair, armchair, sofa] to the unseen classes [swivel chair, ottoman, stool]. The third row shows the generalize from the seen classes [lamp, chandelier] to the unseen class [fan]. Notably, "fan's blade" is novel at the object and part level.

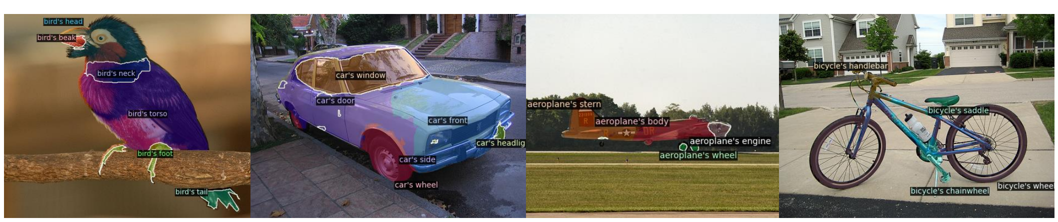

Figure B.10: Qualitative results of few-shot part segmentation on Pascal-Part-116. We display the segmentation map of four classes: "bird", "aeroplane", "car" and "bicycle".

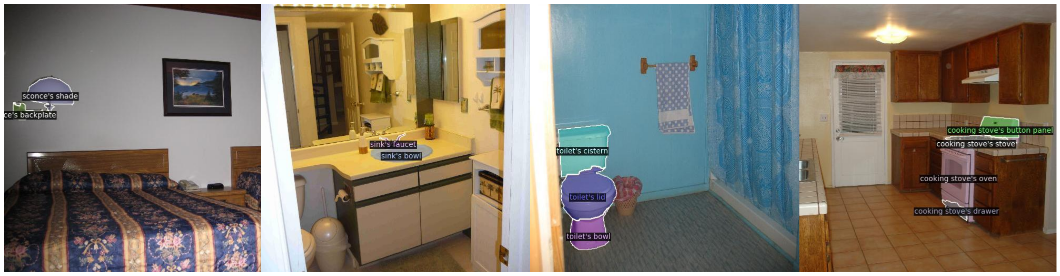

Figure B.11: Qualitative results of few-shot part segmentation on ADE20K-Part-234. We display the segmentation map of four classes: “lamp”, “sink”, “toilet” and “cooking stove”.

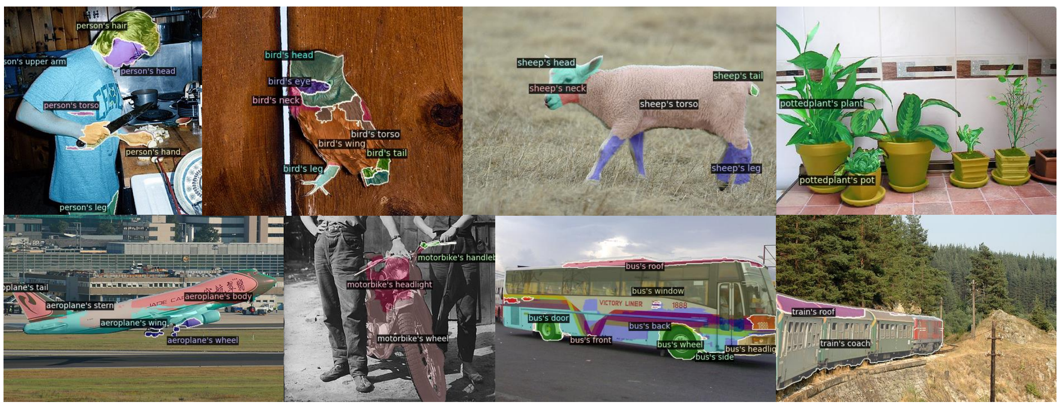

Figure B.12: Qualitative results of cross-dataset part segmentation on Pascal-Part-116. Pascal-Part-116 provides more fine-grained part annotations for the "person" category, such as “hair” and “upper arm”. The model trained on ADE20K-Part-234 demonstrates the ability to recognize “hair” but struggles to generalize from “arm” to “upper arm” and “lower arm” accurately. Moreover, the model exhibits potential in generalizing parts of the “airplane” category. Although ADE20K-Part-234 annotates the parts as “door”, “fuselage”, “landing gear”, “propeller”, “stabilizer”, “turbine engine”, and “wing”, the model can generalize them to Pascal-Part-116’s parts, including “body”, “stern”, “wing”, “tail”, “engine”, and “wheel”, despite the differences in vocabulary and granularity. Notably, ADE20K-Part-234 does not contain related classes to "bird", "sheep", and "potted plant", but the model demonstrates a certain level of generalization ability to segmenting these categories.

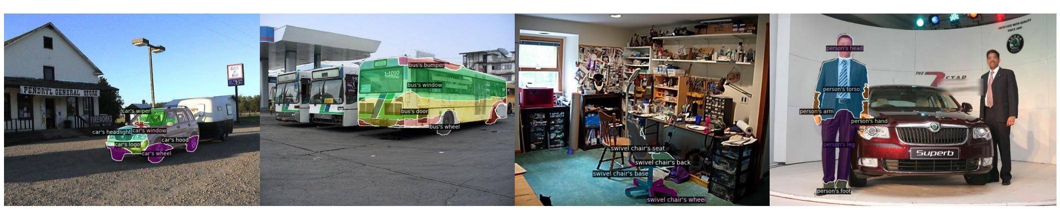

Figure B.13: Qualitative results of cross-dataset part segmentation on ADE20K-Part-234. For the categories “car” and “bus”, the part annotations in Pascal-Part-116 are more coarse-grained. When tested on ADE20K-Part-234, the model trained on Pascal-Part-116 can predict novel parts like “logo”, “wiper”, “hood”, and “bumper”. However, the segments and part labels don’t align accurately. For example, the model still segments the “bus’s roof”, which is annotated in Pascal-Part-116, but wrongly assigns it to “bus’s bumper” in ADE20K-Part-234. This showcases the challenge of generalizing across different granular part definitions. For the novel object “swivel chair”, the model adeptly delineates part boundaries even without relevant objects in Pascal-Part-116. But the category errors are still present. In the case of the “person” category, the model only segments the “upper arm”, which demonstrates the difficulty of generalizing from “upper/lower arm” to “arm”.

misuse of part segmentation technology for malicious purposes, such as creating deepfake images or spreading misinformation. Ensuring security measures and appropriate regulations to prevent such misuse is vital in the development and deployment process.