Contents

A. Limitations & Future Work

Our PartCATSeg framework has shown strong performance but also has some limitations and potential directions for future work. Specifically, it struggles with handling extremely fine-grained part distinctions, where subtle differences between parts are challenging to capture accurately. Additionally, the framework is designed based on semantic segmentation, as it continues to demonstrate superior performance compared to instance segmentation in many scenarios. However, this design choice inherently limits the ability to distinguish individual parts as separate instances. For example, distinguishing between the left and right handles of a bicycle requires an instance-level understanding that the current framework lacks. Achieving this would necessitate the generation or inclusion of instance object masks, which are not currently supported. Expanding the framework to incorporate instance segmentation could address this limitation, enabling more precise and versatile part segmentation for applications that demand detailed instance-level information.

Future work aims to address these limitations by improving fine-grained differentiation through advanced attention mechanisms [23, 47, 68] and adaptive structural priors tailored to specific datasets. Enhancing the framework's ability to distinguish subtle part-level differences could significantly improve its performance in complex scenarios. Additionally, integrating the framework with off-the-shelf open-vocabulary instance segmentation modules offers a promising solution for overcoming the inability to individual parts as separate instances. This integration could enable the model to assign unique labels to similar parts across different instances, further extending its applicability and effectiveness.

B. Experimental Details

B.1. Code & Reproduction

Details can be found in the publicly available code. For additional details, refer to the GitHub repository available at https://github.com/kaist-cvml/part-catseg

B.2. Device Information

All experiments were conducted using eight NVIDIA A6000 GPUs and PyTorch 2.2 for training and evaluation.

B.3. Implementation Details

Our model is based on CAT-Seg [13], a state-of-the-art open-vocabulary semantic segmentation (OVSS) method, redefined to suit open-vocabulary part segmentation (OVPS) tasks in OV-PARTS [70]. It employs a CLIP [57] encoder built on CLIP ViT-B/16 and leverages DINOv2 [54], a pre-trained model, for structural guidance.

We begin by utilizing the pre-trained object-level OVSS models from CAT-Seg and fine-tune them with the datasets described in Appendix F. The model undergoes training with the AdamW [49] optimizer, starting with an initial learning rate of 0.0001, over 20,000 iterations, and a batch size of 8. During training, model checkpoints are saved every 1,000 iterations. The final model is selected based on the highest validation performance. For instance, the best validation score on the Pascal-Part-116 dataset in

the Oracle-Obj setting comes from the checkpoint saved at 12,000 iterations.

B.4. Evaluation Details

For the evaluation protocol, the Pred-All setup of PartCLIPSeg [14] and the Oracle-Obj setup of OV-PARTS [70] were utilized. The Pred-All setup assumes a more challenging scenario in which predictions are made without any prior information. In contrast, the Oracle-Obj setup assumes the availability of object-level masks. As noted in OV-PARTS, the Oracle-Obj setup simulates results achievable when using off-the-shelf open-vocabulary semantic segmentation models.

B.5. Hyperparameters

The model architecture incorporates layers and parameters from CAT-Seg [13]. Furthermore, as outlined in Equation (15), our method defines three key hyperparameters, $\lambda_{\texttt{Obj}}$, $\lambda_{\texttt{Part}}$, and $\lambda_{\texttt{comp}}$, which are associated with two primary loss functions: the disentanglement loss $\mathcal{L}_{\texttt{disen}}$ and the compositional loss $\mathcal{L}_{\texttt{comp}}$. These lambda parameters were fine-tuned through experimental validation on the training set to balance the contributions of the proposed loss functions. The final values were determined as $\lambda_{\texttt{Obj}} = 1.0$, $\lambda_{\texttt{Part}} = 1.0$, and $\lambda_{\texttt{comp}} = 1.0$.

B.6. Baselines

- ZSSeg+ [70, 74]: ZSSeg is a two-stage framework for open-vocabulary semantic segmentation that uses CLIP to classify class-agnostic mask proposals, enabling segmentation of seen and unseen classes. ZSSeg+ extends ZSSeg to support part-level segmentation. We evaluate ZSSeg+ using a ResNet-50 [27] baseline, fine-tuned with Compositional Prompt Tuning based on CoOp [82].
- VLPart [61]: VLPart enables open-vocabulary part segmentation by training on data across multiple granularities (part-level, object-level, and image-level) and segments novel objects into parts through dense correspondences with base objects.
- CLIPSeg [50, 70]: CLIPSeg extends CLIP for segmentation tasks, using a transformer-based decoder to generate segmentation maps conditioned on text or image prompts, supporting tasks like referring expression segmentation and zero-shot segmentation. For evaluation, we fine-tune the FiLM layer, decoder, visual encoder, and language embedding layer in the text encoder, following the approach in [70].
- CAT-Seg [13, 70]: CAT-Seg adapts vision-language models like CLIP by aggregating cosine similarity between image and text embeddings to create cost volumes, enabling segmentation of seen and unseen classes. Additionally, CAT-Seg proposes learning the self-attention heads of CLIP's encoders, achieving effective results.
- PartGLEE [38]: PartGLEE is a part-level segmentation model that uses a unified framework and the Q-Former to model hierarchical relationships between objects and parts, allowing segmentation at any granularity in open-world scenarios.
- PartCLIPSeg [14]: PartCLIPSeg leverages generalized parts and object-level contexts to enhance fine-grained part segmentation, incorporating competitive part relationships and attention mechanisms to improve segmentation accuracy and generalization to unseen vocabularies.

C. Additional Quantitative Evaluation

C.1. Evaluation on Recall

We evaluated various baselines across multiple datasets, as detailed in Table 1, Table 2, and Table 3. Here, we provide an additional analysis of recall performance in zero-shot evaluation on Pascal-Part-116 and ADE20K-Part-234. The recall metric measures a model's ability to correctly identify less frequent or smaller objects, which are often more difficult to segment accurately. As shown in Table A1, our method achieves state-of-the-art recall scores, highlighting its superior capability in identifying and segmenting these challenging parts.

Method	Pascal-Part-116			ADE20K-Part-234		
	Seen	Unseen	h-Recall	Seen	Unseen	h-Recall
ZSSeg+ [74]	<u>65.47</u>	32.13	43.10	<u>55.78</u>	40.71	47.07
CLIPSeg [50, 70]	55.71	43.35	48.76	49.59	48.11	48.84
CAT-Seg [13, 70]	56.00	43.20	48.77	43.48	39.87	41.60
PartCLIPSeg [14]	58.46	<u>47.93</u>	<u>52.67</u>	53.31	<u>51.52</u>	<u>52.40</u>
PartCATSeg (Ours)	**67.15**	**61.02**	**63.94** (+11.27)	**64.81**	**64.22**	**64.52** (+12.12)

[1] The best score is **bold** and the second-best score is <u>underlined</u>.

Table A1. Comparison of zero-shot performance with state-of-the-art methods in terms of **Recall** for **Oracle-Obj** setting on Pascal-Part-116.

C.2. DINO Structural Guidance

We confirmed the effectiveness of DINO's structural guidance for PartCATSeg in Table 6 of the main text. Additionally, the supplementary material examines how DINO's structural guidance impacts the original CAT-Seg. As shown in Table A2, while DINO's structural guidance proves effective, its performance improvement is relatively modest compared to the proposed PartCATSeg. This highlights that the proposed framework for disentangling parts is more effective overall.

Method	Pred-All			Oracle-Obj		
	Seen	Unseen	h-IoU	Seen	Unseen	h-IoU
CAT-Seg	36.80	23.39	28.60	43.81	27.66	33.91
CAT-Seg w/ Structural Guidance	38.84	28.99	33.20	50.37	36.86	42.57
PartCATSeg w/o Structural Guidance	42.29	27.94	33.65	46.44	31.59	37.60
PartCATSeg	**52.62**	**40.51**	**45.77**	**57.49**	**44.88**	**50.41**

[1] The best score is **bold**.

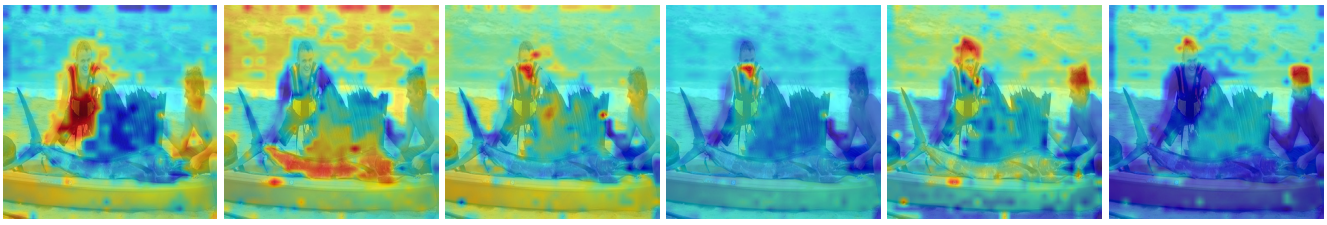
Table A2. Comparison of zero-shot performance with state-of-the-art methods on Pascal-Part-116.

D. Additional Qualitative Results

D.1. Pascal-Part-116 (Oracle-Obj Setting)

The Figure A8 illustrates qualitative evaluations under the Oracle-Obj setting, which assumes that object masks are provided. This setting evaluates the fine-grained part segmentation results using object-level segmentation generated by other off-the-shelf OVSS models.

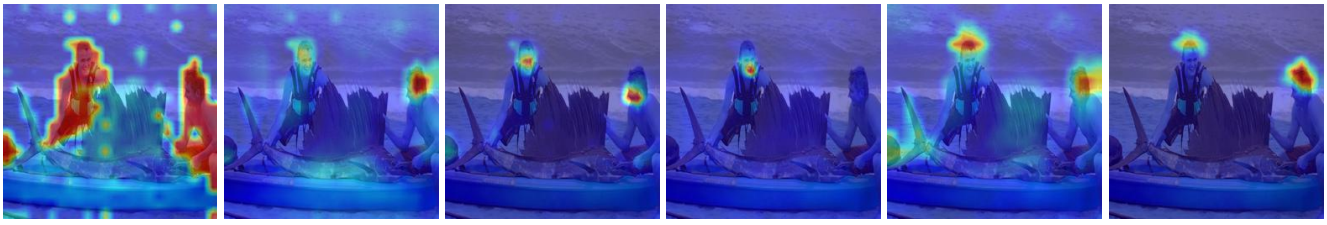
PartCATSeg consistently demonstrates superior segmentation performance compared to other baselines. Notably, it effectively segments small parts, such as arms and eyes, which are often missed by other models.



(a) “person” (b) “person’s eye” (c) “person’s nose” (d) “person’s mouth” (e) “person’s ear” (f) “person’s hair”

(g) “person’s head” (h) “person’s lower arm” (i) “person’s hand” (j) “person’s foot” (k) “person’s leg” (l) “person’s torso”

Figure A1. Image-Text Correspondence Visualization **Before** Training



(a) “person” (b) “person’s eye” (c) “person’s nose” (d) “person’s mouth” (e) “person’s ear” (f) “person’s hair”

(g) “person’s head” (h) “person’s lower arm” (i) “person’s hand” (j) “person’s foot” (k) “person’s leg” (l) “person’s torso”

Figure A2. Image-Text Correspondence Visualization **After** Training

D.2. PartImageNet (Pred-All Setting)

The following Figure A9 illustrates the PartImageNet prediction results of PartCATSeg in Pred-All setup. PartCATSeg demonstrates superior performance compared to other baselines, accurately predicting both appropriate classes and boundaries.

D.3. PartImageNet (Oracle-Obj Setting)

The following Figure A10 illustrates the PartImageNet prediction results of PartCATSeg in Oracle-Obj setup. PartCATSeg demonstrates superior performance compared to other baselines, accurately predicting appropriate boundaries.

D.4. Cost Visualization

The following Figure A2 and Figure A4 present the cost (correspondence) visualization after training. It visualizes the object-specific part correspondence between the caption text and the image. Unlike (pretrained) CLIP Image-Text Similarity Visualization (Figure 2) discussed in Section 1 and Figure A1, as well as Figure A3, the cost volume demonstrates significant improvement in fine-grained alignment after training, as illustrated in Figure A2 and Figure A4.

E. Additional Ablation Study

E.1. Compositional Loss

As detailed in Section 3.5, parts are not only compositional components that constitute an object but also maintain relationships with adjacent parts. Previous methodologies have proposed learning strategies that consider granularity at two levels—the object and its parts. However, they have not focused on the composition of object-specific parts within the object and the relationships between these parts. This limitation often results in small parts, such as “cat’s eye” and “cat’s neck”, being undetected within the

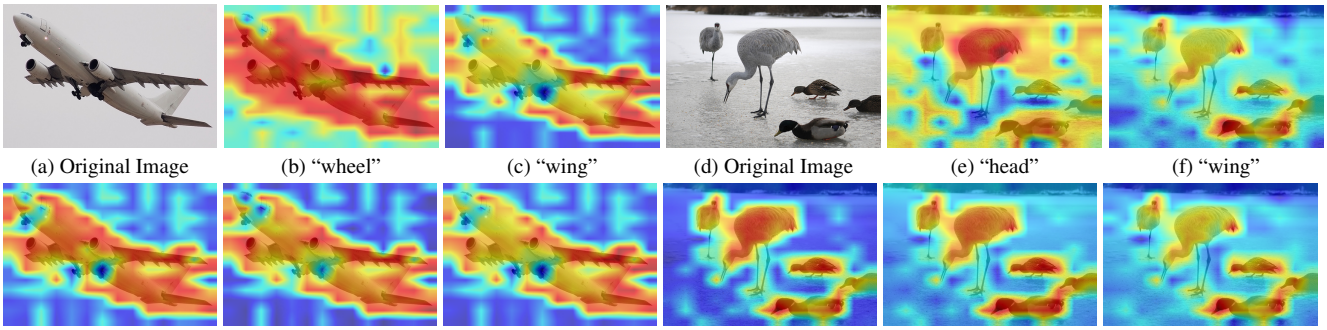


Figure A3. **CLIP Image-Text Similarity Visualization for Object-Level and Part-Level Text.** The visualization compares the frozen CLIP image-text similarity between object-level and part-level text descriptions. (a), (d) show the original images; (b), (c), (e), (f) depict the part-level similarities for terms such as ”wheel” and ”wing” while (g)-(l) show object-specific parts. The stronger activation for object-level text suggests a dominant focus on the entire object rather than individual parts in the image-text correspondence.

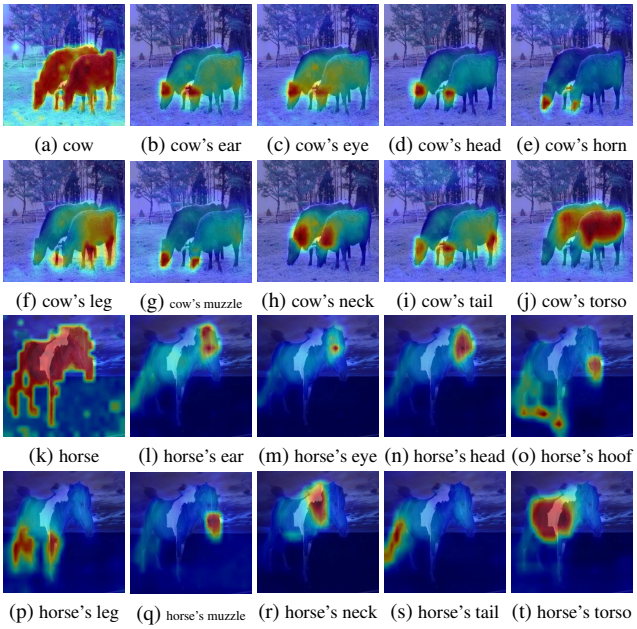


Figure A4. **Cost Volume Visualization After Additional Training.** Cost volume visualization showed that PartCATSeg significantly enhanced fine-grained alignment.

context of a larger object, as shown in Figure A5a. We hypothesize that this issue arises because certain parts become excessively dominant relative to other parts they encompass, leading to a failure to recognize their spatial and compositional relationships. By introducing compositional loss, our model better identifies and segments smaller or more discriminative parts. As illustrated in Figure A5, the inclusion of compositional loss resolves issues of overlapping or diffuse cost volumes, as visualized in Figure A5c and Figure A5d. These visualizations highlight how compositional loss sharpens the focus on specific parts, mitigating the spread of cost volume and ensuring better part-level segmentation.

Furthermore, as shown in Figure A6, compositional loss consistently improves segmentation performance across datasets. For example, in Pascal-Part-116, it enhances the segmentation of challenging parts such as cow’s eye” or cat’s nose,” which are often difficult to detect. Specifically, the comparison demonstrates that softmax normalization in the compositional loss ($\mathcal{L}_{\texttt{comp}}$-SM) outperforms L1 normalization ($\mathcal{L}_{\texttt{comp}}$-L1). Similarly, as shown in Figure A7, compositional loss demonstrates its effectiveness in capturing fine-grained part relationships in PartImageNet, such as “goose’s tail,” “tench’s tail,” and “killer whale’s head.” This improvement underscores the importance of explicitly modeling compositional relationships in part segmentation, particularly for smaller or less distinct parts.

The effectiveness of compositional loss is further validated through quantitative results, as shown in Table A3. The inclusion of $\mathcal{L}_{\texttt{comp}}$ improves performance across both the Pred-All and Oracle-Obj settings on PartImageNet. Notably, it enhances the harmonic IoU for unseen parts, demonstrating its ability to better capture fine-grained compositional relationships and improve segmentation consistency for challenging parts.

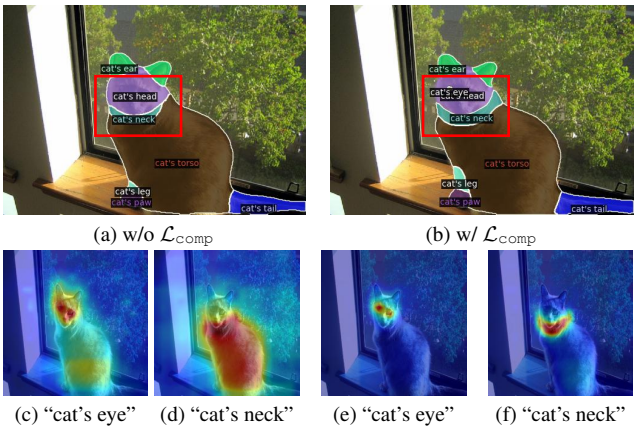
Compositional Loss	Pred-All			Oracle-Obj		
	Seen	Unseen	h-IoU	Seen	Unseen	h-IoU
w/o $\mathcal{L}_{\texttt{comp}}$	**59.21**	50.75	54.66	72.17	68.42	70.24
w/ $\mathcal{L}_{\texttt{comp}}$	57.33	**53.07**	**55.12**	**73.83**	**71.52**	**72.66**

Table A3. Impact of Compositional Loss on PartImageNet

F. Datasets Details

F.1. Pascal-Part-116

Pascal-Part-116 [70] is a modified version of the Pascal-Part [8] dataset specifically designed for Open-Vocabulary Part Segmentation tasks. The dataset includes a mix of base and novel categories, focusing on diverse object-level classes. It features novel classes for the object categories “bird”, “car”, “dog”, “sheep”, and “motorbike”. Additionally, based on object-specific part categories, the dataset comprises 74 base categories and 42 novel categories, offering a comprehensive benchmark for evaluating segmentation



(a) w/o $\mathcal{L}_{\text{comp}}$ (b) w/ $\mathcal{L}_{\text{comp}}$

(c) "cat's eye" (d) "cat's neck" (e) "cat's eye" (f) "cat's neck"

Figure A5. **Ablation on Compositional Loss.** (a) and (b) show segmentation results without and with $\mathcal{L}_{\text{comp}}$, respectively. (c) and (d) show less defined cost volumes without $\mathcal{L}_{\text{comp}}$, while (e) and (f) reveal more exclusive similarities in the cost volumes with $\mathcal{L}_{\text{comp}}$. Notably, "cat's eye" is successfully segmented with the inclusion of $\mathcal{L}_{\text{comp}}$.

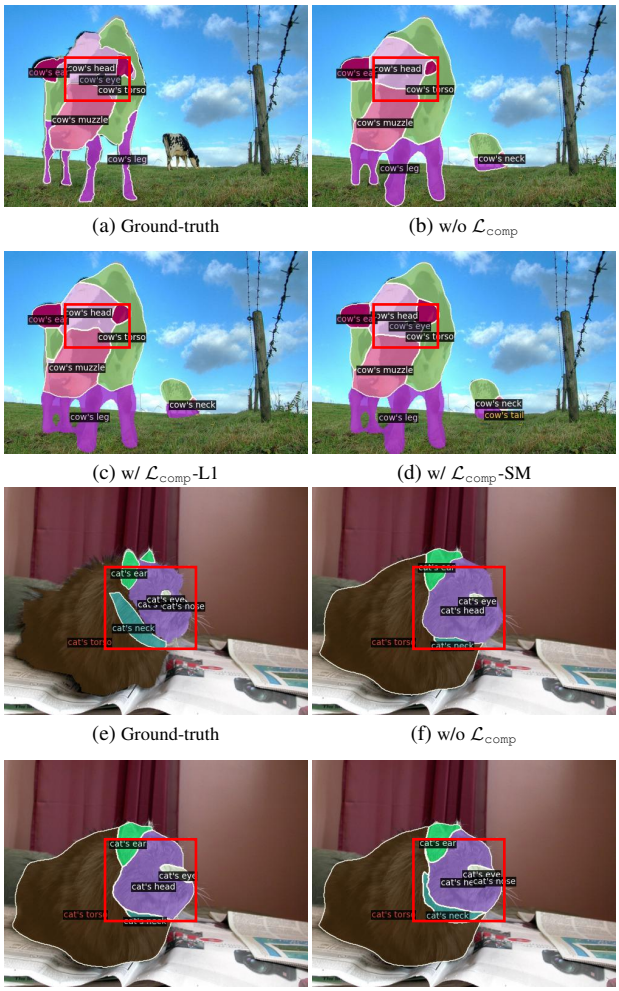
models. Detailed information about the base and novel classes can be found in Table A4.

Pascal-Part-116 Object-specific Part Categories				
Base Categories (74)				
aeroplane's body	aeroplane's stern	aeroplane's wing	aeroplane's tail	aeroplane's engine
aeroplane's wheel	bicycle's wheel	bicycle's saddle	bicycle's handlebar	bicycle's chainwheel
bicycle's headlight	bottle's body	bottle's cap	bus's wheel	bus's headlight
bus's front	bus's side	bus's back	bus's roof	bus's mirror
bus's license plate	bus's door	bus's window	cat's tail	cat's head
cat's eye	cat's torso	cat's neck	cat's leg	cat's nose
cat's paw	cat's ear	cow's tail	cow's head	cow's eye
cow's torso	cow's neck	cow's leg	cow's ear	cow's muzzle
cow's horn	horse's tail	horse's head	horse's eye	horse's torso
horse's neck	horse's leg	horse's ear	horse's muzzle	horse's hoof
person's head	person's eye	person's torso	person's neck	person's leg
person's foot	person's nose	person's ear	person's eyebrow	person's mouth
person's hair	person's lower arm	person's upper arm	person's hand	pottedplant's pot
pottedplant's plant	train's headlight	train's head	train's front	train's side
train's back	train's roof	train's coach	tvmonitor's screen	
Novel Categories (42)				
bird's wing	bird's tail	bird's head	bird's eye	bird's beak
bird's torso	bird's neck	bird's leg	bird's foot	car's wheel
car's headlight	car's front	car's side	car's back	car's roof
car's mirror	car's license plate	car's door	car's window	dog's tail
dog's head	dog's eye	dog's torso	dog's neck	dog's leg
dog's nose	dog's paw	dog's ear	dog's muzzle	motorbike's wheel
motorbike's saddle	motorbike's handlebar	motorbike's headlight	sheep's tail	sheep's head
sheep's eye	sheep's torso	sheep's neck	sheep's leg	sheep's ear
sheep's muzzle	sheep's horn			

Table A4. List of object-specific classes in **Pascal-Part-116**.

F.2. ADE20K-Part-234

ADE20K-Part-234 [70] is an adapted version of the ADE20K dataset [80] tailored for Open-Vocabulary Part Segmentation tasks. The dataset includes a mix of base and novel categories, with a focus on a diverse range of object-level classes. It features novel classes for the object categories "bench", "bus", "fan", "desk", "stool", "truck", "van", "swivel chair", "oven", "ottoman," and "kitchen island". Additionally, based on object-specific part categories, the dataset comprises 176 base categories and 58 novel categories, providing a robust benchmark for evaluating segmentation models. The dataset presents additional challenges due to its diverse categories and the frequent appearance of small parts, which require precise part segmentation.



(a) Ground-truth (b) w/o $\mathcal{L}_{\text{comp}}$

(c) w/ $\mathcal{L}_{\text{comp}}$-L1 (d) w/ $\mathcal{L}_{\text{comp}}$-SM

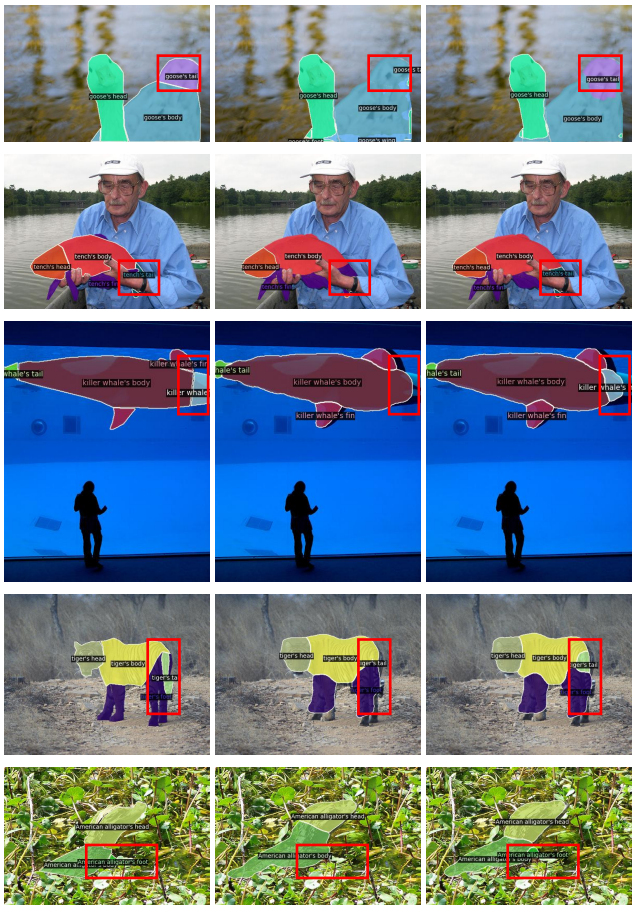
(e) Ground-truth (f) w/o $\mathcal{L}_{\text{comp}}$

(g) w/ $\mathcal{L}_{\text{comp}}$-L1 (h) w/ $\mathcal{L}_{\text{comp}}$-SM

Figure A6. **Qualitative Ablation on Compositional Loss for Pred-All setting in Pascal-Part-116.** Applying compositional loss improves segmentation performance, particularly for challenging parts like "cow's eye" or "cat's nose". Furthermore, using softmax normalization in the compositional loss ($\mathcal{L}_{\text{comp}}$-SM) outperforms L1 normalization ($\mathcal{L}_{\text{comp}}$-L1) by better capturing these fine-grained parts such as "cat's neck".

F.3. PartImageNet

PartImageNet [25] is a dataset adapted from ImageNet [17], comprising approximately 24,000 images across 158 categories, each annotated with detailed part information. These categories are grouped into 11 superclasses, based on the hierarchical taxonomy provided by WordNet [51]. In cross-dataset evaluation settings, PartImageNet categories are not pre-divided into base and novel classes. Therefore, for zero-shot evaluation, we select 40 representative object classes from the dataset, which we further split into 25 base object classes and 15 novel object classes, just as in PartCLIPSeg [14]. Each superclass contains corresponding part classes, allowing us to construct part categories by associating these object classes with their respective part classes within each superclass. This selective splitting ensures that the novel classes



(a) Ground-truth (b) w/o $\mathcal{L}_{\texttt{comp}}$ (c) w/ $\mathcal{L}_{\texttt{comp}}$-SM

Figure A7. **Qualitative Ablation on Compositional Loss in PartImageNet.** The first two rows show results in the Oracle-Obj setting, while the bottom three rows are in the Pred-All setting. Segmentation with $\mathcal{L}_{\texttt{comp}}$ captures finer part relationships, such as "goose's tail," "tench's tail," "killer whale's head," "tiger's tail," and "American alligator's foot," compared to results without $\mathcal{L}_{\texttt{comp}}$.

remain unseen during training, thereby providing a robust evaluation of zero-shot capabilities. Detailed information about the base and novel classes, along with their corresponding superclasses, is presented in Table A6.

F.4. PartImageNet (OOD)

PartImageNet [25] offers an alternative split designed for few-shot learning, which ensures non-overlapping classes across the training and validation sets. This few-shot split includes 109 base object classes in the training set and 19 novel object classes in the validation set. Detailed information about the base and novel classes, along with their corresponding superclasses, is presented in Table A7.

ADE20K-Part-234 Object-specific Part Categories				
Base Categories (176)				
person's arm	person's back	person's foot	person's gaze	person's hand
person's head	person's leg	person's neck	person's torso	door's door frame
door's handle	door's knob	door's panel	clock's face	clock's frame
toilet's bowl	toilet's cistern	toilet's lid	cabinet's door	cabinet's drawer
cabinet's front	cabinet's shelf	cabinet's side	cabinet's skirt	cabinet's top
sink's bowl	sink's faucet	sink's pedestal	sink's tap	sink's top
lamp's arm	lamp's base	lamp's canopy	lamp's column	lamp's cord
lamp's highlight	lamp's light source	lamp's shade	lamp's tube	sconce's arm
sconce's backplate	sconce's highlight	sconce's light source	sconce's shade	chair's apron
chair's arm	chair's back	chair's base	chair's leg	chair's seat
chair's seat cushion	chair's skirt	chair's stretcher	chest of drawers's apron	chest of drawers's door
chest of drawers's drawer	chest of drawers's front	chest of drawers's leg	chandelier's arm	chandelier's bulb
chandelier's canopy	chandelier's chain	chandelier's cord	chandelier's highlight	chandelier's light source
chandelier's shade	bed's footboard	bed's headboard	bed's leg	bed's side rail
table's apron	table's drawer	table's leg	table's shelf	table's top
table's wheel	armchair's apron	armchair's arm	armchair's back	armchair's back pillow
armchair's leg	armchair's seat	armchair's seat base	armchair's seat cushion	shelf's door
shelf's drawer	shelf's front	shelf's shelf	coffee table's leg	coffee table's top
sofa's arm	sofa's back	sofa's back pillow	sofa's leg	sofa's seat base
sofa's seat cushion	sofa's skirt	computer's computer case	computer's keyboard	computer's monitor
computer's mouse	wardrobe's door	wardrobe's drawer	wardrobe's front	wardrobe's leg
wardrobe's mirror	wardrobe's top	car's bumper	car's door	car's headlight
car's hood	car's license plate	car's logo	car's mirror	car's wheel
car's window	car's wiper	cooking stove's burner	cooking stove's button panel	cooking stove's door
cooking stove's drawer	cooking stove's oven	cooking stove's stove	microwave's button panel	microwave's door
microwave's front	microwave's side	microwave's top	microwave's window	refrigerator's button panel
refrigerator's door	refrigerator's drawer	refrigerator's side	dishwasher's button panel	dishwasher's handle
dishwasher's skirt	bookcase's door	bookcase's drawer	bookcase's front	bookcase's side
television receiver's base	television receiver's buttons	television receiver's frame	television receiver's keys	television receiver's screen
television receiver's speaker	glass's base	glass's bowl	glass's opening	glass's stem
pool table's bed	pool table's leg	pool table's pocket	airplane's door	airplane's fuselage
airplane's landing gear	airplane's propeller	airplane's stabilizer	airplane's turbine engine	airplane's wing
minibike's license plate	minibike's mirror	minibike's seat	minibike's wheel	washer's button panel
washer's door	washer's front	washer's side	traffic light's housing	traffic light's pole
light's aperture	light's canopy	light's diffusor	light's highlight	light's light source
light's shade				
Novel Categories (58)				
ottoman's back	ottoman's leg	ottoman's seat	swivel chair's back	swivel chair's base
swivel chair's seat	swivel chair's wheel	fan's blade	fan's canopy	fan's tube
stool's leg	stool's seat	desk's apron	desk's door	desk's drawer
desk's leg	desk's shelf	desk's top	bus's bumper	bus's door
bus's headlight	bus's license plate	bus's logo	bus's mirror	bus's wheel
bus's window	bus's wiper	oven's button panel	oven's door	oven's drawer
oven's top	kitchen island's door	kitchen island's drawer	kitchen island's front	kitchen island's side
kitchen island's top	van's bumper	van's door	van's headlight	van's license plate
van's logo	van's mirror	van's taillight	van's wheel	van's window
van's wiper	truck's bumper	truck's door	truck's headlight	truck's license plate
truck's logo	truck's mirror	truck's wheel	truck's window	bench's arm
bench's back	bench's leg	bench's seat		

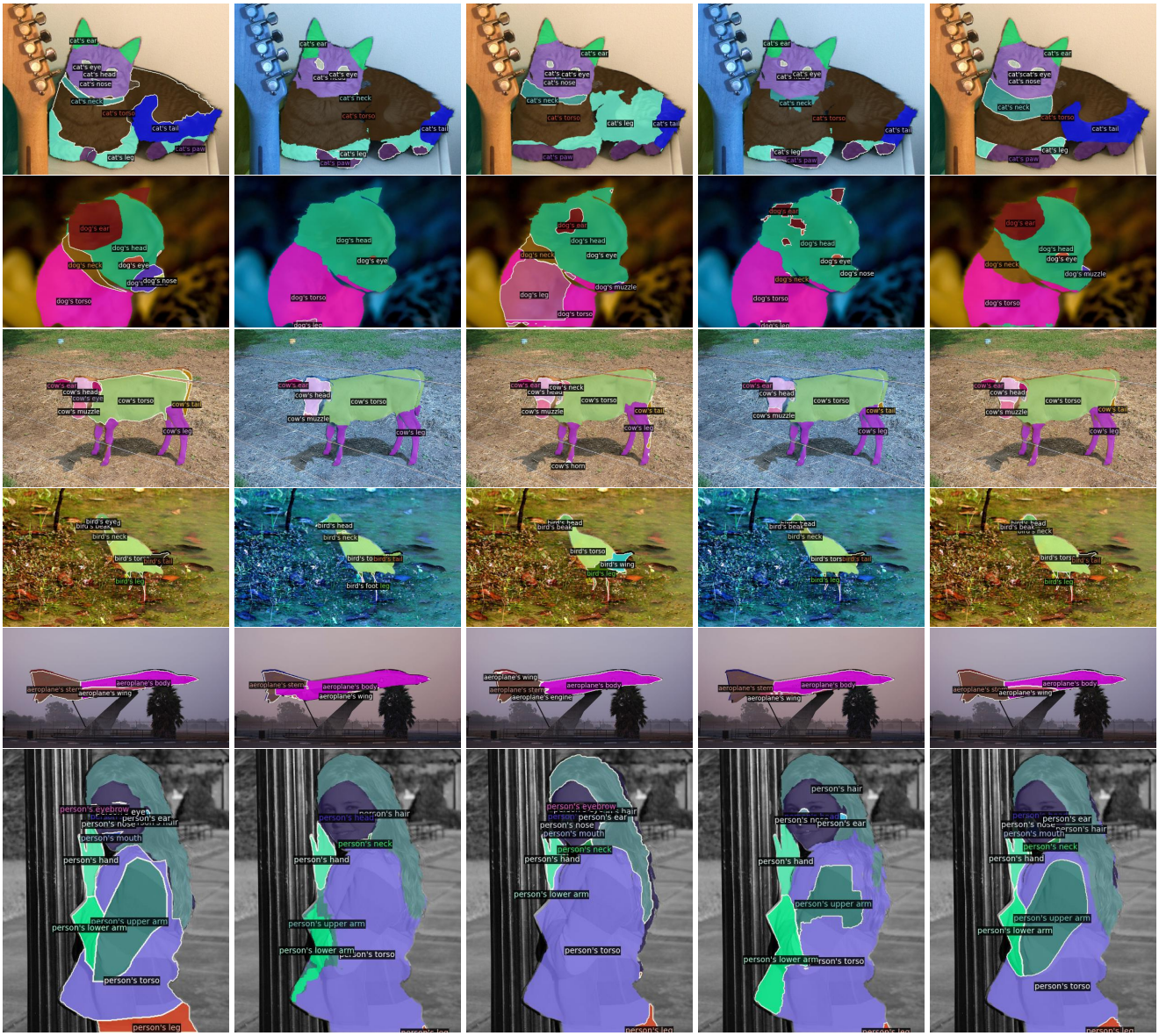
Table A5. List of object-specific classes in **ADE20K-Part-234**.

Superclass	Base Object Categories (25)	Novel Object Categories (15)	Part Classes
Quadruped	tiger, giant panda, leopard, gazelle	ice bear, impala, golden retriever	Head, Body, Foot, Tail
Snake	green mamba	Indian cobra	Head, Body
Reptile	green lizard, Komodo dragon, tree frog	box turtle, American alligator	Head, Body, Foot, Tail
Boat	yawl, pirate	schooner	Body, Sail
Fish	barracouta, goldfish, killer whale	tench	Head, Body, Fin, Tail
Bird	albatross, goose	bald eagle	Head, Body, Wing, Foot, Tail
Car	garbage truck, minibus, ambulance	jeep, school bus	Body, Tier, Side Mirror
Bicycle	mountain bike, moped	motor scooter	Body, Head, Seat, Tier
Biped	gorilla, orangutan	chimpanzee	Head, Body, Hand, Foot, Tail
Bottle	beer bottle, water bottle	wine bottle	Mouth, Body
Aeroplane	warplane	airliner	Head, Body, Engine, Wing, Tail

Table A6. List of selected object classes and their corresponding part classes per superclass from **PartImageNet** [25]. Object categories are categorized into base and novel object classes, with part classes assigned to each respective superclass.

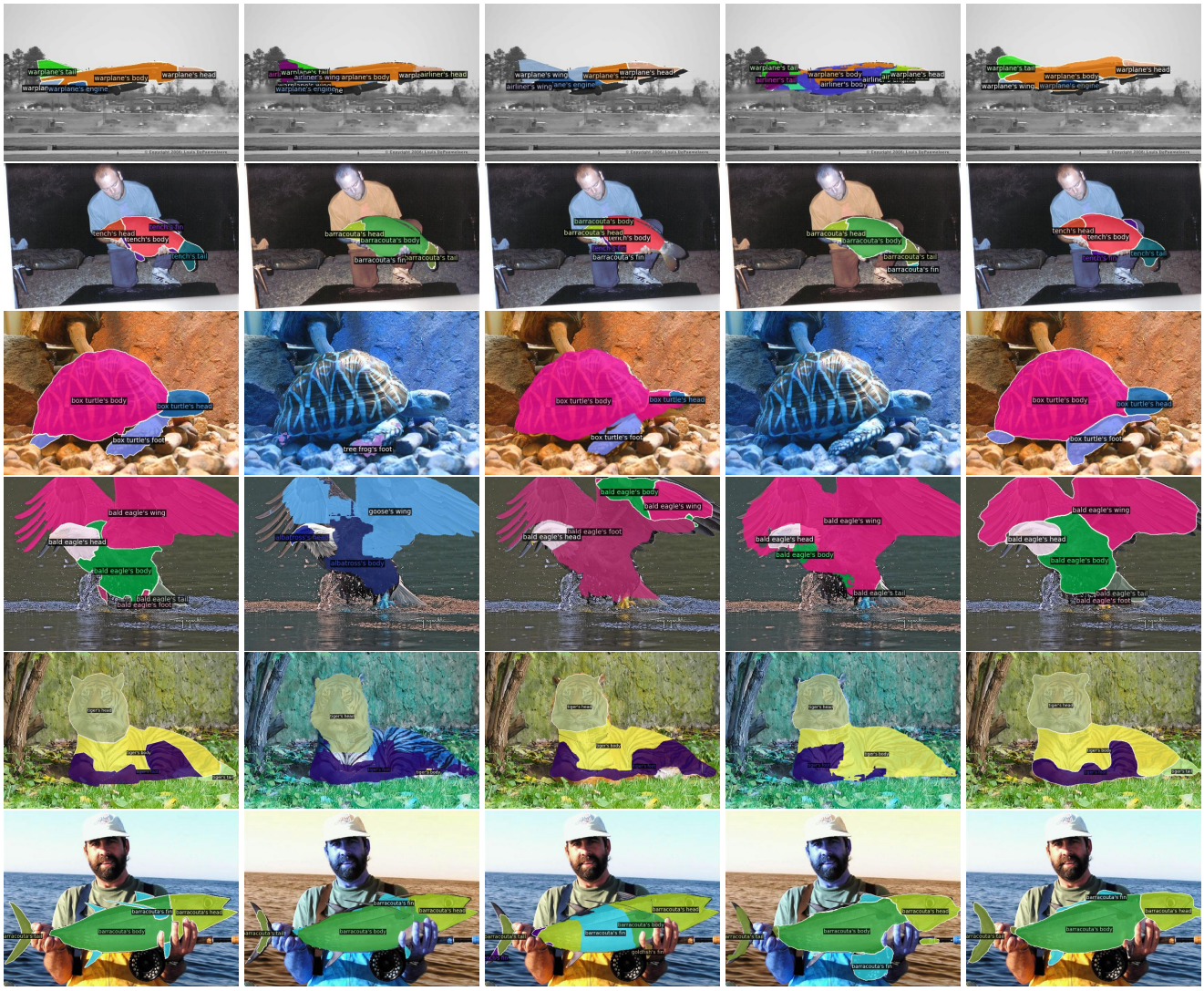
Superclass	Base Object Categories (109)	Novel Object Categories (19)
Quadruped	impala, Egyptian cat, warthog, otter, Tibetan terrier timber wolf, polecat, water buffalo, ox, redbone English springer, tiger, American black bear, leopard, hartebeest vizsla, Brittany spaniel, giant panda, Boston bull, ram cairn, Arabian camel, fox squirrel, Eskimo dog, Irish water spaniel Saluki, Walker hound, cheetah, gazelle, soft-coated wheaten terrier bighorn, brown bear, chow, weasel	golden retriever, cougar, ice bear, mink, Saint Bernard
Snake	night snake, boa constrictor, green mamba, thunder snake, green snake hognose snake, sidewinder, horned viper, diamondback, rock python garter snake, vine snake	Indian cobra
Reptile	Gila monster, common newt, green lizard, bullfrog, American alligator leatherback turtle, spotted salamander, box turtle, tailed frog, African chameleon Komodo dragon, agama, frilled lizard, loggerhead	whiptail, alligator lizard
Boat	yawl, pirate	trimaran
Fish	goldfish, coho, tench, anemone fish, killer whale	barracouta, great white shark
Bird	albatross, spoonbill, black stork, dowitcher, American egret goose, ruddy turnstone, bee eater, kite	little blue heron, bald eagle
Car	garbage truck, minibus, ambulance, snowplow, golfcart police van, minivan, convertible, limousine, recreational vehicle go-kart, tractor, school bus, racer	beach wagon, cab
Bicycle	motor scooter, tricycle, mountain bike	unicycle
Biped	gorilla, gibbon, guenon, macaque, patas howler monkey, chimpanzee, proboscis monkey, spider monkey, baboon colobus, capuchin	siamang, marmoset
Bottle	beer bottle, pop bottle, pill bottle	water bottle
Aeroplane	airliner	

Table A7. **PartImageNet (OOD).** List of selected object classes and their corresponding part classes per superclass from PartImageNet (OOD) [25]. Object categories are divided into base and novel object classes. Detailed associations of part classes with their respective superclasses are provided in Table A6.



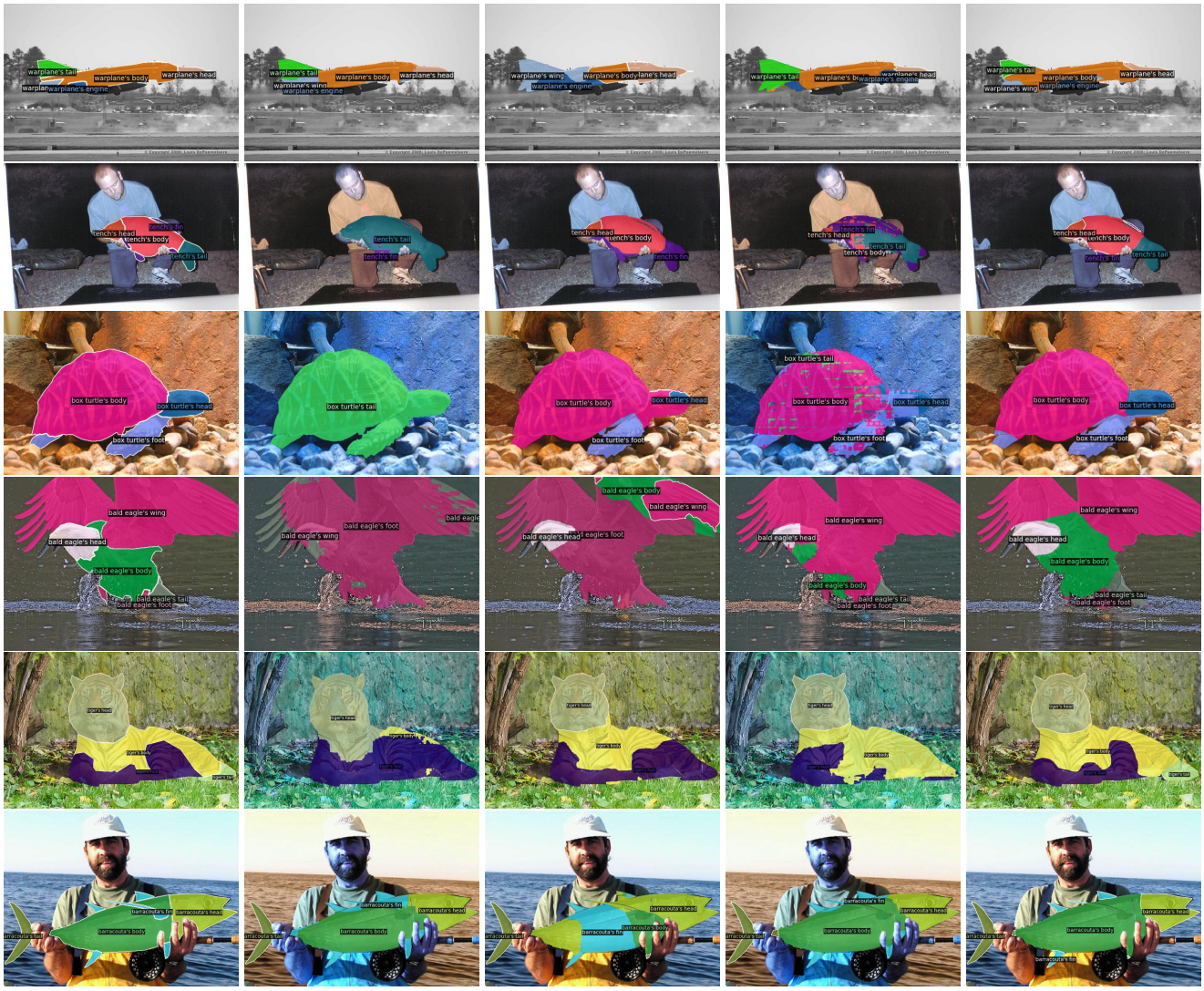
(a) Ground-truth (b) CLIPSeg [57, 70] (c) CAT-Seg [13, 70] (d) PartCLIPSeg [14] (e) PartCATSeg

Figure A8. Qualitative evaluation of zero-shot part segmentation on Pascal-Part-116 in the **Oracle-Obj** configuration. Note that annotations for unseen categories (e.g., bird, cow, dog) are excluded from the training set.



(a) Ground-truth (b) CLIPSeg [57, 70] (c) CAT-Seg [13, 70] (d) PartCLIPSeg [14] (e) PartCATSeg

Figure A9. Qualitative evaluation of zero-shot part segmentation on **PartImageNet** in the **Pred-All** configuration.



(a) Ground-truth (b) CLIPSeg [57, 70] (c) CAT-Seg [13, 70] (d) PartCLIPSeg [14] (e) PartCATSeg

Figure A10. Qualitative evaluation of zero-shot part segmentation on PartImageNet in the **Oracle-Obj** configuration.