Understanding Multi-Granularity for Open-Vocabulary Part Segmentation

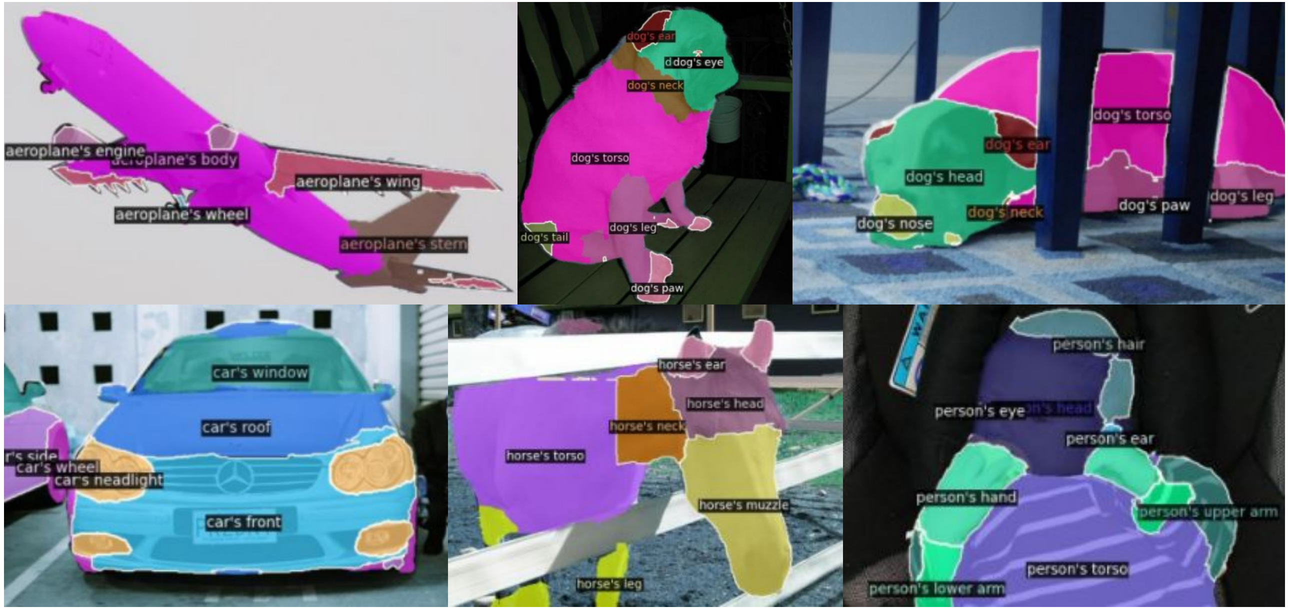
(NeurIPS 2024)

Jiho Choi[1]*, Seonho Lee[1]*, Seungho Lee[2] , Minhyun Lee[2] , Hyunjung Shim[1]

[1] Graduate School of Artificial Intelligence, KAIST, Republic of Korea
[2] School of Integrated Technology, Yonsei University, Republic of Korea

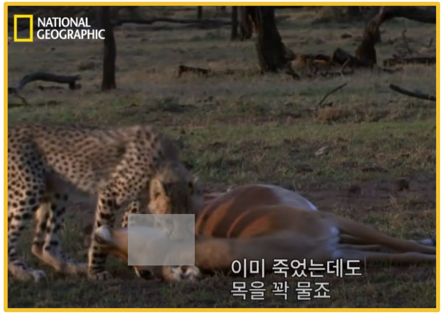
Open-Vocabulary Part Segmentation (OVPS)

- Task of segmenting different **parts** of an object in an image using **textual descriptions**
    - textual descriptions that are not constrained by a fixed set of predefined labels
- Recognizing parts that are more complex and diverse than object



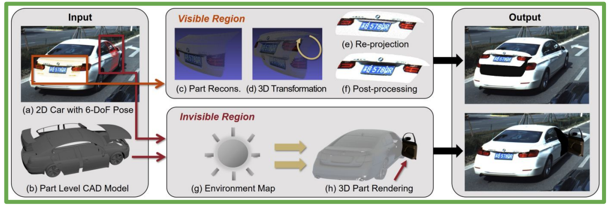
1. Introduction [2/5]

Motivations of OVPS



Why OVPS?

- Practical Usages
  - **Robot instruction**: Handle of tools (knife, ladle, pot)
  - Part Image **Editing / Generation**
- Multi-granularity Understanding
  - (Biomimicry) Mimicking and understanding **animal instincts**
    - e.g. When a “cheetah” hunts an “impala”, it can distinguish the “neck”



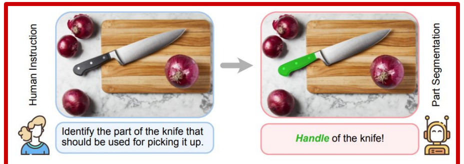


Figure 1: Task Description for InstructPart: Presented with an image observation (left) and a corresponding instruction (in the blue box on the left), the model is required to identify and output the specific part segment (highlighted in the green mask on the right) referenced in the instruction.

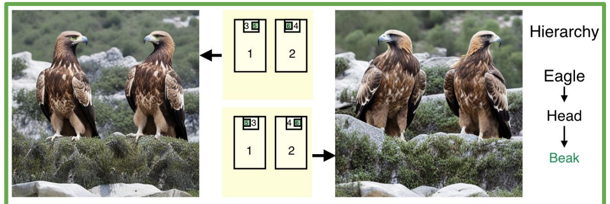


Image Caption: A golden eagle perched on a rugged rock.
Instance Caption: 1-2) A golden eagle; 3-4) Eagle’s head; 5-6) **Eagle’s beak**

d) Image Composition with Whole Instance, Part and Subpart

3D Part Guided Image Editing for Fine-grained Object Understanding (CVPR 2020)
InstructPart: Affordance-based Part Segmentation from Language Instruction (AAAIW 2024)
The Guardian - https://www.theguardian.com/technology/2020/dec/06/the-robot-kitchen-that-will-make-you-dinner-and-wash-up-too
National Geographic - https://youtu.be/xVxMisFY3GY?si=kV92174pFiUsFUdx&t=180
InstanceDiffusion: Instance-level Control for Image Generation (CVRP 2024)

Challenges of OVPS

Perspective on Semantic Segmentation

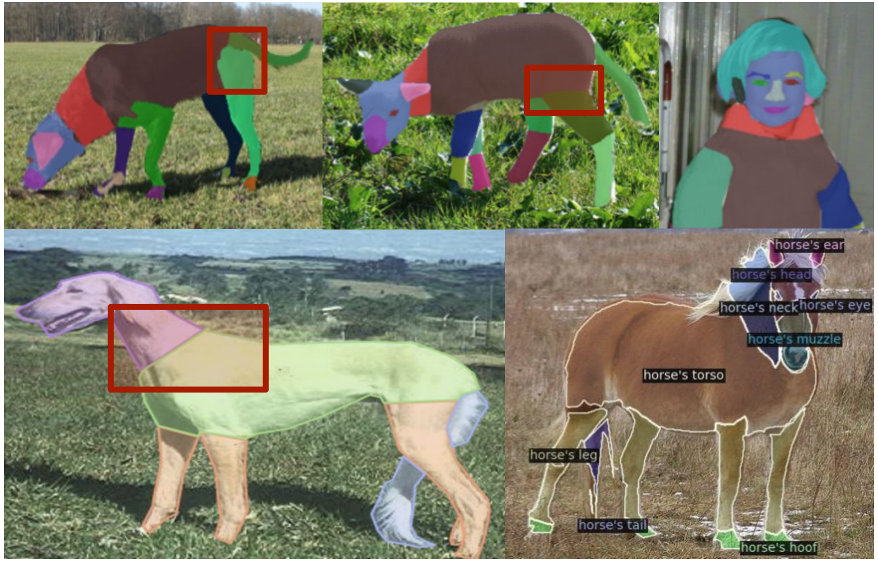
- Parts are **smaller** in size and **diverse** in category compared to objects

'Knowledge-based' characteristics of Parts

- Unlike Semantic, Instance, and Panoptic Segmentation, which can achieve clear answers solely from **visual information** present in nature, Parts reflect a knowledge-based nature, making the alignment with language crucial
- Parts are defined by **linguistic or social consensus**

'Open Granularity' characteristics of Parts:

- A pixel can be labeled as 'nose,' 'face,' or 'head' depending on the annotation
  - can yield **different ground truth** answers
  - analogous to the "ambiguity" discussed in SAM
- Parts are based on **relative and competitive concepts**



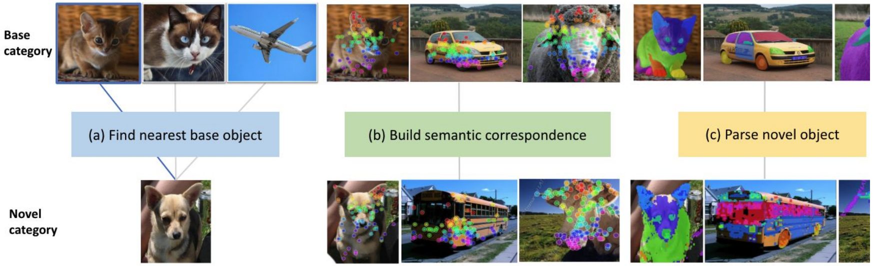
[Top] PASCAL-Part (2014)
[Bottom] Left: PartImageNet (2022) | Right: OV-PARTS (2023)

VLPart: Going Denser with Open-Vocabulary Part Segmentation (ICCV 2023)
OV-PARTS: Towards Open-Vocabulary Part Segmentation (NeurIPS Datasets and Benchmarks Track 2023)
Segment Anything (ICCV 2023)

Previous OVPS Methods

VLPart (ICCV 2023.10)

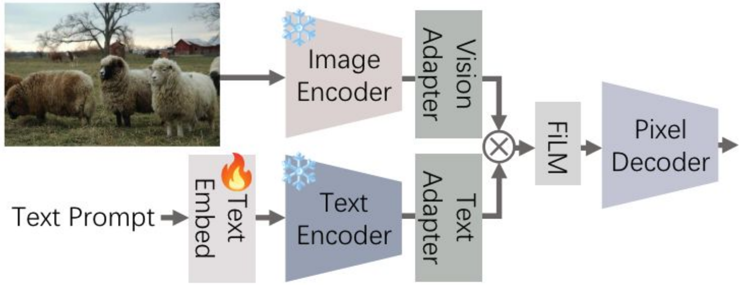
- **DINO** features to **map correspondences** between **base** and **novel** classes and creates **pseudo labels**
- → analogy



Pipeline of VLPart

OV-PARTS (NeurIPS B&D 2023.12)

- introduces **object mask** prompts and transferring knowledge of base class with **few-shot** approach
- Image-Text Alignment
  - FiLM: Feature-wise Linear Modulation



Overall Framework of OV-PARTS

VLPart: Going Denser with Open-Vocabulary Part Segmentation (ICCV 2023)
OV-PARTS: Towards Open-Vocabulary Part Segmentation (NeurIPS Datasets and Benchmarks Track 2023)
FiLM: Visual Reasoning with a General Conditioning Layer (AAAI 2018)

Limitations of Previous Methods

Zero-shot Part Segmentation

Problem Phenomenon 1)
Object-level / Part-level Label Misclassification

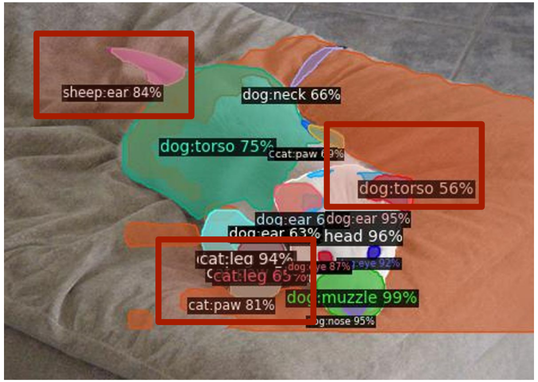
Assumed Cause 1)
Lack of Object-level Context / Lack of Part-level Generalization

Problem Phenomenon 2)
Incomplete / Ambiguous Boundaries

Assumed Cause 2)
Lack of Competitive / Relative Partitioning Inductive Bias

Problem Phenomenon 3)
Missing Labels

Assumed Cause 3)
Small Size / Low-frequency Labels



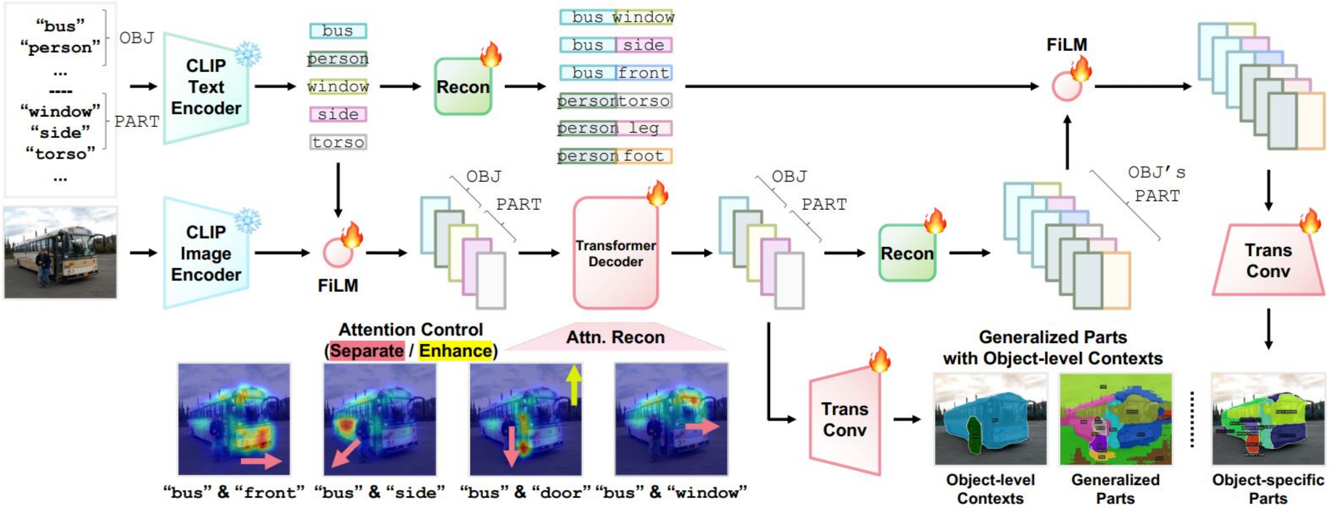
(a) Lack of generalization

(b) Ambiguous boundaries

(c) Missing underrepresented part

PartCLIPSeg [1/1]

- modified the architecture from CLIPSeg
- Generalized Parts with Object-level Contexts
- Attention Control for Ambiguity and Omission



Generalized Parts with Object-level Contexts [1/2]

Object-level (Global) Context / Part-level Generalization

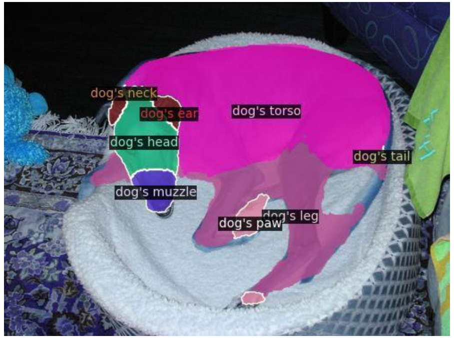
- considers **parts** as **common structural components** across object categories and integrates **object contexts**
  - → Part Information, Boundary Information, Holistic Understanding
- previous approaches
  - VLPart (base/novel correspondence, pseudo-label)
  - OV-PARTS (few-shot samples)

Train (base)

object: cat, cow, horse *etc.*
part: head, ear, neck, torso, leg, paw *etc.*

common structural components

Test (novel)

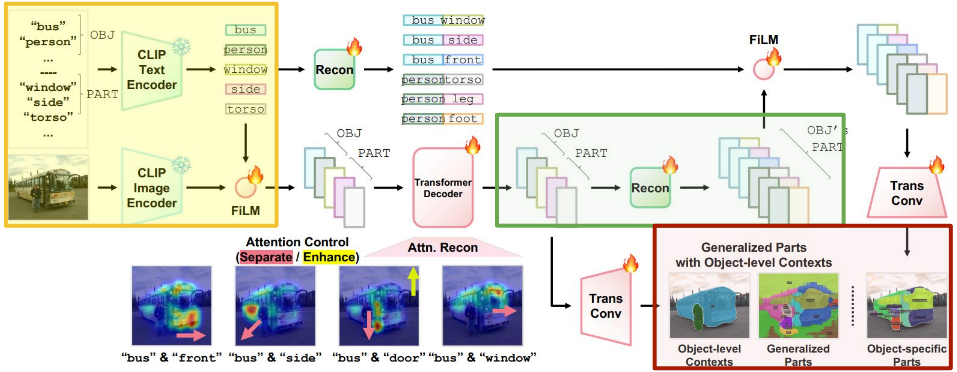


object: dog etc.
part: head, ear, neck, torso, leg, paw etc.

Generalized Parts with Object-level Contexts [2/2]

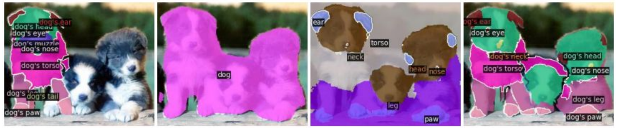
Object and Part Embedding Generation

- feature extraction with CLIP image, text encoders
  - $\mathbf{e}^{\mathcal{T}}_{[\text{obj} \mid \text{part}]} = \text{CLIP}^{*}_{\mathcal{T}}(\mathbf{c}_{[\text{obj} \mid \text{part}]}), \mathbf{e}^{\mathcal{I}} = \text{CLIP}^{*}_{\mathcal{I}}(\mathcal{I})$
- conditioning image features with FiLM
  - FiLM: an adaptive affine transformation
  - $\mathbf{e}^{\mathcal{I}}_{[\text{obj} \mid \text{part}]} = \mathbf{e}^{\mathcal{I}} \oplus \text{FiLM}(\mathbf{e}^{\mathcal{T}}_{[\text{obj} \mid \text{part}]})$



Object-specific Part Construction

- reconstruction
  - $\mathbf{e}^{[\mathcal{T}|\mathcal{I}]}_{\text{obj-part}} = \texttt{Proj}\left(\left[\mathbf{e}^{[\mathcal{T}|\mathcal{I}]}_{\text{obj}} \mid \mathbf{e}^{[\mathcal{T}|\mathcal{I}]}_{\text{part}}\right]\right)$



(a) Ground-truth (b) Object-level Pred. (c) Generalized Parts Pred. (d) Final Pred.

Mask Supervision

- Object, Part, and Object-specific Part
  - $$\mathcal{L}_{\texttt{mask}} = \sum_{i=1}^{|\mathbf{C}_{\text{obj-part}}|+1} \underbrace{\{1 - \text{BCE}(s_i, \hat{s}_i)\}}_{\text{object-specific part}} + \lambda_1 \sum_{i=1}^{|\mathbf{C}_{\text{obj}}|+1} \underbrace{\{1 - \text{BCE}(s^{o}_i, \hat{s}^{o}_i)\}}_{\text{object guidance}} + \lambda_2 \sum_{i=1}^{|\mathbf{C}_{\text{part}}|} \underbrace{\{1 - \text{BCE}(s^{p}_i, \hat{s}^{p}_i)\}}_{\text{generalized part guidance}}$$

* FiLM: Feature-wise Linear Modulation

Learning Transferable Visual Models From Natural Language Supervision (ICML 2021)
FiLM: Visual Reasoning with a General Conditioning Layer (AAAI 2018)

Attention Control for Ambiguity and Omission [1/1]

Self-attention activation maps

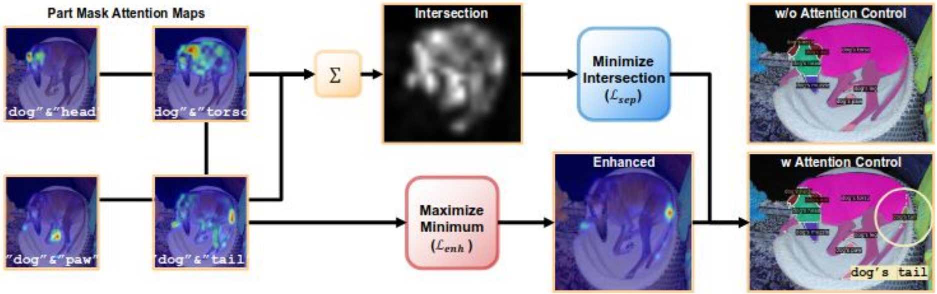
- visual tokens belonging to the same object-specific part mask should exhibit inter-similarity characteristics
- average self-attention map
  - $$\mathcal{A}_{\mathcal{M}_\mathbf{c}} = \frac{1}{|\mathcal{M}_\mathbf{c}|} \sum_{(h,w)\in\mathcal{M}_\mathbf{c}} \left(\mathcal{A}_{\mathbf{c}_{\text{obj}}}[h,w,:,:] + \mathcal{A}_{\mathbf{c}_{\text{part}}}[h,w,:,:]\right)$$

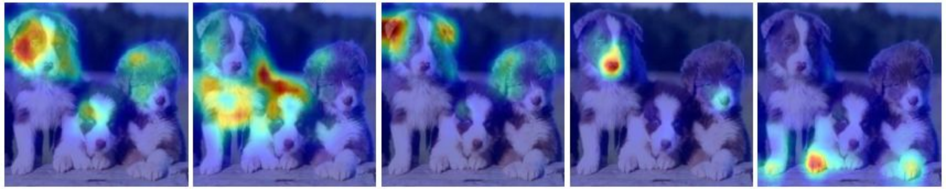
Minimizing Part Overlaps for Ambiguity

- mitigates the ambiguity issue in part boundaries
- parts with minimized intersection (binarize)
  - $$\mathcal{L}_{\text{sep}} = \frac{1}{|\mathbf{C}|} \left| \frac{\{(h,w) \mid \sum_{\mathbf{c}\in\mathbf{C}} \mathcal{B}_{\mathcal{M}_\mathbf{c}}(h,w) > 1\}}{\{(h,w) \mid \sum_{\mathbf{c}\in\mathbf{C}} \mathcal{B}_{\mathcal{M}_\mathbf{c}}(h,w) \geq 1\}} \right|$$

Enhancing Part Activation for Omission

- the minimum activation of the part with the maximum value is enhanced
  - $$\mathcal{L}_{\text{enh}} = 1 - \min_{\mathbf{c}\in\mathbf{C}} \left(\max_{(h,w)\in\mathcal{M}_\mathbf{c}} \mathcal{A}_{\mathcal{M}_\mathbf{c}}[h,w] \right)$$





(e) "head" (f) "torso" (g) "ear" (h) "nose" (i) "paw"

Experimental Setups [1/1]

Datasets:

- Pascal-Part-116 (9.2k)
- ADE20K-Part-234 (8.3k)
- PartImageNet (40 classes from 158 categories)

Tasks:

- Zero-Shot Part Segmentation (Cross-category Part Segmentation)
  - e.g. Unseen: bird, car, dog, sheep, motorbike
- Cross-dataset Part Segmentation

Evaluation Protocols:

- Pred-All (w/o object mask)
- Oracle-Obj (w/ object mask)

인용 (진한 회색 2, Montserrat 8)
PASCAL-Part: Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts (CVPR 2014)

Performance Evaluation [1/3]

Quantitative Results

- Zero-shot Part Segmentation
  - Pascal-Part-116
  - ADE20K-Part-234
  - PartImageNet
- Cross-dataset Part Segmentation
  - PartImageNet → Pascal-Part-116

Table 4: Cross-dataset performance.

Method	Pred-All	Oracle-Obj
PartImageNet → Pascal-Part-116		
CLIPSeg [32, 46]	11.72	14.87
PartCLIPSeg (Ours)	**14.74** (+3.02)	**19.86** (+4.99)
ADE20K-Part-234 → Pascal-Part-116		
CLIPSeg [32, 46]	5.41	17.82
PartCLIPSeg (Ours)	**10.37** (+4.96)	**17.94** (+0.12)

Table 1: Comparison of zero-shot performance with state-of-the-art methods on Pascal-Part-116.

Method	Backbone	Pred-All			Oracle-Obj		
		Seen	Unseen	Harmonic	Seen	Unseen	Harmonic
ZSSeg+ [52]	ResNet-50	<u>38.05</u>	3.38	6.20	**54.43**	19.04	28.21
VLPart [40]	ResNet-50	35.21	9.04	14.39	42.61	18.70	25.99
CLIPSeg [32, 46]	ViT-B/16	27.79	13.27	17.96	48.91	27.54	<u>35.24</u>
CAT-Seg [11, 46]	ViT-B/16	28.17	**25.42**	<u>26.72</u>	36.20	<u>28.72</u>	32.03
PartCLIPSeg (Ours)	ViT-B/16	**43.91**$\pm$0.45	<u>23.56</u>$\pm$0.21	**30.67**$\pm$0.09 (+3.94)	<u>50.02</u>$\pm$0.51	**31.67**$\pm$0.29	**38.79**$\pm$0.13 (+3.55)

[1] The best score is **bold** and the second-best score is <u>underlined</u>. The standard error of an average of 5 results is reported. These are the same for all experiments.

Table 2: Comparison of zero-shot performance with state-of-the-art methods on ADE20K-Part-234.

Method	Backbone	Pred-All			Oracle-Obj		
		Seen	Unseen	Harmonic	Seen	Unseen	Harmonic
ZSSeg+ [52]	ResNet-50	**32.20**	0.89	1.74	**43.19**	27.84	33.85
CLIPSeg [32, 46]	ViT-B/16	3.14	0.55	0.93	38.15	<u>30.92</u>	<u>34.15</u>
CAT-Seg [11, 46]	ViT-B/16	7.02	2.36	3.53	28.01	21.24	24.16
PartCLIPSeg (Ours)	ViT-B/16	<u>14.15</u>$\pm$0.51	**9.52**$\pm$0.13	**11.38**$\pm$0.10 (+7.85)	<u>38.37</u>$\pm$0.14	**38.82**$\pm$0.31	**38.60**$\pm$0.08 (+4.45)

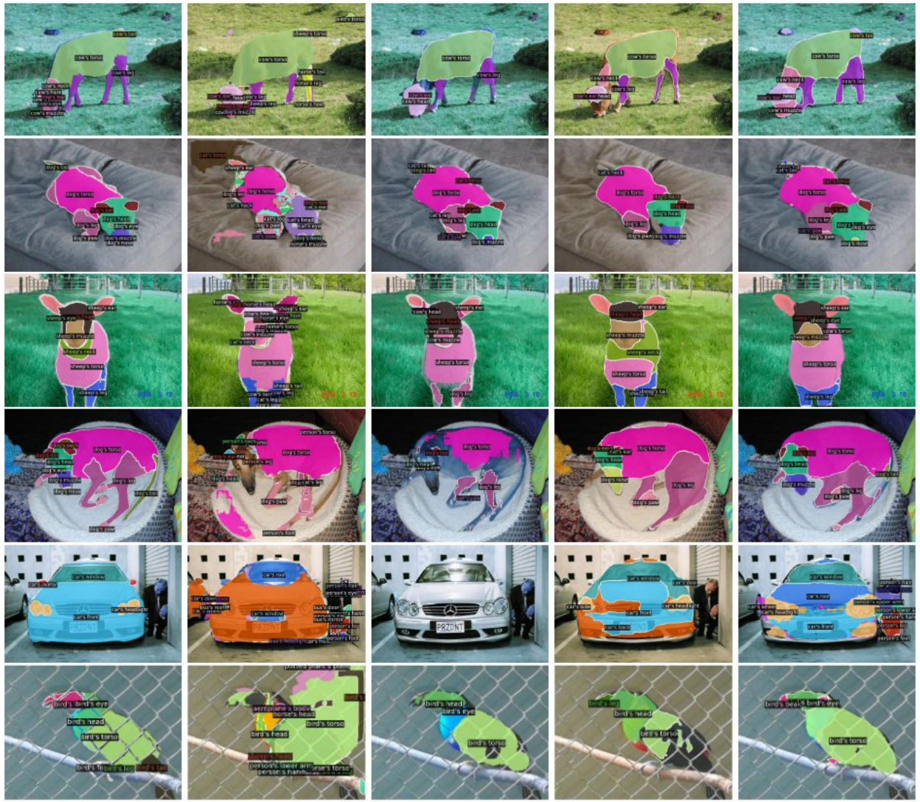
Table 3: Comparison of zero-shot performance with state-of-the-art method on PartImageNet.

Method	Backbone	Pred-All			Oracle-Obj		
		Seen	Unseen	Harmonic	Seen	Unseen	Harmonic
CLIPSeg [32, 46]	ViT-B/16	32.39	12.27	17.80	53.91	37.17	44.00
PartCLIPSeg (Ours)	ViT-B/16	**38.82**$\pm$0.74	**19.47**$\pm$0.45	**25.94**$\pm$0.32 (+8.14)	**56.26**$\pm$0.29	**51.65**$\pm$0.62	**53.85**$\pm$0.37 (+9.85)

Performance Evaluation [2/3]

Qualitative Results [1/2]

- Pred-All



(a) Ground-truth (b) VLPart [40] (c) CLIPSeg [38, 46] (d) CAT-Seg [11, 46] (e) PartCLIPSeg (Ours)

Figure 5: Qualitative results of zero-shot part segmentation on Pascal-Part-116 in **Pred-All** setting. Annotations for unseen categories (bird, car, dog, sheep, etc.) are not included in the train set.

Performance Evaluation [3/3]

Qualitative Results [2/2]

- Oracle-obj

(a) Ground-truth (b) VLPart [40] (c) CLIPSeg [38, 46] (d) CAT-Seg [11, 46] (e) PartCLIPSeg (Ours)

Figure 6: Qualitative results of zero-shot part segmentation on Pascal-Part-116 in **Oracle-Obj** setting.

4. Experiments [5/5]

Ablation Study [1/1]

- Separation & Enhance Losses
  - both separation and enhance losses in improving performance

- Impact of PartCLIPSeg for Underrepresented Parts
  - (small parts)

Figure A2: Comparison of results using only $\mathcal{L}_{sep}$ (top) with both $\mathcal{L}_{sep}$ and $\mathcal{L}_{enh}$ (bottom). The heatmap illustrates attention activation for the "sheep's neck" class.

Table 5: Impact of attention control losses.

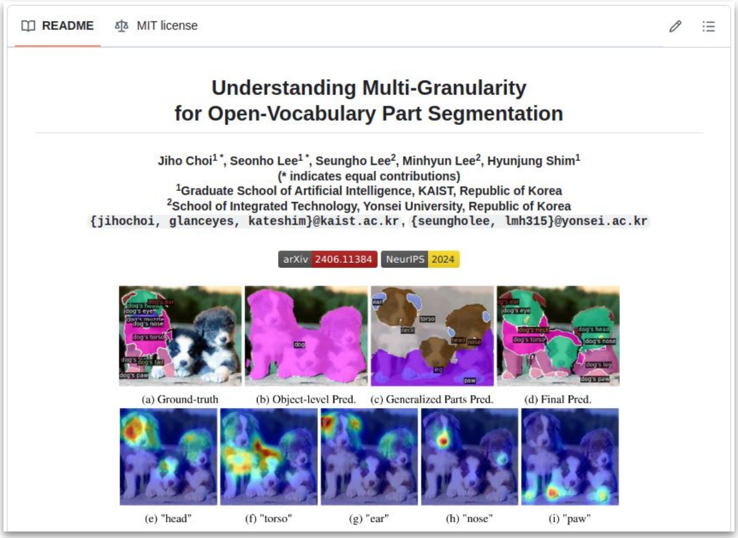
Loss		Pred-All			Oracle-Obj		
$\mathcal{L}_{sep}$	$\mathcal{L}_{enh}$	Seen	Unseen	Harmonic	Seen	Unseen	Harmonic
Pascal-Part-116							
✘	✘	43.86	21.89	29.20	49.09	31.26	38.20
✔	✘	**44.01**	23.18	30.37	**50.37**	31.45	38.72
✔	✔	43.91	**23.56**	**30.67**	50.02	**31.67**	**38.79**
ADE20K-Part-234							
✘	✘	10.86	8.33	9.43	37.39	36.49	36.93
✔	✘	12.78	9.38	10.82	**39.46**	36.04	37.67
✔	✔	**14.15**	**9.52**	**11.38**	38.37	**38.82**	**38.60**

Table 7: Impact of PartCLIPSeg for small parts on Pascal-Part-116 in Oracle-Obj setting. (mIoU)

Part: "eye"	bird	cat	cow	dog	sheep	person
CLIPSeg [32, 46]	**3.33**	18.77	3.65	16.05	0.00	15.30
PartCLIPSeg (Ours)	1.95	**31.01**	**28.16**	**32.79**	**0.67**	**29.16**
Part: "neck"	bird	cat	cow	dog	sheep	person
CLIPSeg [32, 46]	19.09	6.57	0.78	8.12	8.47	30.93
PartCLIPSeg (Ours)	**32.51**	**12.00**	**2.75**	**16.37**	**18.80**	**50.71**
Part: "leg"	bird	cat	cow	dog	sheep	person
CLIPSeg [32, 46]	19.61	38.62	27.85	39.34	52.63	52.67
PartCLIPSeg (Ours)	**31.12**	**44.82**	**63.78**	**41.55**	**54.73**	**55.35**

Conclusion

- introduced SOTA OVPS method: PartCLIPSeg
- utilizes generalized parts and object-level guidance
- separates parts by minimizing their overlaps in attention maps, handling ambiguous part boundaries
- enhanced loss function to improve the detection of underrepresented parts
- Code: https://github.com/kaist-cvml/part-clipseg/


README MIT license

Understanding Multi-Granularity for Open-Vocabulary Part Segmentation

Jiho Choi[1 *], Seonho Lee[1 *], Seungho Lee[2], Minhyun Lee[2], Hyunjung Shim[1]
(* indicates equal contributions)
[1]Graduate School of Artificial Intelligence, KAIST, Republic of Korea
[2]School of Integrated Technology, Yonsei University, Republic of Korea
{jihochoi, glanceyes, kateshim}@kaist.ac.kr, {seungholee, lmh315}@yonsei.ac.kr

arXiv 2406.11384 NeurIPS 2024

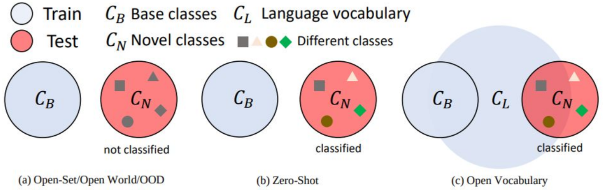
(a) Ground-truth (b) Object-level Pred. (c) Generalized Parts Pred. (d) Final Pred.

(e) "head" (f) "torso" (g) "ear" (h) "nose" (i) "paw"

Appendix

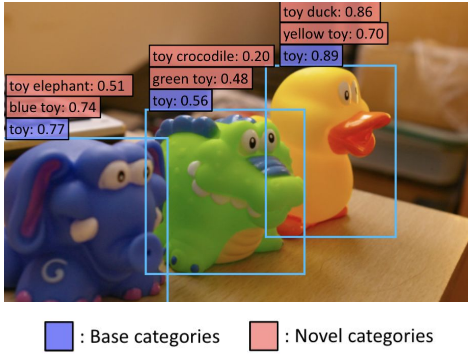
2. Related Work [1/3]

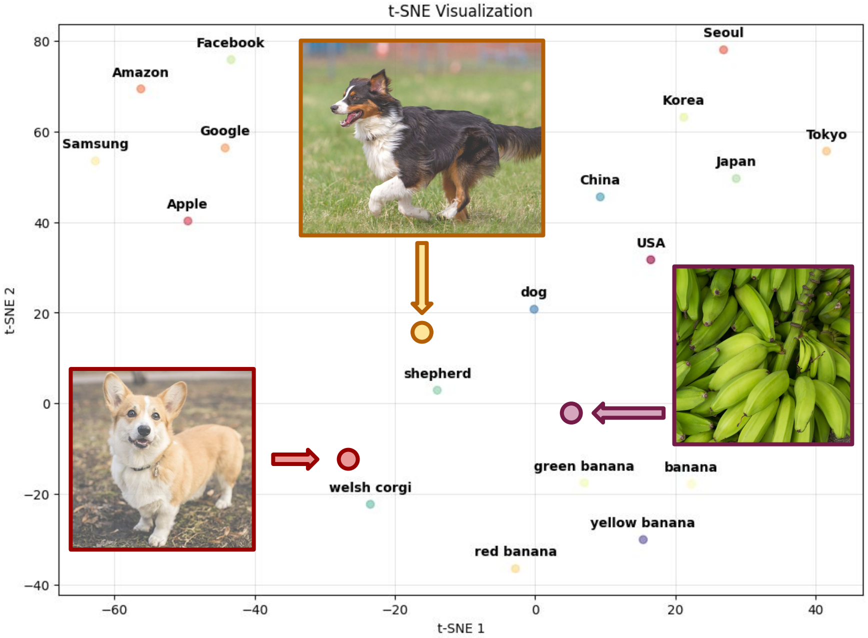
VLM & Open-Vocabulary



Open-Vocabulary & Zero-shot Learning

- Predicting **unseen** (**novel**) category names, not present in the train set (**seen**, **base**), is possible using the embedding space of Vision-Language Models (VLMs) like CLIP





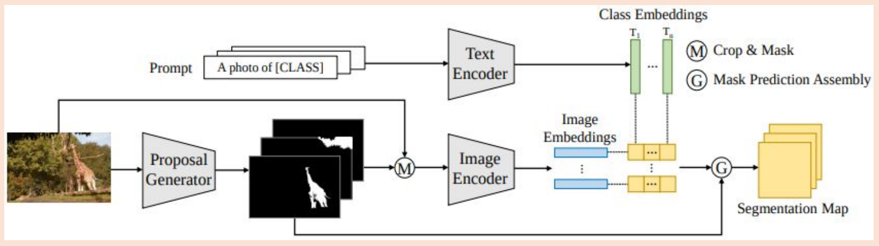
Open-vocabulary Object Detection via Vision and Language Knowledge Distillation (ICLR 2022)
Learning to Generate Text-grounded Mask for Open-world Semantic Segmentation from Only Image-Text Pairs (CVPR 2023)
Towards Open Vocabulary Learning: A Survey (PAMI 2024)

2. Related Work [2/3]

Open-Vocabulary Dense Prediction

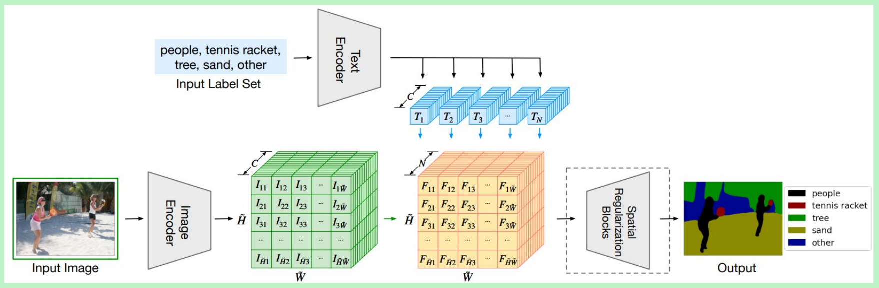
Two-Stage OVSS / OVIS

- ZSSeg (ECCV 2022)
    - **class-agnostic mask proposal**
    - classify proposals with CLIP embedding



One-Stage OVSS

- LSeg (ICLR 2022), MaskCLIP (ECCV 2022)
        - (language-driven semantic segmentation)
    - **image-text similarity (alignment)**
    - ViT attention map as saliency
        - image-text cross-attention



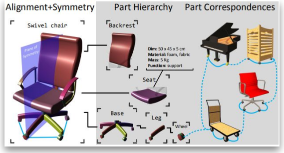
* Open-vocabulary Semantic Segmentation (OVSS)
** Open-vocabulary Instance Segmentation (OVIS)

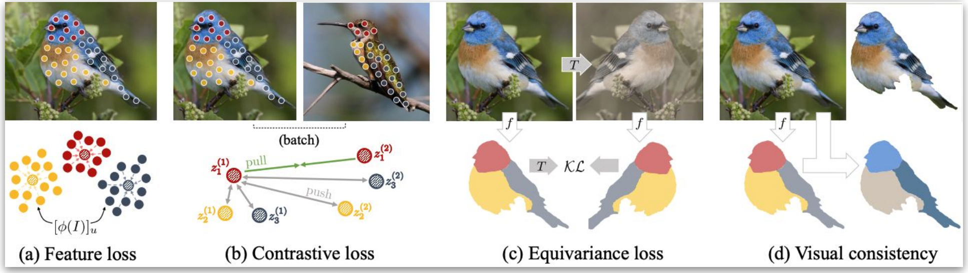
ZSSeg: A Simple Baseline for Open-Vocabulary Semantic Segmentation with Pre-trained Vision-language Model (ECCV 2022)
LSeg: Language-driven Semantic Segmentation (ICLR 2022)
MaskCLIP: Extract Free Dense Labels from CLIP (ECCV 2022)

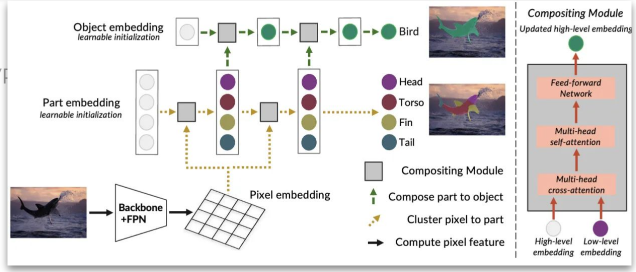
2. Related Work [3/3]

Part Segmentation

- Datasets
  - CUB-200-2011 (2011)
  - PASCAL-Part (CVPR 2014)
  - DeepFashion (CVPR 2016)
  - PartImageNet (ECCV 2022)
  - OV-PARTS (NeurIPS B&D 2023)
  - SubPartImageNet (ECCV 2024)
  - ShapeNet (2015) / PartNet-Mobility (CVPR 2019)
- Tasks (≈)
  - Part Segmentation
  - Fine-grained Segmentation
  - Part Discovery (unsupervised, clustering)
  - 3D Part Segmentation
- → fine-grained understanding







DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations (CVPR 2016)
PASCAL-Part: Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts (CV
PartImageNet: A Large, High-Quality Dataset of Parts (ECCV 2022)
PACO: Parts and Attributes of Common Objects (CVPR 2023)
OV-PARTS: Towards Open-Vocabulary Part Segmentation (NeurIPS Datasets and Benchmarks Track 2023)
SPIN: Hierarchical Segmentation with Subpart Granularity in Natural Images (ECCV 2024)

Unsupervised Part Discovery from Contrastive Reconstruction (NeurIPS 2021)
Compositor: Bottom-up Clustering and Compositing for Robust Part and Object Segmentation (CVPR 2023)

PartCLIPSeg [1/1]