CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation Supplementary Materials

Seokju Cho[1,*] Heeseong Shin[1,*] Sunghwan Hong[1]
Anurag Arnab[2] Paul Hongsuck Seo[1,†] Seungryong Kim[1,†]

[1]Korea University [2]Google Research

{seokju_cho, hsshin98, sung_hwan, phseo, seungryong_kim}@korea.ac.kr
aarnab@google.com

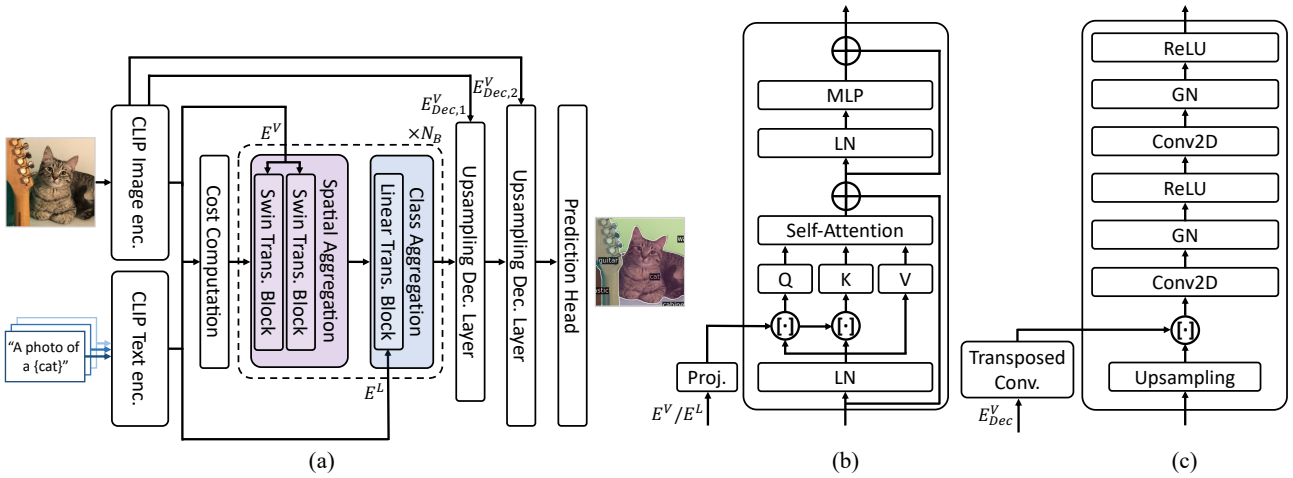

Figure 1. **More architectural details of CAT-Seg:** (a) overall architecture. (b) embedding guidance. Note that a generalized embedding guidance is illustrated to include different attention designs, *i.e.*, shifted window attention [21] or linear attention [17]. (c) upsampling decoder layer. GN: Group Normalization [37]. LN: Layer Normalization [1].

In the following, we provide the full results from MESS [4] in Section A. We further provide implementation details in Section B. We then provide additional experimental results and ablation study in Section C. Finally, we present qualitative results for the benchmarks in Section D and a discussion of limitations in Section E.

A. More Results

Full quantitative results on MESS benchmark. In Table 1, we provide the results of all 22 datasets within MESS [4], including results from Grounded-SAM [13].

B. More Details

B.1. Architectural Details

In the following, we provide more architectural details. Our detailed overall architecture is illustrated in Fig. 1 (a).

Embedding guidance. In this paragraph, we provide more details of embedding guidance, which is designed to facilitate the cost aggregation process by exploiting its rich semantics for a guidance. We first extract visual and text embeddings from CLIP encoders [26]. The embeddings then undergo linear projection and concatenated to the cost volume before query and key projections in aggregation layer. The design is illustrated in Fig. 1 (b).

Upsampling decoder. The detailed architecture is illustrated in Fig. 1(c). In our upsampling decoder, we start by taking high-resolution features from the CLIP ViT model [9]. We then apply a single transposed convolution layer to these extracted features to generate an upsampled feature map. Initially, the extracted feature maps have a resolution of 24×24 pixels. However, after processing them with the transposed convolution operation, we increase their resolution to 48×48 pixels for the first feature map, denoted as $E^V_{Dec,1}$, and to 96×96 pixels for the second feature map, denoted as $E^V_{Dec,2}$.

To obtain $E^V_{Dec,1}$, we utilize the output of the 8th layer

*Equal contribution. †Corresponding authors.

	General						Earth Monitoring					Medical Sciences				Engineering				Agri. and Biology			
	BDD100K	Dark Zurich	MHP v1	FoodSeg103	ATLANTIS	DRAM	iSAID	ISPRS Pots.	WorldFloods	FloodNet	UAVid	Kvasir-Inst.	CHASE DB1	CryoNuSeg	PAXRay-4	Corrosion CS	DeepCrack	PST900	ZeroWaste-f	SUIM	CUB-200	CWFID	Mean
Random (LB)	*1.48*	*1.31*	*1.27*	*0.23*	*0.56*	*2.16*	*0.56*	*8.02*	*18.43*	*3.39*	*5.18*	*27.99*	*27.25*	*31.25*	*31.53*	*9.3*	*26.52*	*4.52*	*6.49*	*5.3*	*0.06*	*13.08*	*10.27*
Best sup. (UB)	*44.8*	*63.9*	*50.0*	*45.1*	*42.22*	*45.71*	*65.3*	*87.56*	*92.71*	*82.22*	*67.8*	*93.7*	*97.05*	*73.45*	*93.77*	*49.92*	*85.9*	*82.3*	*52.5*	*74.0*	*84.6*	*87.23*	*70.99*
ZSSeg-B	32.36	16.86	7.08	8.17	22.19	33.19	3.8	11.57	23.25	20.98	30.27	46.93	37.0	**38.7**	44.66	3.06	25.39	18.76	8.78	30.16	4.35	32.46	22.73
ZegFormer-B	14.14	4.52	4.33	10.01	18.98	29.45	2.68	14.04	25.93	22.74	20.84	27.39	12.47	11.94	18.09	4.78	29.77	19.63	17.52	28.28	16.8	32.26	17.57
X-Decoder-T	47.29	24.16	3.54	2.61	27.51	26.95	2.43	31.47	26.23	8.83	25.65	55.77	10.16	11.94	15.23	1.72	24.65	19.44	15.44	24.75	0.51	29.25	19.8
SAN-B	37.4	24.35	8.87	19.27	36.51	49.68	4.77	37.56	31.75	37.44	**41.65**	69.88	17.85	11.95	19.73	3.13	50.27	19.67	**21.27**	22.64	**16.91**	5.67	26.74
OpenSeeD-T	**47.95**	28.13	2.06	9.0	18.55	29.23	1.45	31.07	30.11	23.14	39.78	59.69	**46.68**	33.76	37.64	13.38	47.84	2.5	2.28	19.45	0.13	11.47	24.33
Gr.-SAM-B	41.58	20.91	**29.38**	10.48	17.33	57.38	12.22	26.68	33.41	19.19	38.34	46.82	23.56	38.06	41.07	**20.88**	**59.02**	**21.39**	16.74	14.13	0.43	38.41	28.52
CAT-Seg-B	46.71	**28.86**	23.74	**26.69**	**40.31**	**65.81**	**19.34**	**45.36**	**35.72**	**37.57**	41.55	48.2	16.99	15.7	31.48	12.29	31.67	19.88	17.52	**44.71**	10.23	**42.77**	**31.96**
OVSeg-L	45.28	22.53	6.24	16.43	33.44	53.33	8.28	31.03	31.48	35.59	38.8	71.13	20.95	13.45	22.06	6.82	16.22	21.89	11.71	38.17	14.0	33.76	26.94
SAN-L	43.81	30.39	9.34	24.46	40.66	68.44	11.77	**51.45**	**48.24**	39.26	**43.41**	**72.18**	7.64	11.94	29.33	6.83	23.65	19.01	18.32	40.01	19.3	1.91	30.06
Gr.-SAM-L	42.69	21.92	28.11	10.76	17.63	60.8	12.38	27.76	33.4	19.28	39.37	47.32	**25.16**	**38.06**	**44.22**	**20.88**	**58.21**	21.23	16.67	14.3	0.43	38.47	29.05
CAT-Seg-L	**47.87**	**34.96**	**32.54**	**33.31**	**45.61**	**73.82**	**20.58**	50.81	46.42	**41.36**	40.79	61.13	3.72	11.94	22.02	11.03	19.9	**22.0**	**27.87**	**53.0**	**22.93**	**39.91**	**34.7**

Table 1. **Full results of quantitative evaluation on MESS [4].**

Figure 2. **Illustration of the patch inference.** During inference, we divide the input image into patches, thereby increasing the effective resolution.

for the ViT-B/16 model, and for the ViT-L/14 model, we use the output of the 16th layer. For the extraction of $E^V_{Dec,2}$, we employ shallower features: the output of the 4th layer for the ViT-B/16 model as a VLM, and the output of the 8th layer for the ViT-L/14 model. These features are employed to enhance cost embeddings with fine details using a U-Net-like architecture [28].

B.2. Other Implementation Details

Training details. A resolution of $H = W = 24$ is used during training for constructing cost volume. The position embeddings of the CLIP image encoder is initialized with bicubic interpolation [33], and we set training resolution as 384×384. For ViT-B and ViT-L variants, we initialize CLIP [26] with official weights of ViT-B/16 and ViT-L/14@336px respectively. All hyperparameters are kept constant across the evaluation datasets.

Text prompt templates. To obtain text embeddings from the text encoder, we form sentences with the class names, such as `"A photo of a {class}"`. We do not explore handcrafted prompts in this work, but it is open for future investigation.

B.3. Patch Inference

The practicality of Vision Transformer (ViT) [9] for high-resolution image processing has been limited due to its quadratic complexity with respect to the sequence length. As our model leverages ViT to extract image embeddings, CAT-Seg may struggle to output to the conventional image resolutions commonly employed in semantic segmentation literature, such as 640×640 [6, 12], without sacrificing some accuracy made by losing some fine-details. Although we can adopt the same approach proposed in [42] to upsample the positional embedding [42], we ought to avoid introducing excessive computational burdens, and thus adopt an effective inference strategy without requiring additional training which is illustrated in Fig. 2.

To this end, we begin by partitioning the input image into overlapping patches of size $\frac{H}{N_P} \times \frac{W}{N_P}$. Intuitively, given an image size of 640×640, we partition the image to sub-images of size 384×384, which matches the image resolution at training phase, and each sub-images has overlapping regions 128×128. Subsequently, we feed these sub-images and the original image that is resized to 384×384 into the model. Given the results for each patches and the image, we merge the obtained prediction, while the overlapping regions are averaged to obtain the final prediction. In practice, we employ $N_P = 2$, while adjusting the overlapping region to match the effective resolution of 640×640.

B.4. More Details of MESS Benchmark

In Table 2, we provide details of the datasets in the MESS benchmark [4].

C. Additional Ablation Study

C.1. Ablation Study of Inference Strategy

Methods	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
CAT-Seg w/ training reso.	14.6	22.1	35.7	60.9	96.3	79.9
Ours	**16.0**	**23.8**	**37.9**	**63.3**	**97.0**	**82.5**

Table 3. **Ablation study of inference strategy.** CLIP with ViT-L is used for ablation.

Table 3 presents effects of different inference strategies for our model. The first row shows the results using the training resolution at inference time. The last row adopts the proposed patch inference strategy. It is shown that our pro-

Dataset	Link	Licence	Split	# of classes	Classes
BDD100K [39]	berkeley.edu	custom	val	19	[road; sidewalk; building; wall; fence; pole; traffic light; traffic sign; ...]
Dark Zurich [29]	ethz.ch	custom	val	20	[unlabeled; road; sidewalk; building; wall; fence; pole; traffic light; ...]
MHP v1 [18]	github.com	custom	test	19	[others; hat; hair; sunglasses; upper clothes; skirt; pants; dress; ...]
FoodSeg103 [36]	github.io	Apache 2.0	test	104	[background; candy; egg tart; french fries; chocolate; biscuit; popcorn; ...]
ATLANTIS [10]	github.com	Flickr (images)	test	56	[bicycle; boat; breakwater; bridge; building; bus; canal; car; ...]
DRAM [7]	ac.il	custom (in download)	test	12	[bird; boat; bottle; cat; chair; cow; dog; horse; ...]
iSAID [34]	github.io	Google Earth (images)	val	16	[others; boat; storage tank; baseball diamond; tennis court; bridge; ...]
ISPRS Potsdam [5]	isprs.org	no licence provided	test	6	[road; building; grass; tree; car; others]
WorldFloods [24]	github.com	CC NC 4.0	test	3	[land; water and flood; cloud]
FloodNet [27]	github.com	custom	test	10	[building-flooded; building-non-flooded; road-flooded; water; tree; ...]
UAVid [22]	uavid.nl	CC BY-NC-SA 4.0	val	8	[others; building; road; tree; grass; moving car; parked car; humans]
Kvasir-Inst. [16]	simula.no	custom	test	2	[others; tool]
CHASE DB1 [11]	kingston.ac.uk	CC BY 4.0	test	2	[others; blood vessels]
CryoNuSeg [23]	kaggle.com	CC BY-NC-SA 4.0	test	2	[others; nuclei in cells]
PAXRay-4 [30]	github.io	custom	test	4x2	[others, lungs], [others, bones], [others, mediastinum], [others, diaphragm]
Corrosion CS [3]	figshare.com	CC0	test	4	[others; steel with fair corrosion; ... poor corrosion; ... severe corrosion]
DeepCrack [20]	github.com	custom	test	2	[concrete or asphalt; crack]
PST900 [31]	github.com	GPL-3.0	test	5	[background; fire extinguisher; backpack; drill; human]
ZeroWaste-f [2]	ai.bu.edu	CC-BY-NC 4.0	test	5	[background or trash; rigid plastic; cardboard; metal; soft plastic]
SUIM [15]	umn.edu	MIT	test	8	[human diver; reefs and invertebrates; fish and vertebrates; ...]
CUB-200 [35]	caltech.edu	custom	test	201	[background; Laysan Albatross; Sooty Albatross; Crested Auklet; ...]
CWFID [14]	github.com	custom	test	3	[ground; crop seedling; weed]

Table 2. **Details of the datasets in the MESS benchmark [4].**

posed approach can bring large performance gains, compared to using the training resolution.

C.2. Ablation on VLM

VLM	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
EVA-02-CLIP-L/14 [32]	16.4	24.5	37.8	62.7	**97.9**	**83.7**
SigLIP-ViT-L/16 [40]	**18.0**	**26.1**	**39.1**	60.9	97.2	80.8
CLIP-ViT-L/14	16.0	23.8	37.9	**63.3**	97.0	82.5

Table 4. **Results on various VLMs.**

Table 4 shows the results with various VLMs. We found that CAT-Seg can be applied to various VLMs, and better results can be obtained when a more powerful model is applied.

D. More Qualitative Results

We provide more qualitative results on A-847 [41] in Fig. 3, PC-459 [25] in Fig. 4, A-150 [41] in Fig. 5, and PC-59 [25] in Fig. 6. We also further compare the results in A-847 [41] with other methods [8, 19, 38] in Fig. 7.

E. Limitations

To evaluate open-vocabulary semantic segmentation results, we follow [12, 19] and compute the metrics using the other segmentation datasets. However, since the ground-truth segmentation maps involve some ambiguities, the reliability of the evaluation dataset is somewhat questionable. Constructing a more reliable dataset including ground-truths accounting for above issue for accurate evaluation is an intriguing topic.

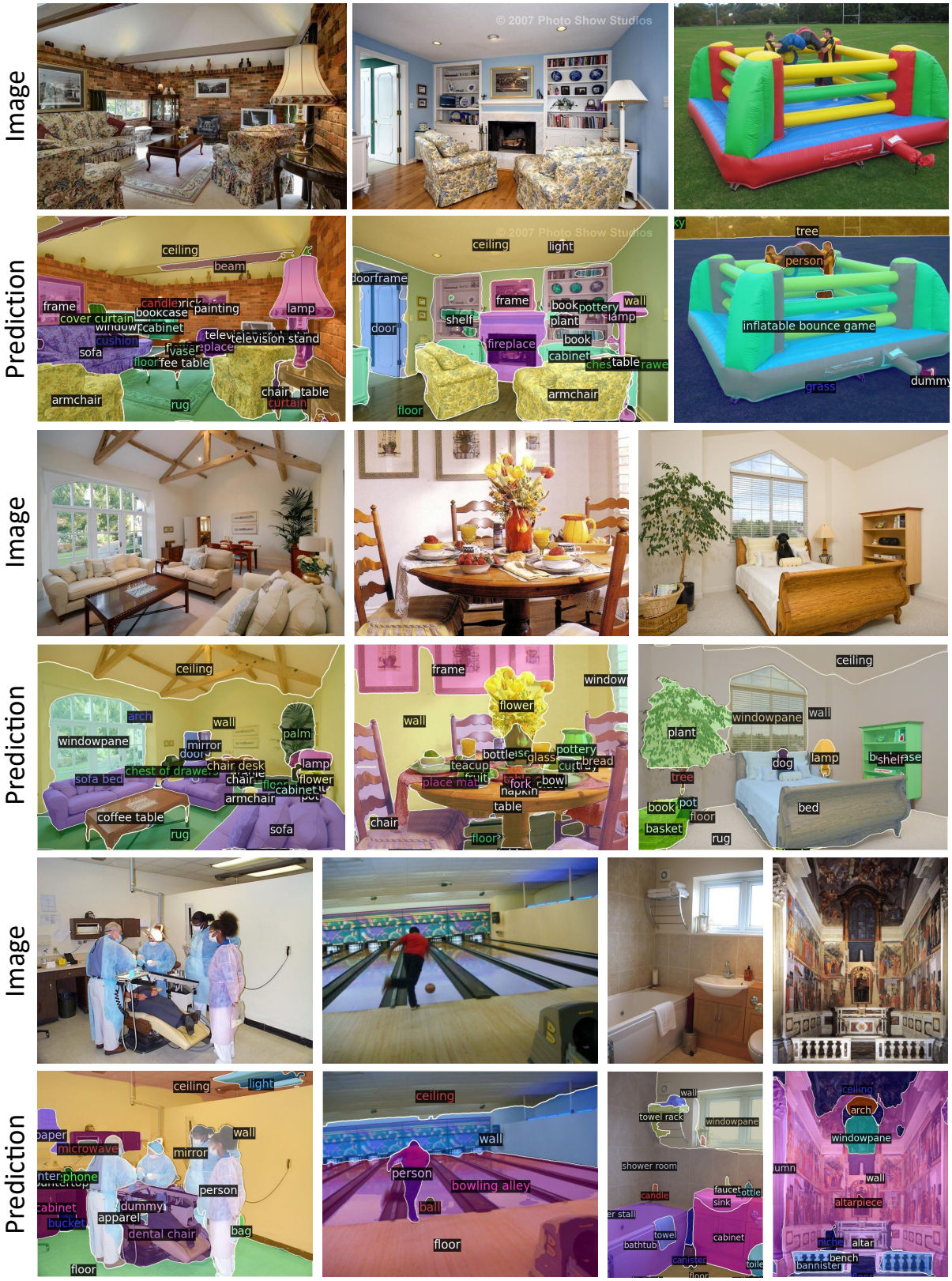

Figure 3. **Qualitative results on ADE20K [41] with 847 categories.**

Figure 4. **Qualitative results on PASCAL Context [25] with 459 categories.**

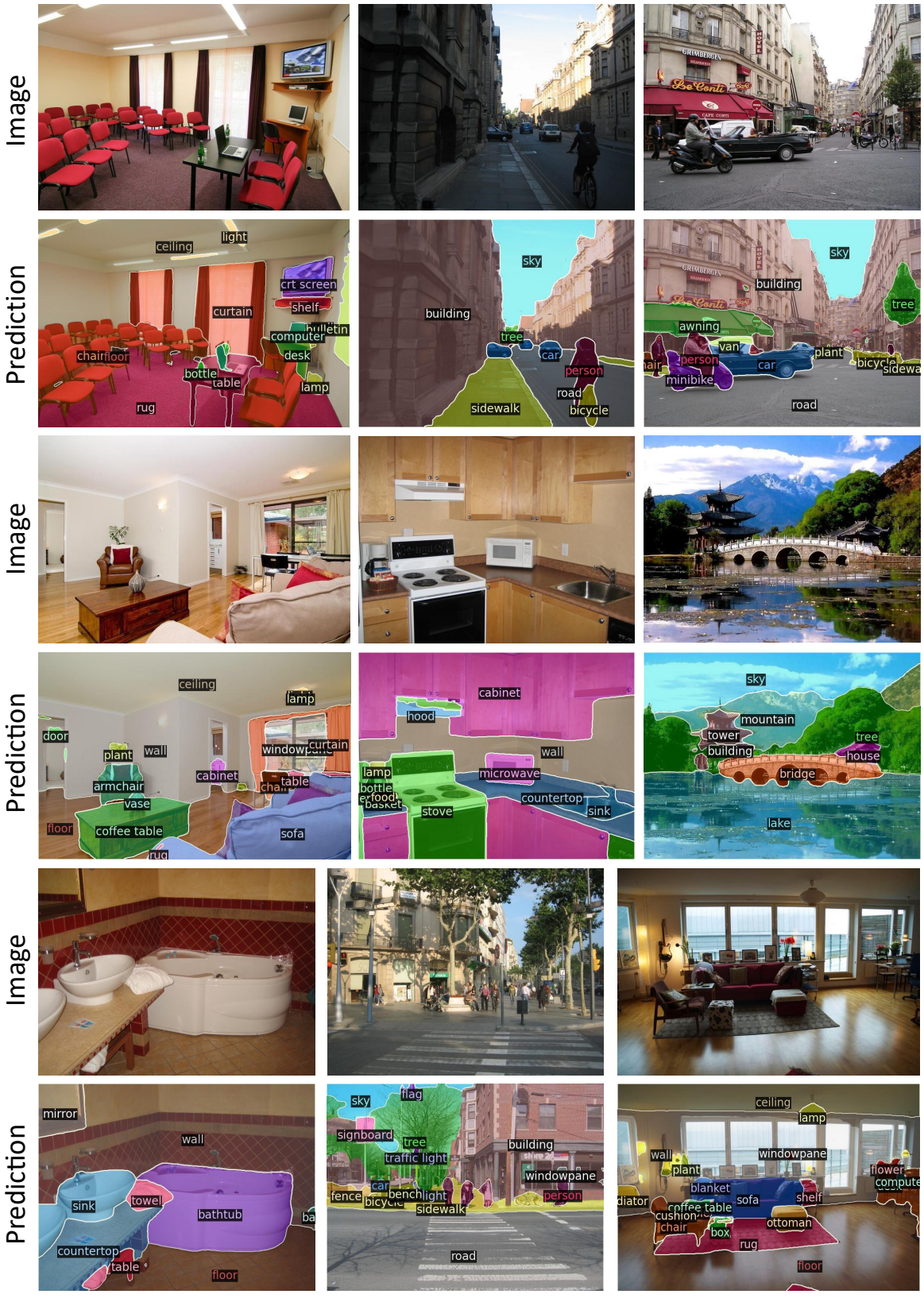

Figure 5. **Qualitative results on ADE20K [41] with 150 categories.**

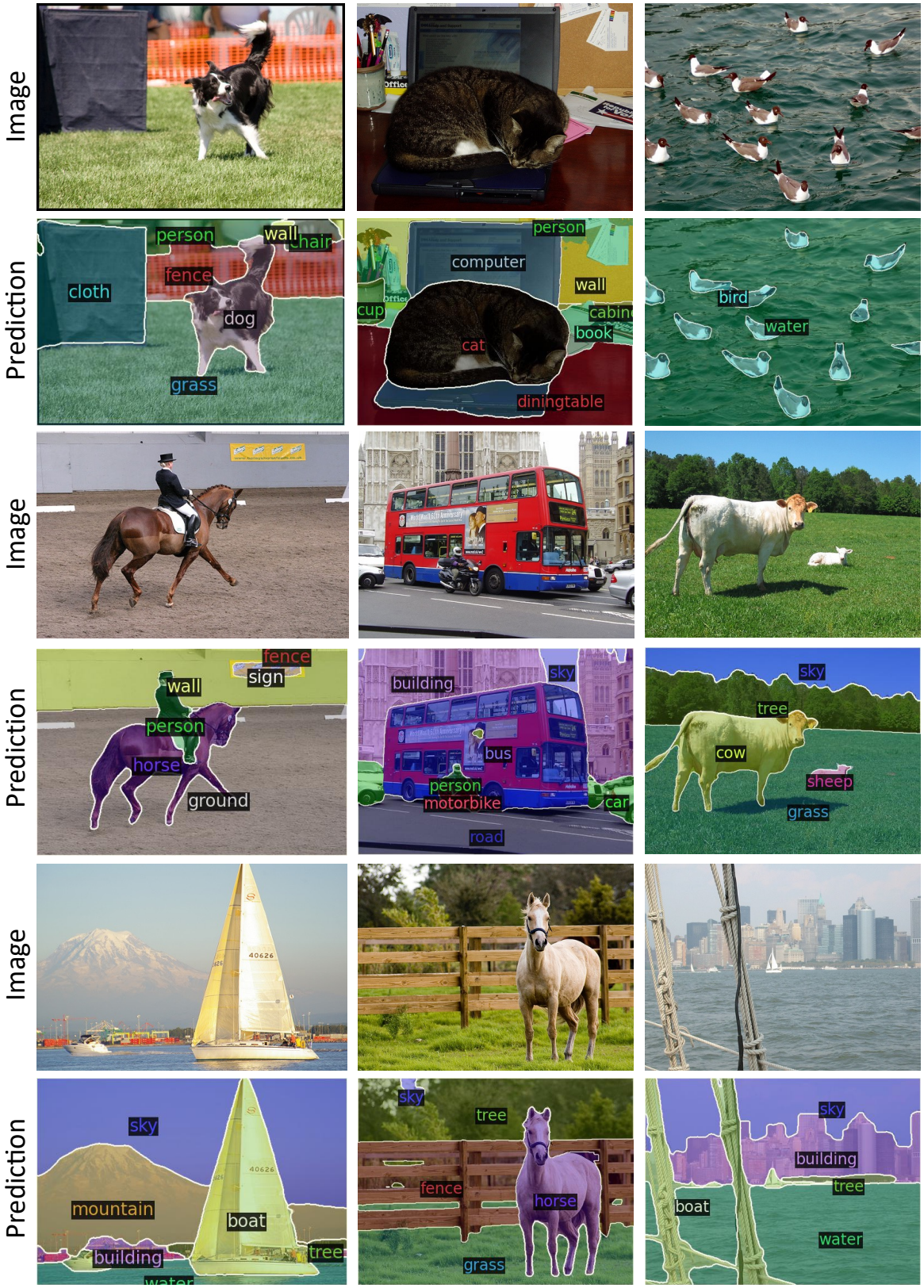

Figure 6. **Qualitative results on PASCAL Context [25] with 59 categories.**

Figure 7. **Comparison of qualitative results on ADE20K [41] with 847 categories.** We compare CAT-Seg with ZegFormer [8], ZSseg [38], and OVSeg [19] on A-847 dataset.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. Zerowaste dataset: towards deformable object segmentation in cluttered scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21147–21157, 2022. 3

[3] Eric Bianchi and Matthew Hebdon. Corrosion condition state semantic segmentation dataset. *University Libraries, Virginia Tech: Blacksburg, VA, USA*, 2021. 3

[4] Benedikt Blumenstiel, Johannes Jakubik, Hilde Kühne, and Michael Vössing. What a mess: Multi-domain evaluation of zero-shot semantic segmentation. *arXiv preprint arXiv:2306.15521*, 2023. 1, 2, 3

[5] BSF Swissphoto. Isprs potsdam dataset within the isprs test project on urban classification, 3d building reconstruction and semantic labeling. https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx, 2012. 3

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2

[7] Nadav Cohen, Yael Newman, and Ariel Shamir. Semantic segmentation in art paintings. In *Computer Graphics Forum*, pages 261–275. Wiley Online Library, 2022. 3

[8] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 3, 8

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[10] Seyed Mohammad Hassan Erfani, Zhenyao Wu, Xinyi Wu, Song Wang, and Erfan Goharian. Atlantis: A benchmark for semantic segmentation of waterbody images. *Environmental Modelling & Software*, 149:105333, 2022. 3

[11] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548, 2012. 3

[12] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 540–557. Springer, 2022. 2, 3

[13] Grounded-SAM Contributors. Grounded-Segment-Anything, 2023. 1

[14] Sebastian Haug and Jörn Ostermann. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part IV 13*, pages 105–116. Springer, 2015. 3

[15] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1769–1776. IEEE, 2020. 3

[16] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A Riegler, Thomas de Lange, Peter T Schmidt, Håvard D Johansen, et al. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27*, pages 218–229. Springer, 2021. 3

[17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020. 1

[18] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017. 3

[19] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. *arXiv preprint arXiv:2210.04150*, 2022. 3, 8

[20] Yahui Liu, Jian Yao, Xiaohu Lu, Renping Xie, and Li Li. Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing*, 338:139–153, 2019. 3

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[22] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020. 3

[23] Amirreza Mahbod, Gerald Schaefer, Benjamin Bancher, Christine Löw, Georg Dorffner, Rupert Ecker, and Isabella Ellinger. Cryonuseg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images. *Computers in biology and medicine*, 132:104349, 2021. 3

[24] Gonzalo Mateo-Garcia, Joshua Veitch-Michaelis, Lewis Smith, Silviu Vlad Oprea, Guy Schumann, Yarin Gal, Atılım Güneş Baydin, and Dietmar Backes. Towards global flood mapping onboard low cost satellites with machine learning. *Scientific reports*, 11(1):7249, 2021. 3

[25] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 3, 5, 7

[26] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[27] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. 3

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[29] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019. 3

[30] Constantin Seibold, Simon Reiß, Saquib Sarfraz, Matthias A Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H Maier-Hein, Jens Kleesiek, and Rainer Stiefelhagen. Detailed annotations of chest x-rays via ct projection for report understanding. *arXiv preprint arXiv:2210.03416*, 2022. 3

[31] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447. IEEE, 2020. 3

[32] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3

[33] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[34] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 3

[35] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 3

[36] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 506–515, 2021. 3

[37] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

[38] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 736–753. Springer, 2022. 3, 8

[39] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 3

[40] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. 3

[41] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 3, 4, 6, 8

[42] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pages 696–712. Springer, 2022. 2