Images as Noisy Labels: Unleashing the Potential of the Diffusion Model for Open-Vocabulary Semantic Segmentation
Supplementary Materials

Fan Li [1] Xuanbin Wang [1] Xuan Wang[1] Zhaoxiang Zhang[1] Yuelei Xu[1]*
[1]Northwestern Polytechnical University

The supplementary material is organized as follows: Section A provides additional implementation details for stable diffusion and a comprehensive overview of the MESS benchmark. Section B extends the comparison with more diffusion-based methods. Section C presents the complete results across all 22 datasets in MESS. Section D explores the quantitative impact of incorporating different vision-language foundation models into our framework. In Section E, we provide further analysis of ablation experiments. Section F shows more visualization results of the proposed method.

A. Implementation Details

A.1. The Details of Stable Diffusion

The pre-trained weight of the stable diffusion model is available at `https://huggingface.co/stabilityai/stable-diffusion-2`.

A.2. More Details of MESS Benchmark

As shown in Table 7, we provide detailed information about the MESS benchmark, which includes 22 datasets comprising 448 classes and 25,079 images. The benchmark spans four distinct data types: visible spectrum, multispectral, microscopic, and electromagnetic. This diversity reflects a broad range of real-world applications, enabling a thorough evaluation of model performance across both general-purpose tasks and specialized domain-specific scenarios.

B. Extended Comparison with Diffusion-based Methods

In the main paper, we compare our method with the previous SOTA approach, ODISE. In Table 1, we extend our comparison to include additional diffusion-based open-vocabulary semantic segmentation approaches, providing a more comprehensive evaluation. Notably, even with these expanded comparisons, our method continues to outperform all previous works by a significant margin, further validating its effectiveness.

Model	PC-459	A-150	PC-59	PAS-20	PAS-20^b
Dataset Diffusion [22]	-	-	-	-	60.2
OVDiff [14]	-	14.1	32.9	80.9	69.0
OVAM [20]	-	-	-	-	82.5
ProxyCLIP [15]	-	22.6	37.7	83.2	60.6
FreeDA [1]	-	23.2	43.5	87.9	-
DEDOS (Ours)	**25.6**	**39.4**	**65.7**	**97.6**	**84.6**

Table 1. Quantitative comparison with previous diffusion-based open-vocabulary semantic segmentation approaches.

VLM	PC-459	A-150	PC-59	PAS-20	PAS-20^b
EVA-02-B [9]	22.1	33.5	60.4	95.1	81.3
EVA-02-L [9]	**26.5**	**40.1**	64.9	97.0	**85.4**
CLIP-ViT-L	25.6	39.4	**65.7**	**97.6**	84.6

Table 2. Comparison of performance using different VLMs as backbones with the proposed method.

C. Full Quantitative Results on MESS

To comprehensively validate the effectiveness of our method, we present the complete test results on the MESS dataset, as shown in Table 3. It can be observed that our method achieves optimal results on most of the datasets, highlighting its adaptability and robustness. Notably, it excels in general domains as well as in agriculture and biology, significantly outperforming all previous state-of-the-art methods. However, in a few specific cases, such as CHASE DB1 and PST900—which consist of microscopic and electromagnetic images, respectively—the performance does not reach optimal levels. We attribute this to the diffusion model's limited prior knowledge of these specialized spectral domains, which poses challenges in capturing the unique characteristics of such images. Despite these isolated cases, the overall results strongly underscore the versatility and effectiveness of our method, showcasing

*Corresponding author.

	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
Random (LB)	*1.5*	*1.3*	*1.3*	*0.2*	*0.6*	*2.2*	*0.6*	*8.0*	*18.4*	*3.4*	*5.2*	*28.0*	*27.3*	*31.3*	*31.5*	*9.3*	*26.5*	*4.5*	*6.5*	*5.3*	*0.1*	*13.1*
Best sup. (UB)	*44.8*	*63.9*	*50.0*	*45.1*	*42.2*	*45.7*	*65.3*	*87.6*	*92.7*	*82.2*	*67.8*	*93.7*	*97.1*	*73.5*	*93.8*	*49.9*	*85.9*	*82.3*	*52.5*	*74.0*	*84.6*	*87.2*
ZSSeg-B	32.4	16.9	7.1	8.2	22.2	33.2	3.8	11.6	23.3	21.0	30.3	46.9	37.0	**38.7**	**44.7**	3.1	25.4	18.8	8.8	30.2	4.4	32.5
ZegFormer-B	14.1	4.5	4.3	10.0	19.0	29.5	2.7	14.0	25.9	22.7	20.8	27.4	12.5	11.9	18.1	4.8	29.8	19.6	17.5	28.3	16.8	32.3
X-Decoder-T	47.3	24.2	3.5	2.6	27.5	27.0	2.4	31.5	26.2	8.8	25.7	55.8	10.2	11.9	15.2	1.7	24.7	19.4	15.4	24.8	0.5	29.3
SAN-B	37.4	24.4	8.9	19.3	36.5	49.7	4.8	37.6	31.8	37.4	41.7	**69.9**	17.9	12.0	19.7	3.1	50.3	19.7	21.3	22.6	16.9	5.7
OpenSeeD-T	48.0	28.1	2.1	9.0	18.6	29.2	1.5	31.1	30.1	23.1	39.8	59.7	**46.7**	33.8	37.6	13.4	47.8	2.5	2.3	19.5	0.1	11.5
Gr.-SAM-B	41.6	20.9	**29.4**	10.5	17.3	57.4	12.2	26.7	33.4	19.2	38.3	46.8	23.6	38.1	41.1	**20.9**	**59.0**	21.4	16.7	14.1	0.4	38.4
CAT-Seg-B	46.7	28.9	23.7	26.7	40.3	65.8	19.3	45.4	35.7	37.6	41.6	48.2	17.0	15.7	31.5	12.3	31.7	19.9	17.5	44.7	10.2	**42.8**
DEDOS-B	**48.1**	**33.4**	29.0	**30.5**	**44.7**	**69.6**	**20.4**	**47.3**	**40.2**	**40.6**	**41.9**	64.0	24.9	31.7	44.5	13.2	29.4	**21.4**	**26.5**	**49.0**	**17.5**	37.7
OVSeg-L	45.3	22.5	6.2	16.4	33.4	53.3	8.3	31.0	31.5	35.6	38.8	71.1	21.0	13.5	22.1	6.8	16.2	21.9	11.7	38.2	14.0	33.8
SAN-L	43.8	30.4	9.3	24.5	40.7	68.4	11.8	51.5	48.2	39.3	**43.4**	**72.2**	7.6	11.9	29.3	6.8	23.7	19.0	18.3	40.0	19.3	1.9
Gr.-SAM-L	42.7	21.9	28.1	10.8	17.6	60.8	12.4	27.8	33.4	19.3	39.4	47.3	**25.2**	**38.1**	44.2	**20.9**	**58.2**	21.2	16.7	14.3	0.4	38.5
CAT-Seg-L	47.9	35.0	32.5	33.3	45.6	73.8	20.6	50.8	46.4	41.4	40.8	61.1	3.7	11.9	22.0	11.0	19.9	**22.0**	27.9	53.0	22.9	39.9
DEDOS-L	**49.8**	**38.3**	**34.9**	**34.1**	**48.6**	**75.6**	**21.5**	**53.1**	**48.3**	**43.7**	42.3	64.7	21.3	29.5	**46.9**	14.5	30.2	20.7	**28.6**	**56.1**	**24.6**	**42.8**

Table 3. mIoU results for all datasets on MESS [4]. MESS covers 5 specific domains with a total of 22 datasets. Random and supervised are provided for reference. The best results are highlighted in bold.

Method	PC-459	A-150	PC-59	PAS-20	PAS-20^b
w/o average and max queries	25.1	39.0	65.2	97.5	84.2
with average and max queries	**25.6**	**39.4**	**65.7**	**97.6**	**84.6**

Table 4. Ablation study on average and max queries.

γ	0.8	1	1.5	1.8	2
mIoU	65.1	65.3	65.7	65.6	65.2

Table 5. Ablation study on the loss weights γ of L_{consist} on the PC-59 dataset.

Function	MSE	Smooth L_1	Cross entropy	KL divergence
mIoU	64.9	65.7	65.1	64.6

Table 6. Quantitative comparison of different loss functions L_{consist} on the PC-59 dataset.

its capability to handle diverse and complex scenarios with remarkable success.

D. Ablation on Other Vision-language Models

As shown in Table 2, we evaluate the performance of various vision-language foundation models (VLMs) integrated into our proposed framework. To ensure a fair comparison, we maintain consistent parameter settings across all experiments, with the decoder based on Mask2Former [6]. The results demonstrate that utilizing a more powerful vision-language foundation model yields better performance, underscoring the robustness and versatility of our approach in the era of foundation models.

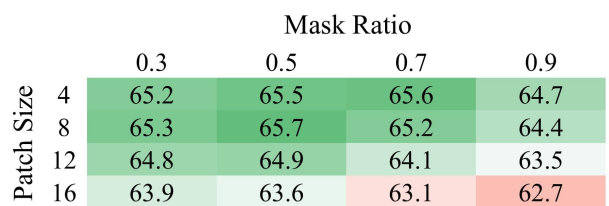

Figure 1. Ablation study of the patch size and the mask ratio of our method on the PC-59 dataset. The color indicates the difference to the CAT-Seg performance of 63.3 mIoU.

E. More Ablation Experiments

E.1. Impact of Average and Max Queries

We provide quantitative results of integrating average and max queries in Table 4. The results show that combining average and max queries yields performance gains across multiple datasets, demonstrating their important role in representation learning. Average queries capture global context, reducing sensitivity to noise, while max queries emphasize salient and discriminative elements, highlighting crucial features. This synergy strengthens segmentation robustness across diverse scenes.

E.2. The Weight γ of L_{consist}

As shown in Table 5, the proposed method exhibits robustness to the choice of the weight coefficient γ, with $\gamma = 1.5$ selected empirically as the default parameter. This observation demonstrates the method's robustness in achieving consistently high performance across different weight configurations, emphasizing its reliability and versatility in varied scenarios.

Dataset	Link	Licence	Sensor type	Number of images	Number of classes
BDD100K [30]	berkeley.edu	custom	Visible spectrum	1000	19
Dark Zurich [24]	ethz.ch	custom	Visible spectrum	50	20
MHP v1 [16]	github.com	custom	Visible spectrum	980	19
FoodSeg103 [29]	github.io	Apache 2.0	Visible spectrum	2135	104
ATLANTIS [8]	github.com	Flickr (images)	Visible spectrum	1295	56
DRAM [7]	ac.il	custom (in download)	Visible spectrum	718	12
iSAID [27]	github.io	Google Earth (images)	Visible spectrum	4055	16
ISPRS Potsdam [5]	isprs.org	no licence provided*	Multispectral	504	6
WorldFloods [21]	github.com	CC NC 4.0	Multispectral	160	3
FloodNet [23]	github.com	custom	Visible spectrum	5571	10
UAVid [18]	uavid.nl	CC BY-NC-SA 4.0	Visible spectrum	840	8
Kvasir-Inst. [13]	simula.no	custom	Visible spectrum	118	2
CHASE DB1 [10]	kingston.ac.uk	CC BY 4.0	Microscopic	20	2
CryoNuSeg [19]	kaggle.com	CC BY-NC-SA 4.0	Microscopic	30	2
PAXRay-4 [25]	github.io	custom	Electromagnetic	180	4x2
Corrosion CS [3]	figshare.com	CC0	Visible spectrum	44	4
DeepCrack [17]	github.com	custom	Visible spectrum	237	2
PST900 [26]	github.com	GPL-3.0	Electromagnetic	288	5
ZeroWaste-f [2]	ai.bu.edu	CC-BY-NC 4.0	Visible spectrum	929	5
SUIM [12]	umn.edu	MIT	Visible spectrum	110	8
CUB-200 [28]	caltech.edu	custom	Visible spectrum	5794	201
CWFID [11]	github.com	custom	Visible spectrum	21	3

Table 7. Details of the datasets in the MESS benchmark [4]. It consists of 22 datasets with 448 categories and 25,079 images covering four different data types: visible spectrum, multispectral, microscopic and electromagnetic.

E.3. Loss Function L_{consist}

As shown in Table 6, we investigate various metrics to define the loss function, including MSE, Smooth L_1, cross-entropy, and KL divergence. The results indicate that the choice of loss function significantly influences model performance, with Smooth L_1 achieving the best results. This can be attributed to its effectiveness in mitigating the influence of outliers and noise, which is essential for learning robust scene distributional representations. As a result, it enhances the model's stability, generalization, and overall reliability.

E.4. Patch Size and Mask Ratio

Figure 1 illustrates the impact of different mask patch sizes and mask ratios on model performance. Our method demonstrates significant improvements with patch sizes ranging from 4 to 8 and mask ratios between 0.3 and 0.7. The optimal performance is achieved when the patch size is set to 8 and the mask ratio is 0.5. This underscores the crucial role of selecting both patch size and mask ratio to enhance the model's performance.

F. Further Qualitative Examples

We present qualitative comparisons with the previous state-of-the-art method, CAT-Seg, showcasing the consistent superiority of our method across a variety of scenarios. Our method demonstrates significant improvements in the completeness of spatial regions. For example, the second and fourth rows of Figure 2 and the fourth row of Figure 3 illustrate more comprehensive spatial coverage. Additionally, our approach delivers a more reasonable spatial distribution and avoids trivial prediction results, as evident in the first and third rows of Figure 2, the second row of Figure 3, and the first and third rows of Figure 4. Furthermore, our method excels in accurately detecting object shapes, as demonstrated in the second row of Figure 5 and the second and third rows of Figure 4. These improvements can be attributed to the model's ability to effectively capture the spatial relationships between scene elements and to learn implicit semantic synergies between different target classes. Consequently, our approach achieves superior segmentation performance in diverse settings.

Figure 2. Qualitative results on the ADE20K validation set. From left to right: visual results predicted by CAT-Seg and Ours, and Ground Truth.

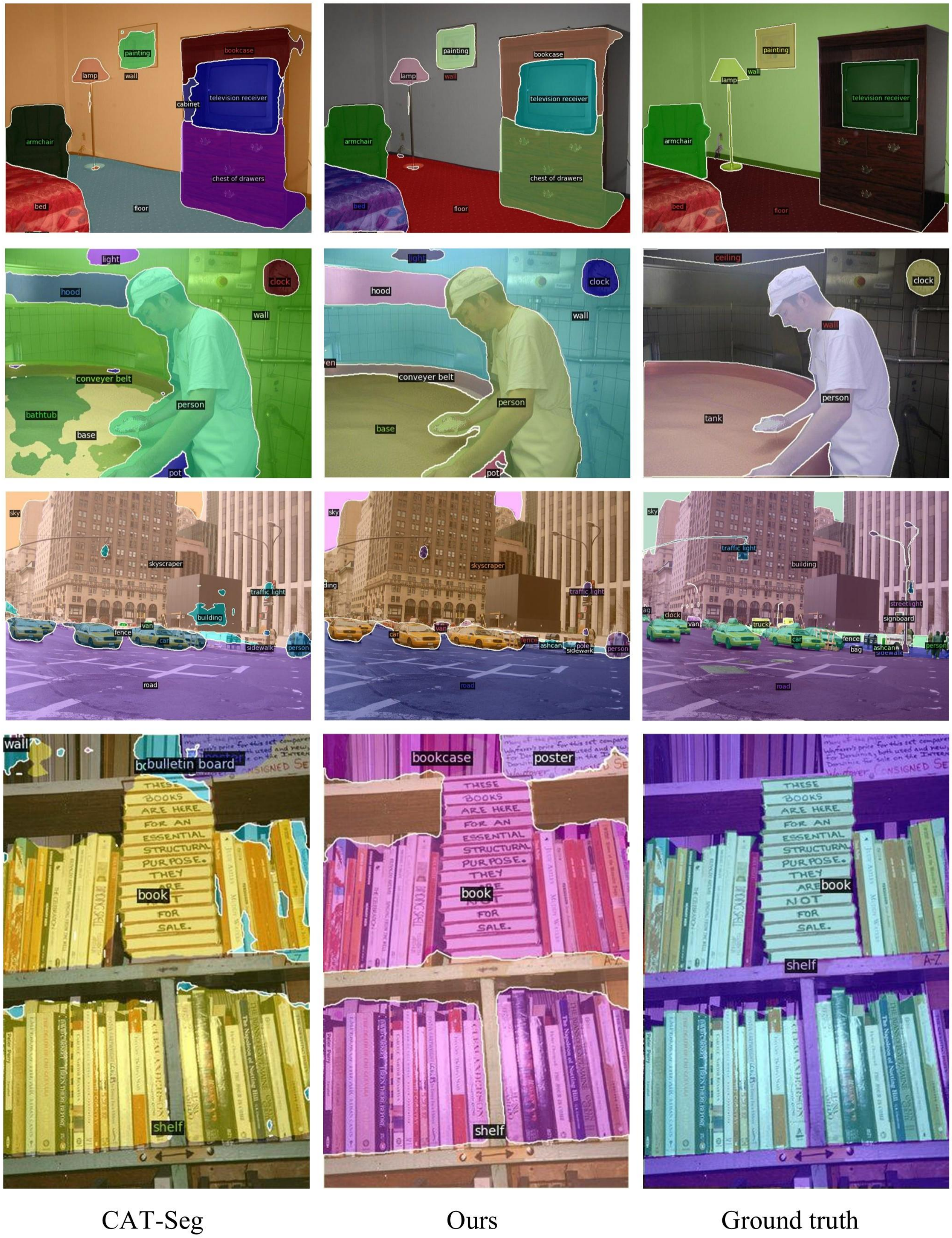

Figure 3. Qualitative results on the ADE20K validation set. From left to right: visual results predicted by CAT-Seg and Ours, and Ground Truth.

CAT-Seg Ours Ground truth

Figure 4. Qualitative results on the ADE20K validation set. From left to right: visual results predicted by CAT-Seg and Ours, and Ground Truth.

CAT-Seg Ours Ground truth

Figure 5. Qualitative results on the ADE20K validation set. From left to right: visual results predicted by CAT-Seg and Ours, and Ground Truth.

CAT-Seg Ours Ground truth

Figure 6. Qualitative results on the ADE20K validation set. From left to right: visual results predicted by CAT-Seg and Ours, and Ground Truth.

References

[1] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented

prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3689–3698, 2024. 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. 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, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 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] 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

[9] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. 1

[10] 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

[11] 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

[12] 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

[13] 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

[14] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 299–317. Springer, 2024. 1

[15] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024. 1

[16] 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

[17] 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

[18] 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

[19] 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

[20] Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C SanMiguel, and Jose M Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9242–9252, 2024. 1

[21] 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

[22] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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

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

[29] 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

[30] 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