AetherVision-Bench: An Open-Vocabulary RGB-Infrared Benchmark for Multi-Angle Segmentation across Aerial and Ground Perspectives

Aniruddh Sikdar* [1], Aditya Gandhamal* [2], Suresh Sundaram[1,3]

[1]Robert Bosch Centre for Cyber Physical Systems, Indian Institute of Science, Bengaluru, India
[2]Kotak IISc AI-ML Centre, Indian Institute of Science, Bengaluru, India
[3]Department of Aerospace Engineering, Indian Institute of Science, Bengaluru, India

Abstract

Open-vocabulary semantic segmentation (OVSS) involves assigning labels to each pixel in an image based on textual descriptions, leveraging world models like CLIP. However, they encounter significant challenges in cross-domain generalization, hindering their practical efficacy in real-world applications. Embodied AI systems are transforming autonomous navigation for ground vehicles and drones by enhancing their perception abilities, and in this study, we present AetherVision-Bench, a benchmark for multi-angle segmentation across aerial, and ground perspectives, which facilitates an extensive evaluation of performance across different viewing angles and sensor modalities. We assess state-of-the-art OVSS models on the proposed benchmark and investigate the key factors that impact the performance of zero-shot transfer models. Our work pioneers the creation of a robustness benchmark, offering valuable insights and establishing a foundation for future research.

1. Introduction

Developing deep learning models with strong cross-domain generalization is essential for real-world computer vision tasks, including surveillance[20, 21], disaster management [11], and robotics across multiple domains [1, 12, 24]. Open-vocabulary semantic segmentation (OVSS) emerges as a promising paradigm, enabling pixel-level classification from an open set of categories through the integration of natural language descriptions as semantic guidance. Recent large-scale vision-language models, pretrained on vast internet-scale data with natural language supervision, demonstrate strong generalization capabilities [10, 18, 19, 34]. Despite being trained on image-text pairs, these models rely on image-level supervision, limiting their pixel-level segmentation performance [5, 37].

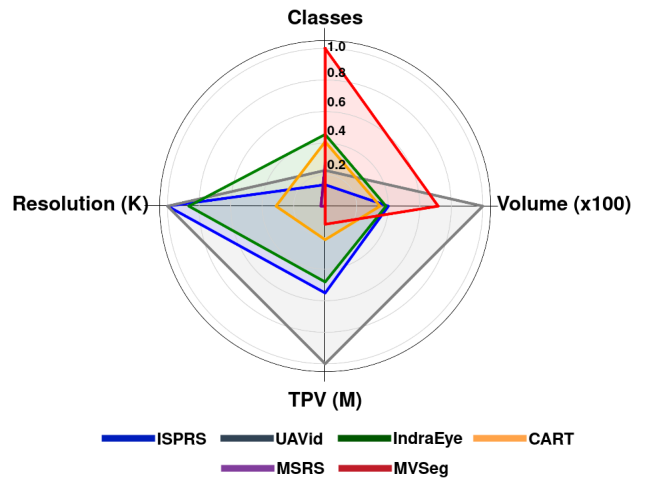Recent studies have improved these models for segmen-

Figure 1. Radar chart comparing test sets across AetherVision-Bench datasets, highlighting the number of classes, image resolution, Total Pixel Volume (TPV), and overall dataset volume. While the maximum number of classes is 26, the data is normalized based on resolution for visualization purposes. The chart illustrates the balanced nature of the benchmark across these key dimensions.

tation tasks through innovative approaches, including (1) mask proposal networks [8, 31–33] and (2) using cost aggregation methods [4]. [2] observes that pixel-level classification performs poorly on unseen classes, particularly in the presence of significant domain shifts. This limitation becomes increasingly pronounced as the number of unseen categories grows [5], with open-vocabulary models exhibiting heightened vulnerability to overfitting during downstream task adaptation [13, 29]. Although OVSS models effectively capture open-set visual concepts for segmentation, their zero-shot generalization performance degrades notably when faced with major category shifts, distribution, or domain shifts between large-scale pretraining data and downstream tasks [35]. This constrains the practical deployment of open-vocabulary models in real-world scenarios.

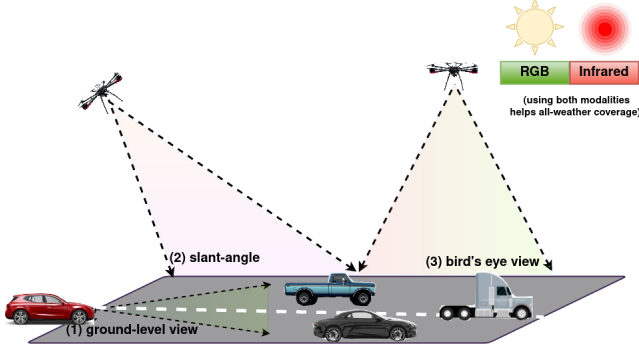Embodied AI systems are revolutionizing autonomous

Figure 2. Comparison of viewpoints across three settings: (1) ground-level, (2) slant-angle, and (3) bird's-eye view. We evaluate methods across these perspectives while accounting for the multi-modal nature of the datasets to ensure robustness under adverse weather conditions.

Dataset	Modality		Viewpoint		
	RGB	IR	BEV	SA	GL
ISPRS	✓	✗	✓	✗	✗
UAVid	✓	✗	✗	✓	✗
IndraEye	✓	✓	✗	✓	✗
CART	✓	✓	✗	✓	✗
MSRS	✓	✓	✗	✗	✓
MVSeg	✓	✓	✗	✗	✓

Table 1. Comparison of datasets with respect to modalities and different viewpoints viz. birds-eye-view (BEV), slant-angle (SA) and ground-level (GL).

navigation for ground vehicles [28] and drones [24] by enabling advanced perception capabilities. These systems should be designed to harness both visiable (RGB) and infrared (IR) sensor data, ensuring resilient performance in adverse weather conditions and low-light environments [23]. They should be robust to sensor modality changes, such as transitions from RGB to IR [22], ensuring reliable performance under diverse conditions including day-night cycles, fog, rain, and glare [25]. However, existing standard benchmarks primarily evaluate in-domain tasks, offering limited perspective on cross-domain performance.This limits understanding of zero-shot segmentation models' generalization to novel domains and their suitability for complex datasets with varied sensors or vocabularies. Existing benchmarks, such as the Multi-domain Evaluation of Semantic Segmentation (MESS) [2], have been developed to assess model performance across a broader range of complex, domain-specific datasets. However, leveraging foundation models like CLIP [18] for advanced perception presents three key challenges that remain unexplored in previous benchmarks: (1) fine-tuning on limited training dataset, for downstream segmentation tasks while preserving pretrained knowledge, (2) handling significant domain shifts arising from out-of-distribution data, and (3) accommodating substantial variations between seen and unseen classes, particularly those arising from differing viewing perspectives. These challenges are integral to ensuring the robustness of models across diverse downstream tasks.

To address this, we introduce AetherVision-Bench: an open-vocabulary RGB-Infrared benchmark for multi-angle segmentation across drones, satellites, and ground perspectives, as shown in Figure 2. We evaluate state-of-the-art OVSS models across six RGB and four IR datasets, carefully chosen to encompass a diverse set of classes and out-of-distribution domains. These datasets are strategically se-

lected to include various object viewpoints, ensuring a comprehensive assessment of model performance under different conditions. We evaluate these models under diverse training regimes, including COCO-Stuff, which offers limited object viewpoint diversity, and RGB datasets featuring both eye-level and slant-angle perspectives, validated comprehensively on the AetherVision benchmark. This underscores the critical need for robust models capable of handling diverse illumination conditions, sensor shifts, noisy sensor inputs, and varying viewing angles. The radar plot in Fig. 1 illustrates the balanced distribution of datasets included in AetherVision-Bench. Our benchmark aims to advance zero-shot segmentation and boost real-world OVSS performance.

Our key contributions are summarized as follows: (1) We develop a taxonomy grounded in quantitative analysis of diverse semantic segmentation datasets and introduce a novel benchmark for out-of-domain OVSS. This benchmark is designed to rigorously assess advanced perception systems across a wide range of scenarios, including ground-level driving scenes, aerial views, and bird's-eye perspectives. (2) We present a comprehensive evaluation of state-of-the-art zero-shot models on AetherVision-Bench, revealing that OVSS performance is heavily influenced by class semantics, textual similarity, and sensor modality.

2. Related Works

Open-vocabulary semantic segmentation (OVSS) Large-scale pre-training has revolutionized computer vision by training models on image-text pairs, enabling representation of visual and textual semantics in a shared space. Open-vocabulary semantic segmentation (OVSS) aims to bridge the gap between image-level and pixel-level understanding by enabling segmentation of unseen classes. OVSS models are generally categorized into two main types: (i) mask-proposal networks [6, 8, 30–33], and (ii) cost-aggregation networks [4]. Mask-proposal approaches generate class-agnostic region proposals, which are then semantically aligned with CLIP-derived textual embeddings to facilitate

Method	RGB							IR				
	ISPRS	UAVid	IndraEye	CART	MSRS	MVSeg	mIoU	IndraEye	CART	MSRS	MVSeg	mIoU
SAN	37.56	<u>41.65</u>	7.84	37.36	20.35	21.96	27.79	4.91	<u>25.35</u>	**19.93**	11.93	<u>15.53</u>
Zegformer	14.04	17.85	12.92	17.18	12.48	11.6	14.34	3.21	3.34	7.38	3.4	4.33
OVSeg	17.52	33.69	13.4	31.67	19.36	14.87	21.75	6.82	16.46	13.91	5.29	10.62
CAT-Seg	<u>45.36</u>	41.55	12.84	<u>39.66</u>	**26.63**	<u>22.27</u>	<u>31.39</u>	6.7	20.02	17.85	<u>12.62</u>	14.30
OVSeg-L	31.03	38.79	**16.44**	35.21	19.98	19.58	26.84	**12.48**	20.7	15.53	9.01	14.43
CAT-Seg-L	**50.92**	**41.9**	<u>16.22</u>	**39.85**	<u>20.77</u>	**26.47**	**32.69**	<u>11.74</u>	**25.84**	<u>19.44</u>	**15.32**	**18.09**

Table 2. Performance comparison of state-of-the-art OVSS methods trained on COCO-Stuff and evaluated on both RGB and IR modalities of the AetherVision-Bench.

open-vocabulary classification at the pixel level. For instance, OpenSeg [8] leverages local image regions to form proposals, whereas models like ZegFormer [5] and ZSseg [31] adopt a two-stage architecture to address this task. CAT-Seg [4] enhances robustness by using similarity scores to combine CLIP's image and text embeddings into a cost volume, rather than relying on direct embeddings. OVSS models are commonly evaluated on standard datasets like ADE20K [36] and Pascal VOC [7], which primarily feature everyday, in-domain images. However, such evaluations fail to capture real-world conditions with inherent domain shifts and environmental diversity.

Related Benchmarks OVSS models are typically evaluated on datasets composed exclusively of in-domain images [7, 36], which are considered the de facto standard for model evaluation [8, 15]. Few works have explored broader dataset diversity. The Segmentation in the Wild (SegInW) benchmark [38] includes 25 datasets, however, most of them still feature everyday, in-domain images, with only two notable exceptions. Inspired by the HELM benchmark [16] for evaluating large language models, the MESS benchmark [2] was subsequently introduced for assessing OVSS models. MESS motivates a comprehensive evaluation across a wide spectrum of domain-specific datasets, including medicine, engineering, earth observation, biology, and agriculture.. However, a key limitation lies in the relatively low number of classes within these datasets, resulting in high random mIoU scores. Additionally, the benchmark empirically demonstrates that current OVSS models struggle to segment objects in infrared imagery, highlighting sensor-shift as a critical challenge. Most state-of-the-art models exhibit suboptimal performance and offer limited support for generalization in open-world settings. To the best of our knowledge, open-vocabulary semantic segmentation (OVSS) models have not been systematically evaluated on datasets that account for sensor shifts and variations in look-angle, as commonly encountered in drones and autonomous vehicles. This work presents the first comprehensive effort to address these challenges, aiming to provide new insights and advance the robustness of OVSS models in practical settings.

3. AetherVision-Benchmark

Problem Formulation Let I represent an image associated with a set of classes $C = \{C_1, C_2, ..., C_N\}$, where each class C_i is defined by a textual label. OVSS models aim to assign a class label C_i for each pixel in an image I, where the candidate class set N can vary at inference and may include novel categories unseen during fine-tuning on the training data.

3.1. AetherVision-Benchmark

Real-world OVSS deployment faces key robustness challenges often overlooked in current research. These challenges highlight the need to improve OVSS model robustness for real-world use. To address this, we propose AetherVision-Bench: An Open-Vocabulary RGB-Infrared Benchmark for Multi-Angle Segmentation Across Drones, Satellites, and Ground Perspectives.

We filter the available datasets based on two exclusion criteria: (1) viewing angle and (2) sensor modality. Guided by the proposed taxonomy, we categorized datasets into three primary look-angle clusters: (1) *ground-level*, (2) *slant-angle*, and (3) *bird's-eye* view, as shown in Table 1. This facilitates the assessment of a model's ability to segment objects under significantly varying viewpoints. Subsequently, we evaluate on both RGB and IR modalities to promote robustness under varying environmental and weather conditions. For each viewpoint category, we evaluate the models' segmentation performance under diverse domain and sensor shifts. For the bird's-eye view, we selected the ISPRS [26] dataset. To capture varying slant-angle perspectives, we included UAVid [17], IndraEye [24], and CART [14] datasets, where both IndraEye and CART also provide corresponding IR imagery. For ground-level viewpoints, we use the MSRS [27] and MVSeg [9] datasets, with their IR modalities. We adopt three evaluation settings: (1) training OVSS models on COCO-Stuff [3], (2) training on the slant-angle RGB IndraEye dataset, and (3) training on the ground-level MSRS dataset. These settings assess the models' capability to generalize across diverse look-angles by training on a specific viewpoint and validating on all three. Notably, settings (2) and (3) introduce an additional

Method	RGB							IR				
	ISPRS	UAVid	Indraeye	CART	MSRS	MVSeg	mIoU	IndraEye	CART	MSRS	MVSeg	mIoU
SAN	10.11	13.86	5.02	12.27	68.63	7.77	19.61	2.29	5.9	30.18	2.2	10.14
Zegformer	4.59	4.76	6.68	3.41	71.49	4.9	15.97	**8.06**	3.76	**34.83**	1.86	12.13
OVSeg	10.45	4.18	5.16	2.81	65.0	4.55	21.29	3.81	2.65	28.66	1.39	11.29
CAT-Seg	**18.34**	**28.34**	**7.06**	**22.87**	**71.93**	**9.24**	**26.30**	4.33	**10.34**	33.59	**4.47**	**13.18**

Table 3. Performance comparison of state-of-the-art OVSS methods trained on MSRS (RGB) and evaluated on both RGB and IR modalities of the AetherVision-Bench.

Method	RGB							IR				
	ISPRS	UAVid	Indraeye	CART	MSRS	MVSeg	mIoU	IndraEye	CART	MSRS	MVSeg	mIoU
SAN	13.33	15.57	**75.84**	6.04	**9.28**	6.12	**21.03**	23.24	3.21	**10.51**	3.12	10.02
Zegformer	7.80	7.32	75.52	3.51	4.72	2.9	16.96	28.05	2.31	2.66	0.48	8.38
OVSeg	12.3	16.23	61.8	9.0	5.68	4.12	18.19	22.41	5.12	2.53	1.55	7.90
CAT-Seg	**17.2**	**17.62**	58.88	**15.3**	4.03	**8.84**	20.31	**33.90**	**8.07**	4.99	**4.94**	**12.97**

Table 4. Performance comparison of state-of-the-art OVSS methods trained on IndraEye (RGB) and evaluated on both RGB and IR modalities of the AetherVision-Bench.

challenge of limited training samples. We selected these datasets based on official test splits, high-quality annotations, and sufficient image resolution. No ethical concerns were raised, and the datasets were used in compliance with their respective licenses. This benchmark is ideal for evaluating model robustness across diverse look angles, with challenges like weather changes, sensor degradation, and complex environments.

4. Experiments

Implementation All models are evaluated using the mean intersection over union (mIoU), using their official implementations, and pretrained weights. We use the AetherVision-Bench to evaluate state-of-the-art OVSS models, including OVSeg [15], Zegformer [5], SAN [32], and CAT-Seg [4], which feature various architectural approaches (e.g., two-stage mask-based, one-stage mask-based, and pixel-based).

4.1. Multi-domain OVSS

Table 2 presents a comparison of OVSS models on the AetherVision-Bench, trained on the COCO-Stuff dataset. Table 3 compares the models when trained on the MSRS dataset's training set, using the RGB modality. As shown in Table 4, training on the slant-angle RGB samples from the IndraEye dataset leads to a significant drop in performance. **In-domain vs multi-domain** Most OVSS models are typically evaluated on in-domain datasets. However, AetherVision-Bench's real-world benchmarking is particularly effective for assessing the robustness under conditions that closely resemble those found in real-world deployments. As noted, the CAT-Seg model performs bet-

ter than other models, when trained on COCO-Stuff and MSRS datasets, but its performance significantly drops when trained on the slant-angle view in the benchmark.
Sensor Domain Shift As seen in all the tables across all settings, the robustness of all models to sensor shifts is poor. The overall performance decreases by nearly half, as demonstrated in Table 2, 3, and 4. Even when models are trained on the MSRS RGB dataset, their performance on co-registered IR pairs drops to less than half. For noisy sensors, as seen with the MVSeg dataset, performance deteriorates even further.
Evaluation under diverse viewing angles Models trained on COCO-Stuff or ground-level datasets demonstrate limited effectiveness on drone-view imagery. Similarly, those trained on slant-angle data such as the IndraEye dataset show poor generalization to ground-level datasets like MSRS and MVSeg. This indicates that OVSS models must depend heavily on their pre-training to generalize across different viewpoints, incase these viewing points are not available during training for downstream tasks.

5. Conclusion

This paper presents a comprehensive benchmark for evaluating the robustness of OVSS models, essential for their real-world deployment. We introduce the AetherVision-Bench, a Benchmark for multi-angle segmentation across various perspectives. This benchmark evaluates models across two challenging scenarios: viewing angle and sensor modality. It is designed to address real-world complexities, making it ideal for embodied AI systems that enhance perception and revolutionize autonomous navigation for ground vehicles and drones.

References

[1] Aayush Agrawal, Aniruddh Sikdar, Rajini Makam, Suresh Sundaram, Suresh Kumar Besai, and Mahesh Gopi. Syn2real domain generalization for underwater mine-like object detection using side-scan sonar. *IEEE Geoscience and Remote Sensing Letters*, 2025. 1

[2] Benedikt Blumenstiel, Johannes Jakubik, Hilde Kühne, and Michael Vössing. What a mess: Multi-domain evaluation of zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 3

[4] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 1, 2, 3, 4

[5] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. De-coupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 1, 3, 4

[6] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation maskclip. 2022. 2

[7] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 3

[8] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 1, 2, 3

[9] Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L Yuille, and Li Cheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1104, 2023. 3

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[11] Josy John, K Harikumar, J Senthilnath, and Suresh Sundaram. An efficient approach with dynamic multiswarm of uavs for forest firefighting. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(5):2860–2871, 2024. 1

[12] Josy John, Shridhar Velhal, and Suresh Sundaram. A resource-efficient decentralized sequential planner for spatiotemporal wildfire mitigation. *IEEE Transactions on Automation Science and Engineering*, 2025. 1

[13] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. 1

[14] Connor Lee, Matthew Anderson, Nikhil Ranganathan, Xingxing Zuo, Kevin Do, Georgia Gkioxari, and Soon-Jo Chung. Caltech aerial rgb-thermal dataset in the wild. In *European Conference on Computer Vision*, pages 236–256. Springer, 2024. 3

[15] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 3, 4

[16] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022. 3

[17] 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

[18] 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

[19] Jia Shi, Gautam Gare, Jinjin Tian, Siqi Chai, Zhiqiu Lin, Arun Vasudevan, Di Feng, Francesco Ferroni, and Shu Kong. Lca-on-the-line: Benchmarking out-of-distribution generalization with class taxonomies. *arXiv preprint arXiv:2407.16067*, 2024. 1

[20] Aniruddh Sikdar, Sumanth Udupa, Suresh Sundaram, and Narasimhan Sundararajan. Fully complex-valued fully convolutional multi-feature fusion network (fc 2 mfn) for building segmentation of insar images. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 581–587. IEEE, 2022. 1

[21] Aniruddh Sikdar, Sumanth Udupa, Prajwal Gurunath, and Suresh Sundaram. Deepmao: Deep multi-scale aware over-complete network for building segmentation in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 487–496, 2023. 1

[22] Aniruddh Sikdar, Qiranul Saadiyean, Prahlad Anand, and Suresh Sundaram. Ssl-rgb2ir: Semi-supervised rgb-to-ir image-to-image translation for enhancing visual task training in semantic segmentation and object detection. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5017–5023. IEEE, 2024. 2

[23] Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. Skd-net: Spectral-based knowledge distillation in low-light thermal imagery for robotic perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9041–9047. IEEE, 2024. 2

[24] Aniruddh Sikdar, Prajwal Gurunath, Sumanth Udupa, Suresh Sundaram, et al. Saga: Semantic-aware gray color augmen-

tation for visible-to-thermal domain adaptation across multi-view drone and ground-based vision systems. *arXiv preprint arXiv:2504.15728*, 2025. 1, 2, 3

[25] Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. Ogpnet: Optical guidance meets pixel-level contrastive distillation for robust multi-modal and missing modality segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6922–6930, 2025. 2

[26] BSF Swissphoto. Isprs potsdam dataset within the isprs test project on urban classification, 3d building reconstruction and semantic labeling, 2012. 3

[27] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022. 3

[28] Sumanth Udupa, Prajwal Gurunath, Aniruddh Sikdar, and Suresh Sundaram. Mrfp: Learning generalizable semantic segmentation from sim-2-real with multi-resolution feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5904–5914, 2024. 2

[29] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 1

[30] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2

[31] 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 *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 1, 3

[32] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 4

[33] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023. 1, 2

[34] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1

[35] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2024. 1

[36] 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

[37] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 1

[38] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15116–15127, 2023. 3