Seeing Beyond: Extrapolative Domain Adaptive Panoramic Segmentation

(Supplementary Material)

Yuanfan Zheng[1] Kunyu Peng[2,3] Xu Zheng[4] Kailun Yang[1,*]

[1]Hunan University [2]Karlsruhe Institute of Technology [3]INSAIT, Sofia University "St. Kliment Ohridski" [4]HKUST(GZ)

In this supplementary material, we provide three sections to complement the main manuscript. **Section A** offers a detailed description of the proposed task setting and compares it with related settings. **Section B** presents additional experiments to demonstrate the effectiveness of the proposed module and includes an analysis of its sensitivity coefficients. **Section C** discusses the limitations of our method and provides an outlook on the societal implications.

Sec. A: Clarification and Discussion

- Task Clarification
- Benchmark Setup
- Implementation Details

Sec. B: Quantitative Comparison

- Further Analysis
- Sensitivity Analysis
- Model Efficiency
- Visualization Analysis

Sec. C: Limitations and Outlook

- Societal Implications
- Future Research Directions
- Limitations and Potential Solutions

Sec. A: Clarification and Discussion

A.1. Task Clarification

Clarification of the setting. As illustrated in Fig. 1, the figure depicts the conceptual differences among three domain adaptation settings: Closed-Set Domain Adaptation, Open-Set Domain Adaptation, and Panoramic Open-Set Domain Adaptation. (a) **Closed-Set Domain Adaptation**: The source domain and the target domain share an identical set of classes (for example, cats and dogs). The objective is to reduce the domain discrepancy under the assumption that all labels are fully shared. (b) **Open-Set Domain Adaptation**: The target domain contains additional unknown categories that do not appear in the source domain. The goal is to properly align the shared classes while identifying and filtering out target samples belonging to unknown categories. (c) **Panoramic Open-Set Domain Adaptation (PODA)**: This setting extends the open-set scenario to a cross-modal context, where the source domain provides pinhole images and the target domain provides panoramic images. The target domain has known and unknown classes, causing challenges due to camera and semantic differences.

Clarification of the domain shift. From Table 1, it is clear that conventional Unsupervised Domain Adaptation (UDA) methods are unable to handle open-set categories or adapt to diverse weather conditions. While Panoramic UDA extends standard UDA by explicitly modeling variations in the field of view (FoV), it still cannot handle unknown categories or environmental shifts. Open-Set Domain Adaptation focuses exclusively on unseen classes, and Open Compound Domain Adaptation aims to handle domain shifts caused by varying weather conditions. However, both neglect the FoV discrepancies inherent to panoramic imagery. In contrast, the proposed **PODA** framework simultaneously addresses open-set recognition, weather-related domain shifts, and FoV variations. By jointly modeling these three critical factors, **PODA** achieves superior generalization and robustness, enabling more reliable cross-domain perception in complex real-world driving scenarios.

A.2. Technical Clarifications

A.3. Benchmark Setup

Dataset and class configuration. To evaluate open-set cross-domain semantic segmentation in a controlled yet diverse manner, we construct a benchmark spanning real-to-real, synthetic-to-real, and synthetic-to-synthetic transfers. For source and target domains, we conduct the **domain-shared (base)** and **domain-specific (private)** classes, where the latter appear only in one domain and thus represent unforeseen categories encoun-

*Corresponding author (e-mail: kailun.yang@hnu.edu.cn).

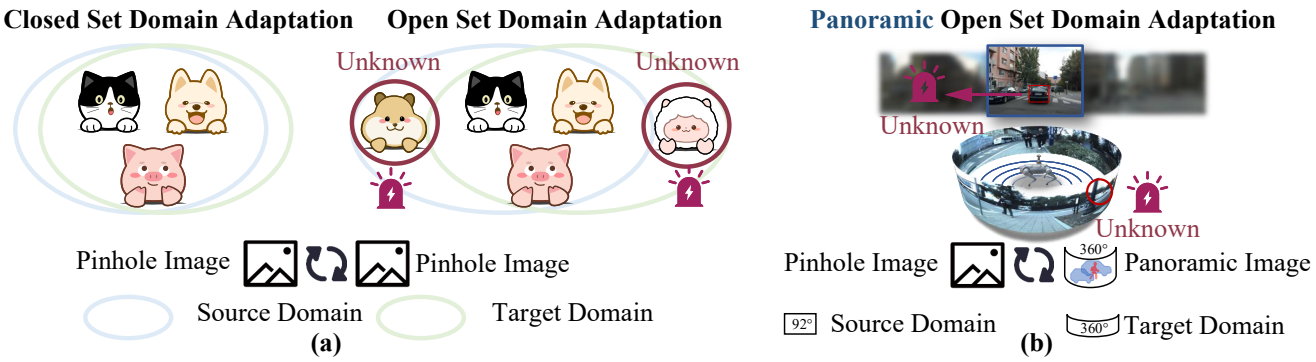

Figure 1. Conceptual comparison of the domain adaptation setting.

tered during deployment. Weather conditions follow the same convention, capturing whether environmental factors are shared or exclusive to a single domain. The benchmark includes four representative transfers as CityScapes→DensePASS, SynPASS→DensePASS, GTA→SynPASS, and SynPASS→ACDC, covering varying degrees of semantic and environmental mismatch. In the real-to-real case, both domains share sunny scenes but include human and vehicle categories that remain private. Synthetic-to-real transfers introduce additional private traffic participants while maintaining shared weather. Synthetic-to-synthetic transfer expands this gap further by adding private meteorological conditions such as cloud, fog, and rain, along with fine-grained static and dynamic categories exclusive to the target. The adverse weather scenario emphasizes extreme condition shifts, where fog and rain remain shared while cloud, sunny, night, and snow are domain-specific. Together, these configurations simultaneously expose category- and weather-level discrepancies, forming a unified and reproducible benchmark for assessing robustness under open-set domain shifts.

Input resolution configuration. For all four open-set benchmark settings, the source and target datasets exhibit diverse image resolutions, reflecting their original collection protocols. To ensure consistency during training, all images are cropped to a unified size of 512×512. Specifically, CityScapes images are originally 1024×512 while DensePASS images are 2048×400, SynPASS images are 2048×1024, GTA images are 1280×720, and ACDC images are 960×540. This unified crop strategy balances the varying aspect ratios and resolutions across datasets, providing the input size for network training while preserving sufficient spatial context for semantic segmentation.

A.4. Implementation Details

We present the details in Tab. 4, including the pipeline, the auxiliary network, and the training configuration.

Pipeline. Our proposed framework, **EDA-PSeg**, follows a two-branch pipeline designed for open-set unsupervised domain adaptive semantic segmentation. It consists of a *source-domain supervised branch* and a *target-domain self-training branch*, both of which are processed under a unified data augmentation and preprocessing scheme. For the *source domain*, each image and its annotation are loaded and resized before being cropped to a fixed resolution of 512×512. Standard data augmentations, such as random horizontal flipping, are applied to enhance robustness. All images are normalized using ImageNet statistics ($mean = [123.675, 116.28, 103.53]$, $std = [58.395, 57.12, 57.375]$) and padded to maintain consistent tensor sizes. For the *target domain*, a similar preprocessing strategy is adopted, but with adaptive scale resizing and random cropping to accommodate scale variation and perspective distortion common in real-world scenes. Both source and target domains share identical normalization and padding configurations to ensure consistent feature distribution across domains. During training, the model alternately samples mini-batches from both domains following the **DACS** [19] (Domain Adaptive Cross-domain Self-training) strategy. The source-domain batches provide supervised signals, while the target-domain batches are assigned pseudo-labels that are progressively refined using our auxiliary MobileSAM module (described below). At inference time, each test image undergoes deterministic evaluation using a single-scale forward pass with resizing and normalization. Multi-scale and flip testing are disabled for computational consistency. This unified pipeline ensures stable and reproducible adaptation performance while preserving domain-level consistency.

Auxiliary Network. Following existing open-set domain adaptation approaches, we adopt **MobileSAM** [23] as an auxiliary network to refine pseudo-labels during training. MobileSAM is a lightweight and parameter-efficient variant of the Segment Anything Model (SAM) [10]. It is employed solely to distinguish foreground and background regions, improving the reliability of pseudo-labels in the target domain. Specifically, for each target image, MobileSAM gen-

Table 1. Comparison of domain adaptation settings with the domain shift.

Domain Adaptation Setting	Label Shift	Weather Shift	FoV Shift
Unsupervised Domain Adaptation [5]	✗	✗	N/A
Panoramic Domain Adaptation [25]	✗	✗	✓
Open-Set Domain Adaptation [4]	✓	✗	N/A
Open Compound Domain Adaptation [13]	✗	✓	N/A
Panoramic Open-Set Domain Adaptation (Ours)	✓	✓	✓

Table 2. Base and private classes used in the PAN2PAN, SYN2SYN, and SYN2REAL experiments. Shared classes (green) are common to both the source and target domains, while private classes (red) appear only in either the source or the target domain. Weather conditions follow the same convention: green indicates shared conditions across both domains, and red indicates conditions specific to a single domain.

Setting	Classes (Base / Private)	Weather
Open-Set CityScapes → DensePASS	'road', 'sidewalk', 'building', 'wall', 'fence', 'traffic light', 'vegetation', 'terrain', 'sky', 'car', 'bus', 'motorcycle', 'bicycle'. CityScapes/DensePASS: 'pole', 'traffic sign', 'person', 'rider', 'truck', 'train'.	Sunny→Sunny
Open-Set SynPASS → DensePASS	'road', 'sidewalk', 'building', 'wall', 'fence', 'pole', 'traffic light', 'traffic sign', 'vegetation', 'terrain', 'sky', 'person', 'car'. SynPASS: 'other', 'ground', 'bridge', 'railtrack', 'groundrail', 'static', 'dynamic', 'water'. DensePASS: 'bus', 'truck', 'train', 'motorcycle', 'bicycle', 'rider'.	Sunny→Sunny
Open-Set GTA → SynPASS	'road', 'sidewalk', 'building', 'wall', 'fence', 'pole', 'traffic light', 'traffic sign', 'vegetation', 'terrain', 'sky', 'person', 'car'. GTA: 'rider', 'truck', 'bus', 'train', 'motorcycle', 'bicycle'. SynPASS: 'other', 'ground', 'bridge', 'railtrack', 'groundrail', 'static', 'dynamic', 'water'.	Sunny→Sunny, Cloud, Fog, Rain, Night
Open-Set SynPASS → ACDC	'road', 'sidewalk', 'building', 'wall', 'fence', 'pole', 'traffic light', 'traffic sign', 'vegetation', 'terrain', 'sky', 'person', 'car'. SynPASS: 'other', 'ground', 'bridge', 'railtrack', 'groundrail', 'static', 'dynamic', 'water'. ACDC: 'rider', 'truck', 'bus', 'train', 'motorcycle', 'bicycle'.	Fog, Rain, Night, Sunny, Cloud→Fog, Rain, Night, Sunny, Snow

Table 3. Image resolutions for different datasets involved in the four open-set settings. Crop size is unified to 512×512 for training.

Setting	Source Image Size (W×H)	Target Image Size (W×H)	Crop Size (W×H)
Open-Set CityScapes → DensePASS	CityScapes: 1024×512	DensePASS: 2048×400	512×512
Open-Set SynPASS → DensePASS	SynPASS: 2048× 1024	DensePASS: 2048×400	512×512
Open-Set GTA → SynPASS	GTA: 1280×720	SynPASS: 2048×1024	512×512
Open-Set SynPASS → ACDC	SynPASS: 2048×1024	ACDC: 960×540	512×512

Table 4. **Key Settings of the Proposed EDA-PSeg Framework**

Component	Setting
Framework	Two-branch open-set UDA: source-supervised + target self-training.
Backbone	**DAFormer** with **MiT-B5** encoder and Graph decoder.
Auxiliary Net	**MobileSAM** for pseudo-mask refinement.
Input Size	512×512 (crop, resize, flip, normalize).
Normalization	ImageNet mean=[123.7,116.3,103.5], std=[58.4,57.1,57.4].
Pseudo-labels	Refined by MobileSAM [23].
Teacher Update	EMA with $\alpha = 0.999$.
Optimizer	AdamW, LR=6×10^{-5} (decoder $\times 10$).
Schedule	Warmup + polynomial decay, 40k iterations.
Sampling	DACS [19] with rare-class focus (`min_pixels=3000`).
Evaluation	H-score & mIoU.

erates segmentation masks corresponding to potential foreground classes. Class frequencies are then computed from these masks, and the dominant category is selected as the valid pseudo-label class.

Training Configuration. We adopt **DAFormer** [6] as the baseline architecture for UDA semantic segmentation. The model employs a **MiT-B5 [20]** backbone combined with a DAFormerHead decoder variant, modified to predict 14 semantic categories (13 closed-set classes plus one open-set/unknown class). Training follows the **DACS** [19] self-training framework with pseudo-labeling and feature consistency losses. The temporal ensembling coefficient is set to $\alpha = 0.999$ to stabilize the teacher-student updates. The feature distance loss is disabled, enabling adaptation to rely primarily on pseudo-label-based consistency rather than ImageNet feature alignment. Pseudo-label generation ignores the top 15 and bottom 120 pixels of the image to reduce label noise near image borders. A rare-class sampling mechanism is incorporated to alleviate class imbalance, using `min_pixels` $= 3000$ and `class_temp` $= 0.01$ to prioritize under-represented categories during source sampling. The model is optimized using the **AdamW** optimizer with a base learning rate of 6×10^{-5}, while the learning rate of the decoder head is multiplied by 10. A linear warm-up followed by polynomial decay is applied for stable convergence. Training runs for $40,000$ iterations, and evaluation is performed every $4,000$ iterations. Performance is evaluated using both the **H-score** and **mIoU**, which together measure segmentation quality and open-set recognition accuracy. The checkpoint achieving the highest **mIoU** on the validation set is selected as the best. All experiments are performed on a single-GPU setup.

Sec. B: Quantitative Comparison

B.1. Further Analysis

As shown in Tab. 5, we evaluate the effectiveness of the EMA module. The core of our design is the Margin Projection, an enhancement to the self-attention mechanism that consistently improves performance. In the initial comparison (experiment ①), progressively adding attention-based components consistently improves both Common and Private results, indicating that attention enhances domain-

Table 5. **Ablation analysis of the EMA module.**

Cityscapes → DensePASS (**C2D**)				
Exp.	**Method**	**Common**	**Private**	**H-Score**
①	Baseline	52.56	8.57	14.74
①	Self-Att	55.45	10.95	18.28
①	Self-Att+Margin Projection	55.64	15.06	23.71
②	EMA w/o Margin Projection	55.03	7.56	13.30
③	EMA w/o Learnable Scale	55.50	13.07	21.16
④	EMA w/o Learnable Bias	55.59	7.35	12.99
⑤	EMA w/o Learnable Magnitude	54.04	5.20	9.48
⑥	EMA (Full)	56.12	13.00	21.11

Table 6. Comparison of the number of parameters (M), FLOPs (G), MACs (G), and test time per image (ms). Experiments are conducted on a single RTX 4090 and AMD Ryzen 9 5950X CPU.

Method	**#Params**	**FLOPs**	**MACs**	**Time**
OSBP [18]	85.15 M	59.45 G	29.73 G	26.86 ms
UAN [22]	85.29 M	59.45 G	29.73 G	26.67 ms
UniOT [9]	85.22 M	59.45 G	29.73 G	27.03 ms
DMLP [24]	90.15 M	59.45 G	29.73 G	27.12 ms
MIC [8]	85.68 M	14.86 G	7.43 G	26.62 ms
DAF [6]	85.15 M	59.45 G	29.73 G	27.76 ms
HRDA [7]	85.69 M	14.86 G	7.43 G	26.25 ms
BUS [2]	85.68 M	14.86 G	7.43 G	26.06 ms
Ours	86.47 M	59.45 G	29.73 G	28.09 ms

Table 7. Sensitivity analysis of the self-attention layer in the EMA module. Note that, since the Euler encoding in EMA requires both real and imaginary components, the dimensionality must be even.

Cityscapes → DensePASS (**C2D**)				
Exp.	**Layer**	**Common**	**Private**	**H-Score**
①	-	52.56	8.57	14.74
②	2	56.12	13.00	21.11
③	4	54.80	9.94	16.83
⑤	8	54.99	6.04	10.88

Table 8. Sensitivity analysis under different threshold values. Results are reported on Cityscapes → DensePASS (**C2D**).

Cityscapes → DensePASS (**C2D**)				
Exp.	**Threshold**	**Common**	**Private**	**H-Score**
①	0.4	53.03	1.66	3.22
②	0.5	56.02	6.17	11.11
③	0.6	56.81	18.86	28.32
④	0.7	54.59	9.66	16.41

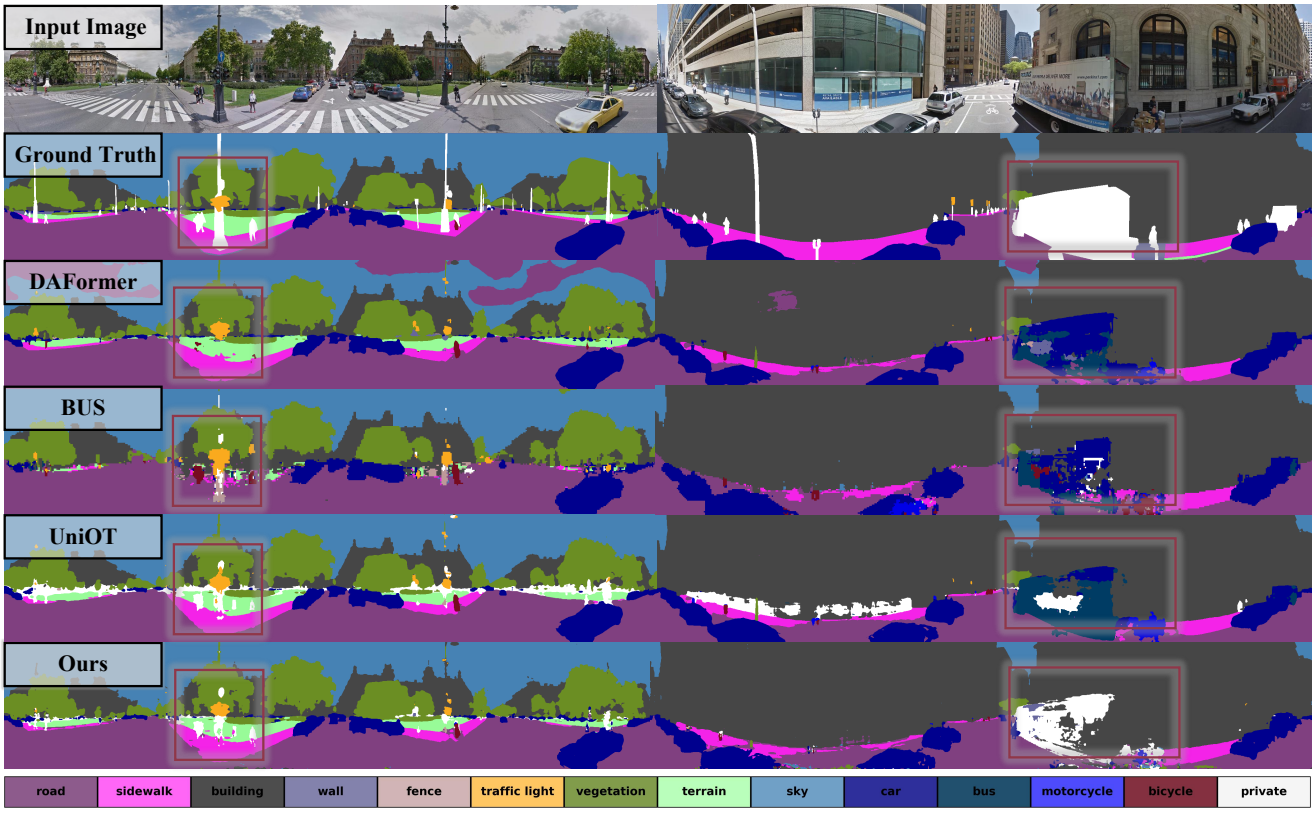

Figure 2. Qualitative comparison between our method and existing OSDA and UDA approaches.

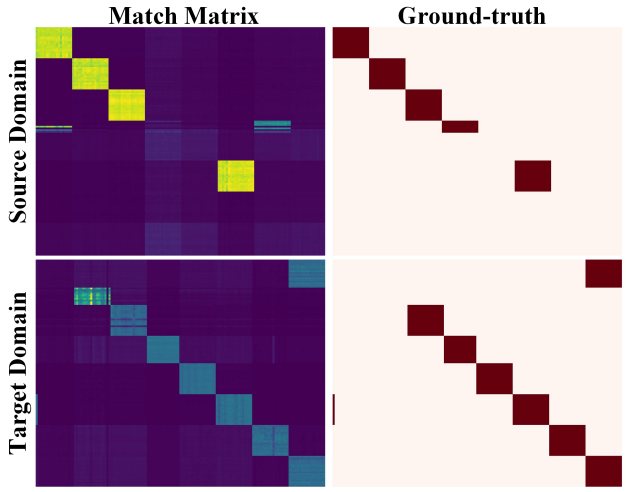

Figure 3. Illustration of the match matrix derived from the graph affinity and the matching ground-truth in the GMA module.

shared representations. Introducing Margin Projection notably increases the Private accuracy and achieves the highest H-Score among non-EMA variants, confirming the benefit of margin-based separation for target-specific learning. In the EMA ablation, removing Margin Projection (②) leads to a clear performance drop, suggesting its importance in maintaining target-domain feature separability. The absence of the Learnable Scale (③) has a minor effect, while omitting the Learnable Bias (④) causes a marked decline, highlighting the need for bias correction to address domain shifts. Finally, removing the Learnable Magnitude (⑤) results in the most severe degradation, demonstrating that the learnable magnitude is crucial for stable feature representation and overall performance.

B.2. Sensitivity Analysis

Sensitivity of attention layer. As shown in Tab. 7, we conduct the sensitivity analysis to evaluate the effect of self-attention layers in the EMA module. A moderate attention depth notably improves adaptation: two layers yield the highest H-Score, with consistent gains across both common and private classes. This configuration strengthens the angle space coupling induced by the Euler encoding, enabling the model to capture domain-invariant and domain-specific patterns more effectively. However, deeper layers lead to overfitting to source-domain priors, reducing target adaptability and overall balance. These results indicate that the EMA module performs best with limited yet sufficient at-

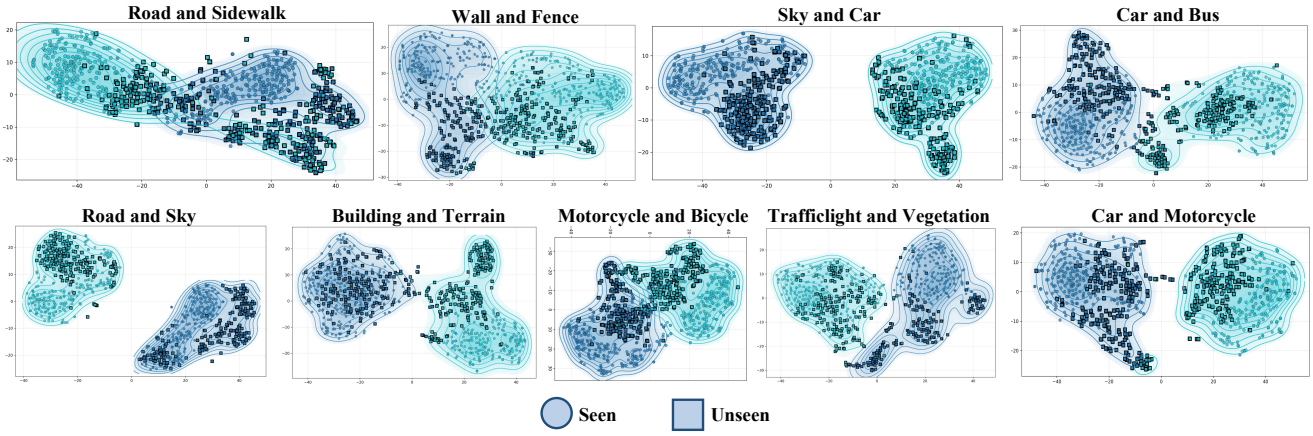

Figure 4. T-SNE visualization of data from seen and unseen viewpoints.

tention depth, while excessive layers hinder generalization.

Sensitivity of threshold. As shown in Tab. 8, we analyze the sensitivity of the threshold used in generating pseudo-labels for private classes during source and target domain mixup training. With the threshold set to 0.4 (①), the model exhibits the lowest mIoU and H-Score for both common and private classes, indicating severe confusion between the two categories. Increasing the threshold to 0.5 (②) leads to a clear improvement in segmentation performance, particularly reflected in a higher H-Score. Further raising the threshold to 0.6 (③) enhances the recall of unknown classes while also improving the accuracy on known class mIoU. However, raising the threshold to 0.7 (④) reduces both common and private mIoU, suggesting that too strict a confidence filter can impair pseudo-label generation.

B.3. Model Efficiency

As shown in Tab. 6, we evaluate the efficiency of the parameters (M), FLOPs (G), MACs (G), and test time per image (ms), and compare these metrics with existing methods. OSBP [18], UAN [22], UniOT [9], and DMLP [24] have about $85M$ to $90M$ parameters and $59.45G$ FLOPs, requiring around $27ms$ per image, indicating relatively high complexity. In contrast, MIC [8], DAF [6], HRDA [7], and BUS [2] are more lightweight, with only 14.86G FLOPs and faster inference. Our method achieves a similar computational scale to the above models, with a slightly higher parameter count ($86.47M$) and inference time ($28.09ms$).

B.4. Visualization Analysis

Qualitative Results. As illustrated in Fig. 2, we present a qualitative comparison of open-set UDA segmentation results using the OSDA [4] methods BUS [2] and UniOT [9], as well as the UDA method DAFormer [6], to evaluate the effectiveness of our proposed approach. Compared with existing methods, our approach delivers improved segmentation performance across both known and unknown classes. In particular, it achieves superior open-set performance relative to DAFormer [6], demonstrates greater robustness to small open-set objects than BUS [2], and substantially mitigates closed-set class and private class foreground objects misidentification errors frequently observed in UniOT [9].

Graph Match Results. As shown in Fig. 3, we visualize the matching matrix from the learnable affinity and its corresponding ground truth within the proposed GMA module for both the source and target domains. The highlighted regions in the ground truth indicate the nodes that should be matched. The predicted matching matrix closely aligns with the ground truth, demonstrating the effectiveness of the proposed GMA module in open-set graph matching.

Seen and Unseen Viewpoints. As shown in Fig. 4, we perform visualization experiments on both the seen dataset Cityscapes [3] and the unseen dataset DensePASS [14]. For the stuff classes [1] (*e.g.*, road, sidewalk, sky) and the thing classes [1] (*e.g.*, car, bus, motorcycle). For stuff classes and thing classes, categories within the same group tend to share certain similarities, which is reflected in their overlapping or intersecting distributions in the t-SNE visualization. In contrast, when comparing categories across stuff and thing classes, their t-SNE embeddings typically show a significant separation, indicating substantial feature-level differences between the two groups. For objects viewed from both seen and unseen perspectives, domain shift occurs across categories due to variations in the field of view.

Qualitative results for EMA module. As shown in Fig. 5, we perform qualitative analysis of the EMA module using Principal Component Analysis (PCA) to identify regions with consistent color or brightness, while the L2 norm highlights areas receiving model attention. For the PCA projection, the proposed EMA method better highlights thing classes such as buildings, maintains smoother and less noisy representations in stuff classes like the sky, and achieves

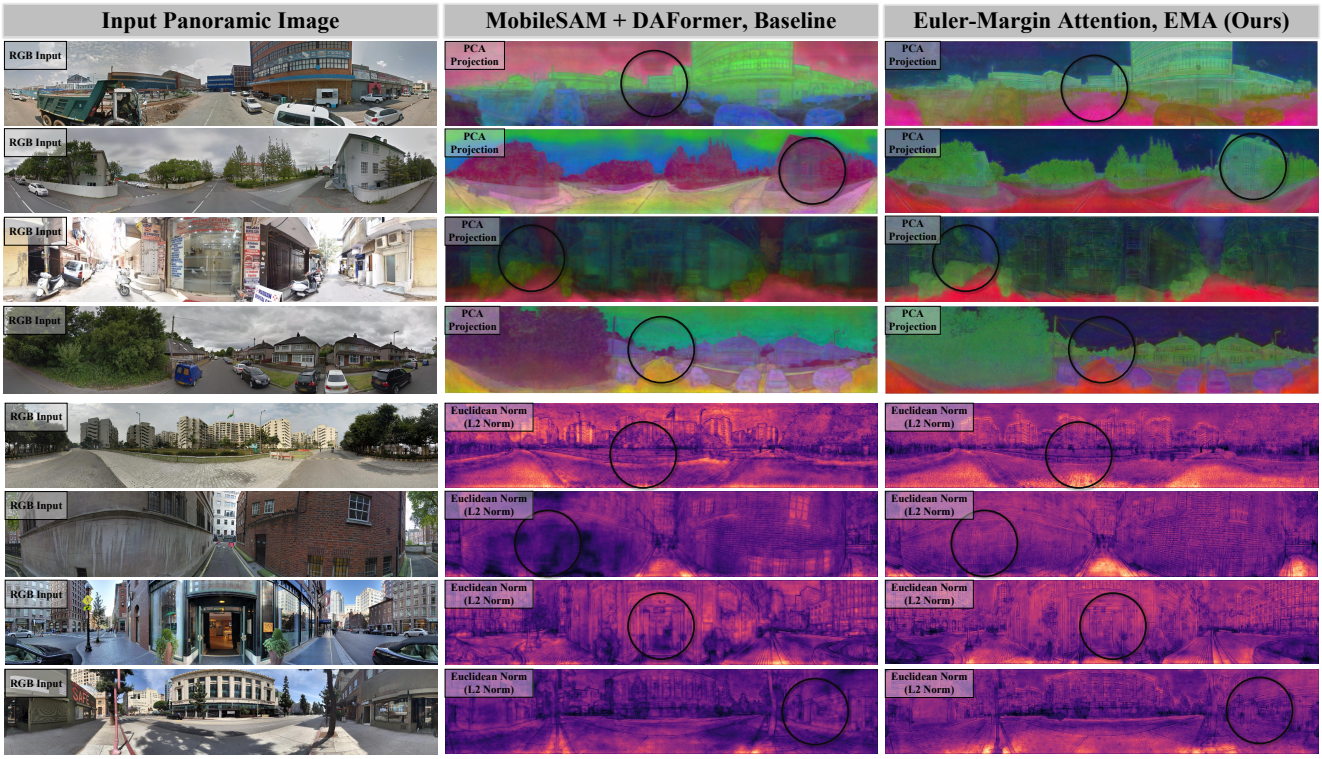

Figure 5. Qualitative results for EMA. We visualize feature representations using Principal Component Analysis (PCA) and the L2 norm.

higher cross-view consistency in vegetation regions. For the Euclidean (L2) norm visualization, the baseline suffers from limited activation and reduced brightness in geometrically distorted regions, whereas our method effectively overcomes this issue and better highlights objects at long ranges and under wide viewpoints.

Sec. C: Limitations and Outlook

C.1. Societal Implications

As illustrated in Fig. 6, our proposed Extrapolative Domain Adaptive Panoramic Segmentation framework delivers significant societal benefits across four domains: quadruped robotics, autonomous driving, assistive navigation for people with visual impairments, and drone-based perception. By enhancing semantic and panoramic perception, our method enables quadruped robots to operate reliably in hazardous environments, supports autonomous vehicles in making safer decisions under diverse conditions, assists visually impaired users with real-time spatial awareness, and strengthens drone perception for environmental monitoring, precision agriculture, and infrastructure inspection.

C.2. Future Research Directions

In the future, we plan to pursue two main research directions. The first focuses on methodological improvements, with the aim of improving the performance, efficiency, and robustness of our current approach. The second direction involves application-oriented extensions, where we intend to adapt and apply our methods to broader or more complex real-world scenarios. Building upon the first research direction, we note that in pseudo-label training, the private class is currently threshold-based. To advance this, future work suggests improving it to a threshold-free private class mechanism, in line with existing research [15, 17, 26]. Despite progress in related areas, test-time adaptation [12, 16] and source-free domain adaptation [11, 21] have received limited attention in open-world vision for panoramic images. In the future, we will expand the applicability of this method further. Experiments show that the technique maintains stable performance under different fields of view, which confirms its adaptability to various imaging conditions. Subsequent work will explore its applications in panoramic, fisheye, wide-angle, and pinhole images and further extend it to real-world scenes.

C.3. Limitations and Potential Solutions

In edge computing scenarios or environments with limited computational resources, our model still exhibits a limit for parameter scalability and inference efficiency. This limitation primarily arises from the adoption of the cross-domain open-set graph matching mechanism and the Euler-Margin

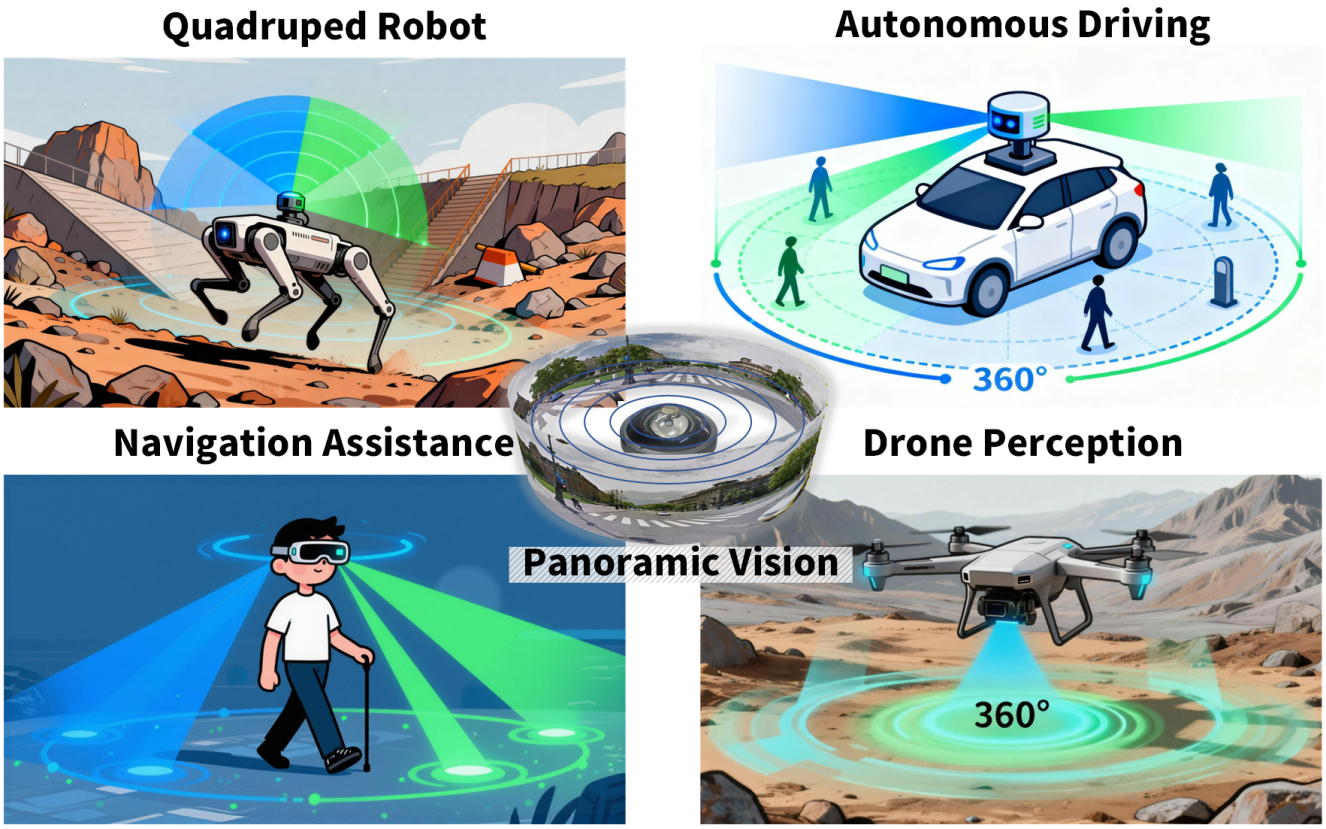

Figure 6. Societal implications of panoramic vision technology.

attention module. Although these components enhance the generalization in unseen views, they inevitably introduce additional parameter overhead and computational costs during inference. We provide two of the above issues. For model architecture, we consider adopting a lightweight self-attention mechanism integrated with our proposed Euler-interval projection and amplitude & phase modulation, aiming to preserve representational expressiveness while reducing parameter complexity. To enhance computational efficiency, we plan to conduct the quantization to accelerate inference and reduce memory consumption.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018. 6

[2] Seun-An Choe, Ah-Hyung Shin, Keon-Hee Park, Jinwoo Choi, and Gyeong-Moon Park. Open-set domain adaptation for semantic segmentation. In *CVPR*, 2024. 4, 6

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 6

[4] Zhen Fang, Jie Lu, Feng Liu, Junyu Xuan, and Guangquan Zhang. Open set domain adaptation: Theoretical bound and algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 3, 6

[5] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 3

[6] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 4, 6

[7] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*, 2022. 4, 6

[8] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: Masked image consistency for context-enhanced domain adaptation. In *CVPR*, 2023. 4, 6

[9] JoonHo Jang, Byeonghu Na, Dong Hyeok Shin, Mingi Ji, Kyungwoo Song, and Il-Chul Moon. Unknown-

aware domain adversarial learning for open-set domain adaptation. In *NeurIPS*, 2022. 4, 6

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *ICCV*, 2023. 2

[11] Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 7

[12] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 2025. 7

[13] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X. Yu, and Boqing Gong. Open compound domain adaptation. In *CVPR*, 2020. 3

[14] Chaoxiang Ma, Jiaming Zhang, Kailun Yang, Alina Roitberg, and Rainer Stiefelhagen. DensePASS: Dense panoramic semantic segmentation via unsupervised domain adaptation with attention-augmented context exchange. In *ITSC*, 2021. 6

[15] Shijie Ma, Fei Zhu, Xu-Yao Zhang, and Cheng-Lin Liu. ProtoGCD: Unified and unbiased prototype learning for generalized category discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 7

[16] Sarthak Maharana, Baoming Zhang, Leonid Karlinsky, Rogerio Feris, and Yunhui Guo. BATCLIP: Bimodal online test-time adaptation for CLIP. In *ICCV*, 2025. 7

[17] Luigi Riz, Cristiano Saltori, Yiming Wang, Elisa Ricci, and Fabio Poiesi. Novel class discovery meets foundation models for 3D semantic segmentation. *International Journal of Computer Vision*, 2025. 7

[18] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, 2018. 4, 6

[19] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021. 2, 4

[20] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José M. Álvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 4

[21] Gezheng Xu, Li Yi, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Changjian Shui, Charles Ling, A. Ian McLeod, and Boyu Wang. Unraveling the mysteries of label noise in source-free domain adaptation: Theory and practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 7

[22] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Universal domain adaptation. In *CVPR*, 2019. 4, 6

[23] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight SAM for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 2, 4

[24] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Philip H. S. Torr, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 4, 6

[25] Xu Zheng, Pengyuan Zhou, Athanasios V. Vasilakos, and Lin Wang. Semantics, distortion, and style matter: Towards source-free UDA for panoramic segmentation. In *CVPR*, 2024. 3

[26] Fei Zhu, Shijie Ma, Zhen Cheng, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. Open-world machine learning: A review and new outlooks. *arXiv preprint arXiv:2403.01759*, 2024. 7