Supplementary Material

Yaowen Chang[1,*] Zhen Cao[1,*] Xu Zheng[2] Xiaoxin Mi[3] Zhen Dong[1,†]
[1]Wuhan University [2]HKUST (Guangzhou) [3]Wuhan University of Technology
{yaowenchang, zhen.cao, dongzhenwhu}@whu.edu.cn, zhengxu128@gmail.com, xiaoxin.mi@whut.edu.cn

1. Overview

This supplementary material provides additional quantitative and qualitative results to further validate the effectiveness of **DAPASS** for panoramic source-free unsupervised domain adaptation. We present extended experiments on both indoor and outdoor settings, including analyses of image resolution and crop size, additional comparisons with representative baselines, and more qualitative visualizations. Overall, these results provide broader empirical support for the effectiveness of DAPASS.

2. More Quantitative Results

Analysis of hyper-parameter down sampling factor s. In the previous ablation study, we have explored the effect of PCGD hyperparameters. And in this section, we will discuss the CRAM hyper-parameter down sampling factor s. In the following, we analyse the underlying principles of DAPASS on C-to-D, starting with the influence of the resolution and crop size on SFUDA. For the comparison, we use the relative crop size sh/H_T, which is normalised by the image height at the corresponding resolution, to disentangle the crop size from the used image resolution. Fig. 1 shows that both an increased resolution and crop size improve the performance for both SFUDA and UDA. The larger crop provides more context clues and improves the performance of all classes, especially the ones that are difficult to adapt such as wall, fence, truck, bus, and train. And a higher input resolution improves the SFUDA performance by a smaller amount as it improves UDA learning. The improvement originates from a higher IoU for small classes such as pole, traffic light, traffic sign, person, motorbike, and bicycle, while some large classes such as road, sidewalk, and terrain have a decreased performance. This supports that large objects are easier to adapt at LR while small objects are easier to adapt at HR, which can be exploited by the multi-resolution fusion of CRAM.

Computational Complexity and Efficiency. To demonstrate the practical deployment potential of our proposed method, we provide a detailed efficiency analysis in Tab.A. Despite the inevitable increase in GFLOPs due to the multi-resolution processing at high resolution, our method maintains a highly competitive inference speed. Specifically, using the Seg-B1 backbone, the model achieves up to **30.0 FPS** on a single consumer-grade NVIDIA RTX 3090 GPU, meeting the requirements for real-time applications.

The parameter overhead introduced by our specialized modules is remarkably marginal ($< 1\%$, $\sim$0.08M), ensuring that the enhanced representation capability does not come at the cost of a bloated model size. Furthermore, the training process is highly efficient, converging in approximately 8 to 12 hours across different datasets. This efficiency stems from our optimized cross-resolution alignment mechanism, which effectively extracts domain-invariant features without redundant heavy-weight operations. These results underscore that our approach strikes a favorable balance between segmentation accuracy and computational footprint, making it well-suited for resource-constrained scenarios in panoramic vision tasks.

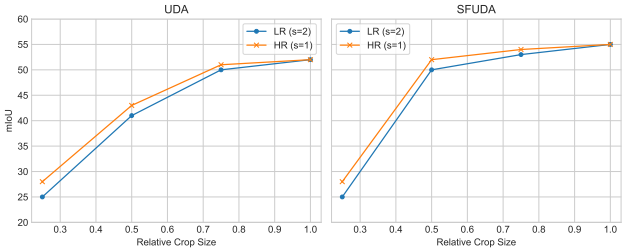

Figure 1. Analysis of the influence of resolution and crop size on UDA and SFUDA performance.

Table A. Efficiency Analysis (Backbone → Ours).

2*Backbone	Params	DensePASS ($400\times2k$)			Stanford2D3D ($1k\times2k$)		
	(M)	GFLOPs	FPS	Tr(h)	GFLOPs	FPS	Tr(h)
Seg-B1	15.2→15.3	57→108	**30.0**	8.0	145→240	14.0	12.0
Seg-B2	26.3→26.3	99→188	**21.0**	10.5	253→418	9.0	16.0

Performance on Minority Classes. Semantic segmenta-

*Equal Contribution
†Corresponding author

tion in panoramic imagery often suffers from extreme class imbalance. As detailed in Tab.B, we categorize the classes based on pixel frequency: **Minority classes** include those with $< 1\%$ pixel scarcity in **DensePASS** (*e.g., Pole, Truck, Bus, Bike*) and $< 10\%$ in **Stanford2D3D** (*e.g., Sofa, Table, Window*).

Standard SFUDA methods often over-fit to majority classes (like *Road* or *Wall*) due to their dominance in the loss function. However, our DAPASS framework significantly outperforms the strong baseline 360SFUDA++ across these challenging minority categories. For instance, with a Seg-B2 backbone, DAPASS achieves a **46.6%** mIoU on DensePASS minority classes, a substantial improvement over the baseline's 44.8%. This gain is primarily attributed to our multi-resolution fusion and alignment strategy, which preserves the fine-grained geometric features of small/rare objects that are typically lost during down-sampling. Following the evaluation protocol of 360SFUDA++, we demonstrate that our model maintains superior generalizability even when supervision is sparse, and we plan to extend this evaluation to even broader, more diverse domain datasets in future work.

Table B. Comparative analysis on Majority vs. Minority class performance. Metrics: SFDA / 360SFUDA++ → **Ours**.

2*Backbone	DensePASS (C-to-D)		Stanford (Spin-Span)	
	Majority	Minority	Majority	Minority
Seg-B1	54.2/59.1→**63.3**	32.4/42.1→**44.0**	72.1/88.8→88.6	44.4/57.5→**58.2**
Seg-B2	54.2/62.1→**64.5**	32.4/44.8→**46.6**	72.1/89.2→89.1	44.4/52.9→**59.2**

Combining Crops from Multiple Resolutions with CRAM. We combine crops from LR and HR using the proposed multi-resolution training strategy for SFUDA, aiming to exploit the complementary information provided by different spatial resolutions. LR crops preserve broader scene context, while HR crops provide richer local details and sharper object boundaries. By jointly leveraging these two types of inputs, CRAM can better balance global semantic understanding and fine-grained structural perception during adaptation. As shown in Tab. C, the multi-resolution setting improves the performance over both LR-only and HR-only baselines by +4.8 mIoU. This result demonstrates that multi-resolution fusion with scale attention enhances feature representation and leads to better domain adaptation.

Table C. Comparison of LR-only, HR-only and combination.

HR	LR	mIoU
$HR_{0.5}$	—	47.5 ± 1.2
—	$LR_{0.5}$	49.2 ± 0.8
$HR_{0.5}$	$LR_{0.5}$	55.0 ± 0.6

3. More Qualitative Results

Spin-to-Span visualisation. We further visualize the indoor segmentation results in Fig. 2. It is shown that our DAPASS not only accurately segments the door under severe distortions (shown on the right side of the figure), but also correctly recognizes the door when distortions are minimal, which illustrates that DAPASS can effectively combine contextual information with fine details to produce robust predictions.

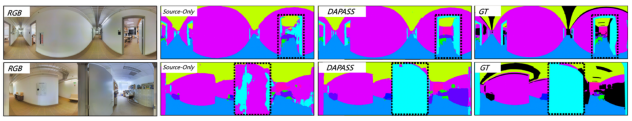

Figure 2. Segmentation visualizations on the Spin-to-Span setting. From left to right are the input image, Source Only, DAPASS, and Ground Truth.

C-to-D visualisation. Fig. 3 provides a qualitative comparison of the C-to-D adaptation scenario. The **Source-Only** model exhibits large coherent errors—e.g., sky is mistakenly labelled as building and road boundaries dissolve into sidewalk—confirming the pronounced domain gap. Incorporating generic SFDA strategies mitigates some of these block-level failures, yet fine structures remain fragmentary and polar-region distortions persist. In contrast, **DAPASS** produces masks that are almost indistinguishable from the ground truth: distant traffic signs and lamp posts are preserved with correct class colours, slender railings retain continuous geometry, and the upper- and lower-pole areas become markedly smoother. The red rectangles highlight these challenging regions; in every case DAPASS either restores objects entirely absent in competing outputs or repairs broken contours. These visual findings substantiate the numerical gains reported in Table 2 and confirm that the proposed PCGD-CRAM pipeline effectively suppresses pseudo-label noise while aligning multi-scale features under severe equirectangular distortion.

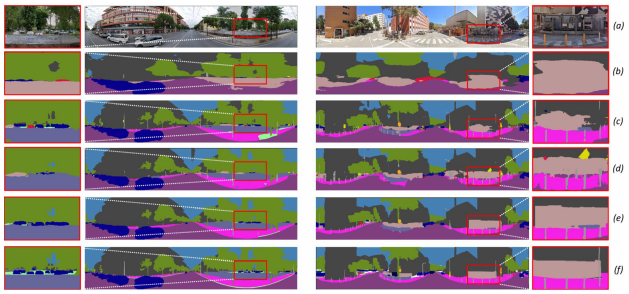

Figure 3. Segmentation visualizations on the C-to-D setting. From top to bottom are the input image, Source Only, SFDA, 360SFUDA++, DAPASS, and Ground Truth.