Supplementary Materials for Occlusion-Aware Seamless Segmentation

Yihong Cao[1,*], Jiaming Zhang[2,*], Hao Shi[3], Kunyu Peng[2], Yuhongxuan Zhang[1], Hui Zhang[1,†], Rainer Stiefelhagen[2], and Kailun Yang[1,†]

[1]Hunan University, [2]Karlsruhe Institute of Technology, [3]Zhejiang University

1 Datasets

Dense annotation of BlendPASS. To enhance the accuracy of dense annotations for objects of the *Thing* class, particularly in representing invisible regions of occluded objects realistically, our labeling process follows the "independent annotation → cross-verification → voting" workflow. All images are densely annotated by three skilled annotators using the EISeg tool [5] for initial segmentation. Specifically, three annotators densely annotate the full region of each object, and occluded objects are independently annotated by three annotators. Subsequently, cross-verification is conducted among the annotators. In cases where there are slight discrepancies in the annotations of occluded objects, the final annotation is determined through majority voting. For annotations that do not reach a consensus, more iterative annotation processes are employed until an agreement is reached. This annotation workflow aims to ensure the accuracy of dense annotations while maximizing the fidelity of annotations for occluded regions to the true object shape. Annotating panoramas with severe distortion is challenging and extremely time-consuming, requiring approximately 210 minutes per person per image. Finally, as illustrated in Tab. S1, 2,960 objects are annotated in the *Thing* class. We establish a finely labeled dataset, BlendPASS, based on panoramic images containing semantic, instance, and amodal instance labels. All annotations are cross-checked to support five tasks simultaneously: semantic segmentation, instance segmentation, amodal segmentation, panoptic segmentation, and amodal panoptic segmentation.

SynPASS. The SynPASS dataset [17] is a panoramic semantic segmentation dataset captured via the Carla simulator [3]. It has four weather conditions including sunny, cloudy, foggy, and rainy scenes, together with daytime and nighttime situations. Overall, it has 9,800 panoramic images with a resolution of 2048×1024 corresponding to a full Field of View (FoV) of 360°×180°, divided into training/validation/testing sets of 5,700/1,690/1,690 images, respectively.

Cityscapes→DensePASS. The Cityscapes→DensePASS benchmark [11] measures the performance of pixel-wise semantic segmentation models learned on 2,979 labeled pinhole images of Cityscapes and transferred to DensePASS, which

* Equal contribution
† Correspondence: zhanghuihby@126.com, kailun.yang@hnu.edu.cn

Table S1: Statistic for occluded and unoccluded objects in different classes.

	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	Total
#Occluded objects	189	6	909	42	18	1	83	38	1286
#Unoccluded objects	613	12	842	38	24	2	71	72	1674
Total	802	18	1751	80	42	3	154	110	**2960**

Table S2: Comparison of the proposed BlendPASS with existing datasets.

	Panoramic Image	Semantic GT	Instance GT	Amodal Instance GT	Cross-checking
KINS [13]	✗	✗	✓	✓	✓
KITTI360-APS [12]	✗	✓	✓	✓	✗
DensePASS [11]	✓	✓	✗	✗	✗
BlendPASS (Ours)	✓	✓	✓	✓	✓

has 100 panoramic images for evaluation. The pinhole images have a resolution of 2048×1024 and the panoramic images have a resolution of 2048×400.

KITTI360-APS→BlendPASS. We carefully reviewed the accessible annotations for KITTI360-APS [12], and 12,320 images containing 89,938 objects of thing classes are available. This dataset serves as the source pinhole domain of our proposed benchmark. For the introduced fresh BlendPASS dataset, considering the scalability of future work, we annotated 100 images containing 2,960 objects of thing classes in the evaluation set, following the format of Cityscapes [2]. This serves as the validation set for the target panoramic domain of the benchmark. However, due to inconsistent classes between the two datasets, we conducted additional manual processing. Specifically, in the KITTI360-APS, there are 11 valid *Stuff* and 7 valid *Thing* classes (while the work [12] claims 10 classes for *Stuff*, our manual verification confirmed that the *traffic light* class is an additional usable annotation class). These 11 *Stuff* classes align with BlendPASS. As for *Thing* classes, we adjusted the annotations of BlendPASS to align with KITTI360-APS, following the corresponding scheme in Tab. S3. This rough alignment further magnifies the challenges of cross-domain in the benchmark. Finally, for the OASS benchmark, there are a total of 11 aligned *Stuff* and 7 aligned *Thing* classes in both domains.

Table S3: Alignment scheme of the *Thing* categories between the KITTI360-APS dataset and our BlendPASS dataset.

Dataset	**Categories**							
BlendPASS	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle
KITTI360-APS [12]	Pedestrains	Cyclists	Car	Truck	Other-vehicles	Van	Two-wheeler	Two-wheeler

2 Evaluation Metrics

In the context of the Occlusion-Aware Seamless Segmentation (OASS) benchmark, we employ five metrics, namely Intersection over Union (IoU), Average Precision (AP), Amodal Average Precision (AAP), Panoptic Quality (PQ), and Amodal Panoptic Quality (APQ), to evaluate the model's performance. We provide a detailed explanation:

IoU. IoU measures the overlap between predicted segment p and ground truth segment g and is calculated as:

$$\text{IoU} = (p \cap g)/(p \cup g). \tag{1}$$

AP. We follow Pascal VOC [4] and COCO [8] and use average precision, which is computed by averaging the ten equally spaced IoU thresholds from 0.5 to 0.95.

AAP. The AAP is an extended metric of AP for Amodal Instance Segmentation, where the ground truth is replaced with amodal segments.

PQ. We adapt the standard panoptic quality metric proposed by [7] and compute it as

$$\text{PQ} = \frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}, \tag{2}$$

where $TP = \{(p,g) \in \boldsymbol{p} \times \boldsymbol{g} : \text{IoU}(p,g) > 0.5\}$ is a set of True Positive matches.

APQ. The APQ is an extended metric of PQ for Amodal Panoptic Segmentation, where the ground truth segments $\boldsymbol{g}$ are replaced with amodal segments.

3 More Results

3.1 More Experiment Details

We use the same data augmentation parameters as DACS [15] and set the RCS temperature to 0.01 to enhance sample frequency for rare classes of the source domain. The EMA decay η is set as 0.999. The threshold τ in pseudo-label weight is set as 0.968. Moreover, the pseudo-labels in the regions 11 pixels above and 88 pixels below the trained patches are ignored, respectively. Moreover, we adopt the ImageNet feature distance loss from DAFormer with a weight of 0.005. For instance and amodal instance branches, we set all loss weights to 1. During inference, the thresholds for instance and amodal instance are set to 0.95. For the retraining of the existing methods [14, 16, 18, 19], we utilize training protocols similar to ours and followed their hyperparameters to ensure fairness. Our experiments are conducted on an NVIDIA Tesla V100 GPU and implemented using PyTorch.

3.2 More OASS Results

In this work, we include a comprehensive set of sub-tasks, including semantic segmentation, instance segmentation, amodal instance segmentation, panoptic segmentation, and amodal panoptic segmentation for OASS. To provide a thorough

Table S4: Panoptic Segmentation results on the KITTI360-APS → BlendPASS benchmark. The per-class results are reported as PQ, and the metric is mPQ.

Task	UDA Method	road	sidewalk	building	wall	fence	pole	traffic-light	traffic-sign	vegetation	terrain	sky	pedestrians	cyclists	car	truck	other-vehicles	van	two-wheeler	Metric
PS	DATR [19]	50.44	09.14	59.92	11.93	11.98	01.95	00.00	03.91	64.60	14.05	70.45	12.15	00.00	38.09	00.00	03.38	00.00	01.29	19.63
	Trans4PASS [16]	53.93	14.12	69.39	19.16	11.77	03.77	00.00	05.15	67.62	16.02	77.41	14.60	04.09	38.23	06.91	00.00	00.00	07.19	22.80
	UniDAPS [18]	65.95	9.48	66.31	17.39	14.28	04.77	00.00	06.14	67.19	16.10	72.68	08.27	00.00	27.25	14.82	09.23	00.00	08.86	22.71
	EDAPS [14]	55.01	17.05	66.84	18.72	14.49	05.76	04.04	04.68	68.21	16.04	72.76	19.56	00.00	37.83	01.82	04.38	00.00	07.89	23.06
	Source-Only	57.84	14.21	73.83	15.49	07.59	00.67	00.00	10.40	58.30	12.39	83.15	14.85	00.00	39.06	05.96	00.00	00.00	07.68	22.30
	UnmaskFormer (Ours)	61.73	24.72	66.80	20.75	15.81	05.22	04.29	03.26	69.02	18.35	79.44	20.90	03.45	42.96	11.26	07.64	00.00	15.98	**26.20**

analysis of these diverse tasks, we present the per-class accuracy results. Tab. S4 further details the per-class accuracy specifically in panoptic segmentation (PS). To benchmark our model, we conduct a comparative study against previous state-of-the-art methods, namely DATR [19], Trans4PASS [16], UniDAPS [18], and EDAPS [14]. As shown in the experimental results, our model demonstrates a notable Mean-PQ of 26.20%, surpassing the performance of the previous best model by a significant margin of +3.14%. A detailed breakdown of per-class performance reveals that our model excels across numerous categories, such as *sidewalk*, *wall*, *pedestrians*, *car*, and *two-wheeler*. It is worth noting that the *van* category, constituting a minor proportion in both source and target datasets, poses a challenge for all models to handle effectively. The superior mPQ achieved by our model proves its effectiveness in addressing the challenges in the panoptic segmentation part of the OASS task.

3.3 More Analysis of Ablation Study

Preliminary of Deformable Patch Embedding. The backbone utilizes a transformer-based structure with a novel arrangement of Deformable Patch Embedding (DPE) [16] layers. DPE aims to capture local geometric variations caused by image distortion. Given an input image or features $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$ (where H and W represent the resolution and C the number of channels), DPE calculates adaptive offsets for each patch:

$$\boldsymbol{\Delta}^{DPE}(i,j) = \left[\min(\max(-\tfrac{H}{r}, g(\boldsymbol{X})_{(i,j)}), \tfrac{H}{r}) \;\; \min(\max(-\tfrac{W}{r}, g(\boldsymbol{X})_{(i,j)}), \tfrac{W}{r})\right], \tag{3}$$

where (i, j) denotes the patch index, $g(\cdot)$ is the offset prediction function, r is a hyperparameter controlling the maximum allowed offset.

Analysis of UA-based Backbone. To improve the modeling capacity of our OASS model, we take not only the deformable patchifying but also the unmasking into account. Specifically, the technical designs involve multiple perspectives. *1) Interleaved DPE arrangement:* Unlike prior work [16] that inserts DPE only in the initial stage, this approach proposes a novel interleaving arrangement. DPE layers are strategically placed within Stages 2 and 4 of the backbone architecture. This design reinforces the model's ability to capture distortion throughout

the processing pipeline. *2) Unmasking Attention (UA) for occlusion handling:* In addition to addressing distortion, the method incorporates an Unmasking Attention (UA) block that is enhanced by adding a simple yet effective pooling layer. *3) Combining self-attention and enhanced pooling:* The UA block leverages both a self-attention layer and an improved pooling mechanism. Having the self-attended pooling feature $\boldsymbol{q'} \in \mathbb{R}^{1\times1\times C}$, a `sigmoid` function $\phi(\cdot)$ is applied to calculate the occlusion-aware mask $\phi(\boldsymbol{q'})$. This mask highlights regions likely affected by occlusion. *4) Incorporating occlusion awareness:* The mask $\phi(\boldsymbol{q'})$ is subsequently used to perform element-wise multiplication with the original feature map, resulting in an occlusion-aware feature. The occlusion-aware feature is further processed by an MLP layer. Based on these crucial designs, our UA-based backbone can address both the image distortion and object occlusion.

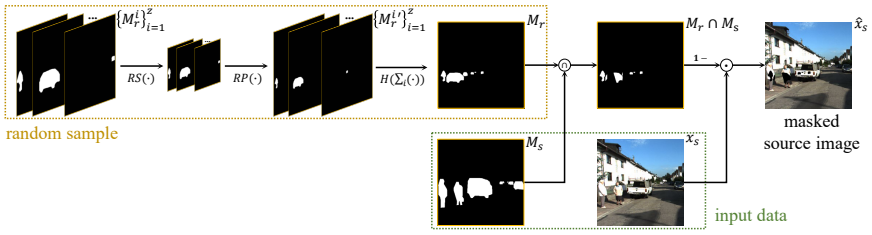

Fig. S1: The process of modeling amodal-oriented masked source images.

Analysis of AoMix. To provide additional amodal-oriented source priors and mixed image samples to enhance the adaptation, we utilize amodal instance masks to mask input images within the AoMix module. An example of the amodal-oriented masked image modeling process for the source image is illustrated in Fig. S1. This method aims to enhance the model's ability to reconstruct object regions obscured by realistic object shapes, enabling it to learn information about invisible parts in the scene. As shown in Tab. 6, we conducted an ablation study on the manner and strategy of AoMix. *1) T for S, T for M vs AoMix:* Using the amodal-oriented masked image in both the source and mixed images enables the model to accurately segment the full regions of occluded objects in the source domain and facilitates better model adaptation to the target panoramic domain. *2) P for S&M vs AoMix:* It is noteworthy that we tried to use random patches instead of amodal instance masks to mask images. Although the mIoU score remained largely unchanged, the mAPQ score witnessed a significant drop of 3.3%, affirming that masks with real shapes provide better guidance for learning occlusion-ignored segmentation ability. *3) W for S&M vs AoMix:* Compared with applying amodal masks to all regions of the image, our method only applies to the *Thing* class region, resulting in superior performance. This is because our method aligns more closely with the real-world scenario where object occlusion of the *Thing* class is often raised by other objects of the *Thing* class.

4 More Visualization Results

4.1 More Visualization Results of OASS

We showcase visualization results for semantic segmentation and panoptic segmentation in Fig. S2 and Fig. S3. These examples show that UnmaskFormer achieves outstanding performance on other tasks within the OASS benchmark. In addition to addressing the occlusion of perspective, UnmaskFormer can unmask the narrow field of view and the gap of domain. As depicted in Fig. S2 for panoptic segmentation, UnmaskFormer surpasses other methods [14, 16, 18, 19] and successfully detects more pedestrians. Compared to the contour-based UniDAPS [18], UnmaskFormer segments the *Thing* objects with more complete and rational shapes. For the category determination of instance masks, existing methods typically rely on fusing results from the semantic branch. Moreover, the visualization results for semantic segmentation in Fig. S3 show that UnmaskFormer achieves more accurate semantic classification results and demonstrates robustness to distortions introduced by wide-FoV panoramic images.

4.2 More Visualization Results on SynPASS

As shown in Fig. S4, we conduct qualitative analyses of panoramic semantic segmentation on the SynPASS dataset. The objective of this analysis is to evaluate the robustness of various methods under diverse weather conditions. Specifically, we examine four distinct weather types, namely *cloudy*, *foggy*, *rainy*, and *sunny*, to assess the adaptability of the models. From top to bottom in Fig. S4, they are the input image, the segmentation results of Trans4PASS [16], the segmentation results of UnmaskFormer, and ground truth. Comparing the performance of UnmaskFormer with the prior state-of-the-art model, Trans4PASS [16], it can be seen that UnmaskFormer obtains notable improvements in handling adverse weather conditions. For instance, in *cloudy* weather, UnmaskFormer demonstrates an enhanced capability to predict and generate more complete *rail track* categories. Moreover, in *foggy* and *rainy* weather, UnmaskFormer can identify *roads* more completely compared to the baseline model. Thanks to the deformable designs, UnmaskFormer can yield better panoramic semantic segmentation results. These observations demonstrate the efficacy of UnmaskFormer in challenging environmental scenarios.

5 Discussion

Future work. The advancements made in Occlusion-Aware Seamless Segmentation (OASS) through our UnmaskFormer open new avenues for future research. In the future, we envision several potential further directions: *1) User-Interactive Image Editing:* We propose exploring user interaction in image editing. Future iterations can integrate user-selected areas to separate amodal layers and employ diffusion processes for completion. These interactive features enhance user-friendliness and expand practical applications. *2) Amodal Optical Flow Fusion:*

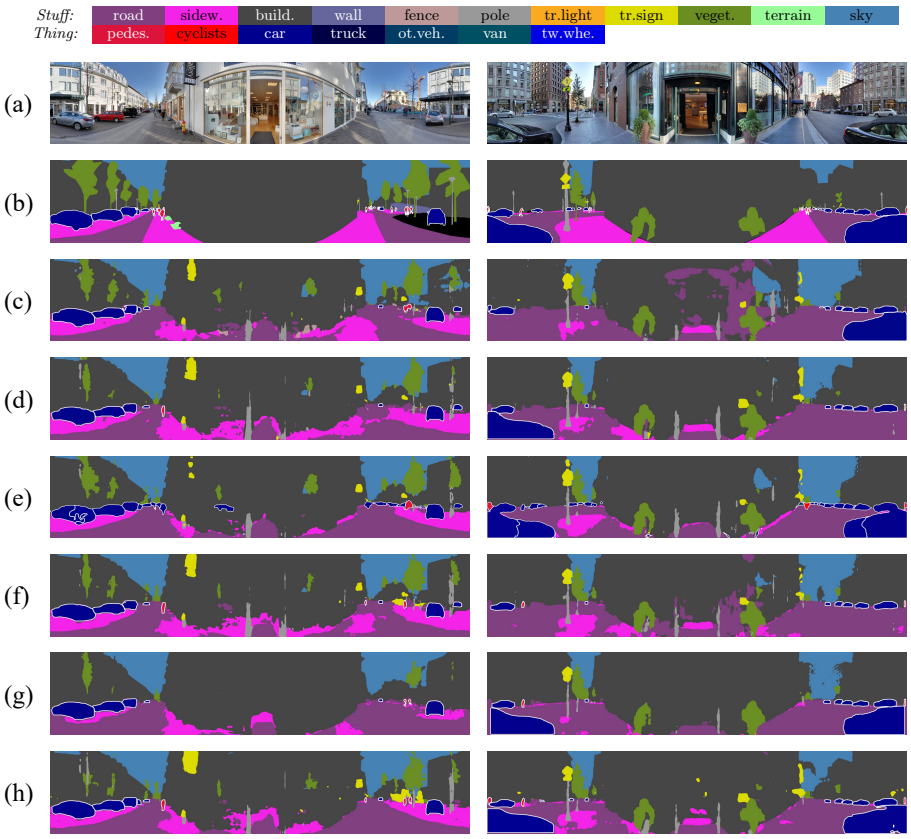

Fig. S2: Visualization results of Panoptic Segmentation. From top to bottom are (a) Image, (b) Ground truth, (c) DATR [19], (d) Trans4PASS [16], (e) UniDAPS [18], (f) EDAPS [14], (g) Source-Only, and (h) UnmaskFormer (Ours).

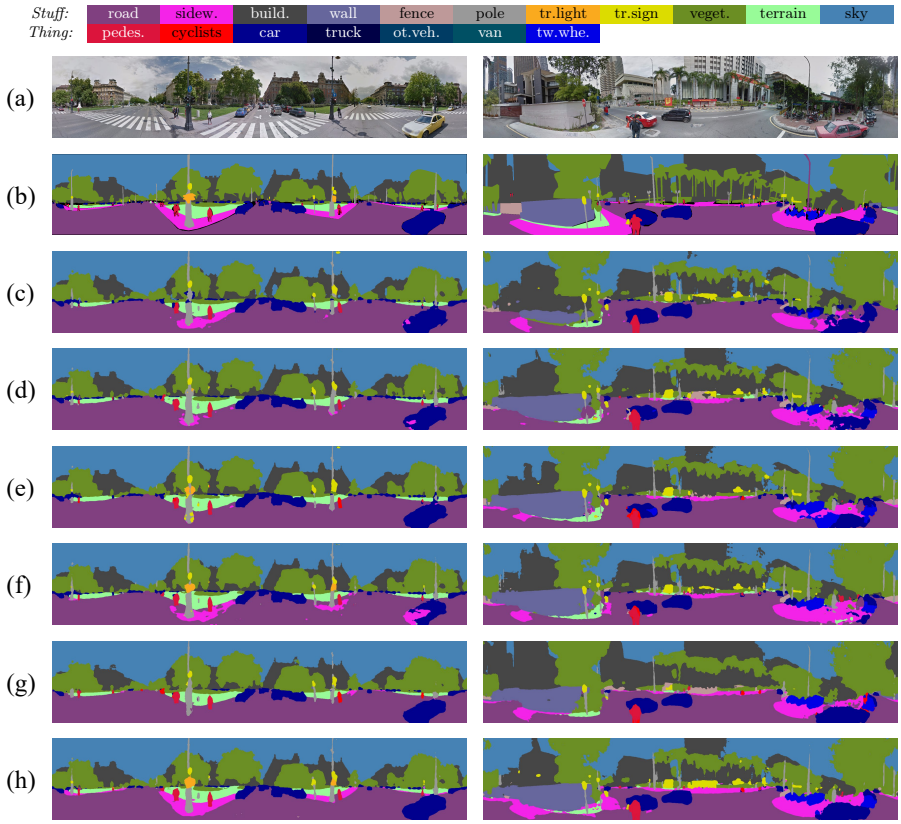

Fig. S3: Visualization results of Semantic Segmentation. From top to bottom are (a) Image, (b) Ground truth, (c) DATR [19], (d) Trans4PASS [16], (e) UniDAPS [18], (f) EDAPS [14], (g) Source-Only, and (h) UnmaskFormer (Ours).

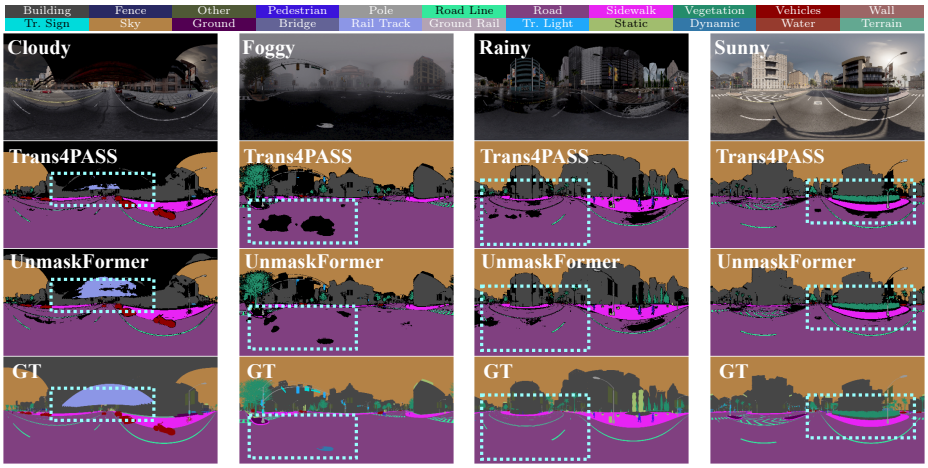

Fig. S4: More visualization results of SynPASS. From top to bottom are: Image, the prediction of Trans4PASS [16], the prediction of our UnmaskFormer, and Ground truth.

We intend to further investigate combining our segmentation approach with amodal optical flow techniques [10] to improve temporal consistency in amodal segmentation. This is useful for dealing with temporal occlusions. Also, it is essential for continuous tracking or monitoring applications, potentially revolutionizing dynamic scene processing.

Limitations and potential solutions. While our UnmaskFormer for OASS showcases significant advancements, we acknowledge certain limitations in our current work, alongside potential approaches to address these challenges: *1) Complex Environment Perception:* The current model may not handle the complexity of in-the-wild scenes. Continuous improvement is vital for panoramic image processing and scene understanding. We could enhance robustness through diverse data, integration with other sensors like LiDAR or event cameras, and depth-aware transferring learning [1] to enhance occlusion reasoning. Additionally, domain generalization techniques [6] can be employed to improve the model's adaptability to new and unseen environments, further bolstering its performance in complex scenarios. *2) Amodal Data Annotation Challenge:* Amodal data annotation for panoramic images is challenging. While our dataset covers diverse images captured in cities located on all continents, we could further explore automated amodal annotation for panoramas, reserving manual annotation for data purification and cleaning purposes. Moreover, semi-supervised amodal instance segmentation methods [9] can also be leveraged to enhance the efficiency of the annotation process. These approaches can help scale up the availability of annotated data for training and validation, addressing the scarcity of labeled data in the amodal segmentation domain.

Societal impacts. In this study, we have introduced a novel task called Occlusion-Aware Seamless Segmentation (OASS) and established a comprehensive bench-

mark incorporating various well-known baseline models. We found that these baseline models exhibit limited performance in the OASS task, primarily due to the intricate nature of occlusion-aware segmentation challenges. To address this, we have developed UnmaskFormer, a solution that significantly enhances performance on the OASS benchmark, outperforming existing domain adaptation panoramic and panoptic segmentation methods and achieving promising state-of-the-art results. Nevertheless, given the criticality of dependability in deep learning systems for Advanced Driver-Assistance Systems (ADAS), it is important to note that UnmaskFormer may still encounter misclassifications in challenging occluded regions and biased content, potentially leading to erroneous predictions with adverse societal implications.

References

1. Chen, M., Zheng, Z., Yang, Y.: Transferring to real-world layouts: A depth-aware framework for scene adaptation. arXiv preprint arXiv:2311.12682 (2023) 9
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) 2
3. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: CoRL (2017) 1
4. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. International Journal of Computer Vision (2010) 3
5. Hao, Y., Liu, Y., Chen, Y., Han, L., Peng, J., Tang, S., Chen, G., Wu, Z., Chen, Z., Lai, B.: EISeg: An efficient interactive segmentation tool based on PaddlePaddle. arXiv preprint arXiv:2210.08788 (2022) 1
6. Jiang, X., Huang, J., Jin, S., Lu, S.: Domain generalization via balancing training difficulty and model capability. In: CVPR. pp. 18993–19003 (2023) 9
7. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: CVPR (2019) 3
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014) 3
9. Liu, Z., Li, Z., Jiang, T.: BLADE: Box-level supervised amodal segmentation through directed expansion. In: AAAI (2024) 9
10. Luz, M., Mohan, R., Sekkat, A.R., Sawade, O., Matthes, E., Brox, T., Valada, A.: Amodal optical flow. arXiv preprint arXiv:2311.07761 (2023) 9
11. Ma, C., Zhang, J., Yang, K., Roitberg, A., Stiefelhagen, R.: DensePASS: Dense panoramic semantic segmentation via unsupervised domain adaptation with attention-augmented context exchange. In: ITSC (2021) 1, 2
12. Mohan, R., Valada, A.: Amodal panoptic segmentation. In: CVPR (2022) 2
13. Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J.: Amodal instance segmentation with KINS dataset. In: CVPR (2019) 2
14. Saha, S., Hoyer, L., Obukhov, A., Dai, D., Van Gool, L.: EDAPS: Enhanced domain-adaptive panoptic segmentation. In: ICCV (2023) 3, 4, 6, 7, 8
15. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: DACS: Domain adaptation via cross-domain mixed sampling. In: WACV (2021) 3

16. Zhang, J., Yang, K., Ma, C., Reiß, S., Peng, K., Stiefelhagen, R.: Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In: CVPR (2022) 3, 4, 6, 7, 8, 9
17. Zhang, J., Yang, K., Shi, H., Reiß, S., Peng, K., Ma, C., Fu, H., Torr, P.H.S., Wang, K., Stiefelhagen, R.: 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) 1
18. Zhang, J., Huang, J., Lu, S.: Hierarchical mask calibration for unified domain adaptive panoptic segmentation. In: CVPR (2023) 3, 4, 6, 7, 8
19. Zheng, X., Pan, T., Luo, Y., Wang, L.: Look at the neighbor: Distortion-aware unsupervised domain adaptation for panoramic semantic segmentation. In: ICCV (2023) 3, 4, 6, 7, 8