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

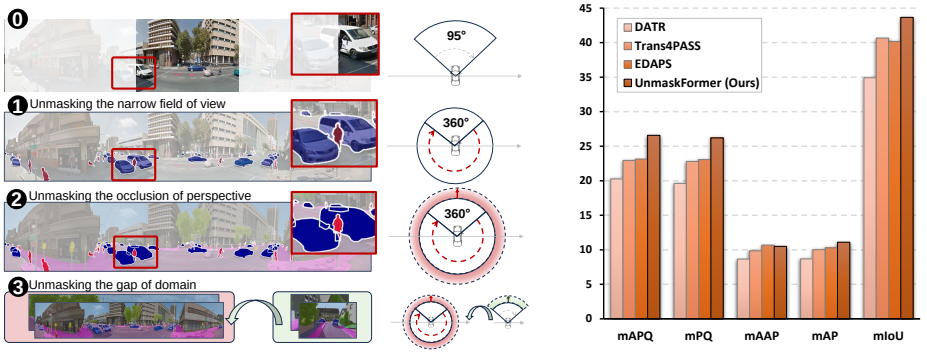
Abstract. Panoramic images can broaden the Field of View (FoV), occlusion-aware prediction can deepen the understanding of the scene, and domain adaptation can transfer across viewing domains. In this work, we introduce a novel task, **Occlusion-Aware Seamless Segmentation (OASS)**, which simultaneously tackles all these three challenges. For benchmarking OASS, we establish a new human-annotated dataset for Blending Panoramic Amodal Seamless Segmentation, *i.e.*, **BlendPASS**. Besides, we propose the first solution **UnmaskFormer**, aiming at unmasking the narrow FoV, occlusions, and domain gaps all at once. Specifically, UnmaskFormer includes the crucial designs of Unmasking Attention (UA) and Amodal-oriented Mix (AoMix). Our method achieves state-of-the-art performance on the BlendPASS dataset, reaching a remarkable mAPQ of 26.58% and mIoU of 43.66%. On public panoramic semantic segmentation datasets, *i.e.*, SynPASS and DensePASS, our method outperforms previous methods and obtains 45.34% and 48.08% in mIoU, respectively. The fresh BlendPASS dataset and our source code are available at `https://github.com/yihong-97/OASS`.

Keywords: Panoramic Scene Understanding · Amodal Segmentation

1 Introduction

Panoramic imaging has advanced significantly [32, 82], allowing the capture of high-quality 360° images with minimalist optical systems [31] suitable for a wide variety of omnidirectional vision applications [1, 17]. Concurrently, panoramic scene understanding has advanced in many areas, such as dense visual prediction [44, 60, 64, 83], holistic scene understanding [65, 87], and panoramic scene segmentation [26, 53, 73, 76]. On the other side, amodal perception [52], a fundamental aspect of human vision that forms the basis of our understanding and interpretation of the world, motivates occlusion-aware amodal prediction [2, 50, 56] aimed at achieving recognition of an object and its complete spatial extent. These diverse research endeavors converge toward the common goal of achieving more comprehensive visual perception and understanding.

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



(a) Occlusion-Aware Seamless Segmentation (OASS) task involves three challenges: (1) unmasking the narrow field of view, (2) unmasking the occlusion of perspective, and (3) unmasking the gap of domain.

(b) Results on our BlendPASS benchmark. Domain adaptive panoptic and panoramic segmentation methods [57, 89, 93] are compared.

Fig. 1: The overview, challenges, and comparison on the proposed Occlusion-Aware Seamless Segmentation (OASS) task.

To unify the aforementioned scene understanding in a seamless form and advance a more comprehensive perception, we introduce **OASS (Occlusion-Aware Seamless Segmentation)**. As illustrated in Fig. 1, the OASS task offers threefold benefits while posing three significant challenges: (1) Panoramic images offer a broader Field of View (FoV), unmasking the narrow FoV of the pinhole imagery, *e.g.*, from 95° to 360°, as compared between ❶ and ❶ of Fig. 1a. However, panoramas introduce severe distortions compared to pinhole images, which will lead to a significant performance degradation [89, 93]. (2) Amodal prediction [2], in contrast to segmentation limited in visible areas [19, 72], can unmask the occlusion in the space perspective. For example, compared between ❶ and ❷ of Fig. 1a, the occluded *pedestrian* can be completely recognized by using amodal prediction. Predicting the complete mask of occluded objects, *i.e.*, occlusion reasoning, is important to enhance the spatial-wise understanding capacity [7,33,37,85]. (3) Unsupervised Domain Adaptation (UDA) [57,91] in ❸ is capable of addressing the need for expensive training data, unmasking domain gaps from label-rich pinhole to label-scare panoramic imagery.

Previous approaches [4,50,56,89,93] are proposed to address the above challenges separately, resulting in sub-optimal and non-seamless solutions. For example, as the comparison depicted in Fig. 2, methods [89, 93] proposed to address panoramic semantic segmentation cannot handle the object occlusion, whereas methods [4, 50, 56] proposed for amodal segmentation cannot generalize well to panoramic imagery and the FoV occlusion remains unsolved. To address the challenges seamlessly, we propose a novel unmasking transformer framework called **UnmaskFormer**. The novelty is three-fold: (1) An *Unmasking Attention (UA)* constructed by a self-attention and an enhanced pooling layer for occlusion prediction, is proposed to unmask the object occlusions within the whole Unmask-

Former framework. (2) We delve deeper into the design of the Deformable Patch Embedding (DPE) [89] and find an alternative yet better solution for addressing the image distortion of panoramas at different stages within the transformer model. (3) We propose an *Amodal-oriented Mix (AoMix)* method that aims to seamlessly integrate the pinhole and panoramic domains, addressing the challenges posed by inconsistent cross-domain scenes. This method also enhances the model's capacity to reconstruct invisible regions of occluded objects, ultimately allowing it to unmask the occlusion of perspective within a scene. Based on these crucial designs, our UnmaskFormer can better handle different panorama-based scene understanding tasks, especially solving object occlusion and image distortion. Fig. 1b shows that UnmaskFormer has striking performances across Semantic (mIoU), Instance (mAP), Amodal Instance (mAAP), Panoptic (mPQ), and Amodal Panoptic Segmentation (mAPQ) tasks by using one model.

To facilitate this evaluation, we spent a large effort to collect and manually annotate a dataset for **Blending Panoramic Amodal Seamless Segmentation (BlendPASS)**. There are 2,000 panoramic images with 360° FoV and a 2048×400 resolution, captured in street scenes. BlendPASS facilitates the learning optimization of segmentation adaptation in an unsupervised fashion, which unfolds as efficient for dense prediction in panoramic imagery. To establish the evaluation benchmark for the OASS task, 100 panoramic images have been manually annotated precisely at the pixel level.

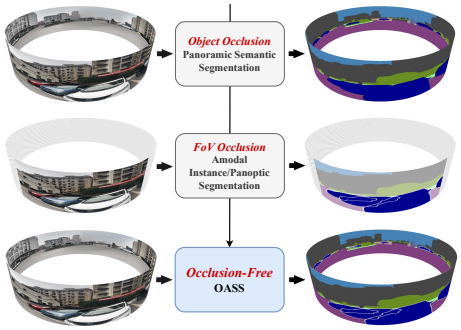


Fig. 2: OASS addresses *object occlusion* and *FoV occlusion* limitations in segmentation tasks.

Extensive experiments are conducted on our proposed BlendPASS dataset and other panoramic datasets. Our UnmaskFormer achieves state-of-the-art performance on BlendPASS, reaching a remarkable mAPQ of 26.58% and mIoU of 43.66%. On the panoramic semantic segmentation datasets SynPASS [90] and DensePASS [48], our method outperforms the previous best method and obtains 45.34% and 48.08% in mIoU, respectively. The significant improvements and results prove the effectiveness of the proposed UnmaskFormer framework in addressing the panorama-based scene understanding.

In this work, we propose contributions as follows:

- We introduce a new panorama-based segmentation task, *i.e.*, **OASS**, aiming at unmasking the narrow field of view, unmasking the occlusion of perspective, and unmasking the gap of domain in a seamless manner.
- To address the OASS task, we propose an **UnmaskFormer** framework with distortion- and occlusion-aware designs in a transformer-based architecture.

- An *Amodal-oriented Mix (AoMix)* method is tailored for improving unsupervised domain adaptation and overcoming the challenges of diverse occlusions in bridging the gap between pinhole and panoramic domains.
- A new panoramic dataset *BlendPASS* is created and manually annotated for benchmarking the blending of panoramic amodal seamless segmentation.

2 Related Work

Domain adaptive panoramic segmentation. Segmentation on wide-FoV fisheye images [11, 58, 61, 80, 81] and panoramic images [20, 39, 75, 78, 95] enables holistic understanding of 360° surroundings [74]. To address the scarcity of annotations, researchers have revisited wide-FoV segmentation from the perspective of UDA [28, 34, 62, 77, 79] by leveraging rich training sources from conventional narrow-FoV data. A variety of adaptation methods including self-training [23–25,70,86,98] and adversarial training [47,48] methods are studied for panoramic segmentation. In this line, P2PDA [88] employs attention-regulated uncertainty-aware adaptation. Trans4PASS [89, 90] introduces distortion-aware transformers and mutual prototypical adaptation with SAM [36]. DPPASS [94] presents a dual-path solution to overcome style and format gaps. DATR [93] captures neighboring distributions without any geometric constraints. Moreover, panoramic panoptic segmentation [15, 30, 49] is investigated by using techniques like contrastive learning for harvesting generalization benefits [29]. In this work, we aim to enable occlusion-informed understanding with both FoV-wise and spatially amodal reasoning of urban scenes.

Amodal scene segmentation. Amodal instance segmentation is a derivative task of instance segmentation [3,10,13,21,42,45,63,66], to predict visible regions of objects and their occluded regions simultaneously. Li *et al*. [38] introduce the concept of amodal instance segmentation, achieved through iterative regression. Qi *et al*. [56] derive the amodal version of KITTI [18], introducing an independent occlusion classification branch. SLN [92] incorporates a semantics-aware distance map to implement amodal segmentation. ORCNN [14] calculates visible masks to infer occlusion masks. Further, shape and contour priors are frequently leveraged for refining amodal segmentation [5, 16, 40, 41, 71]. Aside from amodal instance segmentation, amodal semantic understanding [4, 27, 97] is an important direction. Along this line, [50, 51] propose a fusion of semantic and amodal instance segmentation. Sekkat *et al*. [59] uses the CARLA simulator [12] to create a virtual multi-task amodal perception dataset. Different from these works, we intend to achieve occlusion-aware seamless segmentation, which breaks the occlusion limits in terms of both field-of-view and in-field object occlusions.

3 Established Benchmark

3.1 Overview of the Benchmark

In this work, we establish a novel benchmark specifically designed for OASS. In particular, we address OASS from the perspective of UDA by adapting from the

label-rich pinhole domain to the label-scarce panoramic domain. 360° panoramas have a wide FoV and many small objects in the images, which exaggerate the costs of creating pixel-wise annotations in unconstrained surroundings. Our benchmark overcomes the scarcity of amodal segmentation testbeds in panoramic imagery, while addressing three aforementioned challenges: ❶ *how to address the severe distortions of panoramas when unmasking the narrow field of view*, ❷ *how to predict the full segmentation of objects when unmasking the occlusion of perspective*, and ❸ *how to facilitate optimization of segmentation adaptation when unmasking the gap of domain*.

Our objective with this benchmark is to provide a comprehensive evaluation of methods capable of performing both FoV-wise and spatially occlusion-aware seamless segmentation. We extend the metrics proposed by [35] to the full regions incorporating pixels of the occluded objects. The benchmark metrics cover three aspects: Intersection over Union (IoU) for semantic segmentation, Average Precision (AP) and Panoptic Quality (PQ) for instance and panoptic segmentation, Amodal Average Precision (AAP) and Amodal Panoptic Quality (APQ) for amodal instance and amodal panoptic segmentation.

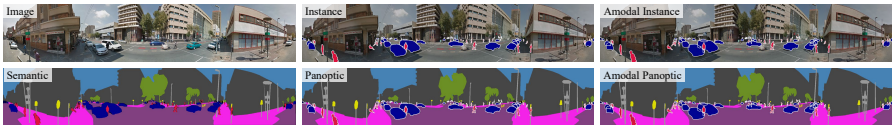


Fig. 3: The established BlendPASS dataset. We provide pixel-level labels for five segmentation tasks related to OASS on the validation set. Zoom in for a better view.

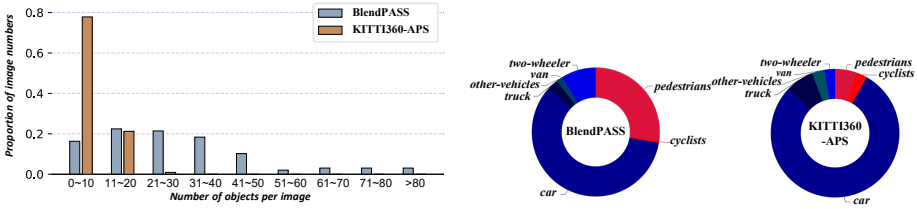
3.2 BlendPASS

We introduce a novel dataset for *Blending Panoramic Amodal Seamless Segmentation* (BlendPASS) tailored for OASS. BlendPASS comprises an unlabeled training set of 2,000 panoramic images for optimizing domain adaptation and a labeled test set of 100 panoramic images. These images are captured from panoramic cameras in driving scenes, all at a resolution of 2048×400 pixels. As depicted in Sec. 3.1, we have provided pixel-wise annotations for five distinct visual segmentation tasks, which greatly extends the semantic labels from DensePASS [48]. Specifically, we have further annotated the instance and amodal instance labels. These labels cover 19 categories that align with the Cityscapes [9] and are further categorized into *Stuff* (*road, sidewalk, building, wall, fence, pole, light, sign, vegetation, terrain, and sky*) and *Thing* (*person, rider, car, truck, bus, train, motorcycle, and bicycle*). To ensure the precision of annotations and the rational handling of occluded object parts, we meticulously conduct manual annotation for the test set, with three annotators following a cross-checking process. Finally, 2,960 objects are annotated in the *Thing* class, with 43% of these objects exhibiting occlusion. More details can be found in the supplementary.

3.3 KITTI360-APS→BlendPASS

For the labeled source in our UDA benchmark for OASS, we employ the available KITTI360-APS dataset [50] designed for amodal panoptic segmentation. This dataset is an extension of KITTI360 [43] and includes annotations for inmodal/amodal instance and panoptic segmentation. The images in KITTI360-APS are captured using pinhole cameras from 9 cities, with a resolution of 1408×376. After our careful examination, a total of 12,320 annotated images are accessed. These annotations encompass 10 *Stuff* (*road, sidewalk, building, wall, fence, pole, traffic sign, vegetation, terrain, and sky*) classes and 7 *Thing* (*car, pedestrians, cyclists, two-wheeler, van, truck, and other vehicles*) classes. In our benchmark, we further process the annotations from BlendPASS to ensure class alignment with KITTI360-APS.

For KITTI360-APS→BlendPASS under OASS, besides the challenges mentioned in Sec. 3.1, as illustrated in Fig. 4, the source-target domains exhibit significant differences in terms of the number of objects per image and the class distribution. These differences present a greater challenge to UDA models.



(a) Statistics of object counts per image. **(b)** Class distribution of the dataset.

Fig. 4: Analysis of test set (BlendPASS) and the training set (KITTI360-APS).

4 Methodology

4.1 UnmaskFormer

Architecture. For the OASS task, the model simultaneously addresses objectives including semantic, instance, amodal instance, panoptic, and amodal panoptic segmentation. Similarly to panoptic segmentation, as illustrated in Fig. 5, we decompose the base model $\mathcal{F}$ into a feature extractor and three branches, *i.e.*, semantic, instance, and amodal instance, to accomplish OASS goals. The semantic branch predicts the per-pixel semantic category of the input image, while the instance and amodal instance branches output class-agnostic object localization predictions. It is noteworthy that, while the instance branch has both top-down [57] and bottom-up [91] decoders, the contour-based bottom-up decoder is impractical for the amodal instance branch in the context of OASS.

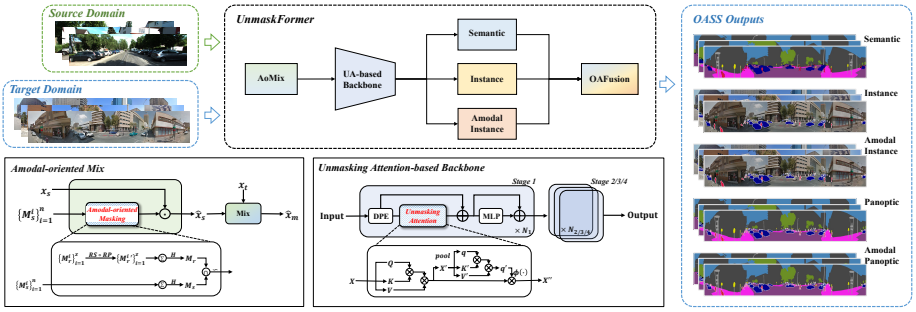


Fig. 5: Overview of the proposed UnmaskFormer framework. The UA-based backbone consists of the DPE [89] and the proposed UA, addressing image distortion and object occlusion in panoramas at once. The AoMix is designed to process input images x_s, x_t, seamlessly integrating pinhole and panoramic domains while enhancing the model's capacity to reconstruct invisible regions of the occluded objects.

This limitation arises from situations where a single pixel is associated with multiple objects simultaneously. In contrast, the proposals-based top-down decoder typically follows Mask R-CNN [22], employing Region Proposal Networks (RPN) to predict candidate objects. This methodology facilitates simultaneous and interference-free segmentation of multiple objects. Hence, in UnmaskFormer, we adopt the top-down decoder for the instance branch and the occlusion-aware amodal instance branch.

For the final output of the UnmaskFormer, we construct an Occlusion-Aware Fusion (OAFusion) module to process outputs from three branches, generating five segmentation maps at once for OASS. The semantic segmentation is directly predicted by the semantic branch. In instance and amodal instance segmentation, the semantic label of class-agnostic instance or amodal instance mask is determined by the majority-voting rule of the predictions from the semantic branch. For amodal segmentation, only regions where the current object does not overlap with other objects are considered. The panoptic and amodal panoptic segmentation are generated by fusing the semantic segmentation with instance and amodal instance segmentation, respectively.

UA-based backbone. One of challenges in OASS is to address both image distortion and object occlusion at the same time. For this end, we construct a UA-based backbone by rearranging previous Deformable Patch Embedding (DPE) and adding an effective pooling layer, as shown in Fig. 5. Specifically, for an input target image or features $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$ between different stages, the original DPE calculates the adaptive offsets and performs the patchifying process. Contrary to inserting DPE in the early stage [89], we perform a new interleaving arrangement of DPE layers in Stages 2 and 4, which can provide reinforced distortion-aware modeling for the whole UnmaskFormer framework. More analysis will be presented in Sec. 5.4.

Apart from the distortion-aware design, we design the UA block by combining the self-attention layer and the enhanced pooling layer. As shown in Fig. 5, after

the self-attention layer, the feature $\boldsymbol{X'}\in\mathbb{R}^{H\times W\times C}$ is forwarded to calculate the pooling query $\boldsymbol{q}$, the key $\boldsymbol{K'}$ and the value $\boldsymbol{V'}$, where $\boldsymbol{q}$=$GAP(\boldsymbol{X'})$ and $GAP(\cdot)$ is the global average pooling operator. After having the self-attended pooling feature $\boldsymbol{q'}\in\mathbb{R}^{1\times 1\times C}$, a `sigmoid` function $\phi(\cdot)$ is applied to calculate the occlusion-aware mask $\phi(\boldsymbol{q'})$. Then, the pool-attended occlusion-aware mask $\phi(\boldsymbol{q'})$ is used to perform a dot product operation with feature $\boldsymbol{X'}$ to obtain the occlusion-aware feature $\boldsymbol{X''}\in\mathbb{R}^{H\times W\times C}$. After the UA block, the occlusion-aware feature $\boldsymbol{X''}$ is further forwarded to an MLP layer as self-attention blocks [72, 89].

The overall UA-based backbone is constructed as the same four-stage architecture as the common segmentation methods [69, 72, 89].

4.2 Cross-Domain Adaptation

UDA strategy. In the OASS setting, the model is assigned to adapt from a labeled source domain $\mathcal{X}_s = \{x_s, y_s\}$ to an unlabeled target domain $\mathcal{X}_t = \{x_t\}$ for multiple segmentation tasks, which $x_s \in \mathbb{R}^{H^s\times W^s\times 3}, x_t \in \mathbb{R}^{H^t\times W^t\times 3}$ are images from pinhole and panoramic cameras respectively. To address the domain gap arising from dissimilar data distributions between the source pinhole and target panoramic domains, a comprehensive loss $\mathcal{L}_{total} = \mathcal{L}_S + \mathcal{L}_T$ is utilized for training the overall network $\mathcal{F}$, where $\mathcal{L}_S$ and $\mathcal{L}_T$ represent the source domain supervision loss and the target domain adaptation loss, respectively. $\mathcal{L}_S$ is utilized for supervision to learn fundamental segmentation capacity through labeled source samples $\{x_s, y_s\}$, whereas the $\mathcal{L}_T$ leverages unlabeled target images x_t to enhance segmentation in the target panoramic scenario. $\mathcal{L}_S$ includes semantic loss, instance loss, and amodal instance loss, each originating from respective branches. The semantic branch employs cross-entropy loss to assign each pixel to its respective category. As for the instance and amodal instance branches, we follow the conventional approach from Mask R-CNN [22], employing supervision using proposals-based bounding boxes and instance-level masks.

To tackle the challenge posed by the absence of labels in the target domain, we employ a self-training strategy to facilitate the model's adaptation from the source to the target domain. Self-training methods [23–25] typically utilize the target predictions p_t as pseudo-labels $\hat{y}_t$ for training. However, due to the noisy predictions in the early stage, resulting in low-confidence pseudo-labels, we adopt a confidence estimation mechanism that assigns weights ω of pseudo-labels-based self-training loss $\mathcal{L}_T$. Additionally, we incorporate the Mean-Teacher framework widely adopted in UDA [23, 57, 67] to further enhance the quality of the pseudo-labels. Thus, $\mathcal{L}_T$ can be expressed as follows:

$$\mathcal{L}_T = -\omega \sum\nolimits_{h,w,c} p_t^{(h,w,c)} \log \hat{y}_t^{(h,w,c)}, \tag{1}$$

where

$$\omega = \frac{1}{H^t \cdot W^t} \sum\nolimits_{h,w} (\max_c \mathcal{T}(x_t)^{(h,w,c)} > \tau), \tag{2}$$

$$\hat{y}_t^{(h,w)} = onehot(\arg\max_c \mathcal{T}(x_t)^{(h,w,c)}). \tag{3}$$

The parameters θ' of the teacher model $\mathcal{T}$ are updated using the parameters θ of the student model $\mathcal{F}$ by the EMA at each iteration.

Amodal-oriented Mix (AoMix). To address the challenges of diverse occlusion cases and inconsistent cross-domain scenes, we propose a new method AoMix, which processes images from two domains alongside self-training. Initially, we reconstruct the masked source images $\hat{x}_s$ based on a random source amodal annotations $\{M_r^i\}_{i=1}^z$. Subsequently, we follow the widely-used class-mix strategy [23, 67] in UDA segmentation tasks to generate masked mixed images $\hat{x}_m$, incorporating information from both the source and target domains.

Specifically, we randomly sample an amodal instance mask sequence $\{M_r^{(i)}\}_{i=1}^z$ of an image from the current batch, and subject it to random scaling $RS(\cdot)$ and peripheral random padding $RP(\cdot)$ to generate a new amodal instance mask sequence $\{M_r^{(i)'}\}_{i=1}^z$ with the object of various positions and sizes. Then, a random binary mask M_r containing multiple objects is produced by summing these masks $\{M_r^{(i)'}\}_{i=1}^z$. This operation can be expressed as follows:

$$M_r = H(\sum\nolimits_i RP(RS(M_r^{(i)}))), \tag{4}$$

where $H(\cdot)$ denoted as a step function. For a source image x_s, we sum the corresponding amodal instance masks $\{M_s^i\}_{i=1}^n$ to obtain a binary mask M_s of *Thing* region. Using the random binary mask M_r and the *Thing* region binary mask M_s, we fill the source image to reconstruct the masked source image $\hat{x}_s = (1 - M_r \cap M_s) \odot x_s$. Amodal-oriented masked image modeling aims to enhance the model's ability to reconstruct object regions obscured by realistic object shapes, enabling it to learn information about invisible parts. To adapt to unlabeled target scenes, we randomly sample half of the semantic classes from a source image and paste the corresponding pixels of the masked source image $\hat{x}_s$ onto a target image x_t based on the source semantic map, creating a new masked mixed image $\hat{x}_m$. This approach effectively transfers the model's capability tailored for distortion- and occlusion-aware to the target panoramic domain.

5 Experiments

5.1 Experiment Setups

Following DAFomer [23], we train UnmaskFomer using AdamW [46]. The learning rate is set to $6{\times}10^{-5}$, with a weight decay of 0.01 and linear warmup for $1.5k$ iterations followed by polynomial decay. The model is trained on a batch size of 4 with a crop size of 376×376 for $40k$ iterations. More details can be found in the supplementary.

5.2 Results of OASS on KITTI360-APS→BlendPASS

We deliver the OASS results on the KITTI360-APS→BlendPASS benchmark in Tab. 1. Representative UDA panoptic segmentation methods UniDAPS [91]

Table 1: Occlusion-Aware Scene Segmentation results on the KITTI360-APS → BlendPASS benchmark. For Semantic Segmentation (**SS**), per-class results are reported as IoU, and the metric is the mIoU. For Amodal Panoptic Segmentation (**APS**), per-class results are reported as APQ for full regions, and the Metric is the mAPQ.

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
SS	DATR [93]	71.87	27.24	70.59	22.76	35.98	23.91	00.00	04.52	77.06	37.14	80.06	51.20	02.24	70.15	08.94	06.00	11.02	27.79	34.91
	Trans4PASS [89]	72.79	33.80	78.38	33.53	37.07	26.56	05.31	05.35	77.43	37.91	84.73	57.48	04.51	76.58	19.04	17.03	12.66	51.70	40.66
	UniDAPS [91]	72.27	29.20	75.78	33.91	38.92	25.94	11.37	07.51	77.61	37.40	81.40	47.56	02.95	75.59	21.29	03.39	01.31	48.85	38.46
	EDAPS [57]	74.41	35.91	76.98	36.45	40.35	28.02	16.08	05.13	78.10	39.50	82.28	55.45	03.25	74.38	06.72	14.09	05.18	50.87	40.17
	Source-Only	72.97	29.14	82.04	31.19	31.37	17.98	00.00	15.97	74.13	33.74	88.66	50.24	04.02	80.42	18.02	11.04	04.57	50.17	38.65
	UnmaskFormer	76.54	37.82	77.06	34.71	44.05	28.27	17.80	02.76	78.70	41.68	84.98	57.34	06.01	80.60	23.47	21.67	18.80	53.56	**43.66**
APS	DATR [93]	51.82	09.15	59.90	11.93	11.98	01.97	00.00	03.91	64.61	14.05	70.40	11.09	00.00	39.30	00.00	03.16	10.07	01.30	20.26
	Trans4PASS [89]	53.91	14.12	69.39	19.15	11.77	03.77	00.00	05.15	67.63	16.02	77.41	15.30	04.24	41.06	06.58	00.00	00.00	07.35	22.94
	EDAPS [57]	54.88	17.04	66.86	18.75	14.47	05.75	04.04	04.64	68.20	16.04	72.76	19.01	00.00	36.73	05.77	04.38	00.00	07.20	23.14
	Source-Only	57.11	14.21	73.58	15.49	07.59	00.67	00.00	10.40	58.30	12.39	83.14	15.23	00.00	40.31	03.83	00.00	00.00	06.08	22.13
	UnmaskFormer	61.84	24.72	66.81	20.77	15.80	05.25	04.29	03.26	69.02	18.35	79.44	20.48	03.14	44.56	12.81	11.25	00.00	16.64	**26.58**

and EDAPS [57] with our amodal extensions are benchmarked. Further, SOTA panoramic semantic segmentation models Trans4PASS [89] and DATR [93] with distortion-adaptive capacities are compared. It is worth noting that UniDAPS does not support occlusion-aware reasoning with a contour-based decoder.

As shown in Tab. 1, compared with the best-performing UDA panoptic segmentation model EDAPS [57], UnmaskFormer outstrips it by clear margins of 3.49% in mIoU and 3.44% in mAPQ. Compared to the panoramic segmentation model Trans4PASS [89], our method also yields significant gains. Delving into the OASS results in semantic- and amodal panoptic segmentation, all benchmarked methods suffer from accurately predicting the full segmentation in the cases considering occlusions, in particular on small objects, evidenced by the unsatisfactory scores in class-wise APQ. Yet, UnmaskFormer harvests great improvements in contrast to the Source-Only model, *e.g.*, on safety-critical *pedestrians*, *truck*, and *two-wheeler*. Finally, under the challenging occlusion scenarios in unstructured surroundings, UnmaskFormer clearly stands out and leads to SOTA OASS scores of 43.66% in mIoU and 26.58% in mAPQ. In Fig. 6, we visualize OASS results produced by our UnmaskFormer and SOTA methods DATR [93], Trans4PASS [89], EDAPS [57]. Compared to them, UnmaskFormer excels in segmenting occluded objects and realistically reconstructing vehicle shapes thanks to our UA-based backbone and AoMix design.

Tab. 2 additionally presents the results in amodal instance segmentation of our comparative experiments on the benchmark of KITTI360-APS→BlendPASS transfer. The instance segmentation performance measured in both AP for visible regions and AAP for full regions are listed. UnmaskFormer reaches a new record of 11.10% in mAP and reaches an on-par mAAP score of 10.50% as EDAPS [57] specifically designed for UDA panoptic segmentation. Compared to EDAPS, our UnmaskFormer has large gains on challenging classes like *trucks* and *cars* frequently under large occlusions in unconstrained scenes.

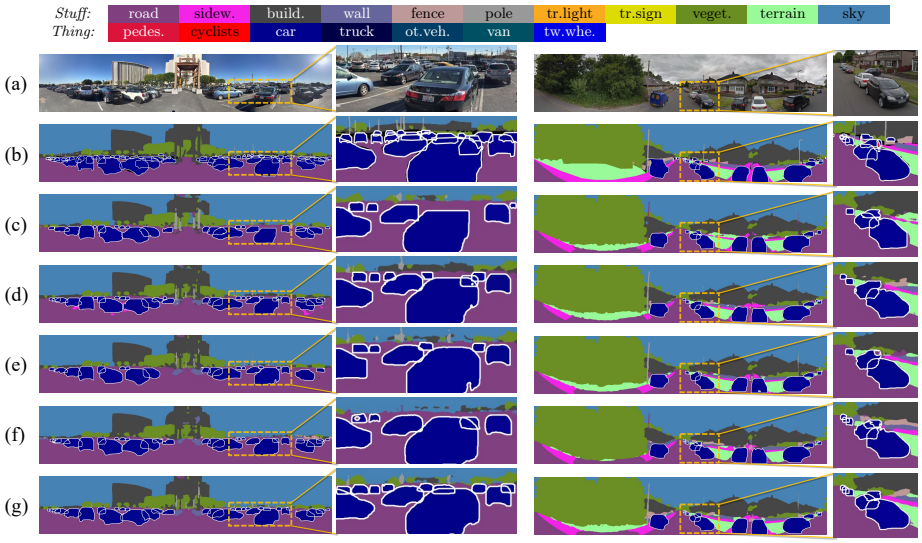


Fig. 6: Visualization results of OASS. From top to bottom are (a) Image, (b) Ground truth, (c) DATR [93], (d) Trans4PASS [89], (e) EDAPS [57], (f) Source-only, and (g) UnmaskFormer (ours).

Table 2: Instance-level Segmentation results on the KITTI360-APS→BlendPASS benchmark. Per-class results are reported as AP for visible regions and AAP for full regions. The Metrics are mAP and mAAP. "A" denotes Amodal.

UDA Method	A	pedestrians	cyclists	car	truck	other-vehicles	van	two-wheeler	Metric
DATR [93]		14.21	00.00	31.15	07.55	03.73	00.39	03.57	08.66
	✓	13.11	00.03	30.60	06.87	04.73	01.67	03.78	08.68
Trans4PASS [89]		16.52	00.03	32.23	10.19	05.31	00.16	05.62	10.01
	✓	15.95	00.21	31.71	08.34	06.01	00.35	06.41	09.85
UniDAPS [91]		02.31	00.00	11.25	06.17	02.80	00.00	01.50	03.43
	✓	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
EDAPS [57]		16.61	00.04	30.83	06.49	11.42	00.37	06.19	10.28
	✓	15.78	00.09	30.02	09.00	12.20	00.36	07.35	**10.68**
Source-Only		15.79	00.02	33.88	13.54	03.58	00.19	06.78	10.54
	✓	15.18	00.01	33.37	12.60	03.28	00.35	06.77	10.22
UnmaskFormer (Ours)		17.55	00.00	35.18	14.18	02.92	00.77	07.14	**11.10**
	✓	16.10	00.07	34.07	12.29	02.19	00.61	08.15	10.50

5.3 Results of Panoramic Semantic Segmentation

We further investigate the generalization capacity of the proposed UnmaskFormer using two panoramic semantic segmentation datasets [48, 90].

SynPASS. In Tab. 3, we conduct a comparison between known convolutional and transformer models for panoramic semantic segmentation on the SynPASS [90]. Here, the models are learned on the training set and evaluated on the validation

set with diverse weather and illumination conditions. UnmaskFormer equipped with the UA and the semantic head, produces SOTA performance in mIoU, which reaches 45.34%. Moreover, in each adverse condition, UnmaskFormer yields robust segmentation of the full 360°×180° panoramas on SynPASS.

Table 3: Panoramic Semantic Segmentation results on the SynPASS benchmark [90], covering four adverse weather conditions, and day- and night-time scenes.

Method	Cloudy	Foggy	Rainy	Sunny	Day	Night	ALL
Fast-SCNN [54]	30.84	22.68	26.16	27.19	29.68	24.75	26.31
DeepLabv3+ (MNv2) [6]	38.94	35.19	35.43	37.73	36.01	30.55	36.72
HRNet (W18Small) [68]	42.92	37.94	37.37	41.45	39.19	32.22	39.80
PVT (Small) [69]	40.75	36.14	34.29	40.14	37.92	28.80	37.47
SegFormer (B2) [72]	46.07	40.99	40.10	44.35	44.08	33.99	42.49
Trans4PASS (Small) [89]	46.74	43.49	43.39	45.94	45.52	37.03	44.80
UnmaskFormer (Ours)	48.00	43.43	43.48	47.06	45.87	36.43	**45.34**

Table 4: Results on the DensePASS benchmark [48]**.** Efficiency is compared in terms of FLOPs (G) and #Parameters (M).

Method	FLOPs	#Parameters	mIoU
SegFormer [72]	13.27	13.66	39.02
DPPASS [94]	*n.a.*	*n.a.*	42.40
DATR [93]	*n.a.*	14.72	42.22
FAN [96]	10.96	13.81	42.54
PoolFormer [84]	09.47	13.47	43.18
SegNeXt [19]	14.03	13.71	43.75
Trans4PASS [89]	12.02	13.93	45.89
UnmaskFormer (Ours)	11.73	13.96	**48.08**

DensePASS. As shown in Tab. 4, we further compare UnmaskFormer against efficient panoramic segmentation transformers including DPPASS [94], DATR [93], and Trans4PASS [90], where the models are trained on Cityscapes [9] and tested on the original DensePASS [48]. Compared to the DATR [93] constituted by distortion-aware attention with $14.72M$ parameters, our UnmaskFormer with fewer parameters ($13.96M$) attains higher segmentation performance. Our UnmaskFormer achieves 48.08% in mIoU, which outperforms its counterparts and maintains as a lightweight segmenter.

5.4 Ablation Study

To better understand the components of the UnmaskFormer, we conduct ablation studies on the KITTI360-APS→BlendPASS benchmark.

Why UA unmasks occlusions? In Tab. 5, we ablate different patch embedding methods [72, 89] and pooling methods [55]. The AvgPool and SimPool are used to replace the GAP operator in UA. Compared to the original PE design and the early-stage DPE, we construct the backbone by using a new interleaving arrangement which improves the distortion-aware modeling, yielding the best mAPQ score 26.58% with a +3.04% and a +2.14% gain respectively. Compared to SimPool [55], our occlusion-aware pooling attention better handles object occlusion and obtains higher results in OASS. An example of feature visualization in Fig. 7 shows that our UA can enhance distortion-tolerant predictions of regions where occlusion occurs. These results show the effectiveness of the UA backbone in addressing image distortion and object occlusion at the same time.

How AoMix boosts amodal prediction? As shown in Tab. 6, we investigate various approaches to masked image modeling. Compared to using masked images only for the source image (T for S) or the mixed image (T for M), our

Table 5: Unmasking Attention (UA) study.

Method	mIoU	mAPQ
PE [72]	41.07	22.00
DPE [89]	40.70	22.90
AvgPool	42.10	23.90
SimPool [55]	43.04	24.74
UA	43.06	25.04

Table 6: Amodal-oriented Mix (AoMix) study.

Method	mIoU	mAPQ
T for S	42.53	23.98
T for M	43.18	24.78
P for S&M	42.52	21.87
W for S&M	41.64	24.12
AoMix	42.39	25.17

Table 7: Ablation study of UnmaskFormer components.

UA	AoMix	OAFusion	mIoU	mAPQ
		✓	41.07	22.00
✓		✓	43.06	25.04
	✓	✓	42.39	25.17
✓	✓		43.66	26.55
✓	✓	✓	43.66	26.58

strategy AoMix, which utilizes both, achieves the best results. Masking the *Thing* class regions (AoMix), as opposed to the entire image (W for S&M) or masked by patches (P for S&M), results in improvements with a gain of +1.05% and +3.3% in mAPQ respectively. With the carefully devised AoMix, the UnmaskFormer earns significant gains by greatly boosting surrounding parsing seamlessly.

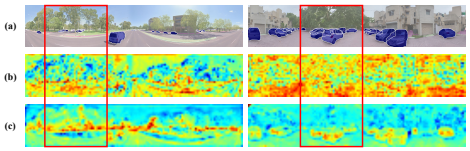


Fig. 7: Features Visualization before and after UA. (a) Input image and ground truth, (b) Features before UA, and (c) Features after UA. Zoom in for a better view.

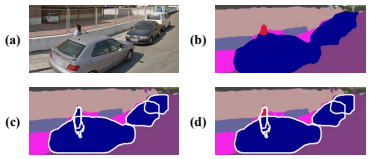


Fig. 8: Visualization of different fusion strategies for APS. (a) Image, (b) Predicted semantic map, (c) Conventional fusion [8, 57], and (d) OAFusion.

What OAFusion fuse? As illustrated in Fig. 8, a substantial portion of the pedestrian area (in **red**) is occluded by the car (in **blue**). Despite the amodal instance branch being capable of predicting the full region of the pedestrian, different fusion approaches yield distinct category classifications. The conventional fusion strategy [8, 57] tailored for panoptic segmentation, given the substantial occlusion of the pedestrian region by the car, misclassifies it as *car* under the majority voting rule. Conversely, our designed OAFusion ensures accurate classification of the full region as the *pedestrian* by disregarding overlapping regions.

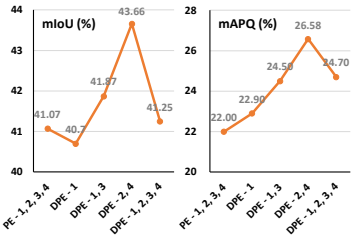
Finally, as shown in Tab. 7, all components cooperate with each other to achieve the best performance. More analyses can be found in the supplementary.

5.5 Analysis of Hyperparameters

We further conduct an analysis of relevant hyperparameters in UnmaskFormer on the KITTI360-APS→BlendPASS benchmark.

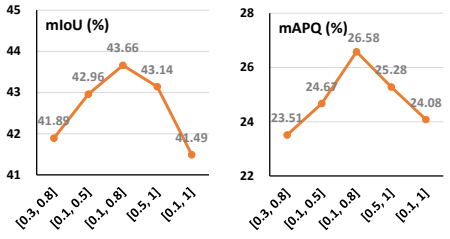
Analysis of deformable designs. In Fig. 9, we conduct experiments to analyze designs of PE [72] and DPE [89] in our UnmaskFormer model. We obtain three

insights: (1) The PE cannot effectively handle the image distortion in OASS, which is in line with the finding in [89]. (2) Using more DPE blocks inside a four-stage model does not ensure optimal performance. (3) Compared with shallow stages, DPE can bring more improvements when acting on deep stages, *e.g.*, using DPE in Stages 2 and 4 (DPE - 2, 4) performs better than in Stages 1 and 3 (DPE - 1, 3). Therefore, we adopt the deformable design of DPE in Stages 2 and 4 to construct the UnmaskFormer as default, which provides the optimal architecture for addressing image distortion and object deformation in OASS.



(a) mIoU **(b)** mAPQ

Fig. 9: Analysis of different designs with PE and DPE blocks in UnmaskFormer.



(a) mIoU **(b)** mAPQ

Fig. 10: Analysis of different scale ranges in the proposed AoMix method.

Analysis of AoMix parameters. In Fig. 10, we conduct experiments to analyze the impact of different scaling parameters in $RS(\cdot)$. Excessively large scaling sizes lead to complete occlusion of objects in images, preventing the model from learning original object information. Overly small scaling sizes result in minimal object occlusion, limiting the model's capacity to make reasonable predictions for occluded regions. Optimal performance is achieved by setting the range in $RS(\cdot)$ to $[0.1, 0.8]$, effectively generating diverse masked source images and masked mixed images. This variety of amodal-oriented samples bridges the gap between the pinhole and panoramic domains, enhancing the model's ability to reconstruct invisible regions of occluded objects and boost seamless segmentation.

6 Conclusion

In this work, we have introduced the task of *Occlusion-Aware Seamless Segmentation (OASS)* for holistic scene understanding. To address OASS, we put forward *UnmaskFormer* for unmasking the narrow field of view, unmasking the occlusion of perspective, and unmasking the gap of domain seamlessly. We establish the *BlendPASS* dataset for facilitating the optimization and evaluation of OASS models, as well as fostering future research in the panoramic and panoptic vision field. Experiments on the fresh BlendPASS, as well as public SynPASS and DensePASS benchmarks, demonstrate the effectiveness of the proposed methods.

Acknowledgements

This work was supported in part by the Major Research Plan of the National Natural Science Foundation of China under Grant 92148204, the National Key RD Program under Grant 2022YFB4701400, the Hunan Leading Talent of Technological Innovation under Grant 2022RC3063, the Top Ten Technical Research Projects of Hunan Province under Grant 2024GK1010, the Key Research Development Program of Hunan Province under Grant 2022GK2011, and in part by Hangzhou SurImage Technology Company Ltd.

References

1. Ai, H., Cao, Z., Zhu, J., Bai, H., Chen, Y., Wang, L.: Deep learning for omnidirectional vision: A survey and new perspectives. arXiv preprint arXiv:2205.10468 (2022)
2. Ao, J., Ke, Q., Ehinger, K.A.: Image amodal completion: A survey. Computer Vision and Image Understanding (2023)
3. Back, S., Lee, J., Kim, T., Noh, S., Kang, R., Bak, S., Lee, K.: Unseen object amodal instance segmentation via hierarchical occlusion modeling. In: ICRA (2022)
4. Breitenstein, J., Fingscheidt, T.: Amodal cityscapes: A new dataset, its generation, and an amodal semantic segmentation challenge baseline. In: IV (2022)
5. Chen, J., Niu, L., Zhang, J., Si, J., Qian, C., Zhang, L.: Amodal instance segmentation via prior-guided expansion. In: AAAI (2023)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
7. Chen, Y., Lin, G., Li, S., Bourahla, O., Wu, Y., Wang, F., Feng, J., Xu, M., Li, X.: BANet: Bidirectional aggregation network with occlusion handling for panoptic segmentation. In: CVPR (2020)
8. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.: Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: CVPR (2020)
9. 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)
10. Dai, J., He, K., Sun, J.: Convolutional feature masking for joint object and stuff segmentation. In: CVPR (2015)
11. Deng, L., Yang, M., Qian, Y., Wang, C., Wang, B.: CNN based semantic segmentation for urban traffic scenes using fisheye camera. In: IV (2017)
12. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: CoRL (2017)
13. Fan, K., Lei, J., Qian, X., Yu, M., Xiao, T., He, T., Zhang, Z., Fu, Y.: Rethinking amodal video segmentation from learning supervised signals with object-centric representation. In: ICCV (2023)
14. Follmann, P., König, R., Härtinger, P., Klostermann, M., Böttger, T.: Learning to see the invisible: End-to-end trainable amodal instance segmentation. In: WACV (2019)
15. Fu, X., Zhang, S., Chen, T., Lu, Y., Zhou, X., Geiger, A., Liao, Y.: PanopticNeRF-360: Panoramic 3D-to-2D label transfer in urban scenes. arXiv preprint arXiv:2309.10815 (2023)

16. Gao, J., Qian, X., Wang, Y., Xiao, T., He, T., Zhang, Z., Fu, Y.: Coarse-to-fine amodal segmentation with shape prior. In: ICCV (2023)
17. Gao, S., Yang, K., Shi, H., Wang, K., Bai, J.: Review on panoramic imaging and its applications in scene understanding. IEEE Transactions on Instrumentation and Measurement (2022)
18. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research (2013)
19. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: SegNeXt: Rethinking convolutional attention design for semantic segmentation. In: NeurIPS (2022)
20. Guttikonda, S., Rambach, J.: Single frame semantic segmentation using multimodal spherical images. In: WACV (2024)
21. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. In: ECCV (2014)
22. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
23. Hoyer, L., Dai, D., Van Gool, L.: DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: CVPR (2022)
24. Hoyer, L., Dai, D., Van Gool, L.: HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In: ECCV (2022)
25. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: MIC: Masked image consistency for context-enhanced domain adaptation. In: CVPR (2023)
26. Hu, X., An, Y., Shao, C., Hu, H.: Distortion convolution module for semantic segmentation of panoramic images based on the image-forming principle. IEEE Transactions on Instrumentation and Measurement (2022)
27. Hu, Y.T., Chen, H.S., Hui, K., Huang, J.B., Schwing, A.G.: SAIL-VOS: Semantic amodal instance level video object segmentation - A synthetic dataset and baselines. In: CVPR (2019)
28. Jang, S., Na, J., Oh, D.: DaDA: Distortion-aware domain adaptation for unsupervised semantic segmentation. In: NeurIPS (2022)
29. Jaus, A., Yang, K., Stiefelhagen, R.: Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning. In: IV (2021)
30. Jaus, A., Yang, K., Stiefelhagen, R.: Panoramic panoptic segmentation: Insights into surrounding parsing for mobile agents via unsupervised contrastive learning. IEEE Transactions on Intelligent Transportation Systems (2023)
31. Jiang, Q., Gao, S., Gao, Y., Yang, K., Yi, Z., Shi, H., Sun, L., Wang, K.: Minimalist and high-quality panoramic imaging with PSF-aware transformers. IEEE Transactions on Image Processing (2024)
32. Jiang, Q., Shi, H., Sun, L., Gao, S., Yang, K., Wang, K.: Annular computational imaging: Capture clear panoramic images through simple lens. IEEE Transactions on Computational Imaging (2022)
33. Ke, L., Tai, Y.W., Tang, C.K.: Deep occlusion-aware instance segmentation with overlapping bilayers. In: CVPR (2021)
34. Kim, J., Jeong, S., Sohn, K.: PASTS: Toward effective distilling transformer for panoramic semantic segmentation. In: ICIP (2022)
35. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: CVPR (2019)
36. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: ICCV (2023)

37. Lazarow, J., Lee, K., Shi, K., Tu, Z.: Learning instance occlusion for panoptic segmentation. In: CVPR (2020)
38. Li, K., Malik, J.: Amodal instance segmentation. In: ECCV (2016)
39. Li, X., Wu, T., Qi, Z., Wang, G., Shan, Y., Li, X.: SGAT4PASS: Spherical geometry-aware transformer for panoramic semantic segmentation. In: IJCAI (2023)
40. Li, Z., Ye, W., Jiang, T., Huang, T.: 2D amodal instance segmentation guided by 3D shape prior. In: ECCV (2022)
41. Li, Z., Ye, W., Jiang, T., Huang, T.: GIN: Generative invariant shape prior for amodal instance segmentation. IEEE Transactions on Multimedia (2023)
42. Li, Z., Ye, W., Terven, J., Bennett, Z., Zheng, Y., Jiang, T., Huang, T.: MUVA: A new large-scale benchmark for multi-view amodal instance segmentation in the shopping scenario. In: ICCV (2023)
43. Liao, Y., Xie, J., Geiger, A.: KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
44. Ling, Z., Xing, Z., Zhou, X., Cao, M., Zhou, G.: PanoSwin: A pano-style swin transformer for panorama understanding. In: CVPR (2023)
45. Liu, Z., Li, Z., Jiang, T.: BLADE: Box-level supervised amodal segmentation through directed expansion. In: AAAI (2024)
46. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
47. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: CVPR (2019)
48. 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)
49. Mei, J., Zhu, A.Z., Yan, X., Yan, H., Qiao, S., Chen, L.C., Kretzschmar, H.: Waymo open dataset: Panoramic video panoptic segmentation. In: ECCV (2022)
50. Mohan, R., Valada, A.: Amodal panoptic segmentation. In: CVPR (2022)
51. Mohan, R., Valada, A.: Perceiving the invisible: Proposal-free amodal panoptic segmentation. IEEE Robotics and Automation Letters (2022)
52. Nanay, B.: The importance of amodal completion in everyday perception. i-Perception (2018)
53. Orhan, S., Bastanlar, Y.: Semantic segmentation of outdoor panoramic images. Signal, Image and Video Processing (2022)
54. Poudel, R.P.K., Liwicki, S., Cipolla, R.: Fast-SCNN: Fast semantic segmentation network. In: BMVC (2019)
55. Psomas, B., Kakogeorgiou, I., Karantzalos, K., Avrithis, Y.: Keep it SimPool: Who said supervised transformers suffer from attention deficit? In: ICCV (2023)
56. Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J.: Amodal instance segmentation with KINS dataset. In: CVPR (2019)
57. Saha, S., Hoyer, L., Obukhov, A., Dai, D., Van Gool, L.: EDAPS: Enhanced domain-adaptive panoptic segmentation. In: ICCV (2023)
58. Sekkat, A.R., Dupuis, Y., Vasseur, P., Honeine, P.: The omniscape dataset. In: ICRA (2020)
59. Sekkat, A.R., Mohan, R., Sawade, O., Matthes, E., Valada, A.: AmodalSynthDrive: A synthetic amodal perception dataset for autonomous driving. arXiv preprint arXiv:2309.06547 (2023)
60. Shen, Z., Lin, C., Liao, K., Nie, L., Zheng, Z., Zhao, Y.: PanoFormer: Panorama transformer for indoor 360° depth estimation. In: ECCV (2022)

61. Shi, H., Li, Y., Yang, K., Zhang, J., Peng, K., Roitberg, A., Ye, Y., Ni, H., Wang, K., Stiefelhagen, R.: FishDreamer: Towards fisheye semantic completion via unified image outpainting and segmentation. In: CVPRW (2023)
62. Shi, Y., Ying, X., Zha, H.: Unsupervised domain adaptation for semantic segmentation of urban street scenes reflected by convex mirrors. IEEE Transactions on Intelligent Transportation Systems (2022)
63. Sun, Y., Kortylewski, A., Yuille, A.: Amodal segmentation through out-of-task and out-of-distribution generalization with a Bayesian model. In: CVPR (2022)
64. Tateno, K., Navab, N., Tombari, F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In: ECCV (2018)
65. Teng, Z., Zhang, J., Yang, K., Peng, K., Shi, H., Reiß, S., Cao, K., Stiefelhagen, R.: 360BEV: Panoramic semantic mapping for indoor bird's-eye view. In: WACV (2024)
66. Tran, M., Vo, K., Yamazaki, K., Fernandes, A., Kidd, M., Le, N.: AISFormer: Amodal instance segmentation with transformer. In: BMVC (2022)
67. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: DACS: Domain adaptation via cross-domain mixed sampling. In: WACV (2021)
68. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
69. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: ICCV (2021)
70. Wang, Z., Yu, M., Wei, Y., Feris, R., Xiong, J., Hwu, W., Huang, T.S., Shi, H.: Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In: CVPR (2020)
71. Xiao, Y., Xu, Y., Zhong, Z., Luo, W., Li, J., Gao, S.: Amodal segmentation based on visible region segmentation and shape prior. In: AAAI (2021)
72. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Álvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS (2021)
73. Xu, Y., Wang, K., Yang, K., Sun, D., Fu, J.: Semantic segmentation of panoramic images using a synthetic dataset. In: SPIE (2019)
74. Yang, K., Hu, X., Bergasa, L.M., Romera, E., Huang, X., Sun, D., Wang, K.: Can we PASS beyond the field of view? Panoramic annular semantic segmentation for real-world surrounding perception. In: IV (2019)
75. Yang, K., Hu, X., Bergasa, L.M., Romera, E., Wang, K.: PASS: Panoramic annular semantic segmentation. IEEE Transactions on Intelligent Transportation Systems (2020)
76. Yang, K., Hu, X., Chen, H., Xiang, K., Wang, K., Stiefelhagen, R.: DS-PASS: Detail-sensitive panoramic annular semantic segmentation through SwaftNet for surrounding sensing. In: IV (2020)
77. Yang, K., Hu, X., Fang, Y., Wang, K., Stiefelhagen, R.: Omnisupervised omnidirectional semantic segmentation. IEEE Transactions on Intelligent Transportation Systems (2022)
78. Yang, K., Hu, X., Stiefelhagen, R.: Is context-aware CNN ready for the surroundings? Panoramic semantic segmentation in the wild. IEEE Transactions on Image Processing (2021)
79. Yang, K., Zhang, J., Reiß, S., Hu, X., Stiefelhagen, R.: Capturing omni-range context for omnidirectional segmentation. In: CVPR (2021)

80. Ye, Y., Yang, K., Xiang, K., Wang, J., Wang, K.: Universal semantic segmentation for fisheye urban driving images. In: SMC (2020)
81. Yogamani, S.K., Witt, C., Rashed, H., Nayak, S., Mansoor, S., Varley, P., Perrotton, X., O'Dea, D., Pérez, P., Hughes, C., Horgan, J., Sistu, G., Chennupati, S., Uricár, M., Milz, S., Simon, M., Amende, K.: WoodScape: A multi-task, multi-camera fisheye dataset for autonomous driving. In: ICCV (2019)
82. Yu, F., Wang, X., Cao, M., Li, G., Shan, Y., Dong, C.: OSRT: Omnidirectional image super-resolution with distortion-aware transformer. In: CVPR (2023)
83. Yu, H., He, L., Jian, B., Feng, W., Liu, S.: PanelNet: Understanding 360 indoor environment via panel representation. In: CVPR (2023)
84. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: MetaFormer is actually what you need for vision. In: CVPR (2022)
85. Yuan, X., Kortylewski, A., Sun, Y., Yuille, A.: Robust instance segmentation through reasoning about multi-object occlusion. In: CVPR (2021)
86. Yue, X., Zheng, Z., Zhang, S., Gao, Y., Darrell, T., Keutzer, K., Vincentelli, A.S.: Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In: CVPR (2021)
87. Zhang, C., Cui, Z., Chen, C., Liu, S., Zeng, B., Bao, H., Zhang, Y.: DeepPanoContext: Panoramic 3D scene understanding with holistic scene context graph and relation-based optimization. In: ICCV (2021)
88. Zhang, J., Ma, C., Yang, K., Roitberg, A., Peng, K., Stiefelhagen, R.: Transfer beyond the field of view: Dense panoramic semantic segmentation via unsupervised domain adaptation. IEEE Transactions on Intelligent Transportation Systems (2022)
89. 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)
90. 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)
91. Zhang, J., Huang, J., Lu, S.: Hierarchical mask calibration for unified domain adaptive panoptic segmentation. In: CVPR (2023)
92. Zhang, Z., Chen, A., Xie, L., Yu, J., Gao, S.: Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In: MM (2019)
93. Zheng, X., Pan, T., Luo, Y., Wang, L.: Look at the neighbor: Distortion-aware unsupervised domain adaptation for panoramic semantic segmentation. In: ICCV (2023)
94. Zheng, X., Zhu, J., Liu, Y., Cao, Z., Fu, C., Wang, L.: Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation. In: CVPR (2023)
95. Zheng, Z., Lin, C., Nie, L., Liao, K., Shen, Z., Zhao, Y.: Complementary bi-directional feature compression for indoor 360° semantic segmentation with self-distillation. In: WACV (2023)
96. Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., Álvarez, J.M.: Understanding the robustness in vision transformers. In: ICML (2022)
97. Zhu, Y., Tian, Y., Metaxas, D., Dollár, P.: Semantic amodal segmentation. In: CVPR (2017)
98. Zou, Y., Yu, Z., Liu, X., Kumar, B.V.K.V., Wang, J.: Confidence regularized self-training. In: ICCV (2019)