GoodSAM: Bridging Domain and Capacity Gaps via Segment Anything Model for Distortion-aware Panoramic Semantic Segmentation –Supplementary Material–

Abstract

Due to the lack of space in the main paper, we provide more details of the proposed methods and experimental results in the supplementary material. Sec. 1 provides more details about the Distortion-Aware Rectification (DAR) Module, while Sec. 2 delves into additional details about the experiments.

1. More details within DAR module

In this section, we provide a more detailed explanation of some methods within the DAR module. In Sec. 1.1, we discuss the reasons for using the sliding window strategy and its effectiveness. Sec . 1.2 provides additional details within the Boundary Enhancement Block. Finally, Sec. 1.3 offers more insights into the Cross-Task Complementary Fusion (CTCF) block, providing a detailed showcase of the process for obtaining ensemble logits.

1.1. Effectiveness of the Sliding Window Strategy

Due to the characteristics of panoramic images with a large field of view (FoV), directly inputting the entire ERP image leads to a degradation in segmentation performance for both SAM and teacher assistant (TA). Therefore, we employ a sliding window strategy to extract local patches from the input ERP images. Since the patches after window-based cropping remain coherent with standard RGB images in terms of FoV, SAM and the TA are not prone to missing the segmentation of small objects due to the oversized FoV. Notably, in the case of SAM employing random point prompts to generate instance masks, the vast FoV might lead to under-sampling and, consequently, omission of instance segmentation for certain small objects. According to [9], we find that horizontal distortion quantifies the differences in pixel distances among various projection types, whereas vertical distortion is more uniformly distributed. As a result, considering limited resources, in this work, we choose the horizontal movement of windows for obtaining patches.

1.2. More details in Boundary Enhancement Block

In the context of semantic segmentation, the presence of long-range context can result in a significant prediction gap between interior and boundary pixel predictions, often leading to blurred boundaries[1]. For panorama images, suffering from distortion and deformation problems makes obtaining accurate boundary predictions more challenging. SAM, having been trained on a large dataset of 1 billion images, can provide high-quality boundary information compared to TA. Therefore, we introduce a Boundary Enhancement Block to combine the boundary predictions from SAM and TA, obtaining a more reliable boundary map. This helps TA and the student model alleviate the distortion problem. To obtain the refined boundary map, we propose a boundary refinement strategy, and the detailed pseudo-code is presented in Alg. 1.

For the boundary refinement strategy, our aim is to reconstruct a more reliable boundary map B^i_{ref} for the current overlapping area O_i by selecting more reliable boundary pixels from B^i_{TA}, B^j_{TA}, and B^i_{SAM}. Specifically, for each pixel P^i_{TA} in B^i_{TA}, we locate the pixels P^j_{TA} and P^i_{SAM} at the same positions in B^j_{TA}, and B^i_{SAM}. We consider two different conditions: firstly, if both P^j_{TA} and P^i_{SAM} are boundary pixels (*i.e.*, the value of corresponding pixels is 1), then P^i_{TA} is considered a reliable boundary pixel.

If the above condition is not satisfied, we consider to find the reliable boundary pixel based on SAM's boundary prediction. We first find the boundary pixel P^v_{SAM} closest to P^i_{SAM} in the vertical direction, and then find the corresponding pixels P^v_{TA} and P^t_{TA} at the same positions in B^i_{TA} and B^j_{TA}, respectively. For each pixel, we consider the softmax value distribution of categories, and if the difference D between the top 2 values in the distributions is below a certain threshold α, it indicates that the pixel conforms to the distribution of boundary pixels. D is formulated as:

$$\mathcal{D} = Dis(P_i^{C_1}) - Dis(P_i^{C_2}), \tag{1}$$

where the Dis denotes the softmax value of pixel distribution from TA, C_i denotes the category channel where the

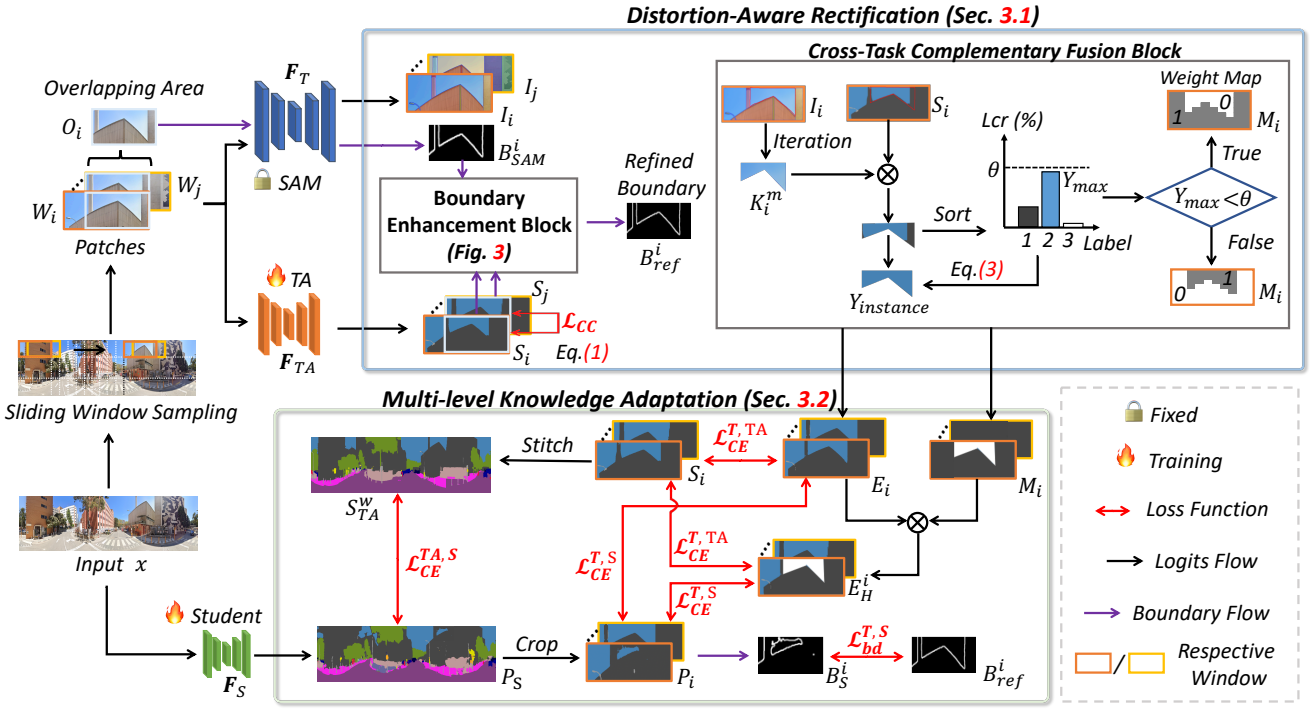

Figure 1. Illustration of the overall framework.

ith largest value in the pixel distribution is located. If either D_i of P^v_{TA} or D_j of P^t_{TA} is less than α, we consider P^i_{SAM} as a more reliable boundary pixel. If above condition is not met, we retain the boundary pixels of B^i_{TA} as reliable pixels. Finally, in B^i_{ref}, we assign a value of 1 to the positions corresponding to reliable boundary pixels, resulting in a refined boundary map that benefits from the combined boundary information provided by SAM.

1.3. More details within CTCF Block

The purpose of the CTCF block is to obtain ensemble logits for the window patches. This is achieved by combining the instance masks output by SAM and the semantic segmentation map output by TA. We provide our overall framework (as shown in Fig. 1) here for better understanding. The fusion mechanism involves providing reliable label assignments to the instance masks output by SAM based on the semantic map from TA. Specifically, we iterate through the instance masks in O_i, considering different area sizes. We define the medium-sized instance masks as those within the area range Γ $[A_{\text{min}}, A_{\text{max}}]$. For medium-sized masks, we set a higher label coverage rate (lcr) threshold θ, while for large-sized and small-sized masks, we set a lower threshold. The specific fusion mechanism is outlined in Alg. 2. By considering masks of various sizes differently and setting different thresholds, our fusion mechanism surpasses existing fusion methods [2, 3].

2. More details in Experiments

Due to space constraints in the main paper, this section provide more details on implementation, comparisons, and ablation studies.

2.1. More datasets and implementation details

Dataset. We leverage two benchmark datasets WildPASS [6] and DensePASS [4] to assess the segmentation performance of the GoodSAM. The WildPASS and DensePASS datasets both comprise the same 2000 unlabeled panorama images gathered from 40 diverse cities for training. For the evaluation of the WildPASS dataset, we use an additional 500 annotated panorama images from 25 cities spanning various continents. For the DensePASS dataset evaluation, we leverage 100 precise annotated images. The resolution of images in both datasets utilized is 400×2048.

Implementation details. We train the proposed framework with PyTorch in 4 NVIDIA A6000 GPUs. In our experimental design, we employ SAM as the teacher model within the framework. We keep SAM frozen during our experiments and utilize it solely for providing instance masks and boundary information. For the TA and student models, we opt for the fine-tuned Segformer [5] series, encompassing B0-B5 variants, which come in six different sizes and exhibit varying performance levels in 2D image semantic segmentation. Specifically, we choose B0, B1, and B2 as our

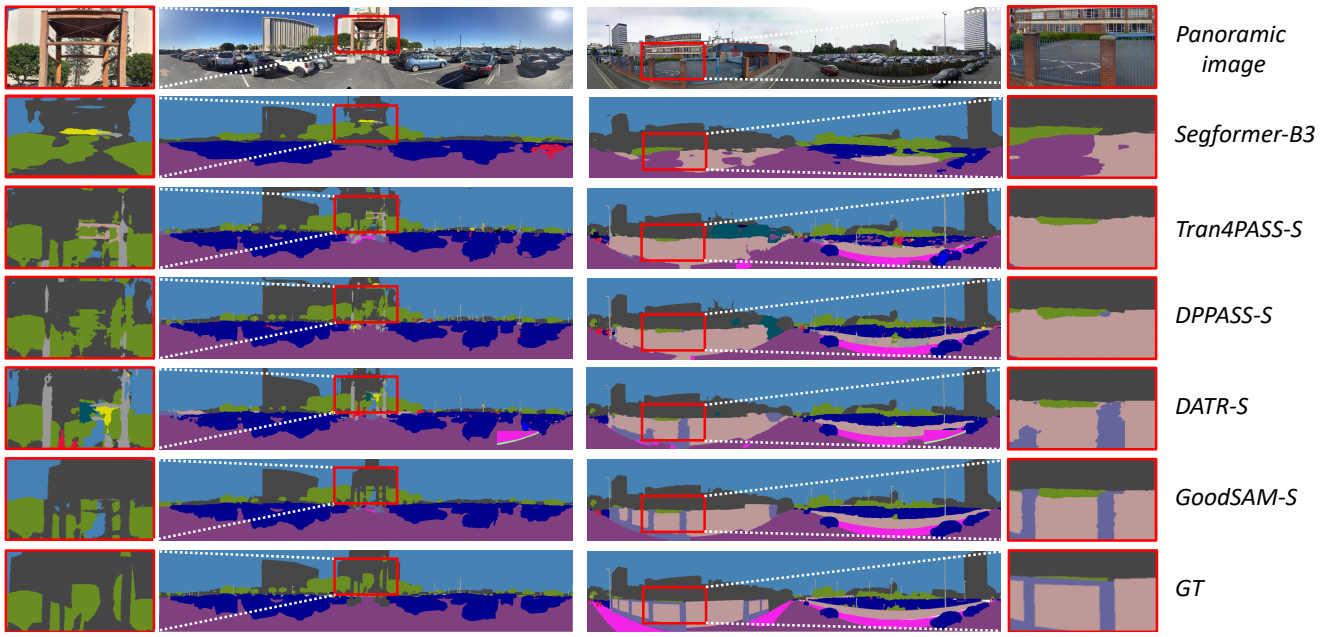

Figure 2. More visual comparisons based on DensePASS validation set.

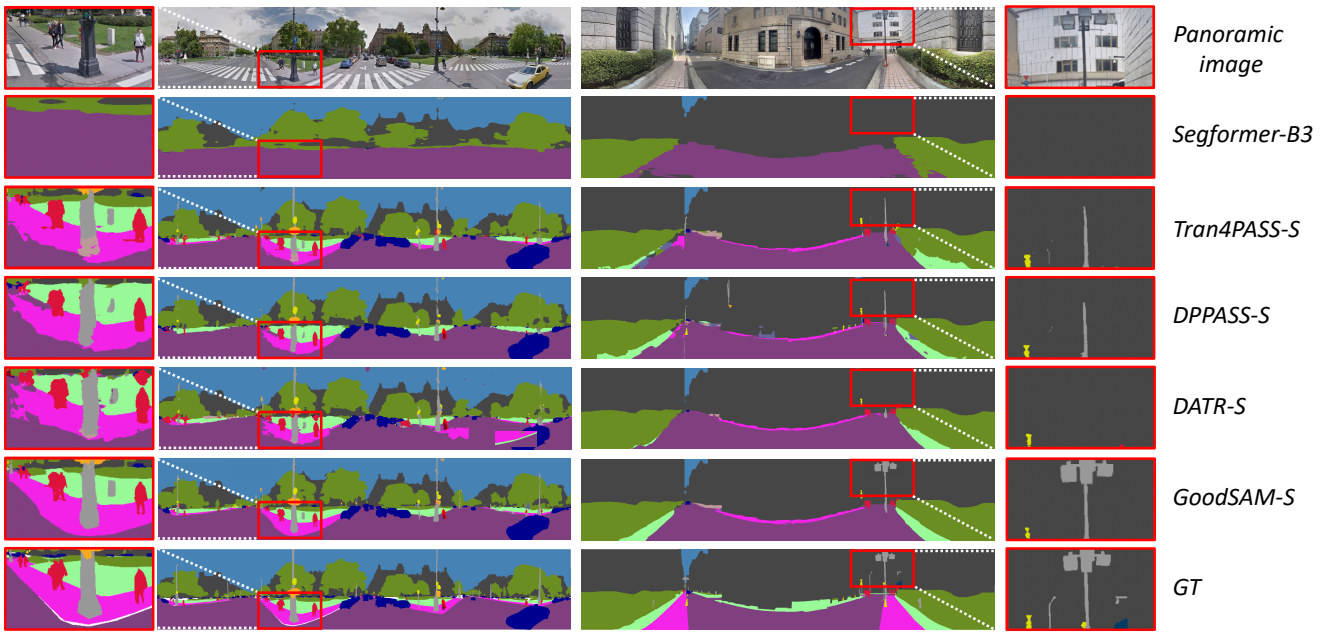

Figure 3. More visual comparisons based on DensePASS validation set.

student models and B3, B4, and B5 as our TA models for the experiment. We employ two AdamW optimizers to update the parameters of the student and TA respectively. The initial learning rate for both optimizers is $1 \times e^{-5}$, with an epsilon setting of $1 \times e^{-8}$, and a weight decay of $5 \times e^{-4}$. We set the window size as 400×512, with stride 256. The hyper-parameter α is set to 0.3. For medium-sized masks, the area range Γ is from 100 to 1000. The thresholds θ for masks of different areas are set to 0.5 and 0.7. The hyper-parameters of weight for reliable masks are set to 0.2. We utilize the mean Intersection-over-Union (mIoU) as the evaluation metric.

2.2. More comparisons with existing works

Figs. 2 and 3 showcase additional visual comparisons based on the DensePASS validation set. We compare mod-

Algorithm 1 Boundary Refinement Strategy

Input: The overlapping area boundary maps from TA predictions B^i_{TA}, B^j_{TA}. The overlapping area boundary maps from SAM B^i_{SAM}.
Output: The refined boundary map B^i_{ref}.
for Each boundary pixel P^i_{TA} in B^i_{TA} **do**
  Find the corresponding pixels P^j_{TA} and P^i_{SAM} at the same position in B^j_{TA}, and B^i_{SAM}.
  if Value(P^j_{TA})==1 & Value(P^i_{SAM})==1 **then**
    The value of the pixel P^i_{ref} at the position corresponding to P^i_{TA} = 1.
  else
    Find the boundary pixel P^v_{SAM} closest to P^i_{SAM} in the vertical direction.
    Find the corresponding pixels P^v_{TA} and P^t_{TA} at the same positions in B^i_{TA} and B^j_{TA}.
    Calculate the softmax value distribution of P^v_{TA} and P^t_{TA}.
    Calculate the D_i of P^v_{TA} or D_j of P^t_{TA} using Eq. 1.
    if $D_i < \alpha \| D_j < \alpha$ **then**
      The value of the pixel P^v_{ref} at the position corresponding to P^v_{SAM} = 1.
    else
      The value of the pixel P^i_{ref} at the position corresponding to P^i_{TA} = 1.
    end if
  end if
end for

θ	0/0	0.5/0.5	0.5/0.7	0.6/0.8
mIoU	47.93	49.06	49.7	49.59
Γ	0/0	100/1000	200/2000	300/3000
mIoU	49.06	49.7	49.63	49.21

Table 1. Ablation about threshold θ and Γ of the fusion module.

els with approximately similar-level parameters, including Segformer-B3 [5], Trans4PASS-S [7], DPPASS-S [10], DATR-S [9], and our GoodSAM-S. It can be observed that our GoodSAM-S outperforms SoTA unsupervised domain adaptation (UDA) methods in ERP semantic segmentation with comparable parameter sizes. Specifically, our GoodSAM-S exhibits superior performance in complex scenes and boundary prediction compared to other methods. This also highlights the reliability of the ensemble logits and refined boundary map obtained through our DAR module.

Algorithm 2 Fusion Mechanism

Input: The overlapping area instance masks I_i from SAM. The overlapping area semantic map S_i from TA.
Output: Ensemble logits E_i for the overlapping area.
for Each instance mask K^m_i in I_i **do**
  Find the same region in S_i as K_{Sem} corresponds to the instance mask K^m_i.
  Calculate the count of each label in K_{Sem} and sort them in descending order.
  Identify the top three labels Y_a (a belongs to the number of categories) in the sorted order.
  if the lcr of most prevalent semantic label $Y_{max} > \theta$ **then**
    $Y_{instance} = Y_{max}$.
  else
    Calculate the Shannon entropy value of Y_a based on S_i.
    Find the $Y_{argmin\{SE(Y_a)\}}$ which the SE value is the smallest.
    $Y_{instance} = Y_{argmin\{SE(Y_a)\}}$.
  end if
end for

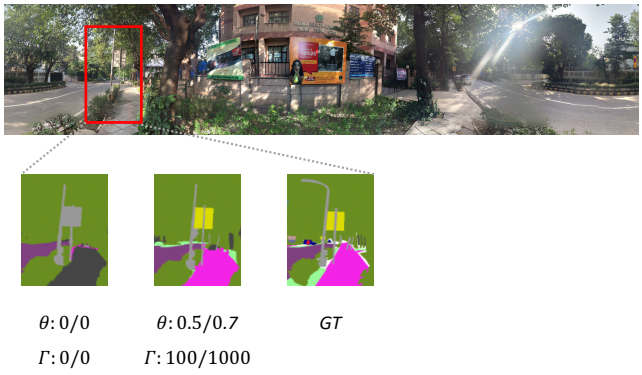

Figure 4. visual comparisons about ablation study of θ and Γ.

2.3. More ablation studies

In the CTCF block, we define the area range Γ for medium-sized masks and lcr thresholds θ for masks of various area levels. Tab. 1 presents the experimental results for different θ and Γ. Fig. 4 presents visual comparisons for the fusion results using θ and Γ.

Lcr thresholds θ. For different lcr thresholds, if we do not set a threshold for instance masks at different area levels, it implies that $Y_{instance}$ is certain to be the most prevalent semantic label Y_{max}. Through data comparison, we find that this setting is likely to result in label assignment errors. However, when we set different lcr thresholds for masks of different areas, we experimentally observe that, compared

Setting	Methods	mIoU(%)
UDA	Trans4PASS+-S (source-only) [8]	51.48
	Trans4PASS+-S (SAM-enhanced) [8]	52.77
	Trans4PASS+-S + SSL [8]	52.35
	Trans4PASS+-S + MPA [8]	55.24
Unsupervised	GoodSAM-S(baseline) + SEPL [3]	54.93
	GoodSAM-S	**60.56**

Table 2. Performance comparison with SAM-enhanced methods.

to the same lcr thresholds, medium-sized masks are more likely to require SE to compute a more reliable label. When using large thresholds, it leads to a decrease in model performance. We speculate that this may be because the SE value calculation for a more reliable label does not apply to all situations. Therefore, we choose to use 0.5/0.7, the model achieves the best performance.

Area range Γ. For different area ranges Γ, we try four different scenarios. First, when we do not set any Γ, we find that when lcr thresholds are both 0.5, the performance is lower because medium-sized masks are more likely to have multiple labels with close lcr. This increases the likelihood of incorrect label assignments. However, when we try three different area ranges, we find that when the area range Γ is 100-1000, our GoodSAM get the best performance.

Analysis of main sources of performance improvement. We further evaluate whether the performance improvement obtained by GoodSAM is more reliant on the performance of SAM or more dependent on our framework design. From Tab. 2, We can find that compared to the previous method [8] of using SAM for pseudo label enhancement, our GoodSAM-S can exceed Trans4PASS+-S+MPA by **5.32** % mIoU on the DensePASS dataset. When training our GoodSAM-S baseline (SAM+TA+SW) with SEPL [3], the performance lags behind that of our GoodSAM-S by **5.63**% IoU. These indicate that the performance of GoodSAM is largely attributed to the contribution of our framework, rather than SAM's capabilities.

References

[1] Da Chen, Jack Spencer, Jean-Marie Mirebeau, Ke Chen, Minglei Shu, and Laurent D Cohen. A generalized asymmetric dual-front model for active contours and image segmentation. *IEEE Transactions on Image Processing*, 30:5056–5071, 2021. 1

[2] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything, 2023. 2

[3] Tianle Chen, Zheda Mai, Ruiwen Li, and Wei-lun Chao. Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *arXiv preprint arXiv:2305.05803*, 2023. 2, 5

[4] 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 *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2766–2772. IEEE, 2021. 2

[5] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2, 4

[6] Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild. *IEEE Transactions on Image Processing*, 30:1866–1881, 2021. 2

[7] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16917–16927, 2022. 4

[8] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Philip HS Torr, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *arXiv preprint arXiv:2207.11860*, 2022. 5

[9] Xu Zheng, Tianbo Pan, Yunhao Luo, and Lin Wang. Look at the neighbor: Distortion-aware unsupervised domain adaptation for panoramic semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18687–18698, 2023. 1, 4

[10] Xu Zheng, Jinjing Zhu, Yexin Liu, Zidong Cao, Chong Fu, and Lin Wang. Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1285–1295, 2023. 4